Evaluacija i izbor modela

Mašinsko učenje 2020/21.

Matematički fakultet
Univerzitet u Beogradu

Evaluacija i izbor modela

- *Evaluacija modela* predstavlja kvantifikaciju njegove sposobnosti predviđanja

Evaluacija i izbor modela

- *Evaluacija modela* predstavlja kvantifikaciju njegove sposobnosti predviđanja
- Ukoliko imamo na raspolaganju konačan broj modela, od kojih je potrebno koristiti jedan, očigledno se postavlja pitanje *izbora modela*, koje se obično rešava tako što se na neki način evaluiraju svi raspoloživi modeli izabere se najbolji

Evaluacija i izbor modela

- *Evaluacija modela* predstavlja kvantifikaciju njegove sposobnosti predviđanja
- Ukoliko imamo na raspolaganju konačan broj modela, od kojih je potrebno koristiti jedan, očigledno se postavlja pitanje *izbora modela*, koje se obično rešava tako što se na neki način evaluiraju svi raspoloživi modeli izabere se najbolji
- Evaluacija modela počiva na *merama kvaliteta modela* i na *tehnikama evaluacije modela*

Evaluacija i izbor modela

- *Evaluacija modela* predstavlja kvantifikaciju njegove sposobnosti predviđanja
- Ukoliko imamo na raspolaganju konačan broj modela, od kojih je potrebno koristiti jedan, očigledno se postavlja pitanje *izbora modela*, koje se obično rešava tako što se na neki način evaluiraju svi raspoloživi modeli izabere se najbolji
- Evaluacija modela počiva na *merama kvaliteta modela* i na *tehnikama evaluacije modela*
- Kako izbor modela počiva na evaluaciji modela, tehnike koje se koriste su slične, što doprinosi konfuziji i potencijalnim greškama

Pregled

Mere kvaliteta modela

- Različite mere za klasifikacione i regresione probleme

Mere kvaliteta modela

- Različite mere za klasifikacione i regresione probleme
- Mogu se osmisliti posebno za konkretan problem

Mere kvaliteta modela

- Različite mere za klasifikacione i regresione probleme
- Mogu se osmisliti posebno za konkretan problem
- Postoje opšte mere koje se najčešće koriste u različitim kontekstima

Pregled

Mere kvaliteta klasifikacionih modela

- Mere koje se najčešće koriste za klasifikaciju

Mere kvaliteta klasifikacionih modela

- Mere koje se najčešće koriste za klasifikaciju
  - *tačnost klasifikacije* (eng. *classification accuracy*)

Mere kvaliteta klasifikacionih modela

- Mere koje se najčešće koriste za klasifikaciju
  - *tačnost klasifikacije* (eng. *classification accuracy*)
  - *preciznost i odziv* (eng. *precision and recall*)

Mere kvaliteta klasifikacionih modela

- Mere koje se najčešće koriste za klasifikaciju
  - *tačnost klasifikacije* (eng. *classification accuracy*)
  - *preciznost i odziv* (eng. *precision and recall*)
  - F_1 mera

Mere kvaliteta klasifikacionih modela

- Mere koje se najčešće koriste za klasifikaciju
  - *tačnost klasifikacije* (eng. *classification accuracy*)
  - *preciznost i odziv* (eng. *precision and recall*)
  - F_1 mera
  - površina ispod ROC krive (eng. *area under the curve – AUC*)

Mere kvaliteta klasifikacionih modela

- Mere koje se najčešće koriste za klasifikaciju
  - *tačnost klasifikacije* (eng. *classification accuracy*)
  - *preciznost i odziv* (eng. *precision and recall*)
  - F_1 mera
  - površina ispod ROC krive (eng. *area under the curve – AUC*)
- Praktično sve ove mere počivaju na *matrici konfuzije* (eng. *confusion matrix*)

Matrica konfuzije

- Matricom konfuzije nazivamo matrica C čije su vrste i kolone označene klasama a element c_{ij} predstavlja broj elemenata klase i koji su klasifikovani u klasu j

Matrica konfuzije

- Matricom konfuzije nazivamo matrica C čije su vrste i kolone označene klasama a element c_{ij} predstavlja broj elemenata klase i koji su klasifikovani u klasu j
- Klasifikacija je očito najbolja kada je ova matrica dijagonalna, što znači da je klasifikacija potpuno ispravna

Matrica konfuzije

- Matricom konfuzije nazivamo matrica C čije su vrste i kolone označene klasama a element c_{ij} predstavlja broj elemenata klase i koji su klasifikovani u klasu j
- Klasifikacija je očito najbolja kada je ova matrica dijagonalna, što znači da je klasifikacija potpuno ispravna
- Nedijagonalni elementi označavaju greške

Matrica konfuzije

- U slučaju binarne klasifikacije, obično se jedna klasa naziva *pozitivnom*, a druga *negativnom*

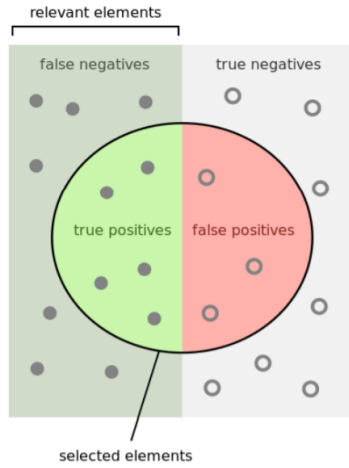
Matrica konfuzije

- U slučaju binarne klasifikacije, obično se jedna klasa naziva *pozitivnom*, a druga *negativnom*
- Tada matrica konfuzije ima specifičan oblik:

Stvarno/Predviđeno	Pozitivno	Negativno
Pozitivno	stvarno pozitivno (TP)	lažno negativno (FN)
Negativno	lažno pozitivno (FP)	stvarno negativno (TN)

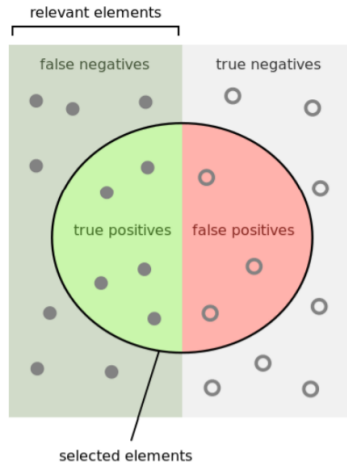
Matrica konfuzije

- Dat je skup instanci od kojih neke pripadaju pozitivnoj klasi (ispunjeni kružići) a neke negativnoj klasi (prazni kružići)



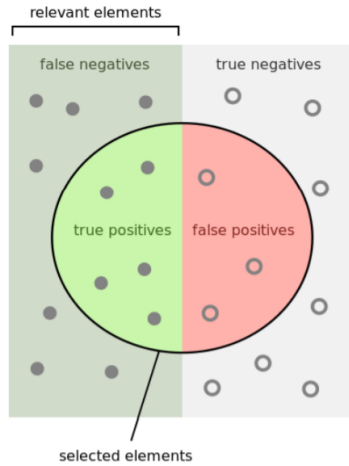
Matrica konfuzije

- Dat je skup instanci od kojih neke pripadaju pozitivnoj klasi (ispunjeni kružići) a neke negativnoj klasi (prazni kružići)
- Dat je model koji instance unutar kružnice klasifikuje kao pozitivne a one izvan kružnice kao negativne



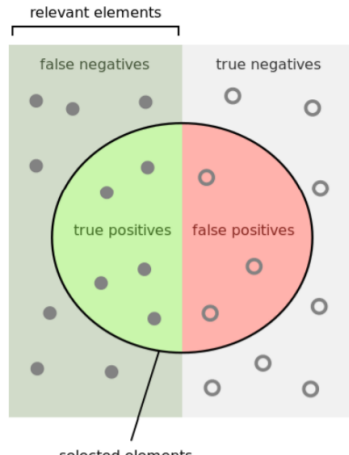
Matrica konfuzije

- *Stvarno pozitivne* (eng. *true positive*, skraćeno TP) instance su pozitivne instance koje su od strane modela prepoznate kao pozitivne



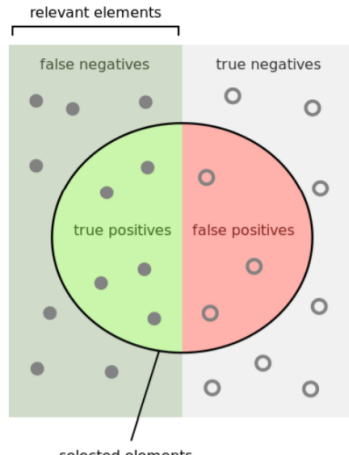
Matrica konfuzije

- *Stvarno pozitivne* (eng. *true positive*, skraćeno TP) instance su pozitivne instance koje su od strane modela prepoznate kao pozitivne
- *Stvarno negativne* (eng. *true negative*, skraćeno TN) instance su negativne instance koje su od strane modela prepoznate kao negativne



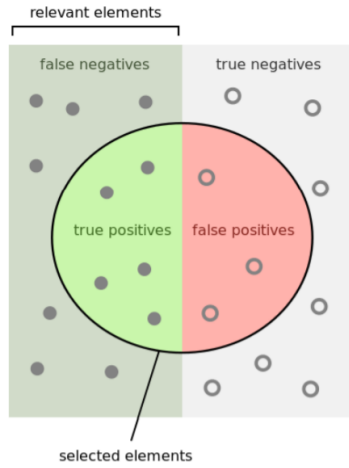
Matrica konfuzije

- *Stvarno pozitivne* (eng. *true positive*, skraćeno TP) instance su pozitivne instance koje su od strane modela prepoznate kao pozitivne
- *Stvarno negativne* (eng. *true negative*, skraćeno TN) instance su negativne instance koje su od strane modela prepoznate kao negativne
- *Lažno pozitivne* (eng. *false positive*, skraćeno FP) instance su negativne instance koje su od strane modela proglašene pozitivnim



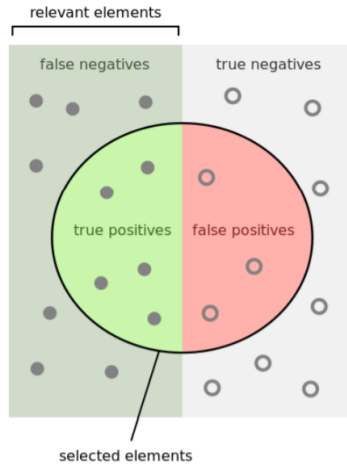
Matrica konfuzije

- *Stvarno pozitivne* (eng. *true positive*, skraćeno TP) instance su pozitivne instance koje su od strane modela prepoznate kao pozitivne
- *Stvarno negativne* (eng. *true negative*, skraćeno TN) instance su negativne instance koje su od strane modela prepoznate kao negativne
- *Lažno pozitivne* (eng. *false positive*, skraćeno FP) instance su negativne instance koje su od strane modela proglašene pozitivnim
- *Lažno negativne* (eng. *false negative*, skraćeno FN) instance su pozitivne instance koje su od strane modela proglašene negativnim



Tačnost

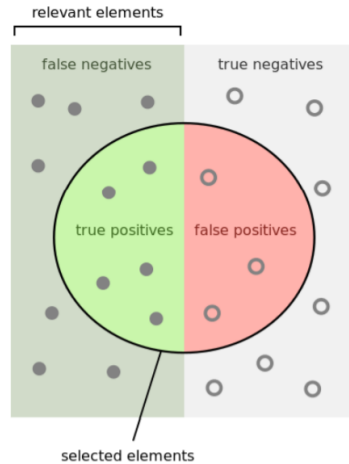
- Tačnost klasifikacije predstavlja udeo tačno klasifikovanih instanci u ukupnom broju instanci



Tačnost

- Tačnost klasifikacije predstavlja udeo tačno klasifikovanih instanci u ukupnom broju instanci
- U slučaju binarne klasifikacije, može se izraziti kao

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

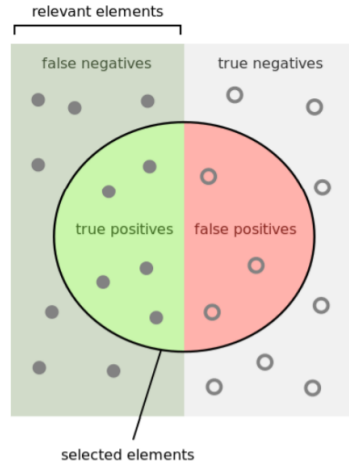


Tačnost

- Tačnost klasifikacije predstavlja udeo tačno klasifikovanih instanci u ukupnom broju instanci
- U slučaju binarne klasifikacije, može se izraziti kao

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

- Iako vrlo intuitivna, tačnost klasifikacije ne mora uvek biti pogodna mera kvaliteta

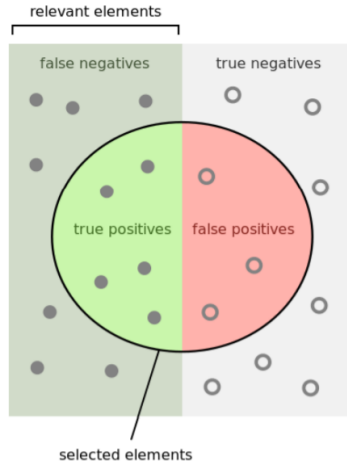


Tačnost

- Tačnost klasifikacije predstavlja udeo tačno klasifikovanih instanci u ukupnom broju instanci
- U slučaju binarne klasifikacije, može se izraziti kao

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

- Iako vrlo intuitivna, tačnost klasifikacije ne mora uvek biti pogodna mera kvaliteta
- Jedan razlog je njena neinformativnost u slučaju neizbalansiranosti klasa

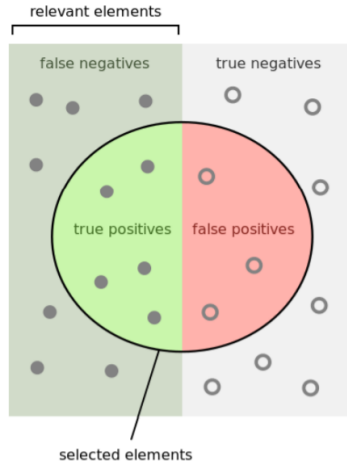


Tačnost

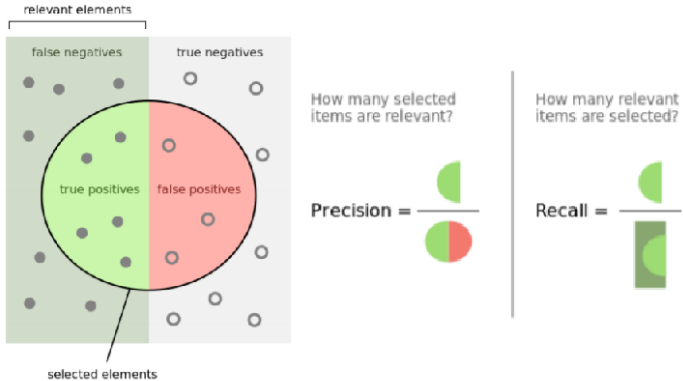
- Tačnost klasifikacije predstavlja udeo tačno klasifikovanih instanci u ukupnom broju instanci
- U slučaju binarne klasifikacije, može se izraziti kao

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

- Iako vrlo intuitivna, tačnost klasifikacije ne mora uvek biti pogodna mera kvaliteta
- Jedan razlog je njena neinformativnost u slučaju neizbalansiranosti klasa
- Na primer, u slučaju detekcije prevara sa kreditnim karticama, detekcije retkih bolesti i slično

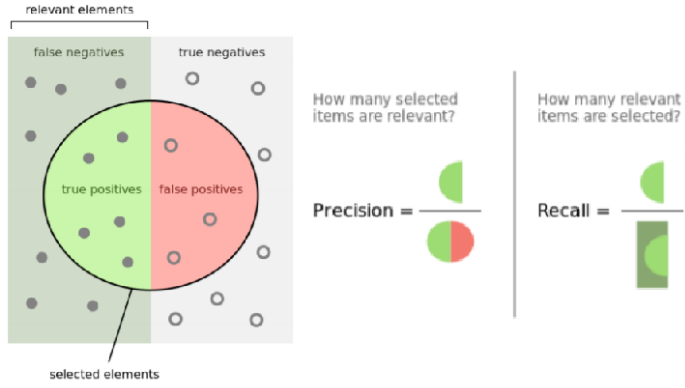


Preciznost i odziv



- Preciznost i odziv se obično koriste pri evaluaciji sistema za pronalaženje informacija, na primer prilikom pretraživanja po nekom pojmu, koliko rezultata je povezano sa tim pojmom

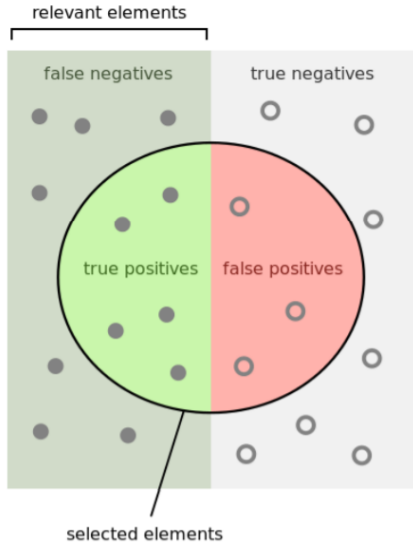
Preciznost i odziv



- Preciznost i odziv se obično koriste pri evaluaciji sistema za pronalaženje informacija, na primer prilikom pretraživanja po nekom pojmu, koliko rezultata je povezano sa tim pojmom
- Taj problem se može smatrati problemom klasifikacije – razdvajanja bitnih od nebitnih informacija pri čemu se bitne prikazuju korisniku

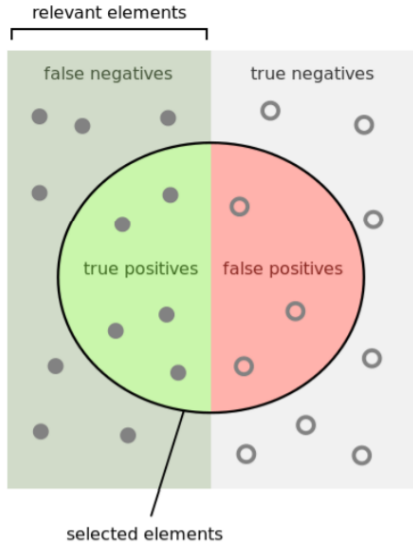
Preciznost i odziv

- Preciznost i odziv se obično koriste pri evaluaciji sistema za pronalaženje informacija, na primer prilikom pretraživanja po nekom pojmu, koliko rezultata je povezano sa tim pojmom



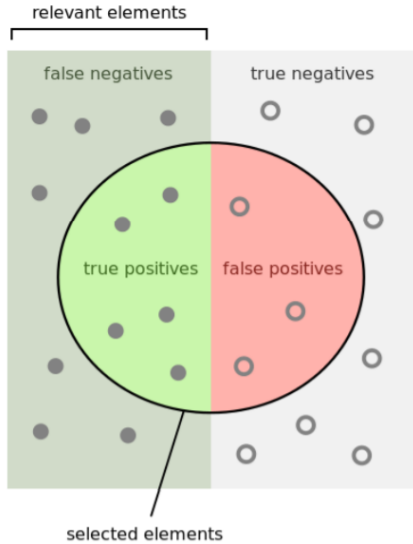
Preciznost i odziv

- Preciznost i odziv se obično koriste pri evaluaciji sistema za pronalaženje informacija, na primer prilikom pretraživanja po nekom pojmu, koliko rezultata je povezano sa tim pojmom
- Taj problem se može smatrati problemom klasifikacije – razdvajanja bitnih od nebitnih informacija pri čemu se bitne prikazuju korisniku



Preciznost i odziv

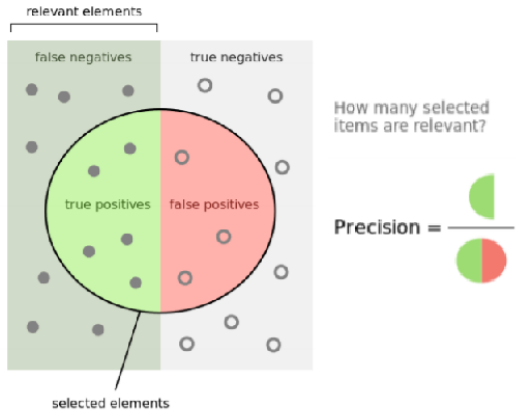
- Preciznost i odziv se obično koriste pri evaluaciji sistema za pronalaženje informacija, na primer prilikom pretraživanja po nekom pojmu, koliko rezultata je povezano sa tim pojmom
- Taj problem se može smatrati problemom klasifikacije – razdvajanja bitnih od nebitnih informacija pri čemu se bitne prikazuju korisniku
- *selected* - rezultati koje je pretraživač odabrao, *relevant* - svi rezultati u vezi sa pretraživanim pojmom



Preciznost i odziv

- Preciznost je udeo pozitivnih instanci u svim instancama koje su proglašene pozitivnim, odnosno

$$Prec = \frac{TP}{TP + FP}$$

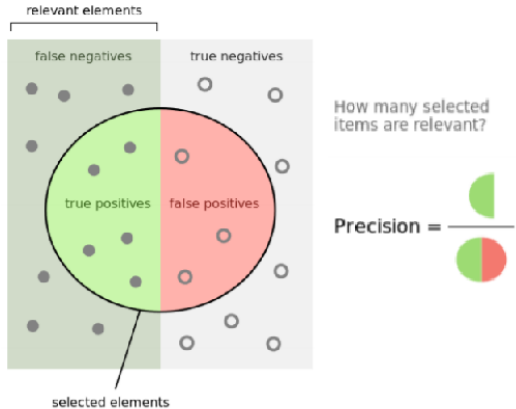


Preciznost i odziv

- Preciznost je udeo pozitivnih instanci u svim instancama koje su proglašene pozitivnim, odnosno

$$Prec = \frac{TP}{TP + FP}$$

- Odgovara na pitanje koliko puta smo bili u pravu kad smo tvrdili da je nešto relevantno

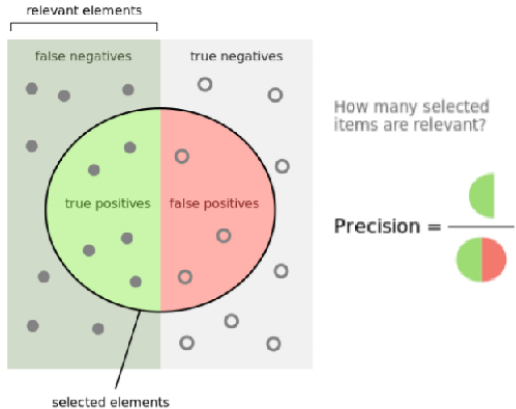


Preciznost i odziv

- Preciznost je udeo pozitivnih instanci u svim instancama koje su proglašene pozitivnim, odnosno

$$Prec = \frac{TP}{TP + FP}$$

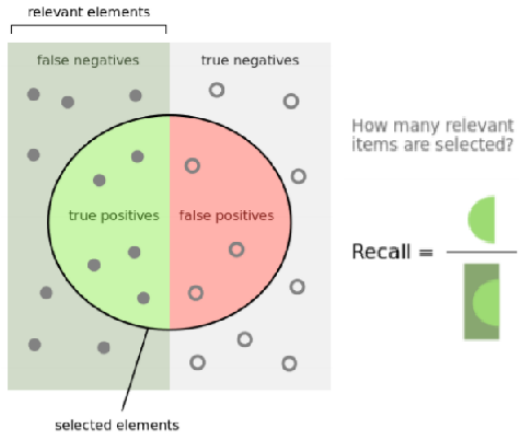
- Odgovara na pitanje koliko puta smo bili u pravu kad smo tvrdili da je nešto relevantno
- Ako je model takav da za dati pojam ne vrati nijednu stranicu (model predviđa da su svi negativni, $FP = 0$), bili smo u pravu svaki put kad smo tvrdili da je nešto relevantno, preciznost je maksimalna



Preciznost i odziv

- Odziv je udeo pronađenih pozitivnih instanci u svim pozitivnim instancama, odnosno

$$Rec = \frac{TP}{TP + FN}$$

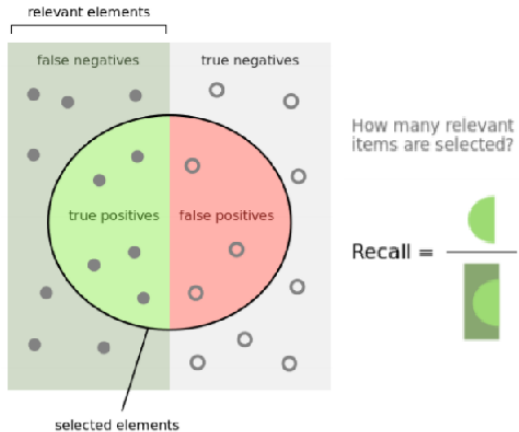


Preciznost i odziv

- Odziv je udeo pronađenih pozitivnih instanci u svim pozitivnim instancama, odnosno

$$Rec = \frac{TP}{TP + FN}$$

- Odgovara na pitanje koliko relevantnih informacija smo našli od svih relevantnih informacija

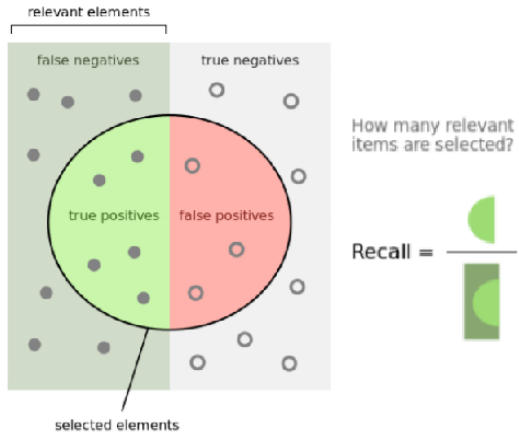


Preciznost i odziv

- Odziv je udeo pronađenih pozitivnih instanci u svim pozitivnim instancama, odnosno

$$Rec = \frac{TP}{TP + FN}$$

- Odgovara na pitanje koliko relevantnih informacija smo našli od svih relevantnih informacija
- Ako je model takav da za dati pojam vrati sve stranice (kružnica se poklapa sa celim pravougaonikom), našli smo sve relevatne informacije (a usput i neke koje nisu relevantne) i odziv je maksimalan



F_1 mera

- Svaku od ovih mera je vrlo lako maksimizovati pojedinačno ali nije lako maksimizovati ih zajedno

F_1 mera

- Svaku od ovih mera je vrlo lako maksimizovati pojedinačno ali nije lako maksimizovati ih zajedno
- Želimo da imamo istovremeno što veću preciznost i što veći odziv

F_1 mera

- Svaku od ovih mera je vrlo lako maksimizovati pojedinačno ali nije lako maksimizovati ih zajedno
- Želimo da imamo istovremeno što veću preciznost i što veći odziv
- Zbog toga ima smisla posmatrati ih samo zajedno

F_1 mera

- Svaku od ovih mera je vrlo lako maksimizovati pojedinačno ali nije lako maksimizovati ih zajedno
- Želimo da imamo istovremeno što veću preciznost i što veći odziv
- Zbog toga ima smisla posmatrati ih samo zajedno
- Način na koji se to najčešće radi je tako što se izračuna F_1 mera – njihova harmonijska sredina

$$F_1 = 2\frac{Prec \cdot Rec}{Prec + Rec}$$

F_1 mera

- Svaku od ovih mera je vrlo lako maksimizovati pojedinačno ali nije lako maksimizovati ih zajedno
- Želimo da imamo istovremeno što veću preciznost i što veći odziv
- Zbog toga ima smisla posmatrati ih samo zajedno
- Način na koji se to najčešće radi je tako što se izračuna F_1 mera – njihova harmonijska sredina

$$F_1 = 2\frac{Prec \cdot Rec}{Prec + Rec}$$

- Maksimizacijom F_1 mere istovremeno maksimizujemo i preciznost i odziv

F_1 mera

- Jedna važna mana F_1 mere je da nije simetrična u odnosu na izbor neke klase kao pozitivne

F_1 mera

- Jedna važna mana F_1 mere je da nije simetrična u odnosu na izbor neke klase kao pozitivne
- Ukoliko bismo imali iste podatke u kojima je jedna klasa proglašena za pozitivnu a druga za negativnu, F_1 mera za takvo okruženje ne bi bila ista kao F_1 mera kada bismo obrnuli klase

F_1 mera

- Jedna važna mana F_1 mere je da nije simetrična u odnosu na izbor neke klase kao pozitivne
- Ukoliko bismo imali iste podatke u kojima je jedna klasa proglašena za pozitivnu a druga za negativnu, F_1 mera za takvo okruženje ne bi bila ista kao F_1 mera kada bismo obrnuli klase
- Ova mana se može prevazići tako što bismo uzeli prosek dobijenih F_1 mera

F_1 mera

- Jedna važna mana F_1 mere je da nije simetrična u odnosu na izbor neke klase kao pozitivne
- Ukoliko bismo imali iste podatke u kojima je jedna klasa proglašena za pozitivnu a druga za negativnu, F_1 mera za takvo okruženje ne bi bila ista kao F_1 mera kada bismo obrnuli klase
- Ova mana se može prevazići tako što bismo uzeli prosek dobijenih F_1 mera
- U praktičnim primenama obično znamo koja klasa je pozitivna pa nema potrebe za obrtanjem

Površina ispod ROC krive

- Jedna mera koja se češće koristi u slučaju binarne klasifikacije sa neizbalanasiranim klasama je površina ispod ROC krive

Površina ispod ROC krive

- Jedna mera koja se češće koristi u slučaju binarne klasifikacije sa neizbalanasiranim klasama je površina ispod ROC krive
- Nećemo definisati šta je ROC kriva (zastareli način pokazivanja ove mere)

Površina ispod ROC krive

- Jedna mera koja se češće koristi u slučaju binarne klasifikacije sa neizbalanasiranim klasama je površina ispod ROC krive
- Nećemo definisati šta je ROC kriva (zastareli način pokazivanja ove mere)
- Intuitivno objašnjenje u terminima verovatnoće

Površina ispod ROC krive

- Jedna mera koja se češće koristi u slučaju binarne klasifikacije sa neizbalanasiranim klasama je površina ispod ROC krive
- Nećemo definisati šta je ROC kriva (zastareli način pokazivanja ove mere)
- Intuitivno objašnjenje u terminima verovatnoće
- Pretpostavlja se da klasifikator zasnovan na nekoj vrsti skora (logistička regresija, metod potpornih vektora) tako što svakoj instanci dodeljuje neki skor

Površina ispod ROC krive

- Jedna mera koja se češće koristi u slučaju binarne klasifikacije sa neizbalanasiranim klasama je površina ispod ROC krive
- Nećemo definisati šta je ROC kriva (zastareli način pokazivanja ove mere)
- Intuitivno objašnjenje u terminima verovatnoće
- Pretpostavlja se da klasifikator zasnovan na nekoj vrsti skora (logistička regresija, metod potpornih vektora) tako što svakoj instanci dodeljuje neki skor
- Ovi skorovi nam omogućavaju da instance sortiramo na osnovu tog skora

Površina ispod ROC krive

- Kako bi izgledao sortirani niz u idealnom slučaju?

Površina ispod ROC krive

- Kako bi izgledao sortirani niz u idealnom slučaju?
- Imali bismo prvo sve instance jedne klase pa onda sve instance druge klase

Površina ispod ROC krive

- Kako bi izgledao sortirani niz u idealnom slučaju?
- Imali bismo prvo sve instance jedne klase pa onda sve instance druge klase
- Međutim, nije realistično da imamo idealno razdvajanje

Površina ispod ROC krive

- Kako bi izgledao sortirani niz u idealnom slučaju?
- Imali bismo prvo sve instance jedne klase pa onda sve instance druge klase
- Međutim, nije realistično da imamo idealno razdvajanje
- Ipak, ako je većina prve klase sa jedne strane prave a većina druge klase sa druge strane prave, možemo reći da se na osnovu skora može klasifikovati i da je klasifikator dobar

Površina ispod ROC krive

- Kako se može interpretirati AUC?

Površina ispod ROC krive

- Kako se može interpretirati AUC?
- Ako bismo iz takvog niza nasumično odabrali jednu instancu iz negativne klase i jednu instancu iz pozitivne klase, šta očekujemo za njihove skorove?

Površina ispod ROC krive

- Kako se može interpretirati AUC?
- Ako bismo iz takvog niza nasumično odabrali jednu instancu iz negativne klase i jednu instancu iz pozitivne klase, šta očekujemo za njihove skorove?
- Očekujemo da je skor instance iz negativne klase manji od skora instance iz pozitivne klase

Površina ispod ROC krive

- Kako se može interpretirati AUC?
- Ako bismo iz takvog niza nasumično odabrali jednu instancu iz negativne klase i jednu instancu iz pozitivne klase, šta očekujemo za njihove skorove?
- Očekujemo da je skor instance iz negativne klase manji od skora instance iz pozitivne klase
- AUC je verovatnoća da je stvarno tako

Površina ispod ROC krive

- Intuitivna mera

Površina ispod ROC krive

- Intuitivna mera
- Može se lako izračunati

Površina ispod ROC krive

- Intuitivna mera
- Može se lako izračunati
- Može da uzme vrednosti između 0 i 1, pri č emu je 1 najbolja vrednost (svi parovi ispunjavaju željenu relaciju), a vrednost 0.5 označava da klasifikator ništa nije naučio već da klasifikuje nasumično

Površina ispod ROC krive

- Intuitivna mera
- Može se lako izračunati
- Može da uzme vrednosti između 0 i 1, pri č emu je 1 najbolja vrednost (svi parovi ispunjavaju željenu relaciju), a vrednost 0.5 označava da klasifikator ništa nije naučio već da klasifikuje nasumično
- Ova mera nije osetljiva na neizbalansiranost klasa jer ocenjuje odnos između parova instanci različitih klasa pa instance iz manje zastupljene klase mogu učestvovati više puta u nasumičnom izboru

Pregled

Mere kvaliteta regresionih modela

- *srednjekvadratna greška* (eng. *mean square error*) i njen koren (eng. *root mean square error*))

Mere kvaliteta regresionih modela

- *srednjekvadratna greška* (eng. *mean square error*) i njen koren (eng. *root mean square error*))
- *srednja relativna greška* izražena u procentima (eng. *mean absolute percentage error*)

Mere kvaliteta regresionih modela

- *srednjekvadratna greška* (eng. *mean square error*) i njen koren (eng. *root mean square error*))
- *srednja relativna greška* izražena u procentima (eng. *mean absolute percentage error*)
- *koeficijent determinacije*, poznatiji pod oznakom R^2

Srednjekvadratna greška i njen koren

- Funkcija greške koju kod regresije najčešće optimizujemo

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

Srednjekvadratna greška i njen koren

- Funkcija greške koju kod regresije najčešće optimizujemo

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

- Kvadrat će učiniti da je jedna greška koja je malo veća (što je slučaj kod odudarajućih podataka) postane mnogo veća

Srednjekvadratna greška i njen koren

- Funkcija greške koju kod regresije najčešće optimizujemo

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

- Kvadrat će učiniti da je jedna greška koja je malo veća (što je slučaj kod odudarajućih podataka) postane mnogo veća
- Ako primenimo koren (RMSE), grešku ćemo izraziti na istoj skali na kojoj je i ciljna promenljiva

Srednjekvadratna greška i njen koren

- Funkcija greške koju kod regresije najčešće optimizujemo

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

- Kvadrat će učiniti da je jedna greška koja je malo veća (što je slučaj kod odudarajućih podataka) postane mnogo veća
- Ako primenimo koren (RMSE), grešku ćemo izraziti na istoj skali na kojoj je i ciljna promenljiva
  - To se često radi u praksi ali treba da imamo na umu da je moguće da su u zbiru koju smo korenovali potencijalno veliki deo imali odudarajući podaci

Srednjekvadratna greška i njen koren

- Funkcija greške koju kod regresije najčešće optimizujemo

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

- Kvadrat će učiniti da je jedna greška koja je malo veća (što je slučaj kod odudarajućih podataka) postane mnogo veća
- Ako primenimo koren (RMSE), grešku ćemo izraziti na istoj skali na kojoj je i ciljna promenljiva
  - To se često radi u praksi ali treba da imamo na umu da je moguće da su u zbiru koju smo korenovali potencijalno veliki deo imali odudarajući podaci
- RMSE se izražava u jedinicama koje su nama relevantne i možemo ih lako interpretirati

Srednja relativna greška

- Zbog jakog uticaja odudarajućih podataka na MSE i RMSE i zbog potrebe da se greška nekada izrazi u odnosu na vrednost ciljne promenljive, koristi se i srednja relativna greška, najčešće izražena u procentima, koja se definiše kao

$$MAPE = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{y_i - f(x_i)}{y_i} \right|$$

Koeficijent determinacije (R^2)

- Hoćemo da merimo koliki je napredak ostvaren učenjem u odnosu na neki trivijalan prediktivni metod koji bi nam bio dostupan i bez učenja

Koeficijent determinacije (R^2)

- Hoćemo da merimo koliki je napredak ostvaren učenjem u odnosu na neki trivijalan prediktivni metod koji bi nam bio dostupan i bez učenja
- Kakav bi to metod mogao biti?

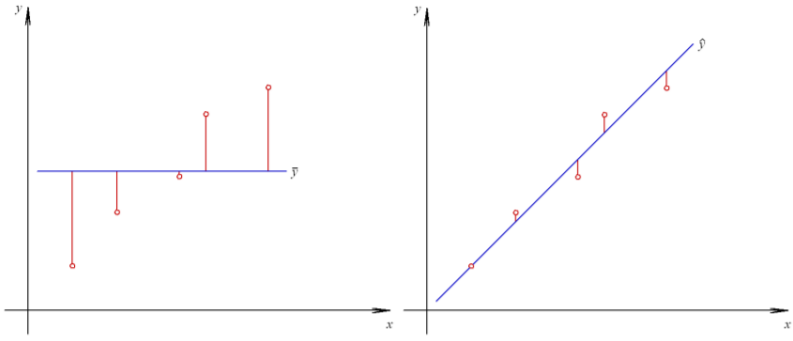
Koeficijent determinacije (R^2)

- Hoćemo da merimo koliki je napredak ostvaren učenjem u odnosu na neki trivijalan prediktivni metod koji bi nam bio dostupan i bez učenja
- Kakav bi to metod mogao biti?
  - Slučajno pogađanje

Koeficijent determinacije (R^2)

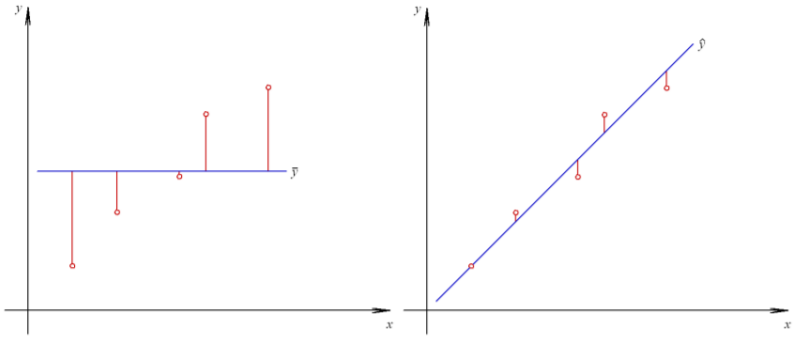
- Hoćemo da merimo koliki je napredak ostvaren učenjem u odnosu na neki trivijalan prediktivni metod koji bi nam bio dostupan i bez učenja
- Kakav bi to metod mogao biti?
  - Slučajno pogađanje
  - Uzimanje prosečne vrednosti ciljne promenljive na nekom uzorku i za svaku instancu predviđati takvu vrednost

Koeficijent determinacije (R^2)



- Posmatrajmo MSE za oba ova modela

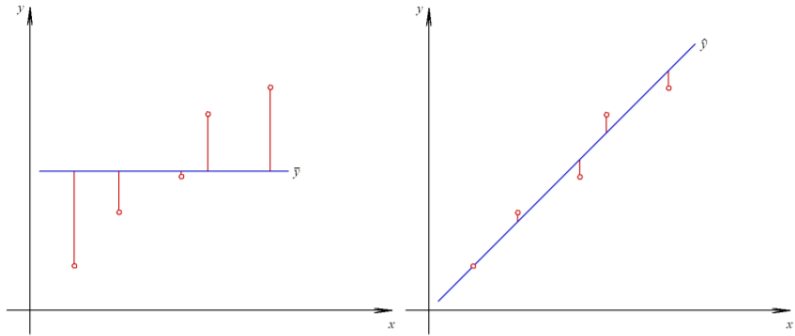
Koeficijent determinacije (R^2)



- Posmatrajmo MSE za oba ova modela
- Setimo se formule za MSE

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

Koeficijent determinacije (R^2)

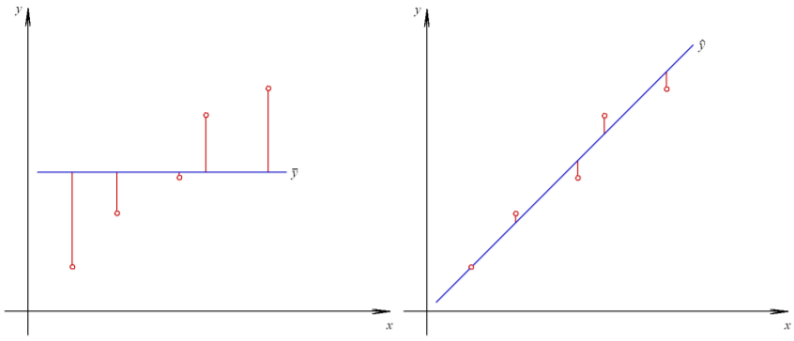


- Posmatrajmo MSE za oba ova modela
- Setimo se formule za MSE

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

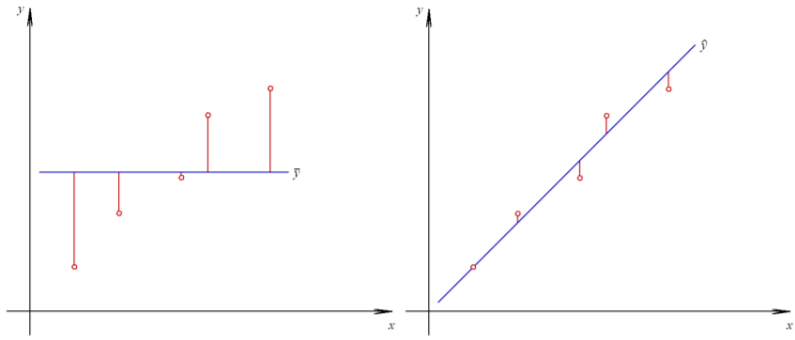
- Ako u formuli za MSE umesto $f(x_i)$ upišemo prosek, šta ćemo dobiti?

Koeficijent determinacije (R^2)



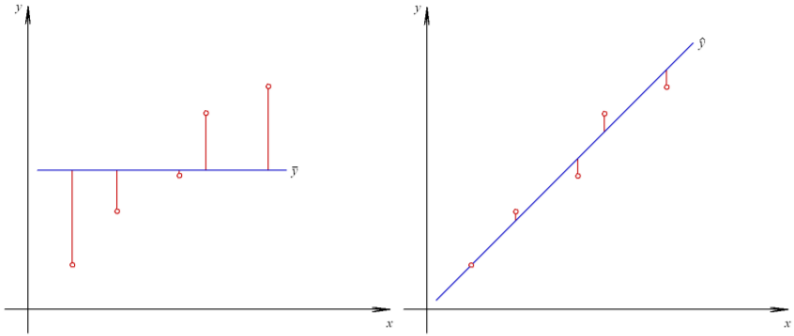
- Ako u formuli za MSE umesto $f(x_i)$ upišemo prosek, šta ćemo dobiti?

Koeficijent determinacije (R^2)



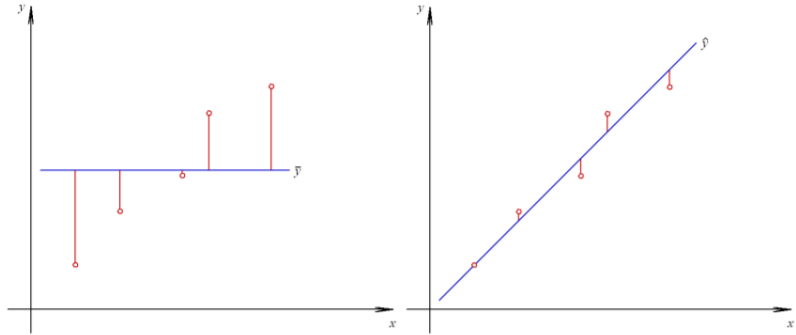
- Ako u formuli za MSE umesto $f(x_i)$ upišemo prosek, šta ćemo dobiti?
- Dobijamo varijansu po y i to ćemo označiti sa var$[y]$

Koeficijent determinacije (R^2)



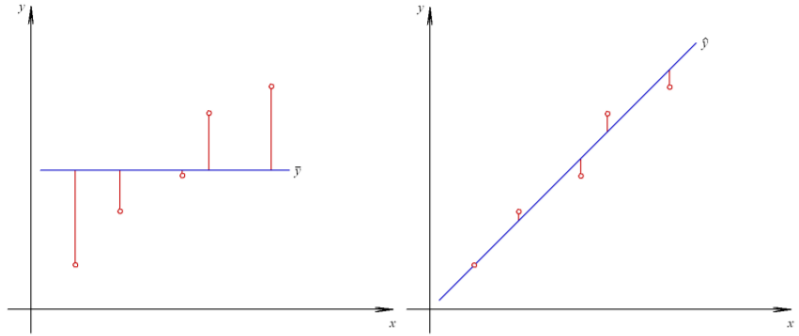
- Neka je sa MSE označena srednjekvadratna greška desnog (pametnog) modela a var[y] nam predstavlja srednjekvadratnu grešku levog (konstantnog) modela

Koeficijent determinacije (R^2)



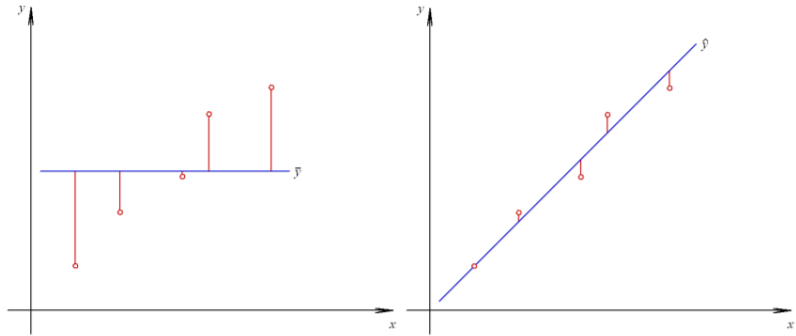
- Neka je sa MSE označena srednjekvadratna greška desnog (pametnog) modela a var[y] nam predstavlja srednjekvadratnu grešku levog (konstantnog) modela
- Očekujemo da će pametni model praviti manju grešku nego konstantni model

Koeficijent determinacije (R^2)



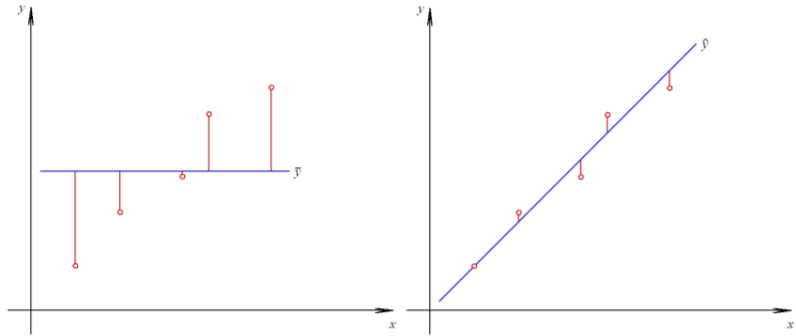
- Neka je sa MSE označena srednjekvadratna greška desnog (pametnog) modela a var[y] nam predstavlja srednjekvadratnu grešku levog (konstantnog) modela
- Očekujemo da će pametni model praviti manju grešku nego konstantni model
- Zbog toga ima smisla posmatrati grešku pametnog modela u odnosu na grešku konstantnog modela: $MSE/\text{var}[y]$

Koeficijent determinacije (R^2)



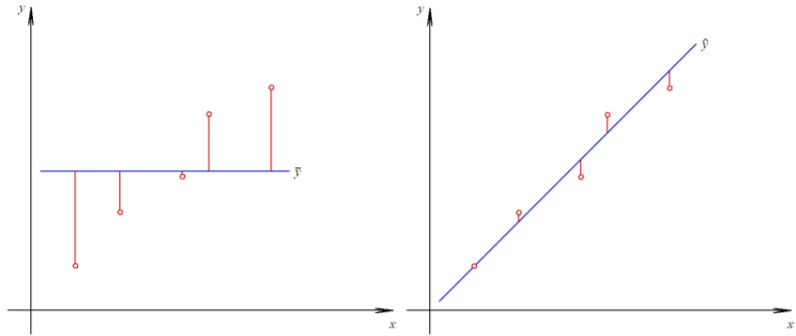
- Posmatrajmo veličinu $\frac{MSE}{\text{var}[y]}$

Koeficijent determinacije (R^2)



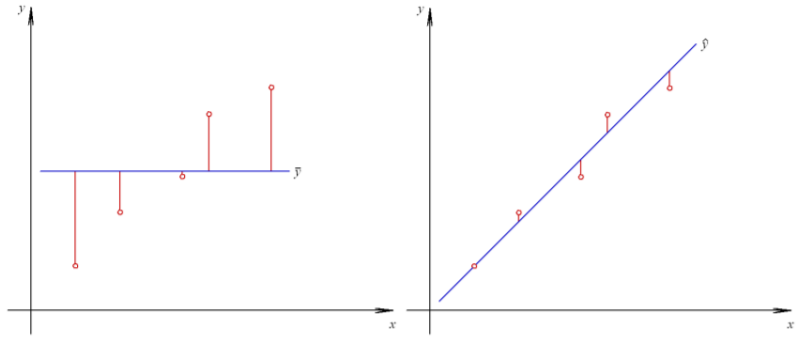
- Posmatrajmo veličinu $\frac{MSE}{\text{var}[y]}$
- Pametni model bolji nego konstantni model ali i dalje pravi neku grešku

Koeficijent determinacije (R^2)



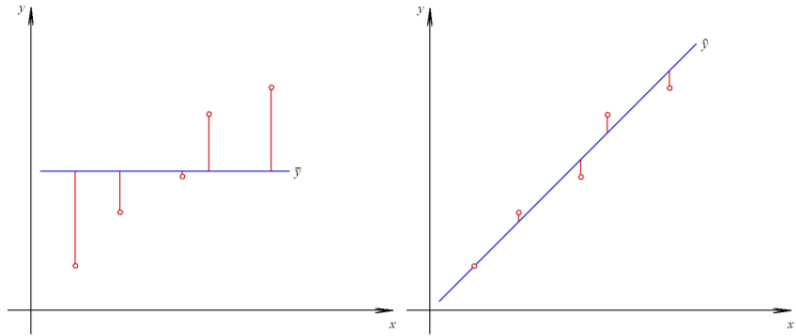
- Posmatrajmo veličinu $\frac{MSE}{\text{var}[y]}$
- Pametni model bolji nego konstantni model ali i dalje pravi neku grešku
- Greška pametnog modela se u ovom kontekstu naziva *preostalom greškom* a njen udeo u odnosu na grešku konstantnog modela *udelom preostale greške*

Koeficijent determinacije (R^2)



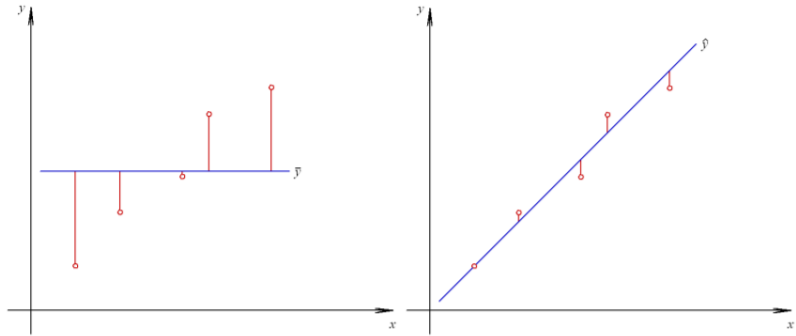
- Posmatrajmo veličinu $1 - \frac{MSE}{\text{var}[y]}$

Koeficijent determinacije (R^2)



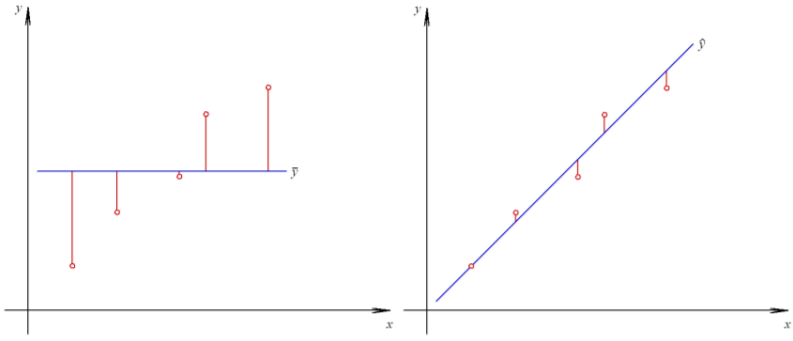
- Posmatrajmo veličinu $1 - \frac{MSE}{\text{var}[y]}$
- Kada od 1 oduzmemo udeo preostale greške, ono što dobijamo je *greška koja je eliminisana*

Koeficijent determinacije (R^2)



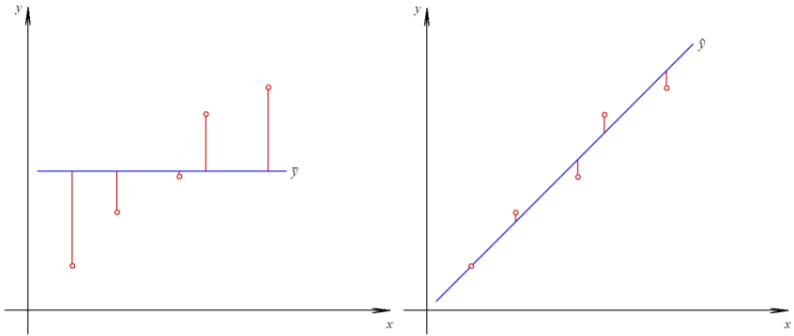
- Posmatrajmo veličinu $1 - \frac{MSE}{\text{var}[y]}$
- Kada od 1 oduzmemo udeo preostale greške, ono što dobijamo je *greška koja je eliminisana*
- Ova veličina se često naziva *udelom objašnjene varijanse*

Koeficijent determinacije (R^2)



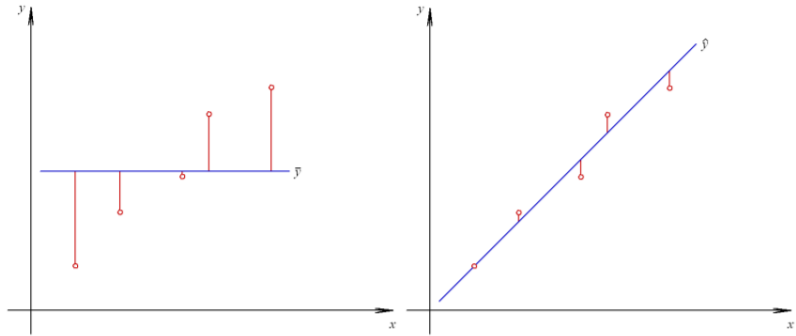
- Objasnimo poreklo pojma *udeo objašnjene varijanse*

Koeficijent determinacije (R^2)



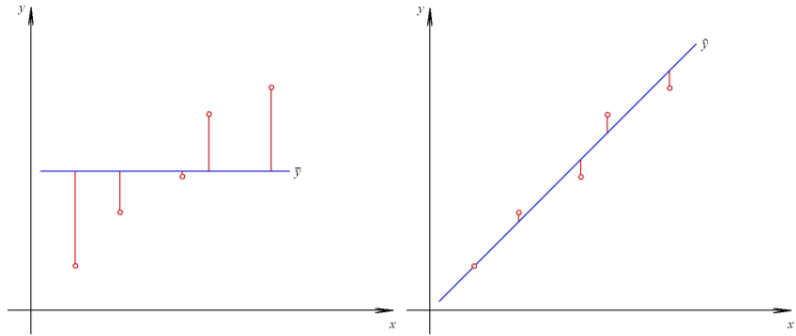
- Objasnimo poreklo pojma *udeo objašnjene varijanse*
- Podsetimo se da modele mašinskog učenja, nalik fizičkim modelima, možemo razumeti kao pokušaj da se objasni variranje neke promenljive na osnovu promena u drugim promenljivim

Koeficijent determinacije (R^2)



- Objasnimo poreklo pojma *udeo objašnjene varijanse*
- Podsetimo se da modele mašinskog učenja, nalik fizičkim modelima, možemo razumeti kao pokušaj da se objasni variranje neke promenljive na osnovu promena u drugim promenljivim
- U fizici, ukoliko dođe do promene temperature gasa, doći će i do promene pritiska

Koeficijent determinacije (R^2)



- Objasnimo poreklo pojma *udeo objašnjene varijanse*
- Podsetimo se da modele mašinskog učenja, nalik fizičkim modelima, možemo razumeti kao pokušaj da se objasni variranje neke promenljive na osnovu promena u drugim promenljivim
- U fizici, ukoliko dođe do promene temperature gasa, doći će i do promene pritiska
- Drugim rečima, promena pritiska je *objašnjena* promenom temperature

Koeficijent determinacije (R^2) - prednosti

$$R^2 = 1 - \frac{MSE}{\text{var}[y]}$$

- Normiranost - R^2 ne može biti veće od 1 dok MSE može biti proizvoljno veliko

Koeficijent determinacije (R^2) - prednosti

$$R^2 = 1 - \frac{MSE}{\text{var}[y]}$$

- Normiranost - R^2 ne može biti veće od 1 dok MSE može biti proizvoljno veliko
- Što smo bliže jedinici, to je model bolji

Koeficijent determinacije (R^2) - prednosti

$$R^2 = 1 - \frac{MSE}{\text{var}[y]}$$

- Normiranost - R^2 ne može biti veće od 1 dok MSE može biti proizvoljno veliko
- Što smo bliže jedinici, to je model bolji
- Ako smo blizu nuli, to znači da nismo bolji od proseka

Koeficijent determinacije (R^2) - prednosti

$$R^2 = 1 - \frac{MSE}{\text{var}[y]}$$

- Normiranost - R^2 ne može biti veće od 1 dok MSE može biti proizvoljno veliko
- Što smo bliže jedinici, to je model bolji
- Ako smo blizu nuli, to znači da nismo bolji od proseka
- Može li R^2 biti negativan?

Koeficijent determinacije (R^2) - prednosti

$$R^2 = 1 - \frac{MSE}{\text{var}[y]}$$

- Normiranost - R^2 ne može biti veće od 1 dok MSE može biti proizvoljno veliko
- Što smo bliže jedinici, to je model bolji
- Ako smo blizu nuli, to znači da nismo bolji od proseka
- Može li R^2 biti negativan?
  - Može

Koeficijent determinacije (R^2) - prednosti

$$R^2 = 1 - \frac{MSE}{\text{var}[y]}$$

- Normiranost - R^2 ne može biti veće od 1 dok MSE može biti proizvoljno veliko
- Što smo bliže jedinici, to je model bolji
- Ako smo blizu nuli, to znači da nismo bolji od proseka
- Može li R^2 biti negativan?
  - Može
  - R^2 računamo na podacima koji nisu skup instanci trening skupa, već izdvojeni podaci koji prethodno nisu korišceni u treniranju i pripadaju takozvanom test skupu

Koeficijent determinacije (R^2) - prednosti

$$R^2 = 1 - \frac{MSE}{\text{var}[y]}$$

- Normiranost - R^2 ne može biti veće od 1 dok MSE može biti proizvoljno veliko
- Što smo bliže jedinici, to je model bolji
- Ako smo blizu nuli, to znači da nismo bolji od proseka
- Može li R^2 biti negativan?
  - Može
  - R^2 računamo na podacima koji nisu skup instanci trening skupa, već izdvojeni podaci koji prethodno nisu korišćeni u treniranju i pripadaju takozvanom test skupu
  - Ako ti podaci iz test skupa imaju mnogo drugačiji prosek, onda MSE može biti veće od var$[y]$ i R^2 može biti negativno

Koeficijent determinacije (R^2) - prednosti

$$R^2 = 1 - \frac{MSE}{\text{var}[y]}$$

- Normiranost - R^2 ne može biti veće od 1 dok MSE može biti proizvoljno veliko
- Što smo bliže jedinici, to je model bolji
- Ako smo blizu nuli, to znači da nismo bolji od proseka
- Može li R^2 biti negativan?
  - Može
  - R^2 računamo na podacima koji nisu skup instanci trening skupa, već izdvojeni podaci koji prethodno nisu korišćeni u treniranju i pripadaju takozvanom test skupu
  - Ako ti podaci iz test skupa imaju mnogo drugačiji prosek, onda MSE može biti veće od var$[y]$ i R^2 može biti negativno
  - To je pokazatelj da podela na trening i test skup nije bila najbolja moguća i da su u test skupu završili sasvim različite instance u odnosu na one iz trening skupa

Pregled

Tehnike evaluacije i izbora modela

- **Podaci korišćeni u evaluaciji modela ni na koji način ne smeju biti korišćeni prilikom njegovog obučavanja**

Pregled

Izbor modela u slučaju nekonfigurabilnog algoritma učenja

- U nedostatku konfigurabilnosti, ne postoje ni mogućnosti izbora, osim po pitanju podataka na kojima se obučavanje vrši

Izbor modela u slučaju nekonfigurabilnog algoritma učenja

- U nedostatku konfigurabilnosti, ne postoje ni mogućnosti izbora, osim po pitanju podataka na kojima se obučavanje vrši
- Svi podaci su vredni i model koji želimo da koristimo u budućnosti treba obučavati na svim dostupnim podacima

Izbor modela u slučaju nekonfigurabilnog algoritma učenja

- U nedostatku konfigurabilnosti, ne postoje ni mogućnosti izbora, osim po pitanju podataka na kojima se obučavanje vrši
- Svi podaci su vredni i model koji želimo da koristimo u budućnosti treba obučavati na svim dostupnim podacima
- Model obučen na svim dostupnim podacima u nastavku ćemo označavati sa M

Evaluacija modela u slučaju nekonfigurabilnog algoritma učenja

- Kako evaluirati model M, a da se u evaluaciji ne koriste podaci na kojima je model obučavan, kad je model obučavan na svim dostupnim podacima?

Evaluacija modela u slučaju nekonfigurabilnog algoritma učenja

- Kako evaluirati model M, a da se u evaluaciji ne koriste podaci na kojima je model obučavan, kad je model obučavan na svim dostupnim podacima?
- Da ne obučavamo model M na svim podacima, već samo na delu podataka, kako bi neki preostali za evaluaciju?

Evaluacija modela u slučaju nekonfigurabilnog algoritma učenja

- Kako evaluirati model M, a da se u evaluaciji ne koriste podaci na kojima je model obučavan, kad je model obučavan na svim dostupnim podacima?
- Da ne obučavamo model M na svim podacima, već samo na delu podataka, kako bi neki preostali za evaluaciju?
- Manje podataka znači manju pouzdanost

Evaluacija modela u slučaju nekonfigurabilnog algoritma učenja

- Kako evaluirati model M, a da se u evaluaciji ne koriste podaci na kojima je model obučavan, kad je model obučavan na svim dostupnim podacima?
- Da ne obučavamo model M na svim podacima, već samo na delu podataka, kako bi neki preostali za evaluaciju?
- Manje podataka znači manju pouzdanost
- Kompromis:

Evaluacija modela u slučaju nekonfigurabilnog algoritma učenja

- Kako evaluirati model M, a da se u evaluaciji ne koriste podaci na kojima je model obučavan, kad je model obučavan na svim dostupnim podacima?
- Da ne obučavamo model M na svim podacima, već samo na delu podataka, kako bi neki preostali za evaluaciju?
- Manje podataka znači manju pouzdanost
- Kompromis:
  - Obučavaćemo model M na *svim* podacima

Evaluacija modela u slučaju nekonfigurabilnog algoritma učenja

- Kako evaluirati model M, a da se u evaluaciji ne koriste podaci na kojima je model obučavan, kad je model obučavan na svim dostupnim podacima?
- Da ne obučavamo model M na svim podacima, već samo na delu podataka, kako bi neki preostali za evaluaciju?
- Manje podataka znači manju pouzdanost
- Kompromis:
  - Obučavaćemo model M na *svim* podacima
  - Pošto ćemo sve podatke iskoristiti za obučavanje, neće preostati podaci koji se mogu iskoristiti za evaluaciju

Evaluacija modela u slučaju nekonfigurabilnog algoritma učenja

- Kako evaluirati model M, a da se u evaluaciji ne koriste podaci na kojima je model obučavan, kad je model obučavan na svim dostupnim podacima?
- Da ne obučavamo model M na svim podacima, već samo na delu podataka, kako bi neki preostali za evaluaciju?
- Manje podataka znači manju pouzdanost
- Kompromis:
  - Obučavaćemo model M na *svim* podacima
  - Pošto ćemo sve podatke iskoristiti za obučavanje, neće preostati podaci koji se mogu iskoristiti za evaluaciju
  - Zato nećemo direktno evaluirati model M nego ćemo evaluirati njegovu aproksimaciju, model M' koji će biti obučavan na *većini* podataka

Evaluacija modela u slučaju nekonfigurabilnog algoritma učenja

- Kako evaluirati model M, a da se u evaluaciji ne koriste podaci na kojima je model obučavan, kad je model obučavan na svim dostupnim podacima?
- Da ne obučavamo model M na svim podacima, već samo na delu podataka, kako bi neki preostali za evaluaciju?
- Manje podataka znači manju pouzdanost
- Kompromis:
  - Obučavaćemo model M na *svim* podacima
  - Pošto ćemo sve podatke iskoristiti za obučavanje, neće preostati podaci koji se mogu iskoristiti za evaluaciju
  - Zato nećemo direktno evaluirati model M nego ćemo evaluirati njegovu aproksimaciju, model M' koji će biti obučavan na *većini* podataka
  - Evaluacijom modela M' aproksimativno se evaluira model M

Evaluacija pomoću skupa za testiranje

- Ukupni podaci se dele na dva skupa – skup za obučavanje i skup za testiranje

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

Tabela: Primer podele podataka na skup za obučavanje (plavo) i skup za testiranje (crveno).

Evaluacija pomoću skupa za testiranje

- Ukupni podaci se dele na dva skupa – skup za obučavanje i skup za testiranje
- Na skupu za obučavanje se određuje model M'

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

Tabela: Primer podele podataka na skup za obučavanje (plavo) i skup za testiranje (crveno).

Evaluacija pomoću skupa za testiranje

- Ukupni podaci se dele na dva skupa – skup za obučavanje i skup za testiranje
- Na skupu za obučavanje se određuje model M'
- Model M' se potom primenjuje na skup za testiranje, čime se dobijaju njegova predviđanja

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

Tabela: Primer podele podataka na skup za obučavanje (plavo) i skup za testiranje (crveno).

Evaluacija pomoću skupa za testiranje

- Ukupni podaci se dele na dva skupa – skup za obučavanje i skup za testiranje
- Na skupu za obučavanje se određuje model M'
- Model M' se potom primenjuje na skup za testiranje, čime se dobijaju njegova predviđanja
- Dobijena predviđanja se nekom merom kvaliteta mogu oceniti u odnosu na tačne vrednosti ciljne promenljive

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

Tabela: Primer podele podataka na skup za obučavanje (plavo) i skup za testiranje (crveno).

Mane tehnike evaluacije pomoću skupa za testiranje

- Od načina na koji je izvršena podela na skup za obučavanje i skup za testiranje zavisiće i ocena kvaliteta.

Mane tehnike evaluacije pomoću skupa za testiranje

- Od načina na koji je izvršena podela na skup za obučavanje i skup za testiranje zavisiće i ocena kvaliteta.
- Greška ocene može biti vrlo velika ako se skup za testiranje pristrasno izabere

Mane tehnike evaluacije pomoću skupa za testiranje

- Od načina na koji je izvršena podela na skup za obučavanje i skup za testiranje zavisiće i ocena kvaliteta.
- Greška ocene može biti vrlo velika ako se skup za testiranje pristrasno izabere
- Odgovor na neke od ovih problema daje tehnika *K-slojne unakrsne validacije* (eng. *K-fold cross-validation*)

K-slojna unakrsna validacija

- Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova, takozvanih *slojeva* (eng. *folds*) $\{S_1, \ldots, S_K\}$

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

Tabela: Prikaz podele podataka pri unakrsnoj validaciji. Različite boje predstavljaju različite slojeve.

K-slojna unakrsna validacija

- Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova, takozvanih *slojeva* (eng. *folds*) $\{S_1, \ldots, S_K\}$
- Za $i = 1, \ldots, K$

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

Tabela: Prikaz podele podataka pri unakrsnoj validaciji. Različite boje predstavljaju različite slojeve.

K-slojna unakrsna validacija

- Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova, takozvanih *slojeva* (eng. *folds*) $\{S_1, \ldots, S_K\}$
- Za $i = 1, \ldots, K$
  - Obučiti model na podacima $\mathcal{D} \setminus S_i$

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

Tabela: Prikaz podele podataka pri unakrsnoj validaciji. Različite boje predstavljaju različite slojeve.

K-slojna unakrsna validacija

- Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova, takozvanih *slojeva* (eng. *folds*) $\{S_1, \ldots, S_K\}$
- Za $i = 1, \ldots, K$
  - Obučiti model na podacima $\mathcal{D} \setminus S_i$
  - Izvršiti predviđanja dobijenim modelom na sloju S_i

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

Tabela: Prikaz podele podataka pri unakrsnoj validaciji. Različite boje predstavljaju različite slojeve.

K-slojna unakrsna validacija

- Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova, takozvanih *slojeva* (eng. *folds*) $\{S_1, \ldots, S_K\}$
- Za $i = 1, \ldots, K$
  - Obučiti model na podacima $\mathcal{D} \setminus S_i$
  - Izvršiti predviđanja dobijenim modelom na sloju S_i
- Izračunati ocenu kvaliteta na osnovu svih predviđanja na celom skupu $\mathcal{D}$

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

Tabela: Prikaz podele podataka pri unakrsnoj validaciji. Različite boje predstavljaju različite slojeve.

K-slojna unakrsna validacija

- Ocena kvaliteta se računa tek nakon što su izračunata sva predviđanja

K-slojna unakrsna validacija

- Ocena kvaliteta se računa tek nakon što su izračunata sva predviđanja
  - Nikako uprosečavanjem ocena kvaliteta na pojedinačnim slojevima!

K-slojna unakrsna validacija

- Ocena kvaliteta se računa tek nakon što su izračunata sva predviđanja
  - Nikako uprosečavanjem ocena kvaliteta na pojedinačnim slojevima!
- Biće kreirano k modela $M'_1, \ldots, M'_K$. Koji od njih biramo za buduće korišćenje? Da li ih nekako kombinujemo?

K-slojna unakrsna validacija

- Ocena kvaliteta se računa tek nakon što su izračunata sva predviđanja
  - Nikako uprosečavanjem ocena kvaliteta na pojedinačnim slojevima!
- Biće kreirano k modela $M'_1, \ldots, M'_K$. Koji od njih biramo za buduće korišćenje? Da li ih nekako kombinujemo?
  - Nijedan i ne kombinujemo ih nikako

K-slojna unakrsna validacija

- Ocena kvaliteta se računa tek nakon što su izračunata sva predviđanja
  - Nikako uprosečavanjem ocena kvaliteta na pojedinačnim slojevima!
- Biće kreirano k modela $M'_1, \ldots, M'_K$. Koji od njih biramo za buduće korišćenje? Da li ih nekako kombinujemo?
  - Nijedan i ne kombinujemo ih nikako
  - Ubuduće koristimo model M treniran na svim raspoloživim podacima a modeli $M'_1, \ldots, M'_K$ su služili za aproksimaciju njegove evaluacije

K-slojna unakrsna validacija

- Prednosti

K-slojna unakrsna validacija

- Prednosti
  - Ocena greške se računa na većoj količini podataka pa je njena varijansa manja

K-slojna unakrsna validacija

- Prednosti
  - Ocena greške se računa na većoj količini podataka pa je njena varijansa manja
- Mane

K-slojna unakrsna validacija

- Prednosti
  - Ocena greške se računa na većoj količini podataka pa je njena varijansa manja
- Mane
  - Model se obučava K puta

K-slojna unakrsna validacija

- Prednosti
  - Ocena greške se računa na većoj količini podataka pa je njena varijansa manja
- Mane
  - Model se obučava K puta
  - Izbor broja K (obično se koriste $K = 5$ ili $K = 10$, kod manjeg broja podataka čak $K = N$)

K-slojna unakrsna validacija

- Prednosti
  - Ocena greške se računa na većoj količini podataka pa je njena varijansa manja
- Mane
  - Model se obučava K puta
  - Izbor broja K (obično se koriste $K = 5$ ili $K = 10$, kod manjeg broja podataka čak $K = N$)
  - Nije rešen problem pristrasnog izbora podskupova

Pregled

Konfigurabilni algoritmi

- Algoritam učenja može biti konfigurabilan po različitim aspektima:

Konfigurabilni algoritmi

- Algoritam učenja može biti konfigurabilan po različitim aspektima:
  - regularizacioni metaparametri

Konfigurabilni algoritmi

- Algoritam učenja može biti konfigurabilan po različitim aspektima:
  - regularizacioni metaparametri
  - parametri tolerancije kod metoda potpornih vektora za regresiju

Konfigurabilni algoritmi

- Algoritam učenja može biti konfigurabilan po različitim aspektima:
  - regularizacioni metaparametri
  - parametri tolerancije kod metoda potpornih vektora za regresiju
  - parametri kernela

Konfigurabilni algoritmi

- Algoritam učenja može biti konfigurabilan po različitim aspektima:
  - regularizacioni metaparametri
  - parametri tolerancije kod metoda potpornih vektora za regresiju
  - parametri kernela
  - izbor podskupa atributa na kom se uči

Konfigurabilni algoritmi

- Algoritam učenja može biti konfigurabilan po različitim aspektima:
  - regularizacioni metaparametri
  - parametri tolerancije kod metoda potpornih vektora za regresiju
  - parametri kernela
  - izbor podskupa atributa na kom se uči
  - kod neuronskih mreža, različite arhitekture

Konfigurabilni algoritmi

- Algoritam učenja može biti konfigurabilan po različitim aspektima:
  - regularizacioni metaparametri
  - parametri tolerancije kod metoda potpornih vektora za regresiju
  - parametri kernela
  - izbor podskupa atributa na kom se uči
  - kod neuronskih mreža, različite arhitekture
- Skup izbora za sve ove aspekte učenja nazivaćemo *konfiguracijom*

Izbor modela u slučaju konfigurabilnog algoritma učenja

- Različite konfiguracije vode različitim modelima

Izbor modela u slučaju konfigurabilnog algoritma učenja

- Različite konfiguracije vode različitim modelima
- Kako izabrati najbolju konfiguraciju, a time i model?

Izbor modela u slučaju konfigurabilnog algoritma učenja

- Različite konfiguracije vode različitim modelima
- Kako izabrati najbolju konfiguraciju, a time i model?
- Ako je $\mathcal{K}$ unapred definisani skup konfiguracija, ceo postupak je sledeći:

Izbor modela u slučaju konfigurabilnog algoritma učenja

- Različite konfiguracije vode različitim modelima
- Kako izabrati najbolju konfiguraciju, a time i model?
- Ako je $\mathcal{K}$ unapred definisani skup konfiguracija, ceo postupak je sledeći:
  - Za svaku konfiguraciju $K \in \mathcal{K}$

Izbor modela u slučaju konfigurabilnog algoritma učenja

- Različite konfiguracije vode različitim modelima
- Kako izabrati najbolju konfiguraciju, a time i model?
- Ako je $\mathcal{K}$ unapred definisani skup konfiguracija, ceo postupak je sledeći:
  - Za svaku konfiguraciju $K \in \mathcal{K}$
    - Uraditi evaluaciju algoritma za konfiguraciju K nekim od metoda koji se koristi u slučaju nekonfigurabilnih algoritama i zapamtiti ocenu kvaliteta.

Izbor modela u slučaju konfigurabilnog algoritma učenja

- Različite konfiguracije vode različitim modelima
- Kako izabrati najbolju konfiguraciju, a time i model?
- Ako je $\mathcal{K}$ unapred definisani skup konfiguracija, ceo postupak je sledeći:
  - Za svaku konfiguraciju $K \in \mathcal{K}$
    - Uraditi evaluaciju algoritma za konfiguraciju K nekim od metoda koji se koristi u slučaju nekonfigurabilnih algoritama i zapamtiti ocenu kvaliteta.
  - Pomoću konfiguracije za koju je ocena kvaliteta najbolja, obučiti model M na svim podacima $\mathcal{D}$.

Evaluacija modela u slučaju konfigurabilnog algoritma učenja

- Kako ćemo reći koliko je dobijeni model dobar?

Evaluacija modela u slučaju konfigurabilnog algoritma učenja

- Kako ćemo reći koliko je dobijeni model dobar?
- Mogli bismo da rezonujemo ovako: podelimo na trening ili test skup, treniramo na trening skupu, testiramo na test skupu i izaberemo konfiguraciju koja je najbolja na test skupu

Evaluacija modela u slučaju konfigurabilnog algoritma učenja

- Kako ćemo reći koliko je dobijeni model dobar?
- Mogli bismo da rezonujemo ovako: podelimo na trening ili test skup, treniramo na trening skupu, testiramo na test skupu i izaberemo konfiguraciju koja je najbolja na test skupu
- U izboru modela je učestvovao test skup

Evaluacija modela u slučaju konfigurabilnog algoritma učenja

- Kako ćemo reći koliko je dobijeni model dobar?
- Mogli bismo da rezonujemo ovako: podelimo na trening ili test skup, treniramo na trening skupu, testiramo na test skupu i izaberemo konfiguraciju koja je najbolja na test skupu
- U izboru modela je učestvovao test skup
- Deo treniranja je i izbor modela što znači da je test skup korišćen prilikom treniranja!

Evaluacija modela u slučaju konfigurabilnog algoritma učenja

- Kako ćemo reći koliko je dobijeni model dobar?
- Mogli bismo da rezonujemo ovako: podelimo na trening ili test skup, treniramo na trening skupu, testiramo na test skupu i izaberemo konfiguraciju koja je najbolja na test skupu
- U izboru modela je učestvovao test skup
- Deo treniranja je i izbor modela što znači da je test skup korišćen prilikom treniranja!
- Procena greške dobijena na ovaj način je optimistična

Evaluacija modela u slučaju konfigurabilnog algoritma učenja

- Kako ćemo reći koliko je dobijeni model dobar?
- Mogli bismo da rezonujemo ovako: podelimo na trening ili test skup, treniramo na trening skupu, testiramo na test skupu i izaberemo konfiguraciju koja je najbolja na test skupu
- U izboru modela je učestvovao test skup
- Deo treniranja je i izbor modela što znači da je test skup korišćen prilikom treniranja!
- Procena greške dobijena na ovaj način je optimistična
- Opisana tehnika je dobar način za izbor modela ali nije dobar način za njegovu evaluaciju

Evaluacija modela u slučaju konfigurabilnog algoritma učenja

- Tehnike evaluacije u slučaju konfigurabilnog algoritma učenja:

Evaluacija modela u slučaju konfigurabilnog algoritma učenja

- Tehnike evaluacije u slučaju konfigurabilnog algoritma učenja:
  - evaluacija pomoću skupova za validaciju i testiranje

Evaluacija modela u slučaju konfigurabilnog algoritma učenja

- Tehnike evaluacije u slučaju konfigurabilnog algoritma učenja:
  - evaluacija pomoću skupova za validaciju i testiranje
  - ugnježdena unakrsna validacija

Evaluacija pomoću skupova za validaciju i testiranje

- Analogon evaluacije pomoću skupa za testiranje

x_1	x_2	x_3	y	
1	9	0	8	trening
0	6	2	1	
1	3	1	5	
4	9	7	6	
1	1	6	7	validacioni
7	2	3	4	
2	9	9	9	
3	3	4	6	test
7	2	1	7	
6	5	1	5	

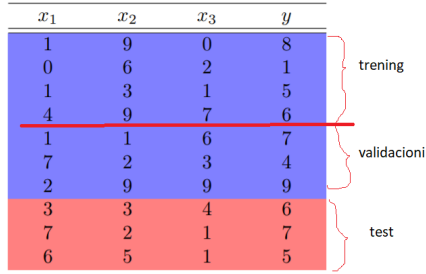
Evaluacija pomoću skupova za validaciju i testiranje

- Analogon evaluacije pomoću skupa za testiranje
- Tehnika se sprovodi kroz naredne korake:

x_1	x_2	x_3	y	
1	9	0	8	trening
0	6	2	1	
1	3	1	5	
4	9	7	6	
1	1	6	7	validacioni
7	2	3	4	
2	9	9	9	
3	3	4	6	test
7	2	1	7	
6	5	1	5	

Evaluacija pomoću skupova za validaciju i testiranje

- Analogon evaluacije pomoću skupa za testiranje
- Tehnika se sprovodi kroz naredne korake:
  - Iz podataka izdvojiti skup za testiranje $\mathcal{T}$



x_1	x_2	x_3	y	
1	9	0	8	trening
0	6	2	1	
1	3	1	5	
4	9	7	6	
1	1	6	7	validacioni
7	2	3	4	
2	9	9	9	
3	3	4	6	test
7	2	1	7	
6	5	1	5	

Evaluacija pomoću skupova za validaciju i testiranje

- Analogon evaluacije pomoću skupa za testiranje
- Tehnika se sprovodi kroz naredne korake:
  - Iz podataka izdvojiti skup za testiranje $\mathcal{T}$
  - Na podacima $\mathcal{D} \setminus \mathcal{T}$ izvršiti izbor modela pomoću skupa za testiranje (deo skupa koji se u ovoj fazi koristi za testiranje se naziva *validacioni skup*) i zapamtiti najbolju konfiguraciju

x_1	x_2	x_3	y	
1	9	0	8	trening
0	6	2	1	
1	3	1	5	
4	9	7	6	
1	1	6	7	validacioni
7	2	3	4	
2	9	9	9	
3	3	4	6	test
7	2	1	7	
6	5	1	5	

Evaluacija pomoću skupova za validaciju i testiranje

- Analogon evaluacije pomoću skupa za testiranje
- Tehnika se sprovodi kroz naredne korake:
  - Iz podataka izdvojiti skup za testiranje $\mathcal{T}$
  - Na podacima $\mathcal{D} \setminus \mathcal{T}$ izvršiti izbor modela pomoću skupa za testiranje (deo skupa koji se u ovoj fazi koristi za testiranje se naziva *validacioni skup*) i zapamtiti najbolju konfiguraciju
  - Pomoću te konfiguracije obučiti model M' na podacima $\mathcal{D} \setminus \mathcal{T}$ i izvršiti predviđanje na podacima iz skupa $\mathcal{T}$

x_1	x_2	x_3	y	
1	9	0	8	trening
0	6	2	1	
1	3	1	5	
4	9	7	6	
1	1	6	7	validacioni
7	2	3	4	
2	9	9	9	
3	3	4	6	test
7	2	1	7	
6	5	1	5	

Evaluacija pomoću skupova za validaciju i testiranje

- Analogon evaluacije pomoću skupa za testiranje
- Tehnika se sprovodi kroz naredne korake:
  - Iz podataka izdvojiti skup za testiranje $\mathcal{T}$
  - Na podacima $\mathcal{D} \setminus \mathcal{T}$ izvršiti izbor modela pomoću skupa za testiranje (deo skupa koji se u ovoj fazi koristi za testiranje se naziva *validacioni skup*) i zapamtiti najbolju konfiguraciju
  - Pomoću te konfiguracije obučiti model M' na podacima $\mathcal{D} \setminus \mathcal{T}$ i izvršiti predviđanje na podacima iz skupa $\mathcal{T}$
  - Oceniti kvalitet predviđanja i prijaviti tu ocenu

x_1	x_2	x_3	y	
1	9	0	8	trening
0	6	2	1	
1	3	1	5	
4	9	7	6	
1	1	6	7	validacioni
7	2	3	4	
2	9	9	9	
3	3	4	6	test
7	2	1	7	
6	5	1	5	

Evaluacija pomoću skupova za validaciju i testiranje

- Analogon evaluacije pomoću skupa za testiranje
- Tehnika se sprovodi kroz naredne korake:
  - Iz podataka izdvojiti skup za testiranje $\mathcal{T}$
  - Na podacima $\mathcal{D} \setminus \mathcal{T}$ izvršiti izbor modela pomoću skupa za testiranje (deo skupa koji se u ovoj fazi koristi za testiranje se naziva *validacioni skup*) i zapamtiti najbolju konfiguraciju
  - Pomoću te konfiguracije obučiti model M' na podacima $\mathcal{D} \setminus \mathcal{T}$ i izvršiti predviđanje na podacima iz skupa $\mathcal{T}$
  - Oceniti kvalitet predviđanja i prijaviti tu ocenu
- Dobijena ocena kvaliteta je ocena modela M

x_1	x_2	x_3	y	
1	9	0	8	trening
0	6	2	1	
1	3	1	5	
4	9	7	6	
1	1	6	7	validacioni
7	2	3	4	
2	9	9	9	
3	3	4	6	test
7	2	1	7	
6	5	1	5	

Ugnježdena unakrsna validacija

- Iste mane kao kod evaluacije pomoću skupa za testiranje - izbor podskupa za validaciju i za testiranje, zašto su baš ovi odabrani povlašćeni?

Ugnježdena unakrsna validacija

- Iste mane kao kod evaluacije pomoću skupa za testiranje - izbor podskupa za validaciju i za testiranje, zašto su baš ovi odabrani povlašćeni?
- Uopštenje unakrsne validacije za konfigurabilne algoritme se naziva *ugnježdena unakrsna validacija*

Ugnježdena unakrsna validacija

- Iste mane kao kod evaluacije pomoću skupa za testiranje - izbor podskupa za validaciju i za testiranje, zašto su baš ovi odabrani povlašćeni?
- Uopštenje unakrsne validacije za konfigurabilne algoritme se naziva *ugnježdena unakrsna validacija*
- Ugnježdena unakrsna validacija se u praksi sprovodi kada imamo mali skup podataka na raspolaganju

Ugnježdena unakrsna validacija

- Iste mane kao kod evaluacije pomoću skupa za testiranje - izbor podskupa za validaciju i za testiranje, zašto su baš ovi odabrani povlašćeni?
- Uopštenje unakrsne validacije za konfigurabilne algoritme se naziva *ugnježdena unakrsna validacija*
- Ugnježdena unakrsna validacija se u praksi sprovodi kada imamo mali skup podataka na raspolaganju
- Kod malog skupa podataka, za različite izbore trening i test skupa dobijamo različite ocene kvaliteta

Ugnježdena unakrsna validacija

- Iste mane kao kod evaluacije pomoću skupa za testiranje - izbor podskupa za validaciju i za testiranje, zašto su baš ovi odabrani povlašćeni?
- Uopštenje unakrsne validacije za konfigurabilne algoritme se naziva *ugnježdena unakrsna validacija*
- Ugnježdena unakrsna validacija se u praksi sprovodi kada imamo mali skup podataka na raspolaganju
- Kod malog skupa podataka, za različite izbore trening i test skupa dobijamo različite ocene kvaliteta
- Zbog velike varijanse u oceni kvaliteta, želimo da uzmemo u obzir *razne* podele na trening i test skup

Ugnježdena unakrsna validacija

- Sprovodi se na sledeći način:

Ugnježdena unakrsna validacija

- Sprovodi se na sledeći način:
  - Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$

Ugnježdena unakrsna validacija

- Sprovodi se na sledeći način:
  - Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$
  - Za $i = 1, \ldots, K$

Ugnježdena unakrsna validacija

- Sprovodi se na sledeći način:
  - Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$
  - Za $i = 1, \ldots, K$
    - Izvršiti izbor modela pomoću unakrsne validacije na podacima $\mathcal{D} \setminus S_i$ i zapamtiti najbolju konfiguraciju

Ugnježdena unakrsna validacija

- Sprovodi se na sledeći način:
  - Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$
  - Za $i = 1, \ldots, K$
    - Izvršiti izbor modela pomoću unakrsne validacije na podacima $\mathcal{D} \setminus S_i$ i zapamtiti najbolju konfiguraciju
    - Pomoću te konfiguracije obučiti model M' na podacima $\mathcal{D} \setminus S_i$ i izvršiti predviđanje na podacima iz skupa S_i

Ugnježdena unakrsna validacija

- Sprovodi se na sledeći način:
  - Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$
  - Za $i = 1, \ldots, K$
    - Izvršiti izbor modela pomoću unakrsne validacije na podacima $\mathcal{D} \setminus S_i$ i zapamtiti najbolju konfiguraciju
    - Pomoću te konfiguracije obučiti model M' na podacima $\mathcal{D} \setminus S_i$ i izvršiti predviđanje na podacima iz skupa S_i
  - Izračunati ocenu kvaliteta na osnovu svih predviđanja na celom skupu $\mathcal{D}$

Ugnježdena unakrsna validacija

- **Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva)** $\{S_1, \ldots, S_K\}$
- Za $i = 1, \ldots, K$
  - Izvršiti izbor modela pomoću unakrsne validacije na podacima $\mathcal{D} \setminus S_i$ i zapamtiti najbolju konfiguraciju
  - Pomoću te konfiguracije obučiti model M' na podacima $\mathcal{D} \setminus S_i$ i izvršiti predviđanje na podacima iz skupa S_i
- Izračunati ocenu kvaliteta na osnovu svih predviđanja na celom skupu $\mathcal{D}$

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

Ugnježdena unakrsna validacija

- Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$
- **Za** $i = 1, \ldots, K$
  - **Izvršiti izbor modela pomoću unakrsne validacije na podacima** $\mathcal{D} \setminus S_i$ **i zapamtiti najbolju konfiguraciju**
  - **Pomoću te konfiguracije obučiti model** M' **na podacima** $\mathcal{D} \setminus S_i$ **i izvršiti predviđanje na podacima iz skupa** S_i
- Izračunati ocenu kvaliteta na osnovu svih predviđanja na celom skupu $\mathcal{D}$

x_1	x_2	x_3	y
1	3	1	5
4	9	7	6
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

trening+validacioni

1	9	0	8
0	6	2	1

test

Ugnježdena unakrsna validacija

- Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$
- **Za** $i = 1, \ldots, K$
  - **Izvršiti izbor modela pomoću unakrsne validacije na podacima $\mathcal{D} \setminus S_i$ i zapamtiti najbolju konfiguraciju**
  - **Pomoću te konfiguracije obučiti model M' na podacima $\mathcal{D} \setminus S_i$ i izvršiti predviđanje na podacima iz skupa S_i**
- Izračunati ocenu kvaliteta na osnovu svih predviđanja na celom skupu $\mathcal{D}$

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	1	6	7
7	2	3	4
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

spojeno trening+validacioni

1	3	1	5	5
4	9	7	6	6

test

Ugnježdena unakrsna validacija

- Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$
- **Za** $i = 1, \ldots, K$
  - **Izvršiti izbor modela pomoću unakrsne validacije na podacima $\mathcal{D} \setminus S_i$ i zapamtiti najbolju konfiguraciju**
  - **Pomoću te konfiguracije obučiti model M' na podacima $\mathcal{D} \setminus S_i$ i izvršiti predviđanje na podacima iz skupa S_i**
- Izračunati ocenu kvaliteta na osnovu svih predviđanja na celom skupu $\mathcal{D}$

x_1	x_2	x_3	y
1	9	0	8
0	6	2	1
1	3	1	5
4	9	7	6
2	9	9	9
3	3	4	6
7	2	1	7
6	5	1	5

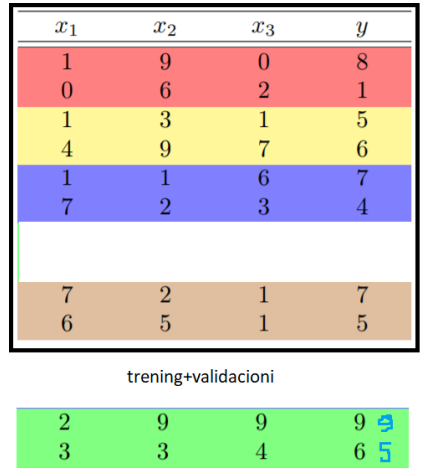
trening+validacioni

1	1	6	7
7	2	3	4

test

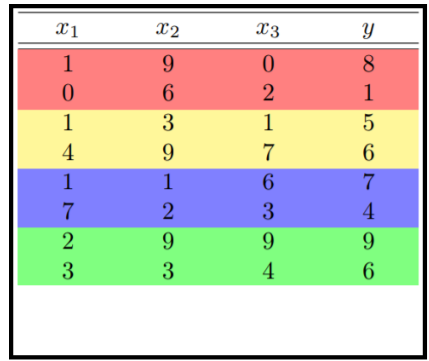
Ugnježdena unakrsna validacija

- Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$
- **Za** $i = 1, \ldots, K$
  - **Izvršiti izbor modela pomoću unakrsne validacije na podacima** $\mathcal{D} \setminus S_i$ **i zapamtiti najbolju konfiguraciju**
  - **Pomoću te konfiguracije obučiti model** M' **na podacima** $\mathcal{D} \setminus S_i$ **i izvršiti predviđanje na podacima iz skupa** S_i
- Izračunati ocenu kvaliteta na osnovu svih predviđanja na celom skupu $\mathcal{D}$



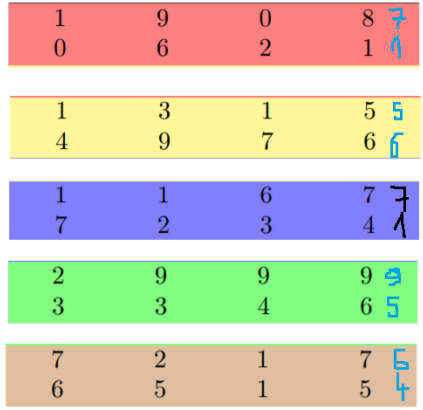
Ugnježdena unakrsna validacija

- Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$
- **Za** $i = 1, \ldots, K$
  - **Izvršiti izbor modela pomoću unakrsne validacije na podacima** $\mathcal{D} \setminus S_i$ **i zapamtiti najbolju konfiguraciju**
  - **Pomoću te konfiguracije obučiti model** M' **na podacima** $\mathcal{D} \setminus S_i$ **i izvršiti predviđanje na podacima iz skupa** S_i
- Izračunati ocenu kvaliteta na osnovu svih predviđanja na celom skupu $\mathcal{D}$



Ugnježdena unakrsna validacija

- Sprovodi se na sledeći način:
  - Podatke $\mathcal{D}$ podeliti na K približno jednakih podskupova (tzv. slojeva) $\{S_1, \ldots, S_K\}$
  - Za $i = 1, \ldots, K$
    - Izvršiti izbor modela pomoću unakrsne validacije na podacima $\mathcal{D} \setminus S_i$ i zapamtiti najbolju konfiguraciju
    - Pomoću te konfiguracije obučiti model M' na podacima $\mathcal{D} \setminus S_i$ i izvršiti predviđanje na podacima iz skupa S_i
  - **Izračunati ocenu kvaliteta na osnovu svih predviđanja na celom skupu $\mathcal{D}$**



Pregled

Modeli sa velikim podacima u praksi

- Za svaku konfiguraciju $K \in \mathcal{K}$

Modeli sa velikim podacima u praksi

- Za svaku konfiguraciju $K \in \mathcal{K}$
  - Obučiti model na skupu za obučavanje

Modeli sa velikim podacima u praksi

- Za svaku konfiguraciju $K \in \mathcal{K}$
  - Obučiti model na skupu za obučavanje
  - Evaluirati ga na validacionom skupu

Modeli sa velikim podacima u praksi

- Za svaku konfiguraciju $K \in \mathcal{K}$
  - Obučiti model na skupu za obučavanje
  - Evaluirati ga na validacionom skupu
- Izabrati najbolji model M od prethodno obučenih

Modeli sa velikim podacima u praksi

- Za svaku konfiguraciju $K \in \mathcal{K}$
  - Obučiti model na skupu za obučavanje
  - Evaluirati ga na validacionom skupu
- Izabrati najbolji model M od prethodno obučenih
- Testirati model M na skupu za testiranje i koristiti kao finalni model

Pregled

Napomene vezane za pretprocesiranje

- Pretpostavimo da rešavamo regresioni problem

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Napomene vezane za pretprocesiranje

- Pretpostavimo da rešavamo regresioni problem
- Podaci su sortirani u odnosu na ciljnu promenljivu i u tom poretku je izvršena podela na K podskupova koji će biti korišćeni za unakrsnu validaciju

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Napomene vezane za pretprocesiranje

- Pretpostavimo da rešavamo regresioni problem
- Podaci su sortirani u odnosu na ciljnu promenljivu i u tom poretku je izvršena podela na K podskupova koji će biti korišćeni za unakrsnu validaciju
- Prvi sloj će sadržati podatke sa najmanjim vrednostima ciljne promenljive koje se neće nalaziti u preostalih $K - 1$ slojeva

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Napomene vezane za pretprocesiranje

- Pretpostavimo da rešavamo regresioni problem
- Podaci su sortirani u odnosu na ciljnu promenljivu i u tom poretku je izvršena podela na K podskupova koji će biti korišćeni za unakrsnu validaciju
- Prvi sloj će sadržati podatke sa najmanjim vrednostima ciljne promenljive koje se neće nalaziti u preostalih $K - 1$ slojeva
- Model treniran na tim podacima će morati da ekstrapolira kada se primeni na prvi sloj i u takvim slučajevima se ne očekuju dobre performanse

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Stratifikacija

- Prilikom bilo kakvih podela podataka često vodi računa o tome da svi podskupovi imaju sličnu raspodelu atributa i ciljne promenljive kao i ukupan skup podataka

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Stratifikacija

- Prilikom bilo kakvih podela podataka često vodi računa o tome da svi podskupovi imaju sličnu raspodelu atributa i ciljne promenljive kao i ukupan skup podataka
- Tehnike koje pokušavaju da zadovolje ovaj zahtev nazivaju se tehnikama *stratifikacije* i predstavljaju vid pretprocesiranja podataka

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Stratifikacija

- Prilikom bilo kakvih podela podataka često vodi računa o tome da svi podskupovi imaju sličnu raspodelu atributa i ciljne promenljive kao i ukupan skup podataka
- Tehnike koje pokušavaju da zadovolje ovaj zahtev nazivaju se tehnikama *stratifikacije* i predstavljaju vid pretprocesiranja podataka
- Pojednostavljena varijanta, koju je uvek moguće izvesti je *stratifikacija po ciljnoj promenljivoj*

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Stratifikacija

- Podela na K delova stratifikovana po ciljnoj promenljivoj sprovodi se na sledeći način:

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Stratifikacija

- Podela na K delova stratifikovana po ciljnoj promenljivoj sprovodi se na sledeći način:
  - Sortirati instance u odnosu na ciljnu promenljivu. Ako je ciljna promenljiva kategorička, to se može uraditi tako što se svakoj njenoj vrednosti pridruži različit broj.

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Stratifikacija

- Podela na K delova stratifikovana po ciljnoj promenljivoj sprovodi se na sledeći način:
  - Sortirati instance u odnosu na ciljnu promenljivu. Ako je ciljna promenljiva kategorička, to se može uraditi tako što se svakoj njenoj vrednosti pridruži različit broj.
  - Za $i = 1, \ldots, K$

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Stratifikacija

- Podela na K delova stratifikovana po ciljnoj promenljivoj sprovodi se na sledeći način:
  - Sortirati instance u odnosu na ciljnu promenljivu. Ako je ciljna promenljiva kategorička, to se može uraditi tako što se svakoj njenoj vrednosti pridruži različit broj.
  - Za $i = 1, \ldots, K$
    - Instance sa indeksima $i + j * K$ za $j = 0, 1, \ldots$, svrstati u podskup P_i.

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Napomene vezane za pretprocesiranje

- *Standardizacija* se sastoji u tome da se od svake vrednosti nekog atributa oduzme prosek svih vrednosti tog atributa, pa da se potom svaka vrednost tog atributa podeli standardnom devijacijom svih vrednosti tog atributa.

Napomene vezane za pretprocesiranje

- *Standardizacija* se sastoji u tome da se od svake vrednosti nekog atributa oduzme prosek svih vrednosti tog atributa, pa da se potom svaka vrednost tog atributa podeli standardnom devijacijom svih vrednosti tog atributa.
- Time se obezbeđuje da svaki atribut ima prosek 0 i standardnu devijaciju 1.

Napomene vezane za pretprocesiranje

- *Standardizacija* se sastoji u tome da se od svake vrednosti nekog atributa oduzme prosek svih vrednosti tog atributa, pa da se potom svaka vrednost tog atributa podeli standardnom devijacijom svih vrednosti tog atributa.
- Time se obezbeđuje da svaki atribut ima prosek 0 i standardnu devijaciju 1.
- Kada imamo podelu skupa na podskupove, postavlja se pitanje na kojim podskupovima treba vršiti standardizaciju

Standardizacija

- Česta greška u praksi mašinskog učenja je u lošoj primeni različitih vidova pretprocesiranja

Standardizacija

- Česta greška u praksi mašinskog učenja je u lošoj primeni različitih vidova pretprocesiranja
- Takva greška je primena standardizacije na ceo skup podataka, pre podele na podskupove koji se koriste u evaluaciji.

Standardizacija

- Česta greška u praksi mašinskog učenja je u lošoj primeni različitih vidova pretprocesiranja
- Takva greška je primena standardizacije na ceo skup podataka, pre podele na podskupove koji se koriste u evaluaciji.
- Prilikom standardizacije, vrednosti atributa na skupu za testiranje utiču na prosek i standardnu devijaciju koji se koriste pri standardizaciji.

Standardizacija

- Česta greška u praksi mašinskog učenja je u lošoj primeni različitih vidova pretprocesiranja
- Takva greška je primena standardizacije na ceo skup podataka, pre podele na podskupove koji se koriste u evaluaciji.
- Prilikom standardizacije, vrednosti atributa na skupu za testiranje utiču na prosek i standardnu devijaciju koji se koriste pri standardizaciji.
- S druge strane, podaci na kojima će se ubuduće model primenjivati nisu dostupni da daju svoj doprinos ovim veličinama.

Standardizacija

- Česta greška u praksi mašinskog učenja je u lošoj primeni različitih vidova pretprocesiranja
- Takva greška je primena standardizacije na ceo skup podataka, pre podele na podskupove koji se koriste u evaluaciji.
- Prilikom standardizacije, vrednosti atributa na skupu za testiranje utiču na prosek i standardnu devijaciju koji se koriste pri standardizaciji.
- S druge strane, podaci na kojima će se ubuduće model primenjivati nisu dostupni da daju svoj doprinos ovim veličinama.
- Veličina podataka za obučavanje je retko dovoljno velika da nas u budućnosti ništa ne može makar malo iznenaditi, a time i dovesti do veće greške modela

Standardizacija

- Česta greška u praksi mašinskog učenja je u lošoj primeni različitih vidova pretprocesiranja
- Takva greška je primena standardizacije na ceo skup podataka, pre podele na podskupove koji se koriste u evaluaciji.
- Prilikom standardizacije, vrednosti atributa na skupu za testiranje utiču na prosek i standardnu devijaciju koji se koriste pri standardizaciji.
- S druge strane, podaci na kojima će se ubuduće model primenjivati nisu dostupni da daju svoj doprinos ovim veličinama.
- Veličina podataka za obučavanje je retko dovoljno velika da nas u budućnosti ništa ne može makar malo iznenaditi, a time i dovesti do veće greške modela
- Izbegavanje zajedničke standardizacije celog skupa podataka omogućava baš taj efekat – daje šansu podacima iz skupa za testiranje da nas iznenade i time ocenu greške modela učine realističnijom