Teorijske osnove nadgledanog učenja

Mašinsko učenje 2020/21

Matematički fakultet
Univerzitet u Beogradu

Pregled

Postavka problema nadgledanog učenja

Princip minimizacije empirijskog rizika

Preprilagođavanje

Regularizacija

Nagodba izmedu sistematskog odstupanja i varijanse

Šta je nadgledano učenje?

- Potrebno je ustanoviti odnos između *atributa* x i *ciljne promenljive* y
- Problemi koje danas razmatramo su često previše kompleksni
- Stoga se pomenuti odnos često aproksimira na osnovu uzorka
- Bilo bi idealno da postoji funkcija f za koju važi $y = f(x)$
- Zavisnosti između promenljivih se modeluju funkcijom koja se naziva *model*
- Model treba da *generalizuje*

Standardni problemi nadgledanog učenja

- Klasifikacija – ciljna promenljiva je kategorička
- Regresija – ciljna promenljiva je neprekidna

Osnovna teorijska postavka nadgledanog učenja

- Odnos između x i y je određen probabilističkim zakonom $p(x, y)$
- Potrebno je odrediti ,,najbolju“ funkciju f takvu da važi $y \approx f(x)$
- *Funkcija greške* (eng. loss) L kvantifikuje odstupanje između y i $f(x)$
- *Funkcional rizika* R formalizuje pojam ,,najboljeg“:

$$R(f) = \mathbb{E}[L(y, f(x))] = \int L(y, f(x))p(x, y)dxdy$$

- Potrebno je odrediti funkciju f za koju je vrednost $R(f)$ najmanja

Ali u praksi...

- Razmatranje svih mogućih modela nije izvodljivo

Ali u praksi...

- Razmatranje svih mogućih modela nije izvodljivo
- Problem je što raspodela $p(x, y)$ nije poznata

Ali u praksi...

- Razmatranje svih mogućih modela nije izvodljivo
- Problem je što raspodela $p(x, y)$ nije poznata
- Potreban je praktičan princip indukcije

Pregled

Postavka problema nadgledanog učenja

Princip minimizacije empirijskog rizika

Preprilagođavanje

Regularizacija

Nagodba izmedu sistematskog odstupanja i varijanse

Ali u praksi...

- Razmatranje svih mogućih modela nije izvodljivo, tako da se pretpostavlja *reprezentacija modela*
- Modeli $f_w(x)$ za neku reprezentaciju zavise od vektora realnih vrednosti w, koje nazivamo *parametrima modela*
- Na primer, linearna reprezentacija

$$f_w(x) = w_0 + \sum_{i=1}^{n} w_i x_i$$

- Funkcional rizika se onda može predstaviti kao funkcija parametara w: $R(w) = R(f_w(x))$

Ali u praksi...

- Problem je što raspodela $p(x, y)$ nije poznata, ali se pretpostavlja da postoji uzorak $\mathcal{D} = \{(x_i, y_i) \mid i = 1, \ldots, N\}$ takav da važi $(x_i, y_i) \sim p(x, y)$

Minimizacija empirijskog rizika

- Empirijski rizik:

$$E(w, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_w(x_i))$$

- Kada nije naveden, argument $\mathcal{D}$ se podrazumeva
- Umesto empirijski rizik, govorićemo prosečna greška ili samo greška

Princip minimizacije empirijskog rizika (ERM)

- funkciju koja minimizuje $E(w, \mathcal{D})$ uzeti za aproksimaciju funkcije koja minimizuje $R(w)$

Princip minimizacije empirijskog rizika (ERM)

- funkciju koja minimizuje $E(w, \mathcal{D})$ uzeti za aproksimaciju funkcije koja minimizuje $R(w)$
- Podsetimo se:

$$E(w, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_w(x_i))$$

$$R(w) = \mathbb{E}[L(y, f_w(x))] = \int L(y, f_w(x)) p(x, y) dx dy$$

Princip minimizacije empirijskog rizika (ERM)

- funkciju koja minimizuje $E(w, \mathcal{D})$ uzeti za aproksimaciju funkcije koja minimizuje $R(w)$
- Podsetimo se:

$$E(w, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_w(x_i))$$

$$R(w) = \mathbb{E}[L(y, f_w(x))] = \int L(y, f_w(x)) p(x, y) dx dy$$

- Princip tvrdi da važi:

$$\arg\min_w E(w, D) \to \arg\min_w R(w) \text{ kada } (|D| \to \infty)$$

Pitanja vezana za ERM

- Ako važi

$$E(w, D) \to R(w) \text{ kada } (|D| \to \infty)$$

ne mora da znači da će da važi

$$\arg\min_w E(w, D) \to \arg\min_w R(w) \text{ kada } (|D| \to \infty)$$

Pitanja vezana za ERM

- Ako važi

$$E(w, D) \to R(w) \text{ kada } (|D| \to \infty)$$

ne mora da znači da će da važi

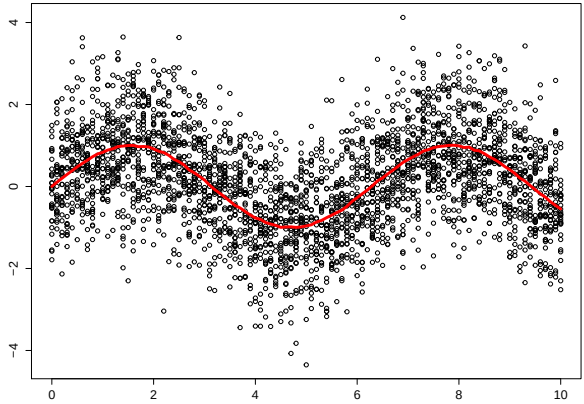
$$\arg\min_w E(w, D) \to \arg\min_w R(w) \text{ kada } (|D| \to \infty)$$

- Prvo svojstvo se odnosi na bilo koju, ali fiksiranu vrednost parametra w, istu i za E i za R, dok se drugo odnosi na potencijalno različite vrednosti parametra w jer E i R ne moraju dostizati minimum u istoj tački

Pitanja vezana za ERM

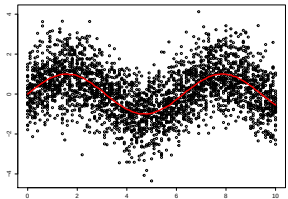
- Da li bi trebalo da radi?
- Odgovor bi trebalo da zavisi od svojstava skupa modela $\{f_w(x)\}$ (*statistička teorija učenja*)

ERM za regresiju



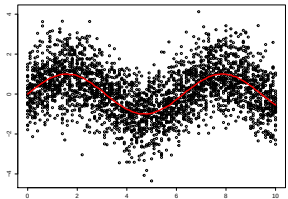
Slika: P. Janičić, M. Nikolić, Veštačka inteligencija, u pripremi.

ERM za regresiju



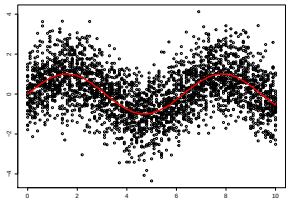
- Jednoj vrednosti atributa ne mora odgovarati jedna vrednost ciljne promenljive pa tako ima smisla govoriti o *raspodeli* vrednosti ciljne promenljive pri datim vrednostima atributa

ERM za regresiju



- Jednoj vrednosti atributa ne mora odgovarati jedna vrednost ciljne promenljive pa tako ima smisla govoriti o *raspodeli* vrednosti ciljne promenljive pri datim vrednostima atributa
- Kako onda za datu vrednost x prilikom predviđanja odrediti vrednost y?

ERM za regresiju



- Jednoj vrednosti atributa ne mora odgovarati jedna vrednost ciljne promenljive pa tako ima smisla govoriti o *raspodeli* vrednosti ciljne promenljive pri datim vrednostima atributa
- Kako onda za datu vrednost x prilikom predviđanja odrediti vrednost y?
- Srednja vrednost tj. prosek, očekivanje
- Regresiona funkcija: $r(x) = \mathbb{E}(y|x) = \int y \ p(y|x)dy$

ERM za regresiju

- Drugačiji pristup dolaska do regresione funkcije

ERM za regresiju

- Drugačiji pristup dolaska do regresione funkcije
- Potrebno je da vrednosti koje daje model što bolje odgovaraju pravim vrednostima, odnosno da je razlika $y - f_w(x)$ što manja.

ERM za regresiju

- Drugačiji pristup dolaska do regresione funkcije
- Potrebno je da vrednosti koje daje model što bolje odgovaraju pravim vrednostima, odnosno da je razlika $y - f_w(x)$ što manja.
- Ako je $y - f_w(x)$ malo, tada će i $(y - f_w(x))^2$ biti malo

ERM za regresiju

- Drugačiji pristup dolaska do regresione funkcije
- Potrebno je da vrednosti koje daje model što bolje odgovaraju pravim vrednostima, odnosno da je razlika $y - f_w(x)$ što manja.
- Ako je $y - f_w(x)$ malo, tada će i $(y - f_w(x))^2$ biti malo
- Kvadrat razlike se stoga može uzeti za funkciju greške i rizik možemo definisati kao

$$\mathbb{E}[(y - f_w(x))^2]$$

ERM za regresiju

- Može se pokazati da važi:

$$\min_w \mathbb{E}[(y - f_w(x))^2] = \min_w \mathbb{E}[(r(x) - f_w(x))^2]$$

ERM za regresiju

- Može se pokazati da važi:

$$\min_w \mathbb{E}[(y - f_w(x))^2] = \min_w \mathbb{E}[(r(x) - f_w(x))^2]$$

- Minimizacija greške je isto što i minimizacija odstupanja predviđanja modela f_w od idealne regresione funkcije r

ERM za regresiju

- Može se pokazati da važi:

$$\min_w \mathbb{E}[(y - f_w(x))^2] = \min_w \mathbb{E}[(r(x) - f_w(x))^2]$$

- Minimizacija greške je isto što i minimizacija odstupanja predviđanja modela f_w od idealne regresione funkcije r
- To znači da:

ERM za regresiju

- Može se pokazati da važi:

$$\min_w \mathbb{E}[(y - f_w(x))^2] = \min_w \mathbb{E}[(r(x) - f_w(x))^2]$$

- Minimizacija greške je isto što i minimizacija odstupanja predviđanja modela f_w od idealne regresione funkcije r
- To znači da:
  - ako je regresiona funkcija $r(x)$ *iz datog skupa modela* (ako važi $r(x) = f_w(x)$ za neko w^*), tada će minimum datog rizika biti postignut upravo u njoj (upravo za w^*)

ERM za regresiju

- Može se pokazati da važi:

$$\min_w \mathbb{E}[(y - f_w(x))^2] = \min_w \mathbb{E}[(r(x) - f_w(x))^2]$$

- Minimizacija greške je isto što i minimizacija odstupanja predviđanja modela f_w od idealne regresione funkcije r
- To znači da:
  - ako je regresiona funkcija $r(x)$ *iz datog skupa modela* (ako važi $r(x) = f_w(x)$ za neko w^*), tada će minimum datog rizika biti postignut upravo u njoj (upravo za w^*)
  - ako regresiona funkcija $r(x)$ *nije iz datog skupa modela*, tada će minimum datog rizika biti postignut za funkciju najbližu funkciji $r(x)$ (najbližu u odnosu na kvadrat razlike)

ERM za regresiju

- Kako je rizik dat izrazom $\mathbb{E}[(y - f_w(x))^2]$, formulacija principa minimizacije empirijskog rizika za regresiju je

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

ERM za regresiju

- Kako je rizik dat izrazom $\mathbb{E}[(y - f_w(x))^2]$, formulacija principa minimizacije empirijskog rizika za regresiju je

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

- Kvadratna greška (*squared loss*):

$$L(u, v) = (u - v)^2$$

ERM za regresiju

- Kako je rizik dat izrazom $\mathbb{E}[(y - f_w(x))^2]$, formulacija principa minimizacije empirijskog rizika za regresiju je

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

- Kvadratna greška (*squared loss*):

$$L(u, v) = (u - v)^2$$

- Srednja greška (srednja vrednost greške) za kvadratnu grešku naziva se srednjekvadratna greška:

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

ERM za regresiju

- Kako je rizik dat izrazom $\mathbb{E}[(y - f_w(x))^2]$, formulacija principa minimizacije empirijskog rizika za regresiju je

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

- Kvadratna greška (*squared loss*):

$$L(u, v) = (u - v)^2$$

- Srednja greška (srednja vrednost greške) za kvadratnu grešku naziva se srednjekvadratna greška:

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

- Moguće su i druge funkcije greške

ERM za klasifikaciju

- Šta treba da minimizujemo?

ERM za klasifikaciju

- Šta treba da minimizujemo?
- Broj grešaka na raspoloživom skupu podataka

ERM za klasifikaciju

- Šta treba da minimizujemo?
- Broj grešaka na raspoloživom skupu podataka
- Indikatorska funkcija:

$$I(F) = \begin{cases} 1 & \text{ako važi } F \\ 0 & \text{ako važi } \neg F \end{cases}$$

- Funkcija greške: $L(u, v) = I(u \neq v)$
- Optimizacioni problem:

$$\min_w \frac{1}{N} \sum_{i=1}^{N} I(y_i \neq f_w(x_i))$$

Pregled

Postavka problema nadgledanog učenja

Princip minimizacije empirijskog rizika

Preprilagođavanje

Regularizacija

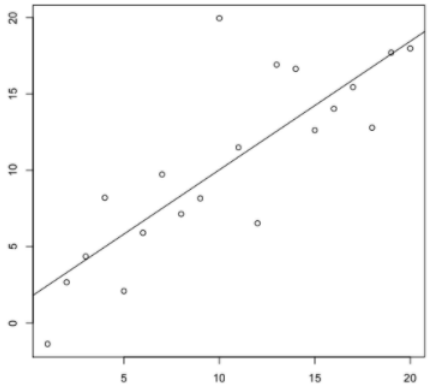
Nagodba izmedu sistematskog odstupanja i varijanse

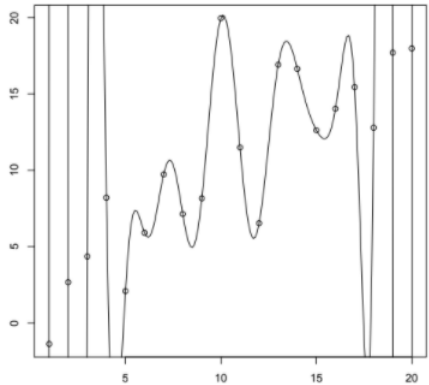
Koliko dobro se model može prilagoditi podacima?

- Imajmo u vidu problem regresije
- Jednostavna linearna regresija: $f_w(x) = w_0 + w_1 x$
- Polinomijalna linearna regresija: $f_w(x) = \sum_{i=0}^{n} w_i x^i$
- Optimizacioni problem: minimizacija *srednje greške*

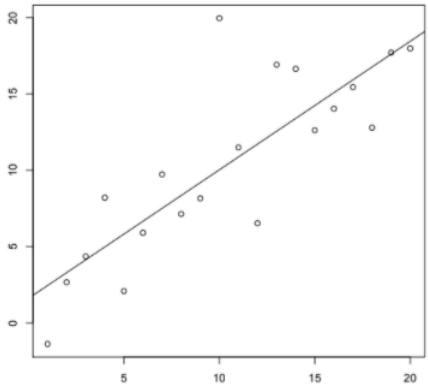
$$\frac{1}{N} \sum_{i=1}^{N} (y_i - f_w(x_i))^2$$

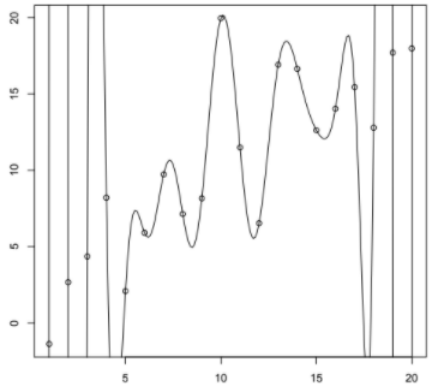
Koji je model bolji?





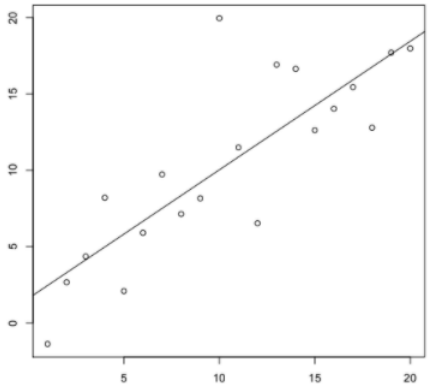
Koji je model bolji?

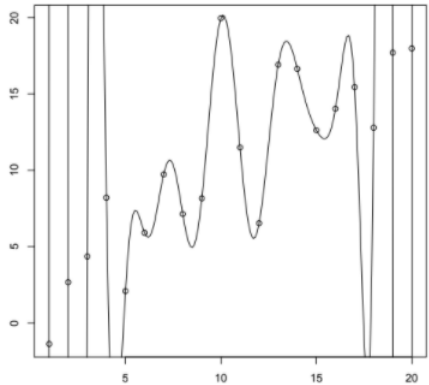




- Polinomijalni model će imati manju vrednost srednje greške

Koji je model bolji?





- Polinomijalni model će imati manju vrednost srednje greške
- Sa druge strane, linearni model ima veću vrednost srednje greške ali *prati trend rasta* koji postoji u podacima

Preprilagođavanje (eng. overfitting)

- Dobra prilagođenost modela trening podacima (mala srednja greška) ne obezbeđuje dobru generalizaciju
- Uporediti sa učenjem napamet
- Uzrok problema je prevelika prilagodljivost modela
- Upravljanje prilagodljivošću modela je od ključnog značaja za dobru generalizaciju!
- Ovo je glavni problem mašinskog učenja i izvor njegove najdublje teorije

Kako učiniti modele manje prilagodljivim?

- Izabrati neprilagodljivu reprezentaciju (npr. linearni modeli)?

Kako učiniti modele manje prilagodljivim?

- Izabrati neprilagodljivu reprezentaciju (npr. linearni modeli)?
- Moguće, ali takav pristup je previše krut i nije podložan finom podešavanju

Kako učiniti modele manje prilagodljivim?

- Izabrati neprilagodljivu reprezentaciju (npr. linearni modeli)?
- Moguće, ali takav pristup je previše krut i nije podložan finom podešavanju
- Da li je moguće fino podešavati prilagodljivost modela nezavisno od izabrane reprezentacije?

Pregled

Postavka problema nadgledanog učenja

Princip minimizacije empirijskog rizika

Preprilagođavanje

Regularizacija

Nagodba izmedu sistematskog odstupanja i varijanse

Regularizacija (1)

- Minimizacija regularizovane greške:

$$\min_w \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_w(x_i)) + \lambda \Omega(w)$$

- Čest izbor *regularizacionog izraza* Ω je kvadrat ℓ_2 norme

$$\Omega(w) = \|w\|_2^2 = \sum_{i=1}^{n} w_i^2$$

Regularizacija (1)

- Minimizacija regularizovane greške:

$$\min_w \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_w(x_i)) + \lambda \Omega(w)$$

- Čest izbor *regularizacionog izraza* Ω je kvadrat ℓ_2 norme

$$\Omega(w) = \|w\|_2^2 = \sum_{i=1}^{n} w_i^2$$

- Regularizacioni izraz kažnjava visoke apsolutne vrednosti parametara, čineći model manje prilagodljivim
- *Regularizacioni parametar* λ služi za fino podešavanje prilagodljivosti modela

Regularizacija (2)

- U slučaju linearnih modela $f_w(x) = \sum_{i=1}^{n} w_i x_i$ važi

$$w = \nabla_x f_w(x)$$

Regularizacija (2)

- U slučaju linearnih modela $f_w(x) = \sum_{i=1}^{n} w_i x_i$ važi

$$w = \nabla_x f_w(x)$$

- Koji je efekat regularizacije?

Regularizacija (2)

- U slučaju linearnih modela $f_w(x) = \sum_{i=1}^{n} w_i x_i$ važi

$$w = \nabla_x f_w(x)$$

- Koji je efekat regularizacije?
- Ograničavanje gradijenta ograničava brzinu promene funkcije

Regularizacija (2)

- U slučaju linearnih modela $f_w(x) = \sum_{i=1}^{n} w_i x_i$ važi

$$w = \nabla_x f_w(x)$$

- Koji je efekat regularizacije?
- Ograničavanje gradijenta ograničava brzinu promene funkcije
- Funkcija koja ima relativno malu vrednost gradijenta ne može brzo rasti i opadati i stoga se ne može lako prilagoditi bilo kakvim podacima.

Regularizacija (2)

- U slučaju linearnih modela $f_w(x) = \sum_{i=1}^{n} w_i x_i$ važi

$$w = \nabla_x f_w(x)$$

- Koji je efekat regularizacije?
- Ograničavanje gradijenta ograničava brzinu promene funkcije
- Funkcija koja ima relativno malu vrednost gradijenta ne može brzo rasti i opadati i stoga se ne može lako prilagoditi bilo kakvim podacima.
- Na primer, kada usled jakog šuma u podacima ciljna promenljiva u bliskim tačkama uzima čas visoke čas niske vrednosti, bilo bi potrebno da funkcija brzo raste ili opada kako bi savršeno odgovarala podacima.

Regularizacija (3)

- Regularizacija nije nužno uvek od koristi

Regularizacija (3)

- Regularizacija nije nužno uvek od koristi
- Značajna u slučaju kada je količina podataka za obučavanje mala u odnosu na prilagodljivost modela

Regularizacija (3)

- Regularizacija nije nužno uvek od koristi
- Značajna u slučaju kada je količina podataka za obučavanje mala u odnosu na prilagodljivost modela
- Kada je količina podataka velika, regularizacija nije neophodna

Regularizacija (3)

- Regularizacija nije nužno uvek od koristi
- Značajna u slučaju kada je količina podataka za obučavanje mala u odnosu na prilagodljivost modela
- Kada je količina podataka velika, regularizacija nije neophodna
- Povećanje količine podataka je najbolja regularizacija ali to može dovesti do drugih problema

Regularizacija (4)

- U opštijem smislu, regularizacijom se naziva bilo koja modifikacija optimizacionog problema koja ograničava prilagodljivost modela i čini ga manje podložnim preprilagođavanju

Regularizacija (4)

- U opštijem smislu, regularizacijom se naziva bilo koja modifikacija optimizacionog problema koja ograničava prilagodljivost modela i čini ga manje podložnim preprilagođavanju
- U još opštijem smislu, regularizacija je bilo kakva modifikacija matematičkog problema koja ga čini bolje uslovljenim

Primer regularizacije – klasifikacioni model

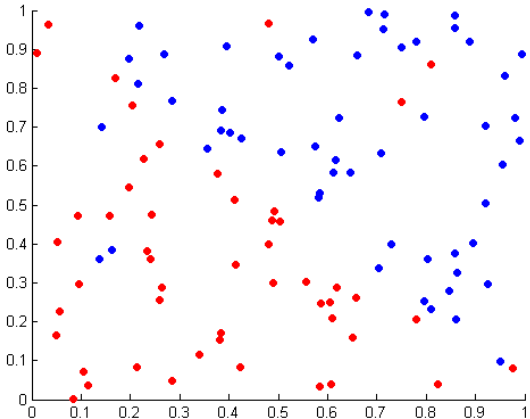
- Linearni klasifikacioni model:

$$f_w(x) = w_0 + w_1 x_1 + w_2 x_2$$

- Polinomijalni klasifikacioni model:

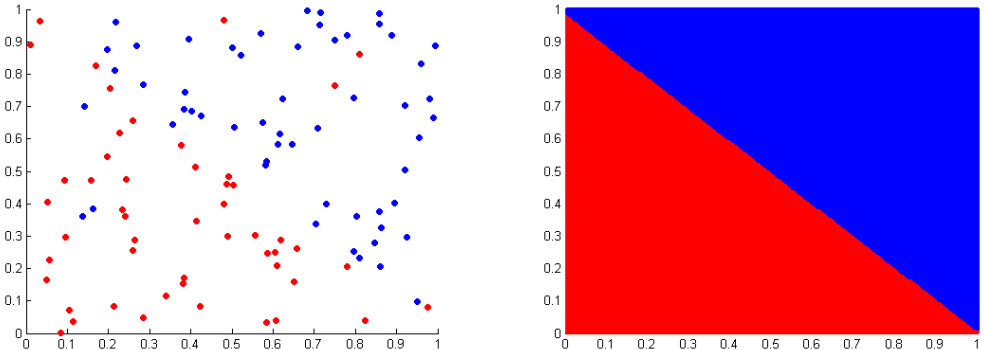
$$f_w(x) = \sum_{i=0}^{n} \sum_{j=0}^{i} w_{ij} x_1^j x_2^{i-j}$$

Primer regularizacije – podaci



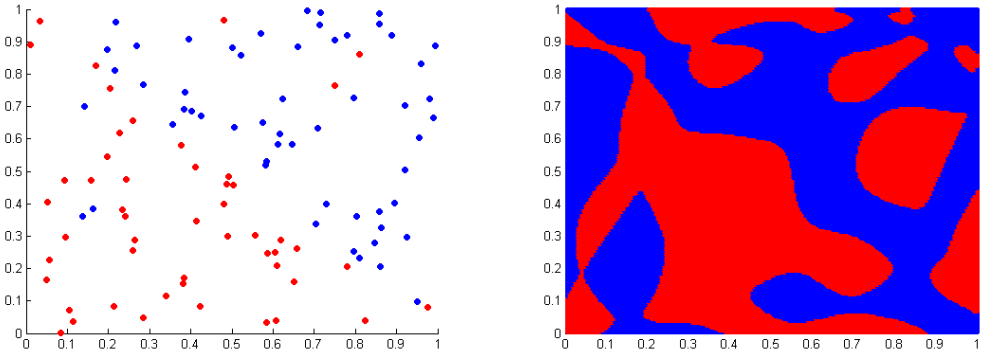
Slika: P. Janičić, M. Nikolić, Veštačka inteligencija, u pripremi.

Primer regularizacije – linearni klasifikator



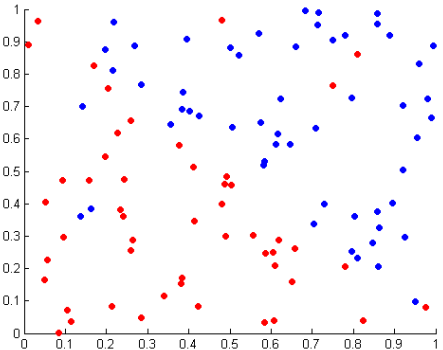
Slika: P. Janičić, M. Nikolić, Veštačka inteligencija, u pripremi.

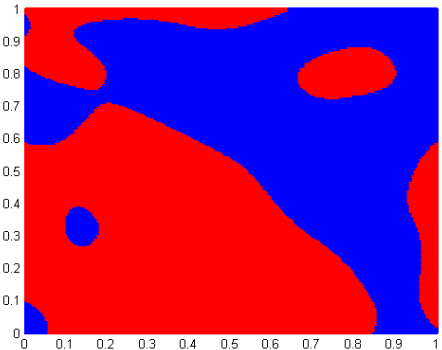
Primer regularizacije – polinomijalni klasifikator



Slika: P. Janičić, M. Nikolić, Veštačka inteligencija, u pripremi.

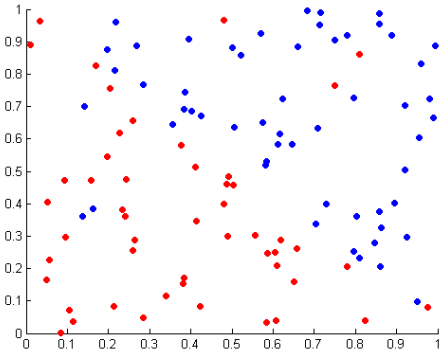
Primer regularizacije – regularizovani polinomijalni klasifikator

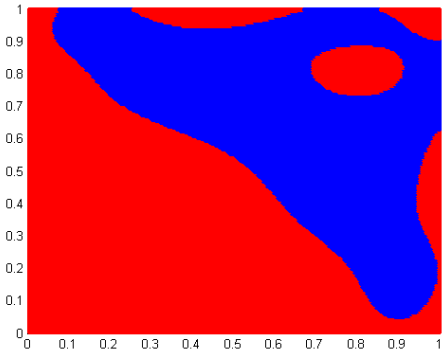




Slika: P. Janičić, M. Nikolić, Veštačka inteligencija, u pripremi.

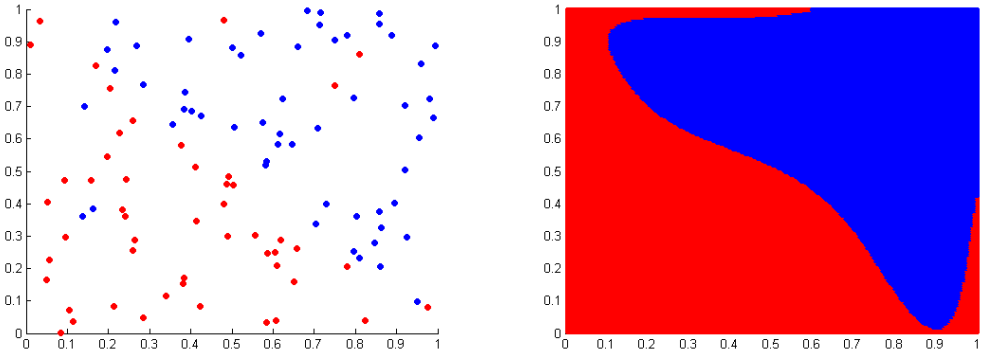
Primer regularizacije – regularizovani polinomijalni klasifikator





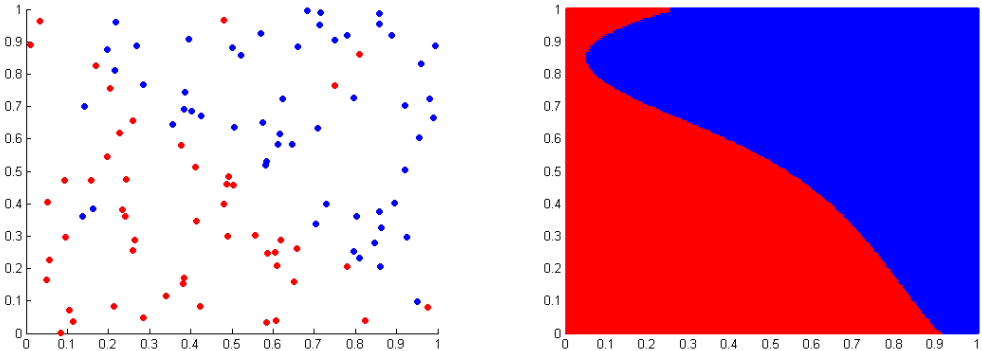
Slika: P. Janičić, M. Nikolić, Veštačka inteligencija, u pripremi.

Primer regularizacije – regularizovani polinomijalni klasifikator



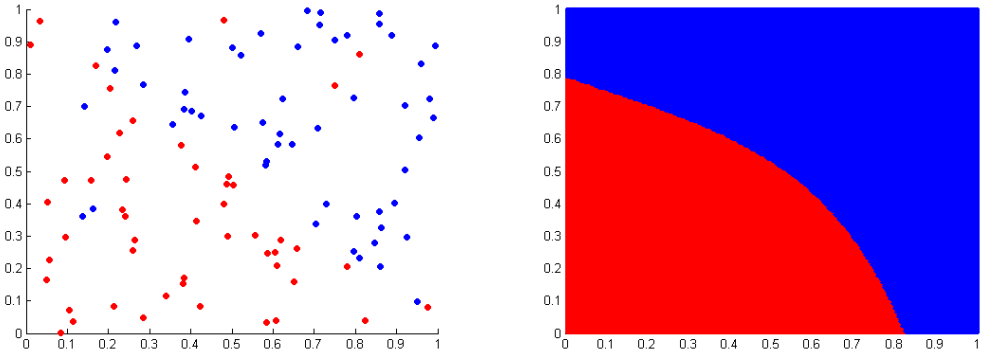
Slika: P. Janičić, M. Nikolić, Veštačka inteligencija, u pripremi.

Primer regularizacije – regularizovani polinomijalni klasifikator



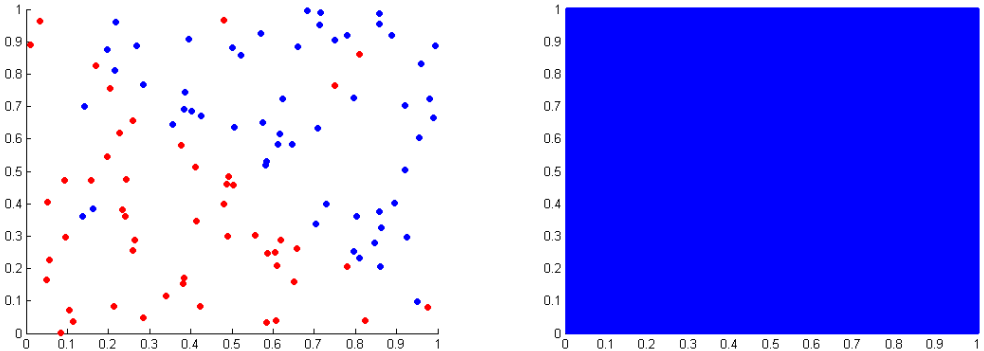
Slika: P. Janičić, M. Nikolić, Veštačka inteligencija, u pripremi.

Primer regularizacije – regularizovani polinomijalni klasifikator



Slika: P. Janičić, M. Nikolić, Veštačka inteligencija, u pripremi.

Primer regularizacije – regularizovani polinomijalni klasifikator



Slika: P. Janičić, M. Nikolić, Veštačka inteligencija, u pripremi.

Pregled

Postavka problema nadgledanog učenja

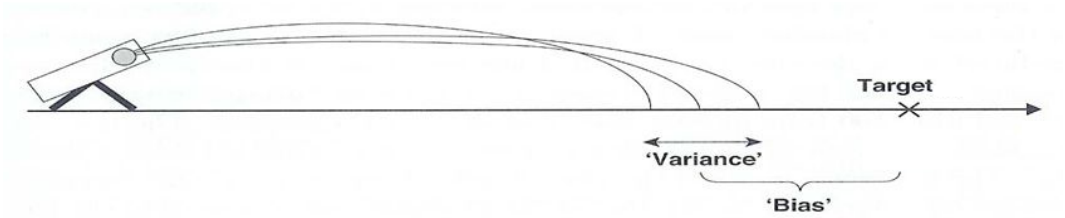
Princip minimizacije empirijskog rizika

Preprilagođavanje

Regularizacija

Nagodba izmedu sistematskog odstupanja i varijanse

Sistematsko odstupanje i varijansa



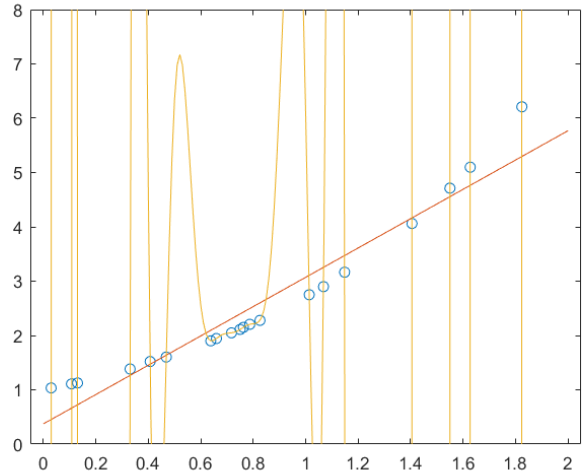
Slika: P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining. Modified.

Nagodba između sistematskog odstupanja i varijanse

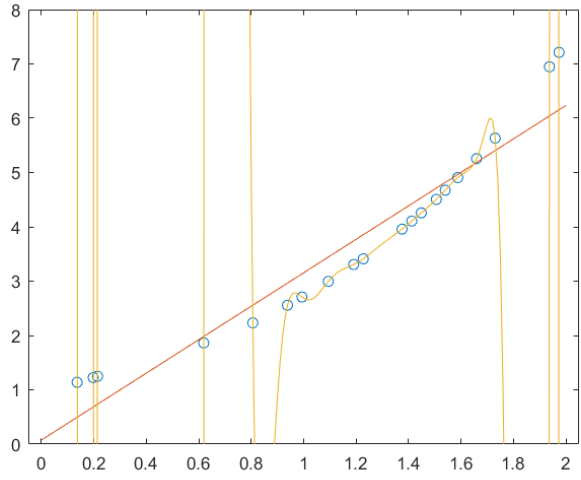
$$\begin{aligned}\mathbb{E}[(f_w(x) - r(x))^2] &= \mathbb{E}[(f_w(x) - \mathbb{E}[f_w(x)] + \mathbb{E}[f_w(x)] - r(x))^2] \\ &= \underbrace{(\mathbb{E}[f_w(x)] - r(x))^2}_{\textit{sistematsko odstupanje}^2} + \underbrace{\mathbb{E}[(f_w(x) - \mathbb{E}[f_w(x)])^2]}_{\textit{varijansa}}\end{aligned}$$

- gde je očekivanje po mogućim uzorcima podataka

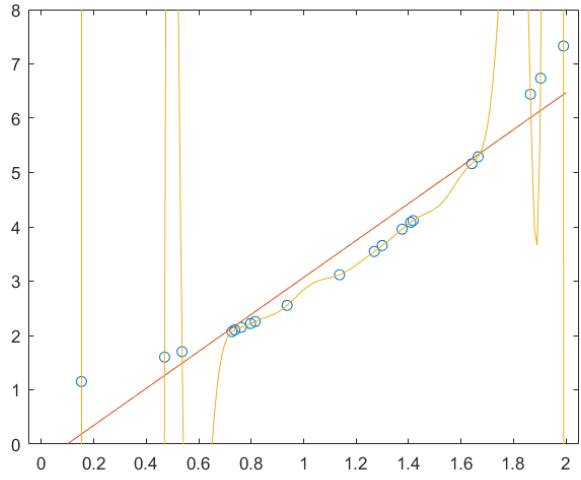
Sistematsko odstupanje i varijansa



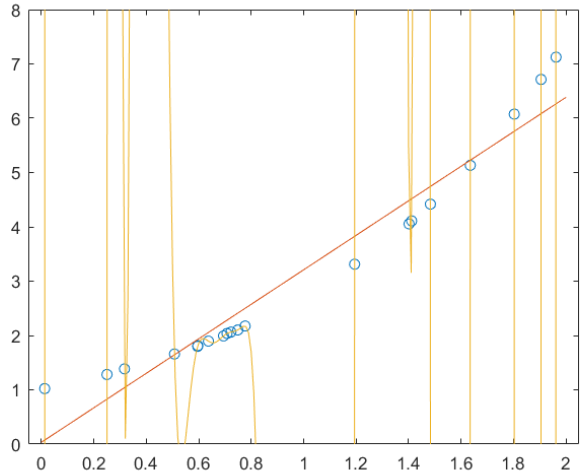
Sistematsko odstupanje i varijansa



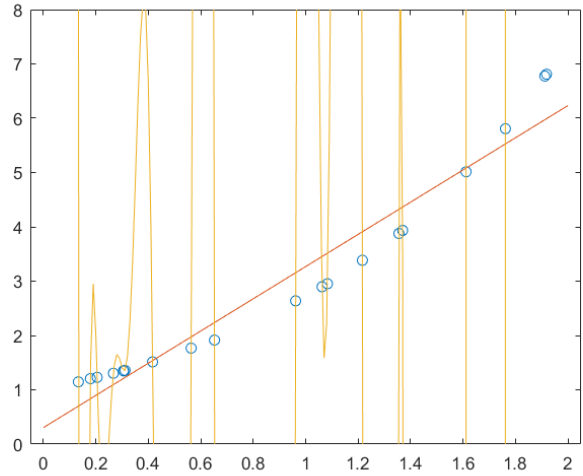
Sistematsko odstupanje i varijansa



Sistematsko odstupanje i varijansa



Sistematsko odstupanje i varijansa



Nagodba između sistematskog odstupanja i varijanse i uslovljenost

- U slučaju fleksibilnih modela, naučeni model dramatično varira u zavisnosti od nebitnih promena u podacima
- To znači visoku varijansu predviđanja, kao i da je problem učenja loše uslovljen

Nagodba između sistematskog odstupanja i varijanse i regularizacija

- Regularizacijom se menja optimizacioni problem, tako što se smanjuje fleksibilnost modela
- Ovim se unosi sistematsko odstupanje u rešenje, ali se varijansa smanjuje više nego što se sistematsko odstupanje povećava!
- To objašnjava smanjenje greške u slučaju regularizovanih modela

Nagodba između sistematskog odstupanja i varijanse