Projektovanje IoT sistema

Uvod u mašinsko učenje

Vladimir Rajović, prema

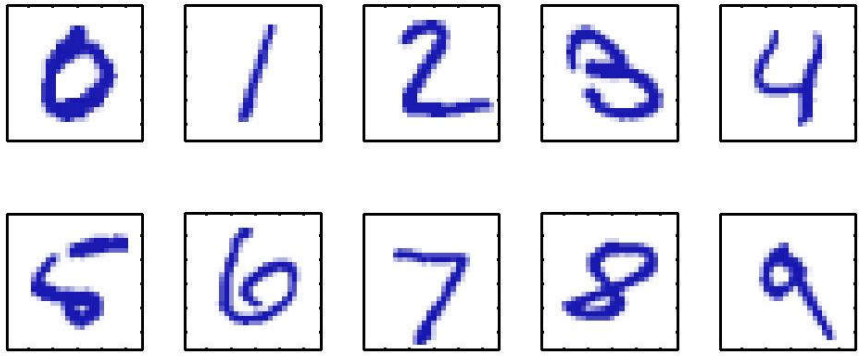
C.M.Bishop *Pattern Recognition and Machine Learning*

Uvod u mašinsko učenje

- Oblast istraživanja koja daje računarima mogućnost da uče bez da budu eksplicitno programirani
- Tehnička definicija: računarski problem uči na osnovu iskustva *E* u odnosu na neki zadatak *T* i neku meru performansi *P*, ako se njegove performanse u izvršavanju *T*, merene sa *P*, povećavaju sa iskustvom *E*.

Uvod u mašinsko učenje

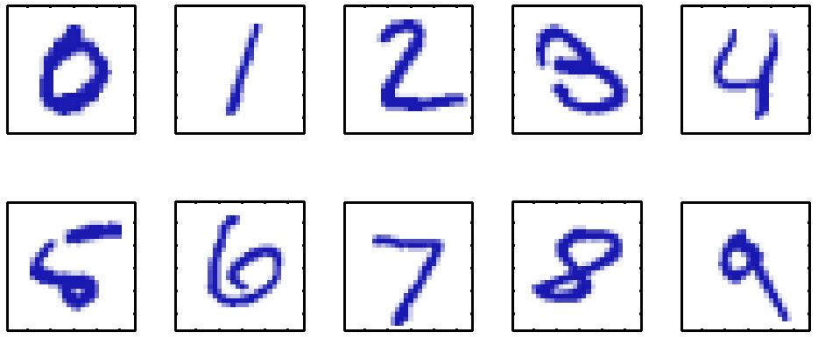
- Primer prepoznavanje pisanih cifara



  - Svaka cifra 28x28 tačaka, vektor $\mathbf{x}$ 784 realnih brojeva
  - Cilj napraviti mašinu koja će na osnovu ulaza $\mathbf{x}$ na izlazu dati koja je cifra u pitanju

Uvod u mašinsko učenje

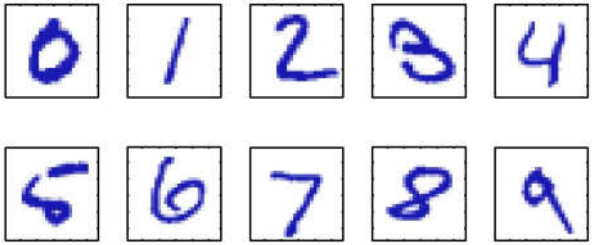
- Primer prepoznavanje pisanih cifara



  - Pristup mašinskog učenja – koristi se veliki skup *N* cifara $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ kao obučavajući skup za podešavanje parametara adaptivnog modela.
  - Kategorije u obučavajućem skupu su unapred poznate.
    - Kategorije se označavaju ciljnim vektorom $\mathbf{t}$ koji predstavlja indentitete odgovarajućih cifara

Uvod u mašinsko učenje

- Primer prepoznavanje pisanih cifara



  - Rezultat algoritma mašinskog učenja može da se predstavi kao fja $\mathbf{y}(\mathbf{x})$ koja se određuje tokom faze obučavanja (učenja) na osnovu obučavajućih podataka.
  - Nakon što je model obučen, može određivati identitet novih slika cifara, koje predstavljaju skup za testiranje
  - Ova mogućnost se naziva *uopštavanje*.
    - U praktičnim primenama obučavajući skup je samo mali podskup svih mogućih podataka i uopštavanje je glavni cilj.

Uvod u mašinsko učenje

- U najvećem broju primena ulazni podaci se *predobrađuju*
  - Transformacija u novi prostor podataka u kome će biti lakše ispuniti zadatak
    - U primeru sa prepoznavanjem cifara ulazni podaci se transliraju i skaliraju tako da je svaka cifra u pravougaoniku iste veličine – to drastično umanjuje različitost u okviru klasa cifara, tako da je algoritmima za prepoznavanje lakše da razlikuju različite cifre.
      - Ovo se nekad naziva *izdvajanje odlika* (feature extraction)

Uvod u mašinsko učenje

- U najvećem broju primena ulazni podaci se *predobrađuju*
  - Predobrada se koristi i za ubrzavanje izračunavanja, primer detekcije lica u sekvenci slika visoke rezolucije. Direktan rad sa velikom količinom podataka je problematičan.
    - Pronalaze se korisne odlike koje se lako računaju, a čuvaju korisne informacije za izdvajanje lica, koje se potom koriste u samom algoritmu.
    - Broj takvih odlika je manji od broja tačaka -> ovo predstavlja oblik smanjenja broja dimenzija.
    - Pažljivo, da ne dođe do gubitka bitnih informacija.

Uvod u mašinsko učenje

- Nadgledano učenje
  - Obučavajući skup sadrži primere ulaznih podataka sa njihovim ciljnim vektorima (labelama, oznakama).
    - Ako je skup mogućih izlaza konačan u pitanju je *klasifikacija* (primer prepoznavanje cifara)
    - Ako je beskonačan u pitanju je *regresija* (primer predviđanje prinosa hemijskog postrojenja na osnovu koncentracije reagenasa, temperature i pritiska).

Uvod u mašinsko učenje

- Nenadgledano učenje
  - Obučavajući skup sadrži primere ulaznih podataka ali bez njihovih oznaka.
    - Grupisanje (*clustering*) je pronalaženje grupe sličnih podataka u skupu
    - Određivanje raspodele podataka (procena gustine, d*ensity estimation*)
    - Vizuelizacija – transformacija podataka iz više dimenzija u dve ili tri

Uvod u mašinsko učenje

- Učenje podsticanjem (*reinforcement*)
  - Naći odgovarajuće postupke u datoj situaciji kako bi se maksimimizirala nagrada
  - Algoritam ne dobija primere optimalnog izlaza, već ih sam nalazi pokušajima i pogreškama
  - Postoji sekvenca stanja i postupaka kojima algoritam deluje na okolinu.
    - Trenutna akcija najčešće ima uticaj na sva naredna stanja, ne samo na sledeće

Uvod u mašinsko učenje

- Učenje podsticanjem (*reinforcement*)
  - Kod ove vrste učenja pravi se kompromis između korišćenja (*exploitation* – korišćenje dokazanih akcija) i probanja (*exploration* – proba novih akcija)

Uvod u mašinsko učenje

- Razni zadaci mašinskog učenja imaju svoje metode i alate, ali postoje koncepti koji su im zajednički

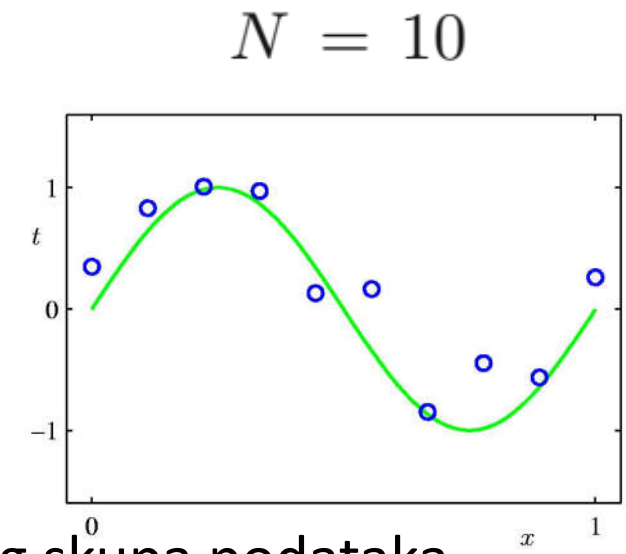
Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Jednostavan regresioni problem: ulazna realna promenljiva x, na osnovu koje se predviđa vrednost izlaza t
    - Veštački primer, jer se zna tačno proces koji je generisao podatke: $\sin(2\pi x)$ uz dodat slučajni šum
    - Obučavajući skup od N obzervacija (merenja, uzoraka, odbiraka) x, $\mathbf{x} \equiv (x_1, \ldots, x_N)^{\mathrm{T}}$ sa odgovarajućim obzervacijama t, $\mathbf{t} \equiv (t_1, \ldots, t_N)^{\mathrm{T}}$

Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Predvideti $\widehat{t}$ za neko $\widehat{x}$
  - Implicitno, prepoznati

$$\sin(2\pi x)$$

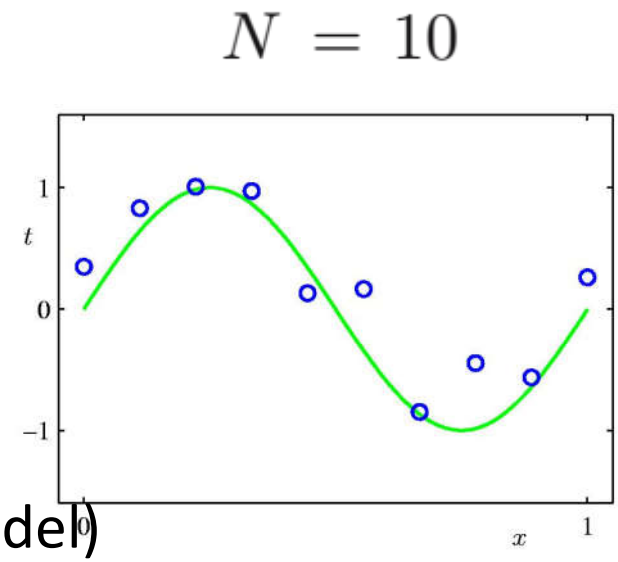


  - Problemi:
    - Uopštavanje na osnovu konačnog skupa podataka
    - Podaci zagađeni šumom, tako da postoji nesigurnost
      - Teorija verovatnoće daje okvir za merenje te nesigurnosti, a teorija odlučivanja koristi ta merenja za optimalna predviđanja prema nekom kriterijumu

Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Jednostavan primer fitovanja

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



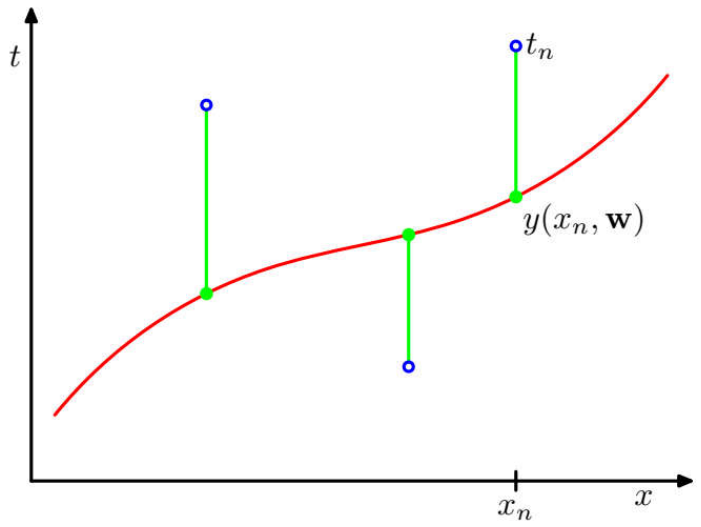
    - Fja nelinearna po x, ali linearna po koeficijentima $\mathbf{w}$ (linearni model)
  - Koeficijenti se određuju fitovanjem polinoma, npr minimizacijom funkcije greške između $y(x, \mathbf{w})$ i obučavajućeg skupa

Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Primer fje greške

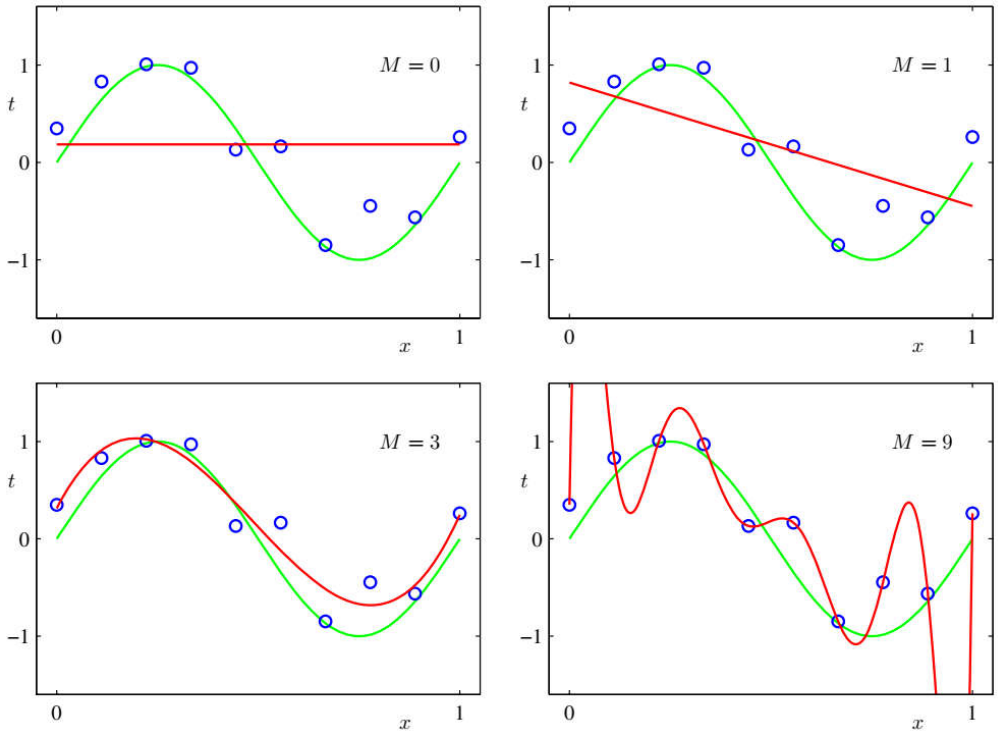
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

    - Nula samo ako predviđanje tačno prolazi kroz obučavajuće podatke
  - Problem se rešava nalaženjem $\mathbf{w}^\star$ za koje je greška najmanja moguća
  - Moguće je u zatvorenoj formi, greška zavisi kvadratno od koeficijenata, lako rešavanje



Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Mnogo veći problem naći red polinoma M
    - preveliko fitovanje za $M = 9$, fja prolazi kroz sve tačke, ali loše predstavlja pravu fju

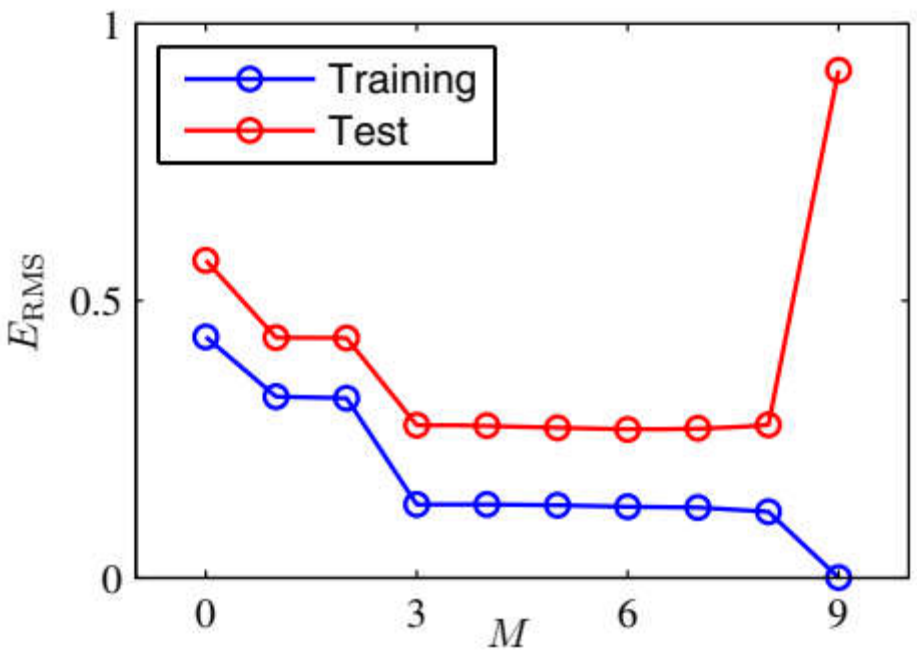


Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Skup za testiranje od 100 članova

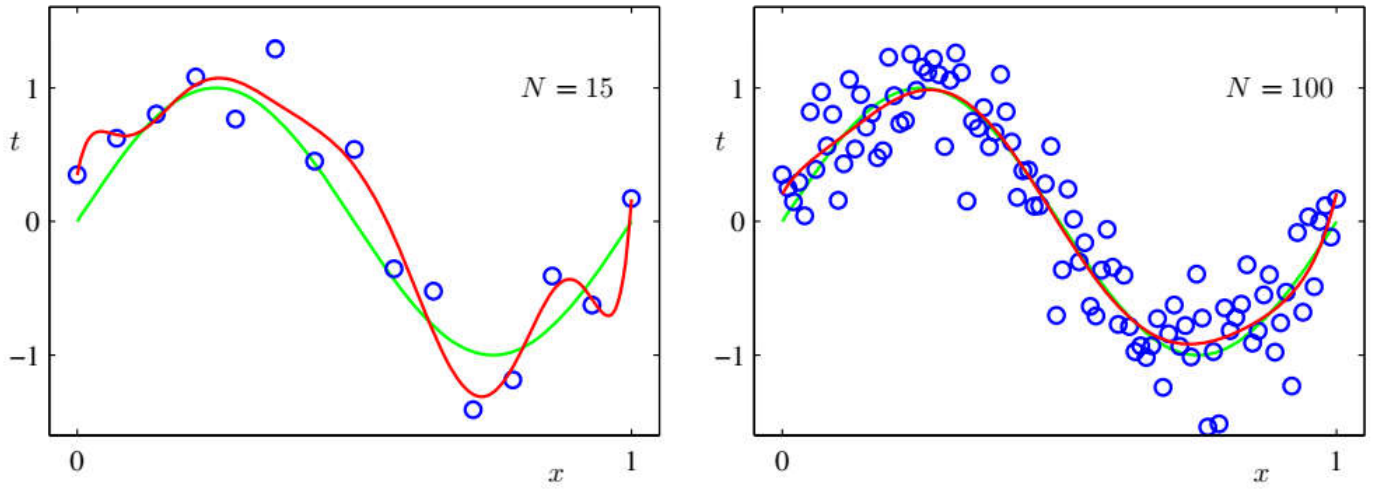
$$E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$$

  - Greška na skupu za testiranje dobra mera koliko je dobro predviđanje
  - Za $M = 9$ polinom postaje podešen na šum!



Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Kako veličina obučavajućeg skupa utiče na fitovanje, primer za $M = 9$



  - Za datu složenost modela, prefitovanje postaje manje ukoliko ima više obučavajućih podataka

Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Empirija kaže da broj podataka treba da bude umnožak (5-10) broja adaptivnih parametara
  - Ograničenje broja parametara u zavisnosti od veličine obučavajućeg skupa je nezadovoljavajuće, razumnije je da kompleksnost modela zavisi od komplesnosti problema (ovo rešava bajesovski pristup)

Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Da bi koristili složene modele sa ograničenim skupom obučavajućih podataka, jedna od tehnika je *regularizacija*, dodavanje kaznenog člana funkciji greške kako bi se sprečilo da koeficijenti dobijaju prevelike vrednosti i tako se podešavaju prema šumu.
    - Najprostiji primer

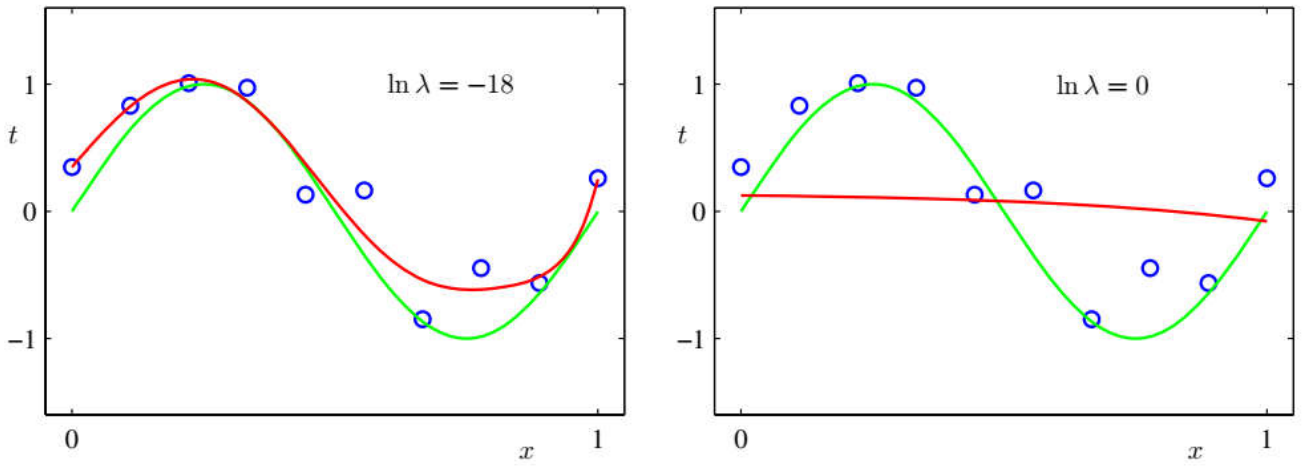
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^{\mathrm{T}}\mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2$$

      - » Koeficijent λ definiše relativnu važnost regularizacionog člana

Uvod u mašinsko učenje

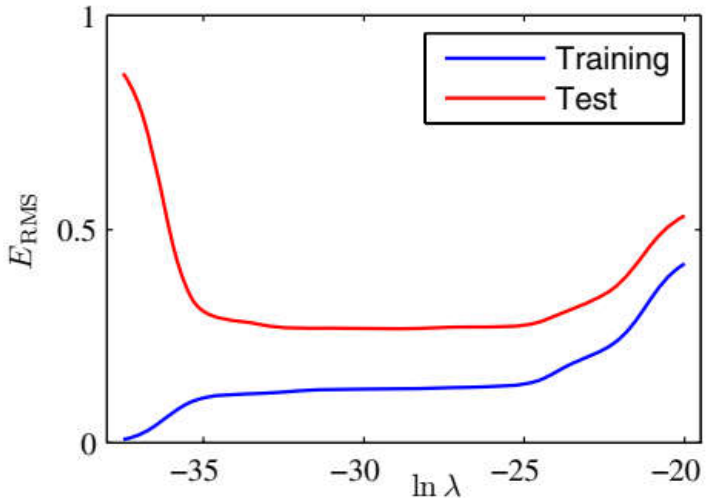
- Fitovanje krive polinomom
  - Regularizacija, $M = 9$, $N = 10$



  - Prevelika regularizacija uzrokuje loše fitovanje

Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Regularizacija, $M = 9,\ N = 10$



  - Efektivna složenost modela i količina prefitovanja kontrolisana koeficijentom λ

Uvod u mašinsko učenje

- Fitovanje krive polinomom
  - Ako bi se konkretan problem rešavao minimizacijom funkcije greške, potrebno je naći odgovarajuću vrednost složenosti modela.
    - Jednostavan način je deljenje dostupnih podataka na obučavajući i validacioni skup
      - » Obučavajući skup se koristi za određivanje koeficijenata
      - » Validacioni skup za optimizaciju složenosti
    - U mnogim primenama je ovo traćenje vrednih podataka, potrebno je nešto pametnije...

Uvod u mašinsko učenje

- Teorija verovatnoće
  - Pravila

$$p(X) = \sum_{Y} p(X, Y)$$

$$p(X, Y) = p(Y|X)p(X)$$

    - $p(X, Y)$ zajednička verovatnoća
    - $p(Y|X)$ uslovna verovatnoća
    - $p(X)$ marginalna verovatnoća
  - Ova dva pravila osnova za dalja razmatranja

Uvod u mašinsko učenje

- Teorija verovatnoće
  - Posledice pravila

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$ Bajesova teorema

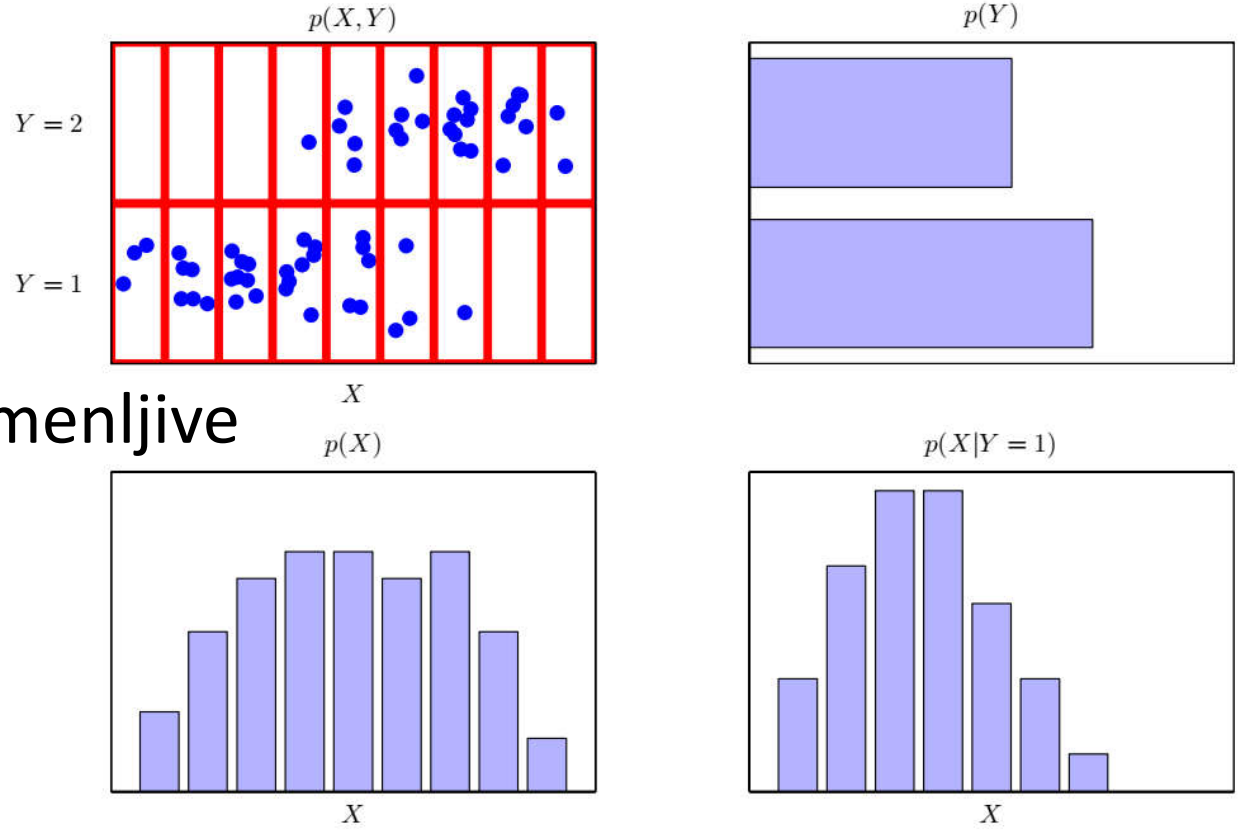
$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

Uvod u mašinsko učenje

- Teorija verovatnoće
  - Ako važi

$$p(X,Y) = p(X)p(Y)$$

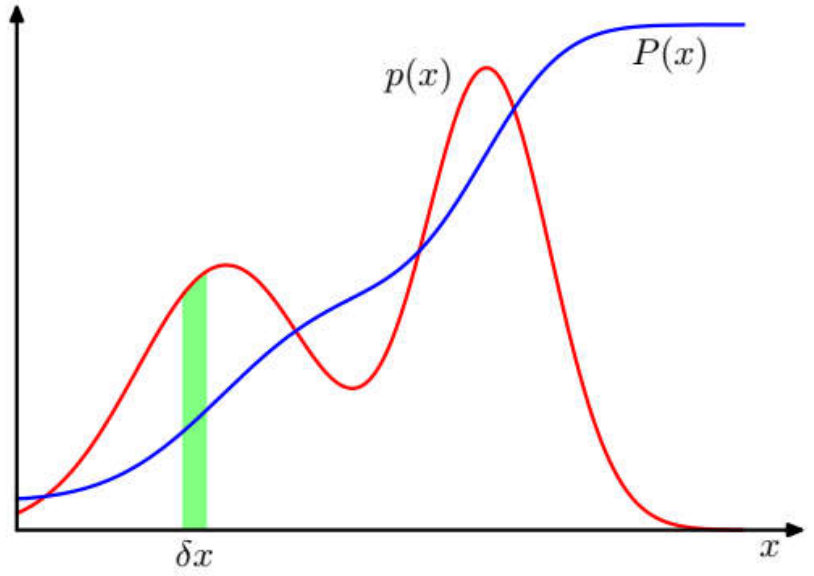
slučajne promenljive su nezavisne



Uvod u mašinsko učenje

- Teorija verovatnoće
  - Ako su u putanju kontinualne slučajne promenljive $p(x)$ se naziva gustina verovatnoće

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$



  - gustina mora da zadovoljava uslove

$$\begin{aligned} p(x) &\geqslant 0 \\ \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x &= 1 \end{aligned}$$

  - Kumulativna raspodela verovatnoće

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

Uvod u mašinsko učenje

- Teorija verovatnoće
  - U slučaju više kontinualnih promenljivih $x_1, \ldots, x_D$ (označavamo kao vektor $\mathbf{x}$), zajednička raspodela gustine $p(\mathbf{x}) = p(x_1, \ldots, x_D)$, za koju takođe mora da važi

$$\begin{aligned} p(\mathbf{x}) &\geqslant 0 \\ \int p(\mathbf{x})\,\mathrm{d}\mathbf{x} &= 1 \end{aligned}$$

  - Važe pravila

$$\begin{aligned} p(x) &= \int p(x,y)\,\mathrm{d}y \\ p(x,y) &= p(y|x)p(x). \end{aligned}$$

  - Ekvivalentno za kombinaciju diskretnih i kontinualnih promenljivih

Uvod u mašinsko učenje

- Teorija verovatnoće
  - Usrednjena vrednost fje po njenoj raspodeli *očekivanje* fje

$$\mathbb{E}[f] = \sum_x p(x) f(x) \qquad \mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x.$$

    - Ako imamo *N* podataka aproksimacija

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

  - Uslovno učekivanje $\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$

Uvod u mašinsko učenje

- Teorija verovatnoće
  - *Varijansa* f-je pokazuje koliko je promenljivosti u fji oko njene srednje vrednosti

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$$

    - Varijansa same promenljive $\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$
  - Za dve slučajne promenljive *kovarijansa* pokazuje koliko se dve promenljive zajedno menjaju

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

    - Ako su nezavisne, kovarijansa nestaje

Uvod u mašinsko učenje

- Teorija verovatnoće
  - U slučaju dva vektora slučajnih promenljivih, kovarijansa

$$\begin{aligned} \mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]. \end{aligned}$$

Uvod u mašinsko učenje

- Teorija verovatnoće
  - **Bajesovska verovatnoća**
    - Klasična ili frekvencijska verovatnoća se bavi učestanostima slučajnih ponovljivih događaja.
    - Kod Bajevskog pristupa verovatnoće daju meru nesigurnosti.
  - Da li će se polarna kapa otopiti do kraja veka?
    - Događaj koji se ne ponavlja
      - Postoji ideja o tome koliko brzo se led topi
      - Novi podaci dovode do revidiranja predviđanja
        - » Preduzimaju se neke akcije kako bi se topljenje leda smanjilo
  - Potrebno je imati način da se izrazi nesigurnost, da se ona revidira kada imamo nove podatke, i da se kao posledica donesu neke odluke i preduzmu neke akcije
    - Bajesova interpretacija verovatnoće to omogućava

Uvod u mašinsko učenje

- Teorija verovatnoće
  - **Bajesovska verovatnoća**
    - Kod fitovanja krivih klasična verovatnoća ima smisla kod slučajnih vrednosti podataka koje merimo/dobijamo
      - Ali NE pomaže kod nesigurnosti oko parametara modela (koeficijenata) ili samog modela
  - Bajesova teorema konvertuje prethodnu (a prori) verovatnoću u sledeću (a posteriori) na osnovu novih podataka

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

    - Slično se može koristiti za zaključivanje o parametrima kao što su koeficijenti kod fitovanja krivih

Uvod u mašinsko učenje

- Teorija verovatnoće
  - **Bajesovska verovatnoća**
    - Pretpostavke o koeficijentima su prethodna verovatnoća $p(\mathbf{w})$
    - Uslovna verovatnoća $p(\mathcal{D}|\mathbf{w})$ uključuje efekat dobijenih podataka $\mathcal{D} = \{t_1, \ldots, t_N\}$
    - Procena nesigurnosti koeficijenata predstavljena je narednom verovatnoćom $p(\mathbf{w}|\mathcal{D})$, korišćenjem Bajesove teoreme

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

Uvod u mašinsko učenje

- Teorija verovatnoće
  - **Bajesovska verovatnoća**

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

    - $p(\mathcal{D}|\mathbf{w})$ se može posmatrati kao fja parametara, i tada se naziva fja verovatnoće (likelihood function).
      - Koliko je verovatan izmereni skup podataka za date parametre. Ovo nije distribucija verovatnoće na $\mathbf{w}$, integral po $\mathbf{w}$ ne mora biti 1.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

      - Sve veličine funkcije $\mathbf{w}$

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}$$

Uvod u mašinsko učenje

- Teorija verovatnoće
  - **Bajesovska verovatnoća**

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

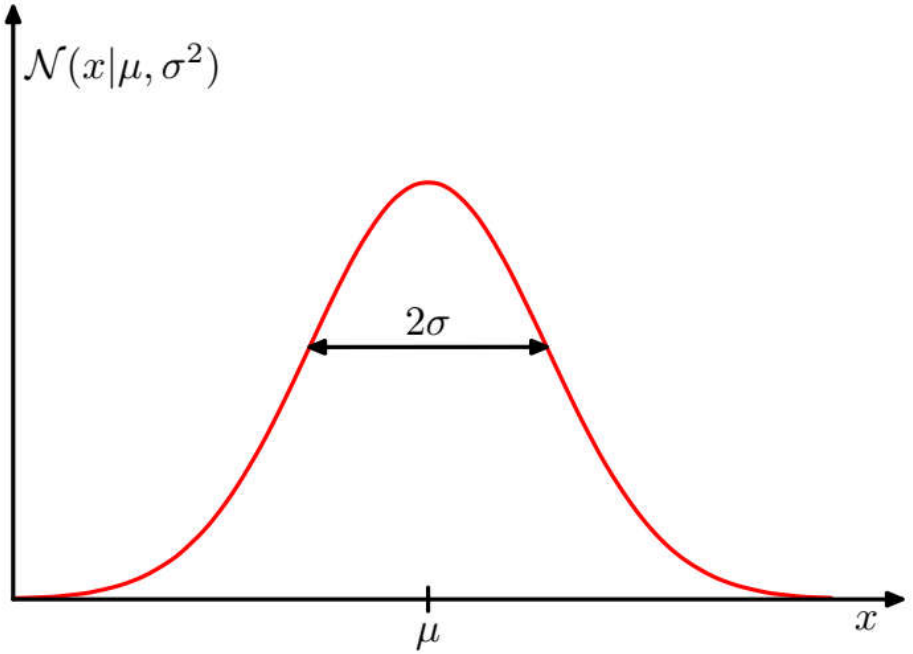
    - $p(\mathcal{D}|\mathbf{w})$ ima centralno mesto u oba pristupa verovatnoći, ali se koristi na različite načine
      - U klasičnom pristupu $\mathbf{w}$ je fiksni parametar određen nekim estimatorom, i greške se razmatraju na svim mogućim skupovima podataka.
      - U Bajesovom pristupu postoji samo jedan skup podataka (koji se trenutno posmatra), a nesigurnost parametara se izražava preko distribucije verovatnoće parametara $\mathbf{w}$

Uvod u mašinsko učenje

- Teorija verovatnoće
  - **Gausova raspodela** $\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$

    μ srednja vrednost, σ^2 varijansa

    σ standardna devijacija, $\beta = 1/\sigma^2$ preciznost



Uvod u mašinsko učenje

- Teorija verovatnoće
  - **Gausova raspodela**
    - Vektor *D* slučajnih promenljivih

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

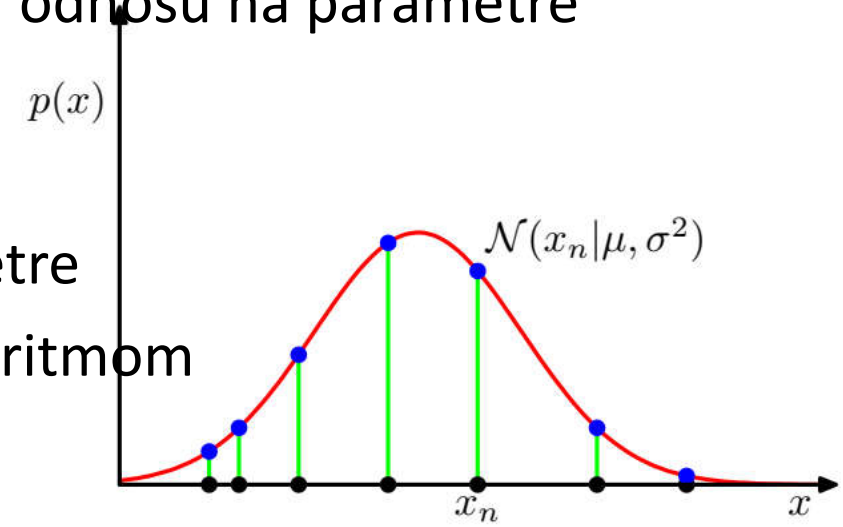
  - Skup opažanja skalarne promenljive $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$
    - Dobijena nezavisnim opažanjima iz Gausove raspodele nepoznatih parametara
      - Namera je odrediti parametre na osnovu podataka
    - Verovatnoća skupa opažanja $p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N}\mathcal{N}\left(x_n|\mu,\sigma^2\right)$

Uvod u mašinsko učenje

- Teorija verovatnoće
  - Skup opažanja skalarne promenljive

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

- Ovo je funkcija verovatnoće (likelihood) za Gausovu distribuciju kada se gleda u odnosu na parametre
- Maksimizacijom ove fje moguće je odrediti parametre
- U praksi lakše raditi sa logaritmom



Uvod u mašinsko učenje

- Teorija verovatnoće
  - U praksi lakše raditi sa logaritmom

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

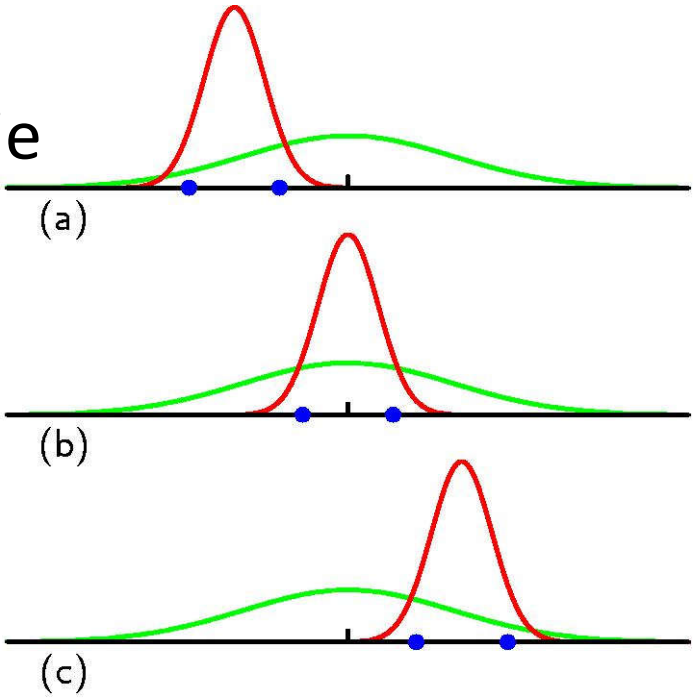
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

  - Ovaj pristup sistemski potcenjuje varijansu (pomeraj/bias)

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \quad \mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

  - Procena varijanse bez pomeraja

$$\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$



Uvod u mašinsko učenje

- Fitovanje krive 2
  - Iz perspektive verovatnoće
  - Cilj predvideti vrednost izlazne promenljive *t* na osnovu nove vrednosti ulazne promenljive *x* i skupa obučavajućih podataka od *N* opažanja i njihovih oznaka, $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$ i $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$
    - Za dato *x*, *t* ima Gausovu raspodelu (pp) sa srednjom vrednošću $y(x, \mathbf{w})$

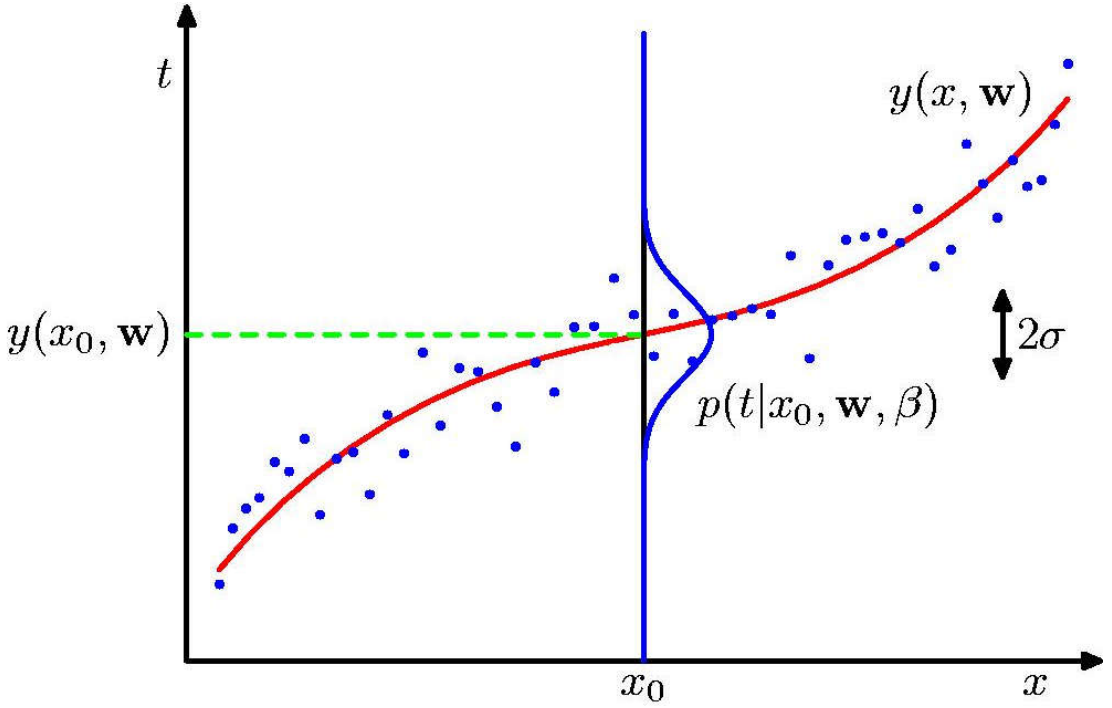
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$$

Uvod u mašinsko učenje

- Fitovanje krive 2

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$$



Uvod u mašinsko učenje

- Fitovanje krive 2
  - Koriste se obučavajući podaci $\{\mathbf{x}, \mathbf{t}\}$ radi maksimizacije fje verovatnoće, odnosno njenog logaritma

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right)$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2$$

Uvod u mašinsko učenje

- Fitovanje krive 2
  - Sada je moguće praviti predviđanja kada dobijemo nove vrednosti *x*
    - Kako sada imamo probabilistički model, predviđanja se izražavaju kao prediktivna distribucija koja daje distribuciju verovatnoće *t*, a ne jednu tačku

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$

Uvod u mašinsko učenje

- Fitovanje krive 2
  - Za koeficijente, uvodimo a priori raspodelu za koeficijente polinoma, neka bude Gausova sa preciznošću α i *M*+1 koeficijenata

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

  - A posteriori raspodela proporcionalna a priori raspodeli i funkciji verovatnoće

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha).$$

Uvod u mašinsko učenje

- Fitovanje krive 2
  - Sada se može naći najverovatnija vrednost koeficijenata maksimiziranjem a posteriori raspodele (tehnika *maximum posterior – MAP*), što je minimum izraza

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

    - Ovo je već viđeno, kod regularizacije prilikom fitovanja krive, sa parametrom regularizacije $\lambda=\alpha/\beta$

Uvod u mašinsko učenje

- Bajesovo fitovanje
  - Pretpostavimo da su parametri raspodele poznati i fiksni i evaluiramo raspodelu

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{w}$$

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$$

  - $p(\mathbf{w}|\mathbf{x}, \mathbf{t})$ a posteriori distribucija parametara, dobija se normalizacijom $p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$
    - Gausova i može se analitički odrediti

Uvod u mašinsko učenje

- Bajesovo fitovanje
  - Dobija se

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)$$

$$m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x).$$

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^{\mathrm{T}}$$

$$\boldsymbol{\phi}(x_n) = \left(x_n^0, \ldots, x_n^M\right)^{\mathrm{T}}$$

Uvod u mašinsko učenje

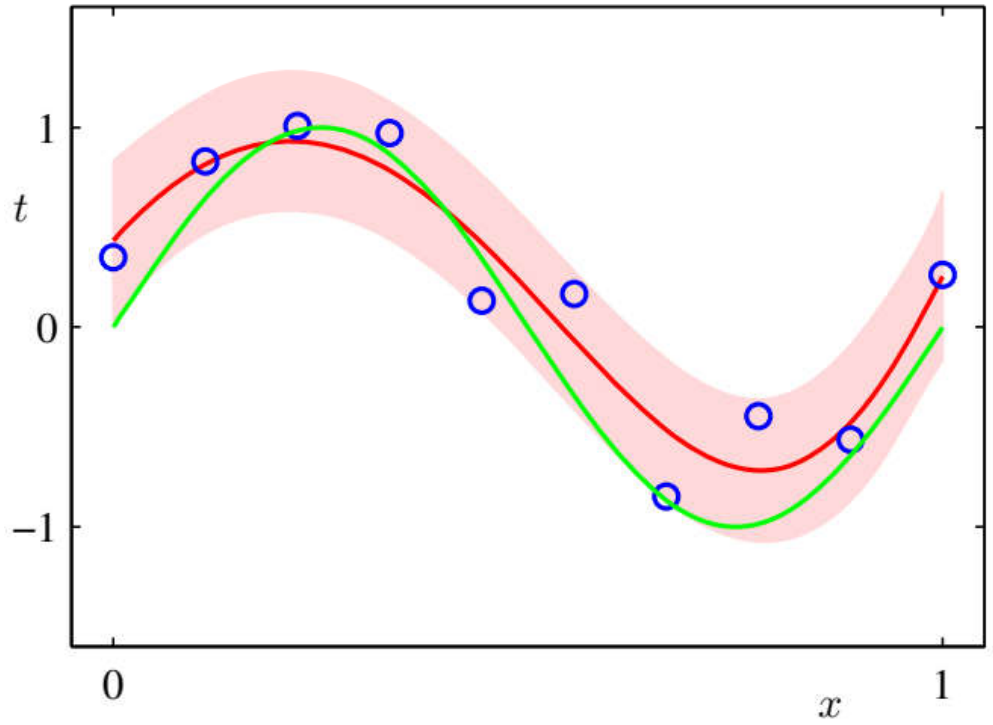
- Bajesovo fitovanje
  - Varijansa

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x).$$

  - Prvi član posledica šuma u podacima, drugi nesigurnost parametara

Uvod u mašinsko učenje

- Bajesovo fitovanje
  - Prediktivna raspodela $M = 9 \; \alpha = 5 \times 10^{-3} \; \beta = 11.1$



Uvod u mašinsko učenje

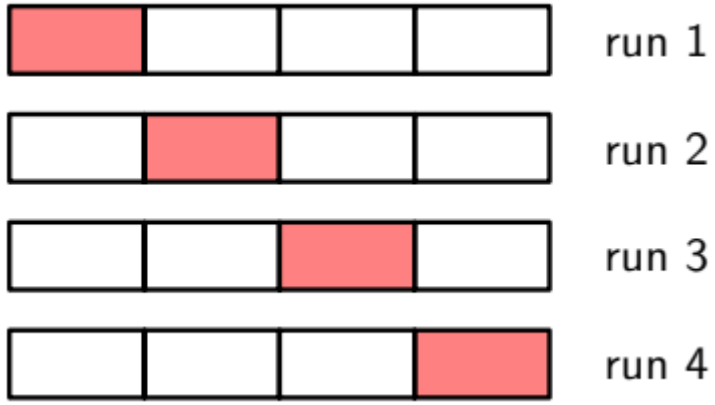
- **Izbor modela**
  - U praksi je potrebno odrediti parametre modela, sa ciljem da se postigne što bolje predviđanje novih podataka
  - Ali, nekad je potrebno i izabrati najbolji model za datu primenu

Uvod u mašinsko učenje

- **Izbor modela**
  - Učinak predviđanja na obučavajućem skupu nije dobar pokazatelj kakvo je predviđanje, zbog prefitovanja (već viđeno)
  - Ako ima dosta podataka: neke podatke iskoristiti kao obučavajući skup za više modela, ili za više parametara jednog modela, i onda koristiti validacioni skup za poređenje i izabrati najbolji
  - Može doći i do prefitovanja sa validacionim skupom, tako da se ponekad izdvaja treći skup, za testiranje i procenu izabranog modela.

Uvod u mašinsko učenje

- **Izbor modela**
  - Često podataka nema puno, i namera je da se što više podataka iskoristi za obučavanje.
    - Ali ako je validacioni skup mali, to dovodi do loše procene učinka predviđanja
  - Zato se može koristiti S-to struka unakrsna validacija
    - Podaci se dele na *S* grupa, (*S*-1) grupa se koristi za obučavanje, a onda evaluira na preostaloj grupi. Usrednjavanje rezultata
      - Ako ima malo podatata, onda *S*=*N* (leave one out)

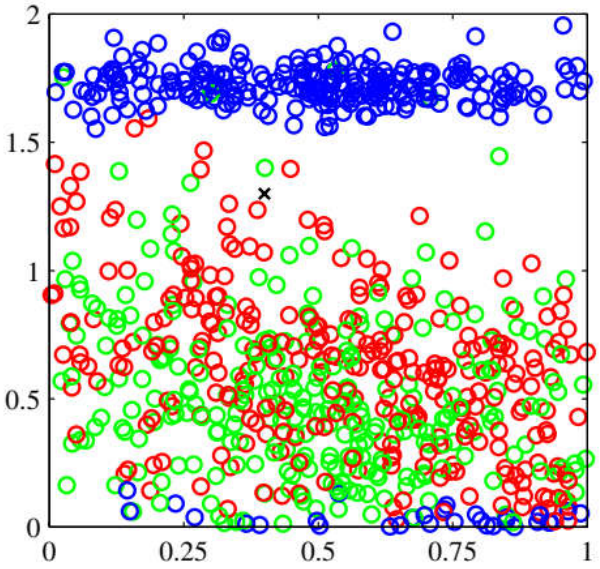


Uvod u mašinsko učenje

- **Izbor modela**
  - Problem S-tostruke unakrsne validacije je što broj prolazaka raste sa *S* ako je obučavanje računarski zahtevno
  - Ako u modelu imamo više parametara, onda broj prolazaka eksponencijalno zavisi od broja parametara
  - Potreban pristup u kome se idealno u jednom prolasku porede modeli i parametri na osnovu podataka za obučavanje, i treba naći meru učinka koja zavisi samo od podataka za obuku i ne trpi zbog pomeraja usled prefitovanja, npr maksimizacija $\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M$
    (fja verovatnoće i broj parametara)

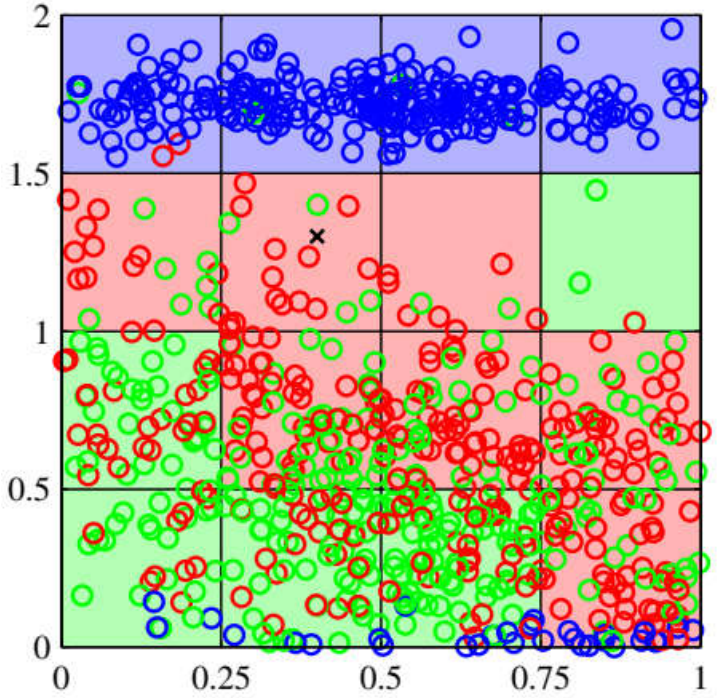
Uvod u mašinsko učenje

- **Više dimenzija**
  - U praksi se često radi sa višedimenzionim podacima – ozbiljan izazov
  - Primer, iz problema koji ima 12 ulaznih veličina i tri moguća izlaza, posmatraju se sada dva ulaza
    - Novi podatak okružen crvenim, ali ima prilično i zelenih – intuicija



Uvod u mašinsko učenje

- **Više dimenzija**
  - Kako intuiciju preneti u algoritam?
    - Jedan pristup je da se ulazni prostor podeli na ćelije
      - Ulazni podatak je istog tipa kao i većina podataka u ćeliji u kojoj se nalazi
    - Najočigledniji problem sa ovim pristupom je povećanje broja dimenzija ulaza, koji uzrokuje exp porast broja ćelija
    - Potrebno je exp više podataka kako bi osigurali da nema praznih ćelija



Uvod u mašinsko učenje

- **Više dimenzija**
  - Fitovanje krive polinomom III reda u slučaju *D* ulaznih promenljivih

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

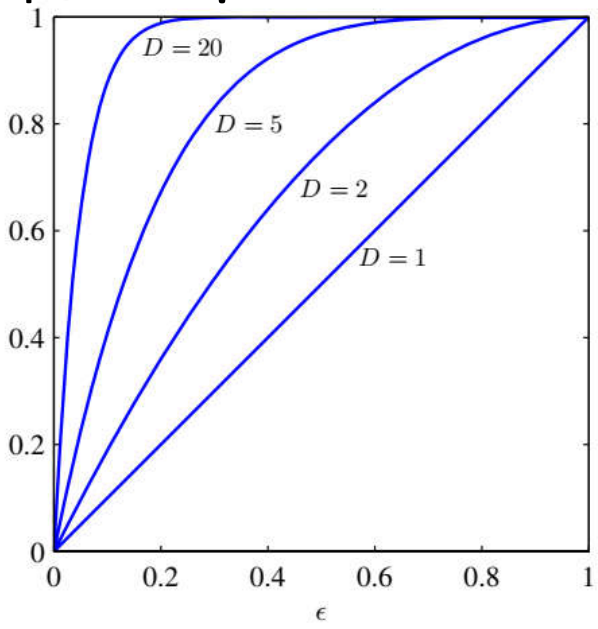
  - Kako raste *D*, broj nepoznatih koeficijenata raste sa D^3
  - Za polinomom reda *M* raste sa D^M
    - Nije eksponencijalna zavisnost, ali je opet praktična upotrebljivost metoda dovedena u pitanje

Uvod u mašinsko učenje

- **Više dimenzija**
  - Geometrijska intucija, zasnovana na 3D, pada u vodu na više dimenzija
    - *D*-dimenzionalna sfera poluprečnika *r* = 1; koji deo zapremine sfere leži između $r = 1-\epsilon$ i $r = 1$?

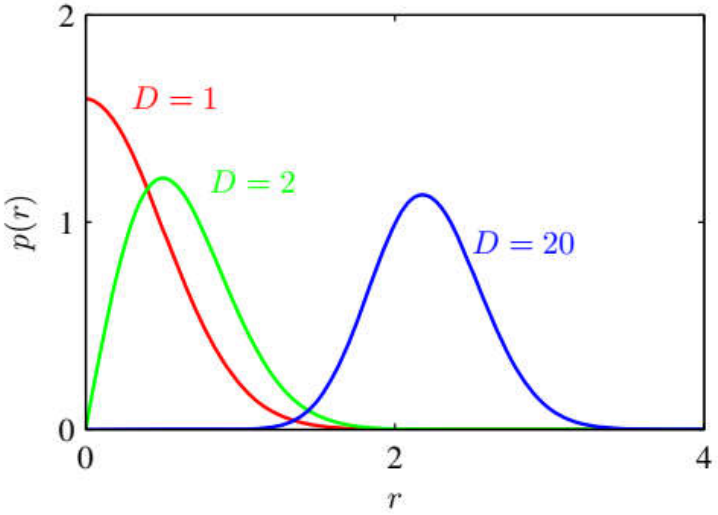
$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$

      - U više dimenzija, veći deo zapremine sfere se nalazi uz spoljašnju površinu!



Uvod u mašinsko učenje

- **Više dimenzija**
  - Višedimenziona Gausova raspodela
    - Prelazak na polarne koordinate i integracija po svim uglovima tako da ostane samo zavisnost od radijusa
    - $p(r)\delta r$ je verovatnoća u uskoj ljusci na radijusu *r*
      - Skoncentrisana verovatnoća



Uvod u mašinsko učenje

- **Više dimenzija**
  - Ne mora biti problem
    - Realni podaci će često biti ograničeni na deo prostora koji ima manje dimenzija, dok su nekad i pravci promena bitnih promenljivih takođe ograničeni
    - Realni podaci često imaju osobine glatkosti (barek lokalno), tj male promene ulaza uzrokuju male promene izlaza -> tada se mogu koristiti tehnike slične lokalnoj interpolaciji za predviđanja promene izlaznih promenljivih usled promene ulaznih promenljivih.
      - Primer za oba, identični objekti na pokretnoj traci, različito postavljeni. Svaka slika je tačka u u prostoru koji ima dimenzija koliko ima piksela. Postoje tri stepena slobode (pozicija na slici i orijentacija) između slika, ali je samo jedan bitan - orijentacija.

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - Vektor ulaza $\mathbf{x}$ i vektor izlaza $\mathbf{t}$, cilj je predvideti izlaz $\mathbf{t}$ za novi ulaz $\mathbf{x}$
  - Zajednička raspodela $p(\mathbf{x}, \mathbf{t})$ daje nesigurnost koja prati ove promenljive
  - Određivanje raspodele na osnovu podataka za obučavanje je primer *zaključivanja*
    - Praktično je potrebno predvideti $\mathbf{t}$ odnosno preduzeti neku akciju na osnovu vrednosti koju $\mathbf{t}$ verovatno ima – teorija odlučivanja

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - Rendgen snimak na osnovu koga treba odrediti da li pacijent ima rak.
    - Ulaz $\mathbf{x}$ skup intenziteta piksela, izlazna promenljiva t određuje da li pacijent ima rak (klasa $\mathcal{C}_1$) ili ne (klasa $\mathcal{C}_2$)
    - Opšti problem zaključivanja uključuje određivanje zajedničke raspodele $p(\mathbf{x}, \mathcal{C}_k)$ (tj $p(\mathbf{x}, t)$), koja nam daje probabilistički opis situacije
    - Ali, na kraju se mora odlučiti da li da se pacijentu uradi neka intervencija ili ne, i ovaj izbor bi trebalo da bude optimalan u nekom smislu
    - *Odluka* – predmet teorije odlučivanja je kako donositi optimalne odluke na osnovu verovatnoća
      - Kada se reši problem zaključivanja, odluka je u opštem slučaju trivijalna

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - Kako to verovatnoće imaju veze sa donošenjem odluka?
  - Kada se dobije snimak $\mathbf{x}$ novog pacijenta, potrebno je odrediti kojoj klasi snimak pripada
  - Bitne su verovatnoće $p(\mathcal{C}_k|\mathbf{x})$, koje određujemo preko Bajesove teoreme

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

  - Ako hoćemo da minimiziramo mogućnost da snimak loše interpretiramo, intuitivno biramo klasu koja ima veću a posteriornu verovatnoću

Uvod u mašinsko učenje

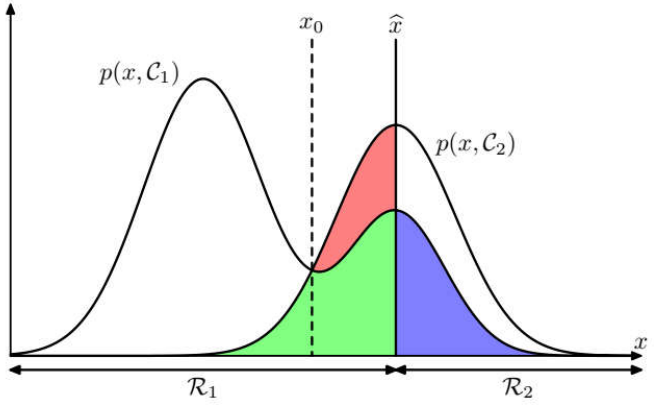
- **Teorija odlučivanja**
  - **Minimizacija loše klasifikacije**
    - Cilj je što manje moguće loših klasifikacija
      - Potrebno pravilo koje dodeljuje $\mathbf{x}$ nekoj od klasa/mogućnosti
      - Pravilo deli ulazni prostor na oblasti odlučivanja $\mathcal{R}_k$ koje odgovaraju klasama $\mathcal{C}_k$ i ne moraju biti neprekidne
        - » Granice između oblasti *granice* ili *površi odlučivanja*
      - Primer sa dve klase
        - » Greška se javlja kada se ulazni vektor dodeli pogrešnoj klasi, sa verovatnoćom

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x} \end{aligned}$$

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Minimizacija loše klasifikacije**
    - Da bi se minimizirala greška, pravilo odlučivanja: ako

$p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2)$ za neki ulaz, onda taj ulaz dodeljujemo klasi $\mathcal{C}_1$, drugim rečima ulaz dodeljujemo klasi za koju je $p(\mathcal{C}_k|\mathbf{x})$ veće



Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Minimizacija loše klasifikacije**
    - Ako imamo *K* klasa, nešto je jednostavnije maksimizirati verovatnoću tačne klasifikacije

$$p(\text{correct}) = \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k)$$
$$= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k)\, \mathrm{d}\mathbf{x}$$

    - Odlučivanje ekvivalentno, ulaz se dodeljuje klasi koja ima najveće $p(\mathcal{C}_k|\mathbf{x})$

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Minimizacija očekivanog gubitka**
    - Nije svaka greška ista, zato se uvodi pojam fje gubitka (loss) ili troška (cost) koja daje meru gubitka prilikom bilo koje od mogućih odluka

$$\begin{array}{c} \\ \text{cancer} \\ \text{normal} \end{array} \begin{array}{c} \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

    - Nekom ulazu $\mathbf{x}$ čija je klasa $\mathcal{C}_k$ dodeljujemo klasu $\mathcal{C}_j$ sa gubitkom L_{kj}
    - Optimalno rešenje minimizuje fju gubitka
      - Ona zavisi od prave klase, koja je nepoznata

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Minimizacija očekivanog gubitka**
    - Optimalno rešenje minimizuje fju gubitka
      - Ona zavisi od prave klase, koja je nepoznata
        - » Ta nesigurnost je opisana sa $p(\mathbf{x}, \mathcal{C}_k)$
        - » Minimizira se usrednjeni gubitak

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

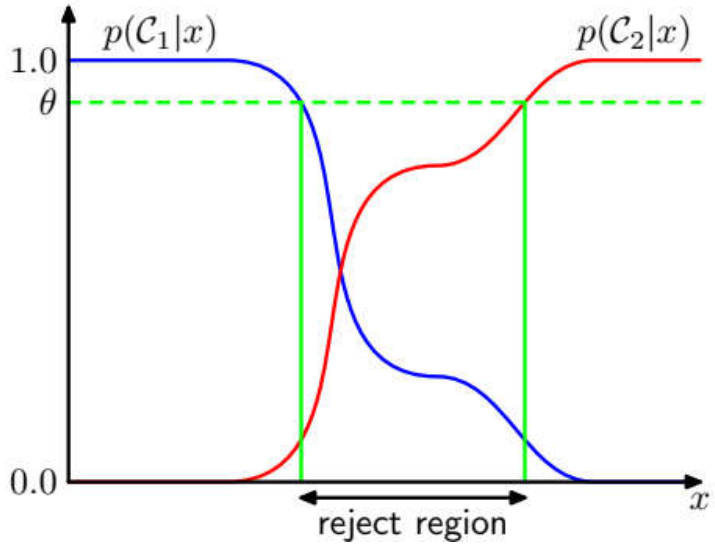
        - » Oblasti odlučivanja se biraju tako da za svaki novi ulaz bude minimalno $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Odbacivanje**
    - Greške klasifikacije u oblastima gde je najveća posteriori verovatnoća $p(\mathcal{C}_k|\mathbf{x})$ dosta manja od jedinice, odnosno gde su zajedničke verovatnoće $p(\mathbf{x}, \mathcal{C}_k)$ uporedive
      - Oblasti u kojima smo relativno nesigurni oko pripadnosti klasama
    - U nekim primenama u ovakvim slučajevima može se odustati od donošenja odluke, kako bi se smanjila greška na ulazima za koje je greška doneta.
      - Kod primera sa rendgen snimcima automatski sistem može u nekim slučajevima odustati od odlučivanja i prepustiti čoveku da donese odluku

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Odbacivanje**
    - Ovo se postiže uvođenjem praga θ i onda odustajati od odlučivanja za onaj ulaz za koji je najveća od posteriornih verovatnoća $p(\mathcal{C}_k|\mathbf{x})$ manja od praga
    - Za $\theta = 1$ nema odbacivanja, za $\theta < 1/K$ i *K* klasa odbacuju se svi ulazi



Uvod u mašinsko učenje

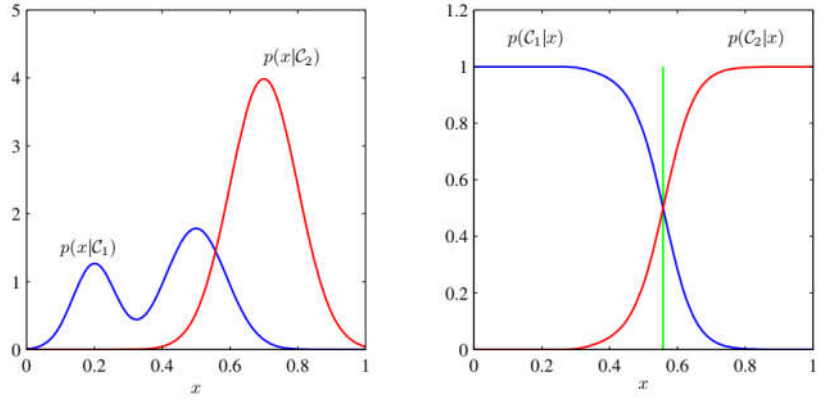
- **Teorija odlučivanja**
  - **Zaključivanje i odlučivanje**
    - Problem klasifikacije se deli na dva koraka
      - Zaključivanje kako bi se na osnovu obučavajućeg skupa odredio model $p(\mathcal{C}_k|\mathbf{x})$
      - Odlučivanje u kome se posteriorne verovatnoće koriste za optimalnu klasifikaciju
    - Alternativno se oba koraka izvršavaju istovremeno i uči fja koja ulaze direktno preslikava na odluke (*diskriminanta*)

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Zaključivanje i odlučivanje, 3 pristupa**
    - Generativni modeli
      - Najpre se za svaku klasu pojedinačno odrede $p(\mathbf{x}|\mathcal{C}_k)$ i $p(\mathcal{C}_k)$. Zatim Bajesova teorema $p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$ uz $p(\mathbf{x}) = \sum p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$, ili se posteriorne verovatnoće dobijaju iz procene $p(\mathbf{x}, \mathcal{C}_k)$.
      - Na osnovu posteriornih verovatnoća donose se odluke
    - Diskriminativni modeli
      - Direktno se određuju posteriorne verovatnoće $p(\mathcal{C}_k|\mathbf{x})$ i na osnovu njih donose odluke
    - Određivanje diskriminante $f(\mathbf{x})$
      - Mapira ulaz direktno u klasu; ovde verovatnoće nemaju ulogu

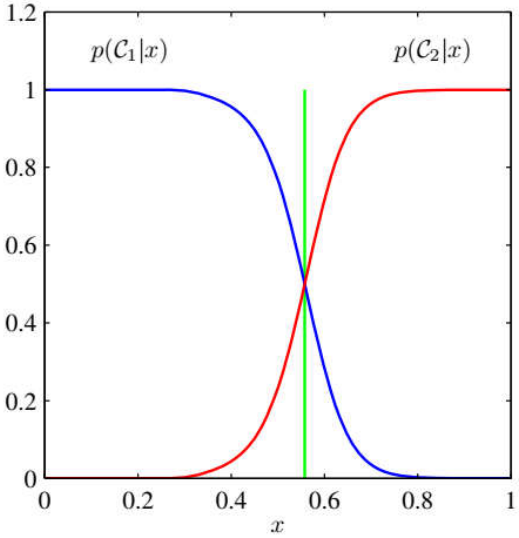
Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Zaključivanje i odlučivanje, 3 pristupa**
    - Prvi pristup najzahtevniji, ulaz u mnogim primenama ima više dimenzija i potreban je veliki obučavajući skup.
      - Ali, omogućava da se nađe $p(\mathbf{x})$, što omogućava detekciju ulaza koji imaju malu verovatnoću i moguću malu tačnost predviđanja (*outlier/novelty*)
    - Ako se traži samo klasifikacija, određivanje distribucije $p(\mathbf{x}, \mathcal{C}_k)$ je traćenje resursa ako je potrebno samo $p(\mathcal{C}_k|\mathbf{x})$



Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Zaključivanje i odlučivanje, 3 pristupa**
    - Treći pristup još jednostavniji.
      - » Samo odrediti granicu ->
    - Ali sada nemamo info o posteriornim verovatnoćama $p(\mathcal{C}_k|\mathbf{x})$, a one su korisne



Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Zaključivanje i odlučivanje, važnost posteriornih verovatnoća**
    - Minimizacija gubitka
      - Ako je potrebno s vremena na vreme ažurirati matricu gubitka, trivijalna revizija kriterijuma odlučivanja ako se znaju ove verovatnoće $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$
      - Ako imamo samo diskriminantu, mora se raditi ispočetka
    - Odbacivanje
      - Moguće ako znamo posteriorne verovatnoće

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Zaključivanje i odlučivanje, važnost posteriornih verovatnoća**
    - Kompenzacija apriori verovatnoća klasa
      - Ako se uzmu rendgen snimci opšte populacije, 1 u 1000 ima rak. Apriori verovatnoća klase jako mala i obučavanje na skupu opšteg tipa je slabo.
        - » Zato se uzima balansirani skup za obučavanje sa približno istim brojem primera obe klase i određuje

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

        - » Zatim podeliti sa udelom date klase u obučavajućem skupu pa pomnožiti sa udelom date klase u opštoj populaciji. Na kraju normalizacija.
        - » Ovo ne može bez posteriornih verovatnoća

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Zaključivanje i odlučivanje, važnost posteriornih verovatnoća**
    - Kombinacija modela
      - U složenoj primeni umesto jednog supersloženog modela možemo imati dva manja modela i kombinovati njihove rezultate
        - » Testiranje na rak: rendgen i krvna slika – dva posebna modela i kombinacija (pretpostavka da su za svaku klasu snimci i krvna slika nezavisni – uslovna nezavisnost)

$$p(\mathbf{x}_{\mathrm{I}}, \mathbf{x}_{\mathrm{B}}|\mathcal{C}_k) = p(\mathbf{x}_{\mathrm{I}}|\mathcal{C}_k)p(\mathbf{x}_{\mathrm{B}}|\mathcal{C}_k)$$

$$\begin{aligned} p(\mathcal{C}_k|\mathbf{x}_{\mathrm{I}}, \mathbf{x}_{\mathrm{B}}) &\propto p(\mathbf{x}_{\mathrm{I}}, \mathbf{x}_{\mathrm{B}}|\mathcal{C}_k)p(\mathcal{C}_k) \\ &\propto p(\mathbf{x}_{\mathrm{I}}|\mathcal{C}_k)p(\mathbf{x}_{\mathrm{B}}|\mathcal{C}_k)p(\mathcal{C}_k) \\ &\propto \frac{p(\mathcal{C}_k|\mathbf{x}_{\mathrm{I}})p(\mathcal{C}_k|\mathbf{x}_{\mathrm{B}})}{p(\mathcal{C}_k)} \end{aligned}$$

Uvod u mašinsko učenje

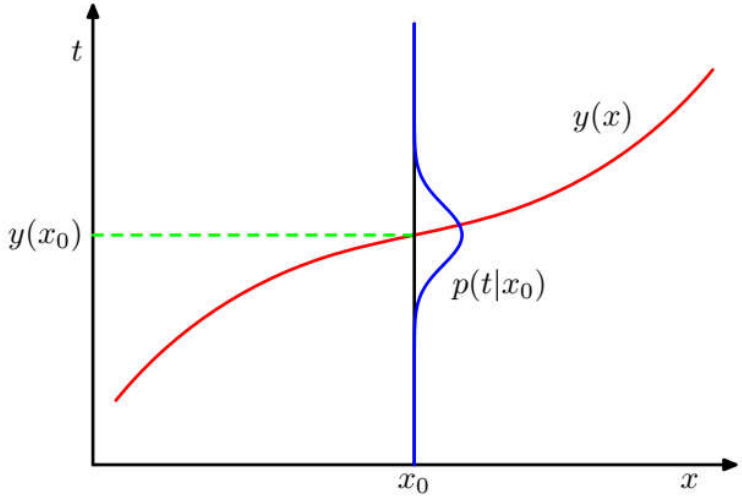
- **Teorija odlučivanja**
  - **Funkcija gubitka za regresiju**
    - Opet primer fitovanja
      - Određuje se procena vrednosti t, $y(\mathbf{x})$, za svaki ulaz $\mathbf{x}$. Pri tome se trpi gubitak $L(t, y(\mathbf{x}))$
      - Usrednjeni gubitak $\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$
      - Čest izbor fje gubitka kod regresije $L(t, y(\mathbf{x})) = \{y(\mathbf{x}) - t\}^2$
      - Očekivani gubitak tada $\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Funkcija gubitka za regresiju**
    - Opet primer fitovanja
      - Potrebno izabrati $y(\mathbf{x})$ koje minimizira gubitak

$$\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} = 2\int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t)\, \mathrm{d}t = 0 \quad y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t)\, \mathrm{d}t}{p(\mathbf{x})} = \int t p(t|\mathbf{x})\, \mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}]$$

      - $\mathbb{E}_t[\mathbf{t}|\mathbf{x}]$ uslovna srednja vrednost t u odnosu na $\mathbf{x}$ (regresiona fja)



Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Funkcija gubitka za regresiju**
    - Fja gubitka može da se razvije i onda opet usrednji

$$\{y(\mathbf{x}) - t\}^2 = \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2$$
$$= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

    - Drugi član predstavlja minimim funkcije gubitka

Uvod u mašinsko učenje

- **Teorija odlučivanja**
  - **Funkcija gubitka za regresiju**
    - Opet tri pristupa

$$y(\mathbf{x}) = \frac{\int tp(\mathbf{x}, t)\,\mathrm{d}t}{p(\mathbf{x})} = \int tp(t|\mathbf{x})\,\mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}]$$

      - Naći $p(\mathbf{x}, t)$, pa $p(t|\mathbf{x})$, na kraju $y(\mathbf{x})$
      - Odmah naći $p(t|\mathbf{x})$, zatim $y(\mathbf{x})$
      - Direktno odrediti $y(\mathbf{x})$ iz obučavajućih podataka

Uvod u mašinsko učenje

- **Teorija informacija**
  - Količina informacija koja se dobija na osnovu opažanja slučajne promenljive

$$h(x) = -\log_2 p(x)$$

  - Usrednjena informacija, odnosno entropija

$$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x)$$

Uvod u mašinsko učenje

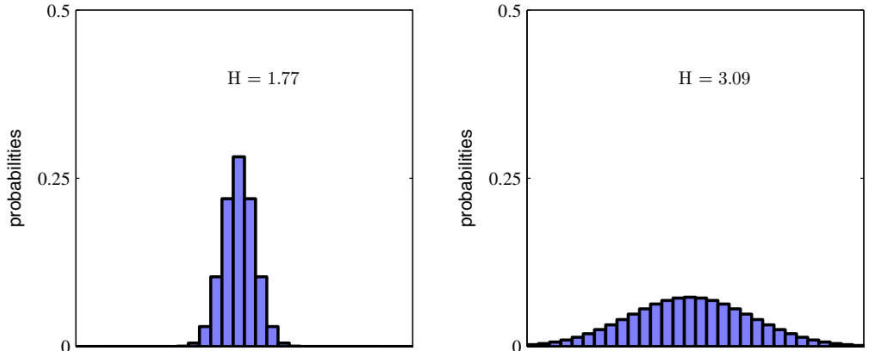
- **Teorija informacija**
  - Entropija diskretne slučajne promenljive *X*

$$\mathrm{H}[p] = -\sum_i p(x_i) \ln p(x_i)$$

- Raspodele $p(x_i)$ koje imaju izražene pikove imaće manju entropiju od onih koje su “ravnomerno raspoređene”
- Entropija nula ako se sve verovatnoće osim jedne nula.
- Maksimalnu entropiju ima uniformna raspodela

$$p(x_i) = 1/M$$

$$\mathrm{H} = \ln M$$



Uvod u mašinsko učenje

- **Teorija informacija**
  - (Diferencijalna) entropija
    - kontinualne promenljive
    - kontinualnih promenljivih

$$-\int p(x)\ln p(x)\,\mathrm{d}x$$

$$\mathrm{H}[\mathbf{x}] = -\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

  - Maksimalnu diferencijalnu entropiju ima Gausova distribucija

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$\mathrm{H}[x] = \frac{1}{2}\left\{1+\ln(2\pi\sigma^2)\right\}$$

    - Raste kako se distribucija širi, odnosno raste varijansa

Uvod u mašinsko učenje

- **Teorija informacija**
  - Dve slučajne promenljive sa zajedničkom raspodelom $p(\mathbf{x}, \mathbf{y})$
    - Ako se zna $\mathbf{x}$, dodatna informacija potrebna da se zna i $\mathbf{y}$ je $-\ln p(\mathbf{y}|\mathbf{x})$
    - Uslovna entropija $\mathbf{y}$ u odnosu na $\mathbf{x}$

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}$$

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

Uvod u mašinsko učenje

- **Teorija informacija**
  - **Relativna entropija i zajednička informacija**

  - Srednja dodatna količina informacija potrebna da se izrazi $\mathbf{x}$ sa nepoznatom raspodelom $p(\mathbf{x})$, ako raspodelu aproksimiramo sa $q(\mathbf{x})$

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x}.\end{aligned}$$

  - Relativna entropija ili Kullback-Leibler divergencija

$$\mathrm{KL}(p\|q) \geqslant 0$$

Uvod u mašinsko učenje

- **Teorija informacija**
  - **Relativna entropija i zajednička informacija**

    - Nepoznata raspodela $p(\mathbf{x})$ koju je potrebno modelovati
      - Recimo aproksimacija parametarskom raspodelom $q(\mathbf{x}|\boldsymbol{\theta})$
      - Jedan način je minimizirati relativnu entropiju ove dve raspodele po $\boldsymbol{\theta}$, ali ne znamo $p(\mathbf{x})$
      - Ali, ako imamo skup podataka za obučavanje $\mathbf{x}_n$ iz nepoznate raspodele

$$\mathrm{KL}(p\|q) \simeq \sum_{n=1}^{N} \{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

        - » Desni član ne zavisi od parametra, a levi je negativni logaritam funkcije verovatnoće za parametar pod raspodelom $q(\mathbf{x}|\boldsymbol{\theta})$
        - » Minimizacija KL-a ekvivalentna maximizaciji fje verovatnoće

Uvod u mašinsko učenje

- **Teorija informacija**
  - **Relativna entropija i zajednička informacija**

  - Zajednička distribucija $p(\mathbf{x}, \mathbf{y})$
    - Ako su skupovi $\mathbf{x}$ i $\mathbf{y}$ nezavisni, $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$
    - Ako nisu, pomoću KL možemo zaključiti koliko imi "fali" da budu

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \end{aligned}$$

    - Ovo je zajednička informacija $\mathbf{x}$ i $\mathbf{y}$.
      - » Uvek je veća od nule, a nula ako su nezavisni

Uvod u mašinsko učenje

- **Teorija informacija**
  - **Relativna entropija i zajednička informacija**

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$$

  - Zajednička informacija: smanjenje nesigurnosti u vezi sa $\mathbf{x}$ ako znamo $\mathbf{y}$ i obrnuto