Pronalaženje skrivenog znanja

Elektrotehnički fakultet Univerziteta u Beogradu

Master akademske studije

modul Računarska tehnika i informatika

2018/2019

Logistička regresija

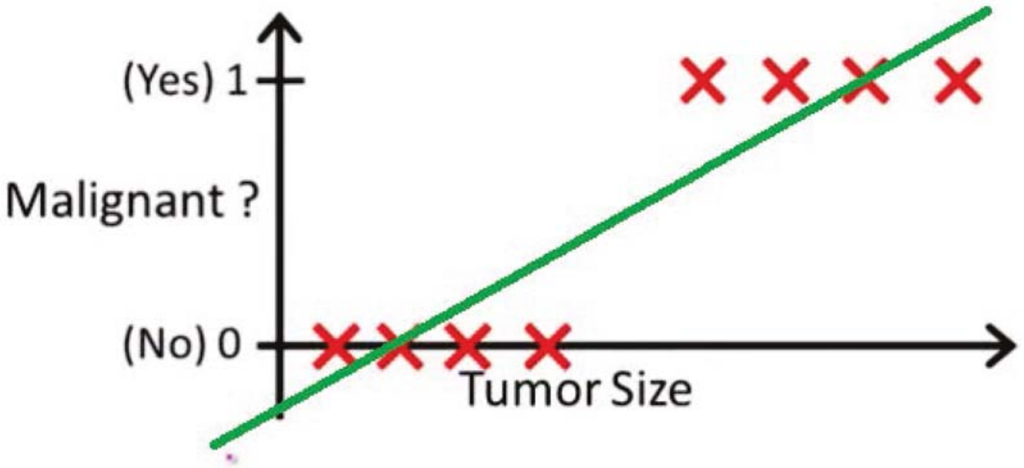
Vuk Batanović, Elektrotehnički fakultet Univerziteta u Beogradu

Korišćenje linearne regresije za klasifikaciju?

- Pretpostavka – rešavamo problem binarne klasifikacije
- Izlazna vrednost y može biti 0 ili 1
- Pitanje – da li se linearna regresija može koristiti u ovakvoj situaciji?
- Odgovor – može, klasa se predviđa tako što se proverava:

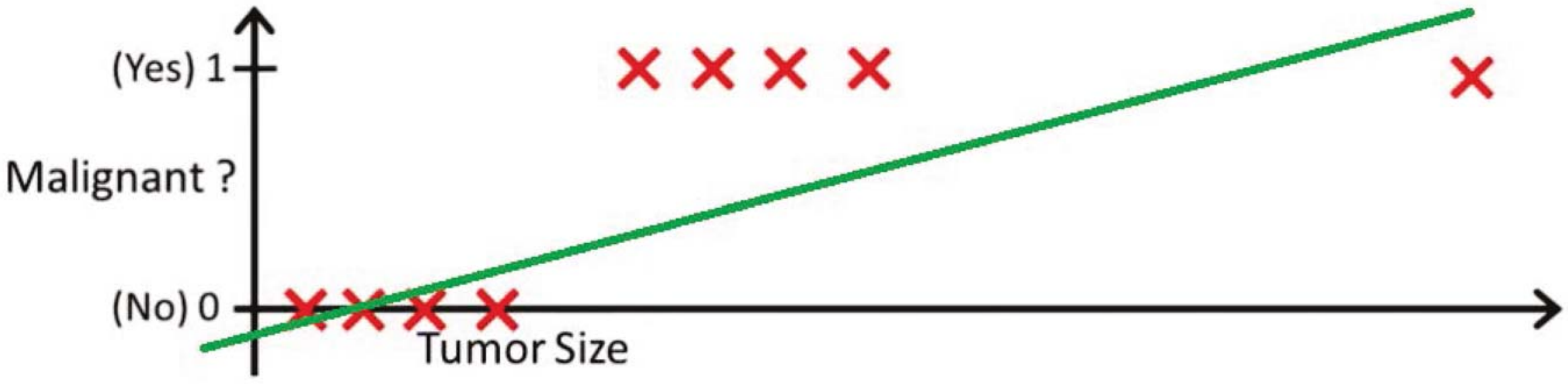
$$h(x) \geq 0.5$$

  - Ako je nejednakost tačna, predviđa se klasa 1
  - Ako je nejednakost netačna, predviđa se klasa 0
- Problemi
  - *Outlier*-i mogu da dramatično poremete kvalitet predikcije
  - $h(x)$ može biti dramatično veće od 1 ili manje od 0



Primer klasifikacije maligniteta tumora pomoću linearne regresije

Slika preuzeta sa: Andrew Ng, Machine Learning, Coursera



Primer uticaja *outlier*-a na klasifikaciju pomoću linearne regresije

Slika preuzeta sa: Andrew Ng, Machine Learning, Coursera

Logistička regresija

- Jedan od najpopularnijih algoritama *klasifikacije*
- Osnovna verzija algoritma služi za binarnu klasifikaciju
  - Najčešće se jedna klasa označava sa $y = 1$ a druga sa $y = 0$
- Predstavio ga statističar Dejvid Koks (*David Cox*) 1958. godine
- Široko rasprostranjena u ekonomskim, sociološkim i medicinskim analizama
- Sastavni element mnogih arhitektura neuralnih mreža

Logistička regresija

- Probabilistički klasifikator – izlaz modela je $P(y|x)$, a ne samo klasifikaciona odluka
- Pošto direktno modeluje $P(y|x)$, logistička regresija spada u diskriminativne modele
- Za razliku od Naïve Bayes algoritma, logistička regresija ne pretpostavlja statističku nezavisnost odlika

Logistička regresija

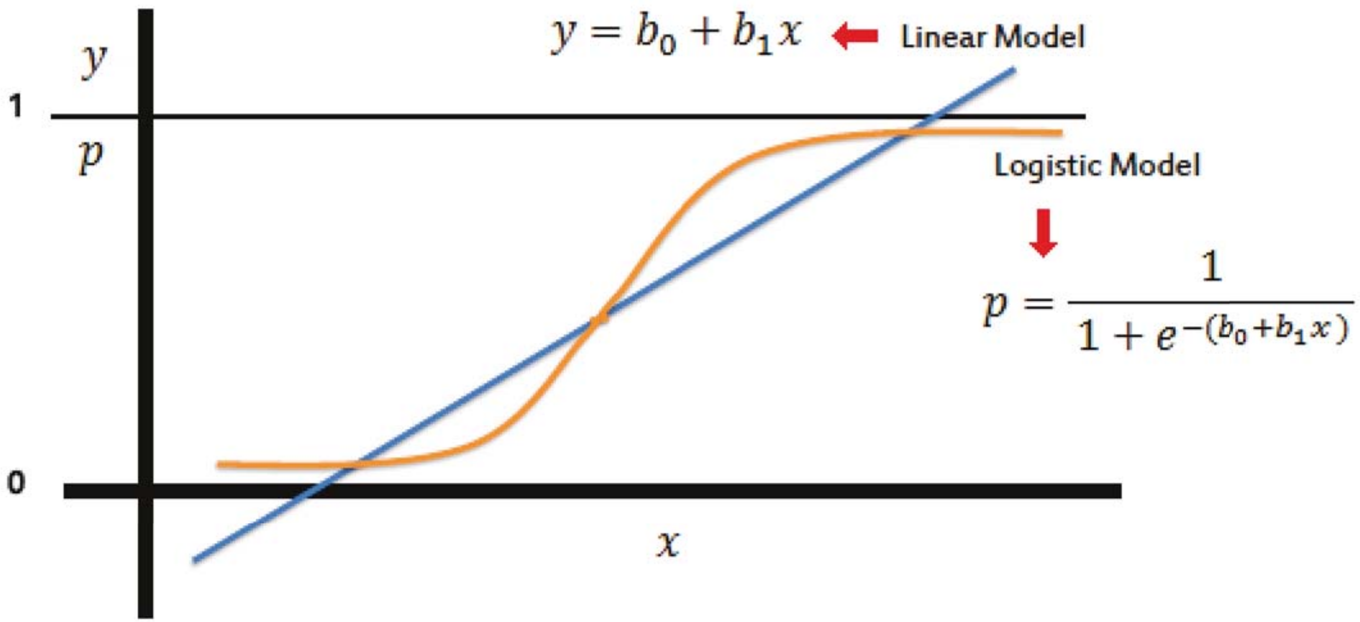
- Ime dobila po logističkoj/sigmoidnoj funkciji koja se koristi u hipotezi i ima oblik:

$$g(z) = \frac{1}{1 + e^{-z}}$$

- Logistička funkcija sabija interval $(-\infty, +\infty)$ na opseg $[0,1]$
- To omogućuje da se na izlazu dobije vrednost koja predstavlja verovatnoću da je $y = 1$
- Hipoteza logističke regresije je:

$$h(x) = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n)}}$$

- gde je n broj odlika koje se koriste u modelu



Razlika u izgledu linearne i logističke funkcije

Slika preuzeta sa: http://www.saedsayad.com/logistic_regression.htm

Hipoteza logističke regresije

- Zbog kraće notacije obično se uvodi fiktivna odlika $x_0 = 1$, tako da hipoteza postaje:

$$h(x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n)}}$$

$$= \frac{1}{1 + e^{-\sum_{i=0}^{n} w_i x_i}} = \frac{1}{1 + e^{-W \cdot X}} = \frac{e^{W \cdot X}}{e^{W \cdot X} + 1} = \frac{e^{W^T X}}{e^{W^T X} + 1}$$

- gde je W vektor svih težinskih parametara, X vektor svih vrednosti odlika, a $W \cdot X = W^T X$ njihov skalarni proizvod

Hipoteza logističke regresije

- Vrednost $h(x)$ predstavlja verovatnoću da je $y = 1$:

$$P(y = 1|x) = h(x) = \frac{e^{W^T X}}{e^{W^T X} + 1}$$

- Verovatnoća za klasu $y = 0$ se dobija kao:

$$P(y = 0|x) = 1 - P(y = 1|x) = 1 - h(x) = 1 - \frac{e^{W^T X}}{e^{W^T X} + 1} = \frac{1}{e^{W^T X} + 1}$$

- Nov podatak se klasifikuje u klasu koja je za njega verovatnija
- To znači da se podatak klasifikuje u klasu $y = 1$ za $h(x) > 0.5$, a u klasu $y = 0$ za $h(x) < 0.5$

Hiperravan razdvajanja

- Granica razdvajanja između klasa $y = 1$ i $y = 0$ je na $h(x) = 0.5$
- Sledi da za granicu razdvajanja važi:

$$h(x) = \frac{e^{W^T X}}{e^{W^T X} + 1} = 0.5$$

$$e^{W^T X} = 1$$

$$W^T X = W \cdot X = 0$$

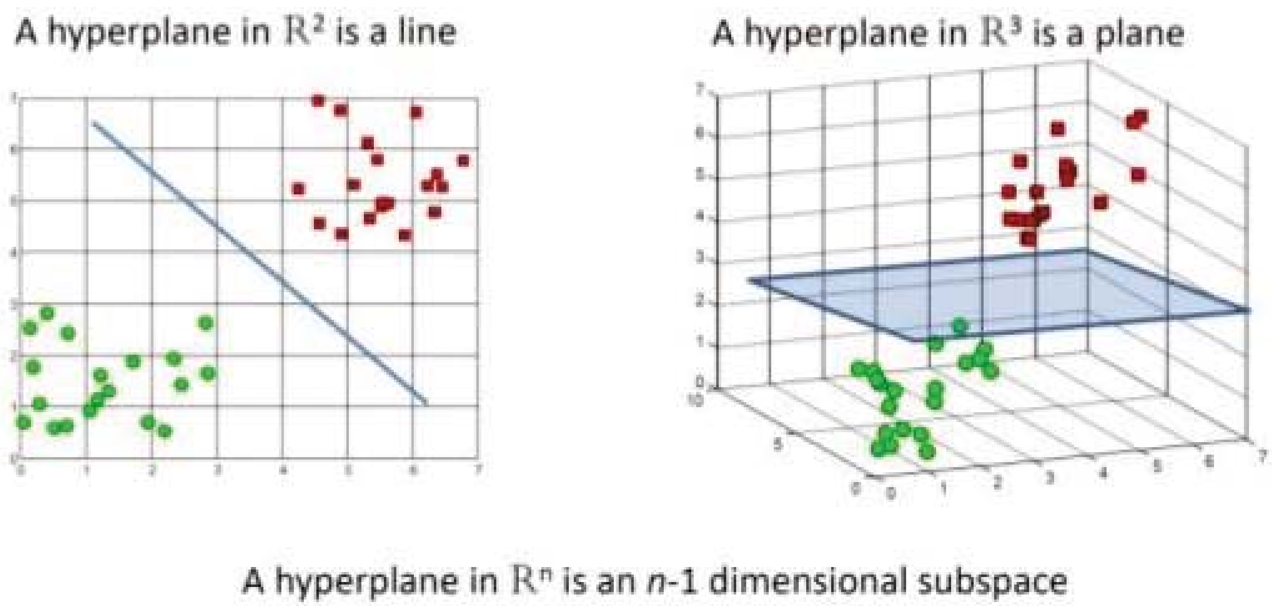
$$\sum_{i=0}^{n} w_i x_i = 0$$

Hiperravan razdvajanja

- Jednačina $\sum_{i=0}^{n} w_i x_i = 0$ definiše hiperravan razdvajanja između klasa (engl. *separating hyperplane*)
  - U slučaju samo jedne ulazne promenljive x, hiperravan je zapravo prava:

$$w_0 + w_1 x = 0$$

  - U opštem slučaju, hiperravan u prostoru dimenzije $n + 1$ je potprostor dimenzije n
- Sve tačke sa jedne strane hiperravni razdvajanja će biti klasifikovane u jednu klasu, a sve tačke sa druge strane u drugu klasu
- Što je neka tačka udaljenija od hiperravni razdvajanja, to je veća verovatnoća da ona pripada klasi koja se nalazi sa te strane hiperravni
- Za tačke na samoj hiperravni razdvajanja se uzima da pripadaju jednoj od klasa, obično $y = 1$



Izgled hiperravni razdvajanja u dvodimenzionalnom i trodimenzionalnom prostoru

Slika preuzeta sa: http://kgpdag.wordpress.com/2015/08/12/svm-simplified/

Nelinearne granice razdvajanja

- Hiperravan razdvajanja predstavlja linearnu granicu razdvajanja
- Korišćenjem polinomijalnih izraza kao argumenta logističke funkcije dobijaju se nelinearne granice razdvajanja
- Na primer, ako je hipoteza oblika:

$$h(x) = \frac{e^{x_1^2+x_2^2-1}}{e^{x_1^2+x_2^2-1}+1}$$

- tada je $y = 1$ za:

$$\frac{e^{x_1^2+x_2^2-1}}{e^{x_1^2+x_2^2-1}+1} \geq 0.5$$

Nelinearne granice razdvajanja

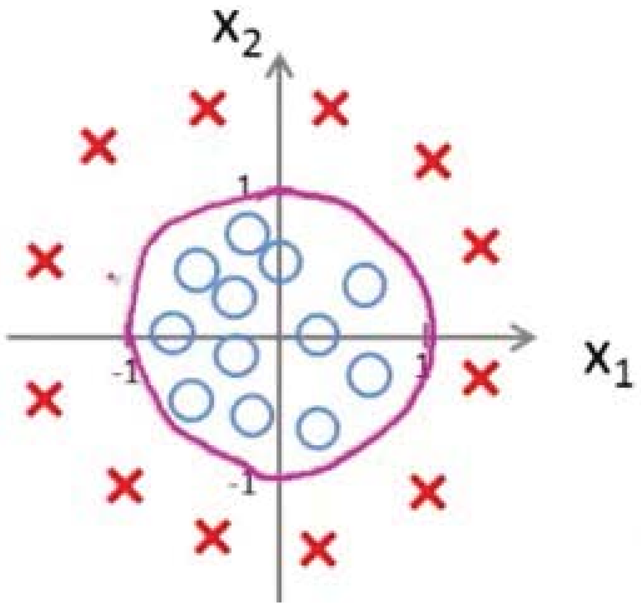
- Iz prethodne nejednakosti sledi:

$$e^{x_1^2+x_2^2-1} \geq 1$$

$$x_1^2 + x_2^2 - 1 \geq 0$$

$$x_1^2 + x_2^2 \geq 1$$

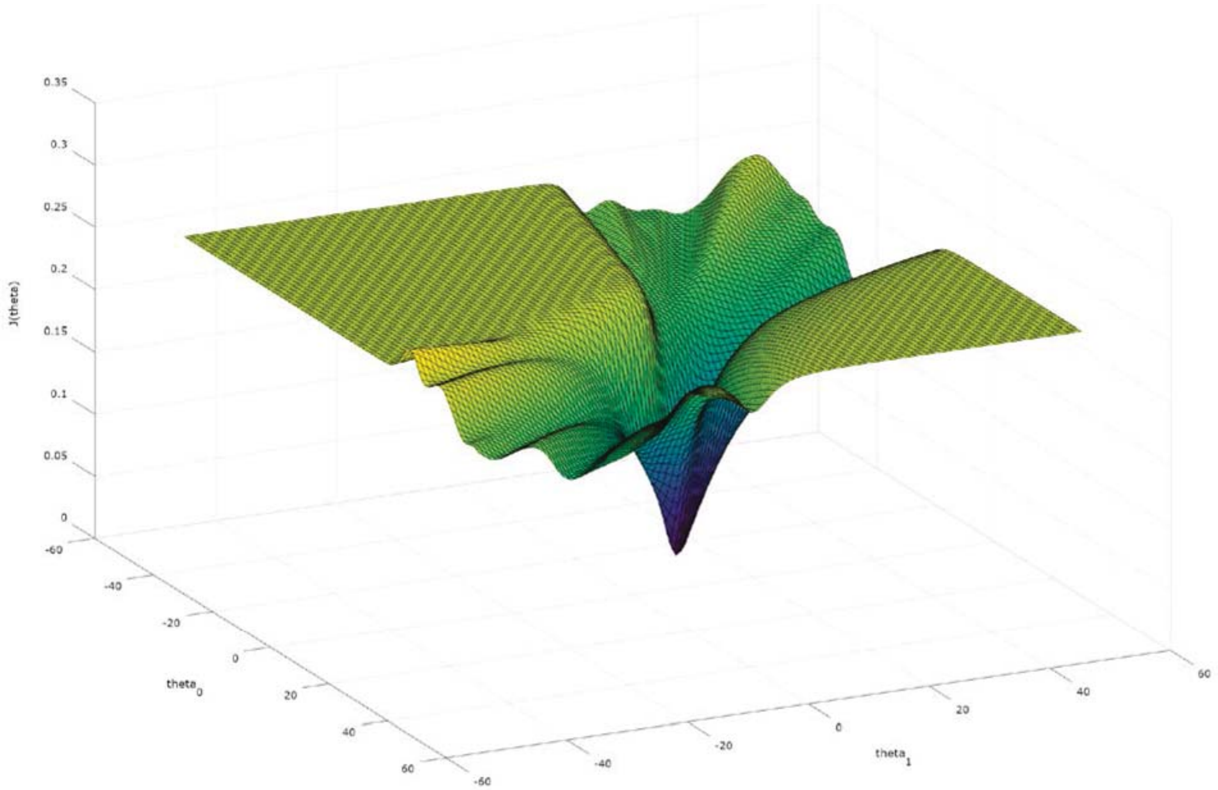
- Za navedeni oblik hipoteze, granica razdvajanja ima oblik kruga poluprečnika 1



Slika preuzeta sa: Andrew Ng, Machine Learning, Coursera

Funkcija greške logističke regresije

- Potrebno je izabrati odgovarajuću funkciju greške, tako da ona bude konveksna
- Da bi funkcija greške bila konveksna, funkcija gubitka mora da bude konveksna
- Kvadrat odstupanja $h(x)$ od y nije pogodna funkcija gubitka za logističku regresiju, jer je zbog logističke funkcije u okviru hipoteze takva funkcija gubitka nekonveksna



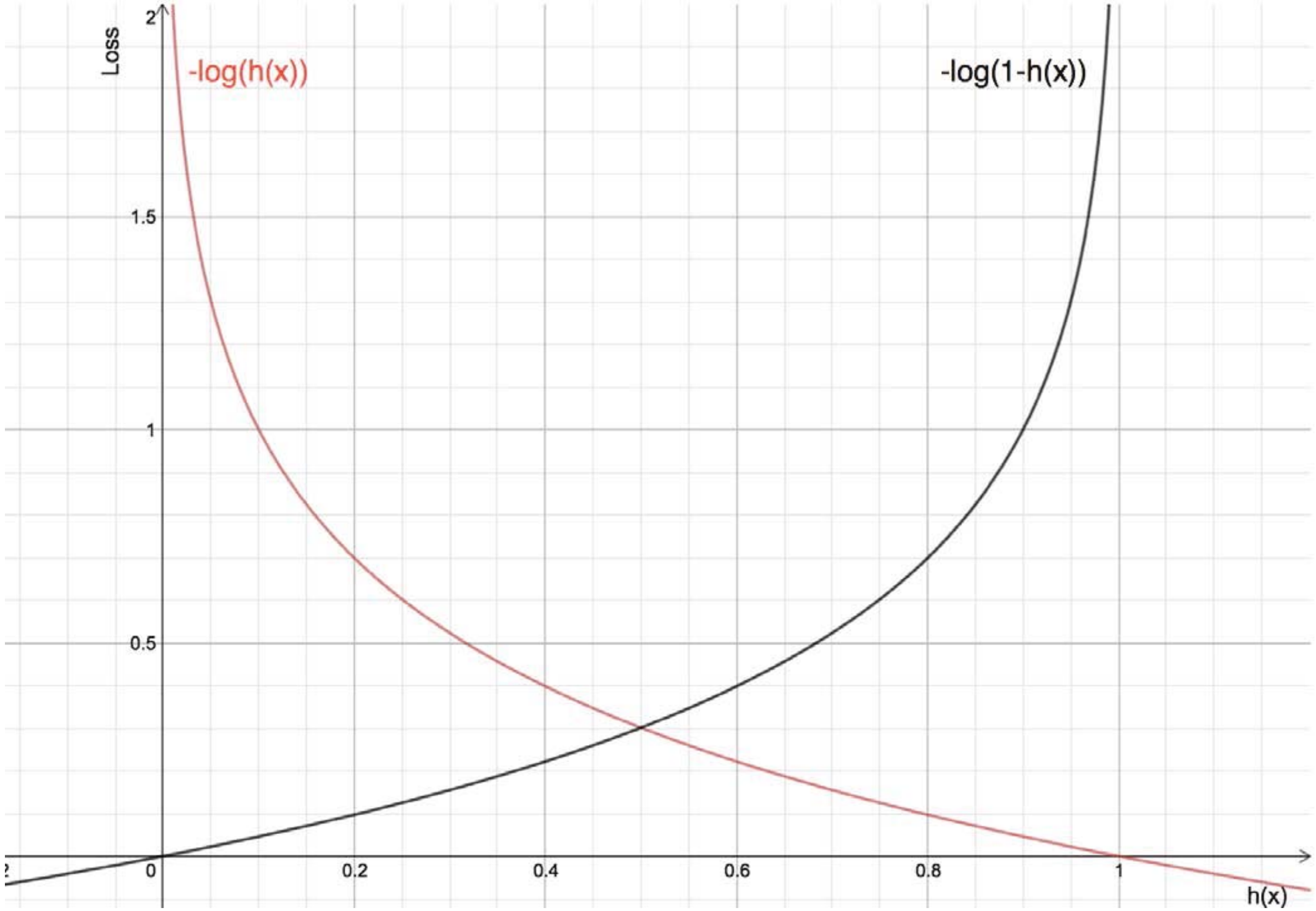
Primer izgleda kvadratne funkcije gubitka u logističkoj regresiji

Slika preuzeta sa: http://stats.stackexchange.com/questions/267400/logistic-regression-cost-surface-not-convex

Funkcija gubitka logističke regresije

- Pored konveksnosti, funkcija gubitka za logističku regresiju bi trebalo da ispoljava sledeće ponašanje:
  - Kada je $y = 1$ potrebno je da greška bude velika kada je vrednost $h(x)$ blizu nuli, a mala kada je vrednost $h(x)$ blizu jedinici
  - Kada je $y = 0$ potrebno je da greška bude velika kada je vrednost $h(x)$ blizu jedinici, a mala kada je vrednost $h(x)$ blizu nuli
- Sledeća funkcija ispunjava tražene uslove:

$$L(h(x), y) = -\ln\big(P(y|x)\big) = \begin{cases} -\ln(P(y=1|x)) = -\ln\big(h(x)\big), & y = 1 \\ -\ln(P(y=0|x)) = -\ln\big(1 - h(x)\big), & y = 0 \end{cases}$$



Oblik funkcije gubitka logističke regresije

Slika preuzeta sa: https://houxianxu.github.io/2015/04/23/logistic-softmax-regression/

Funkcija gubitka / gubitak unakrsne entropije

- Navedena funkcija gubitka se može zapisati kao:

$$L(h(x), y) = -y \ln h(x) - (1 - y) \ln(1 - h(x))$$

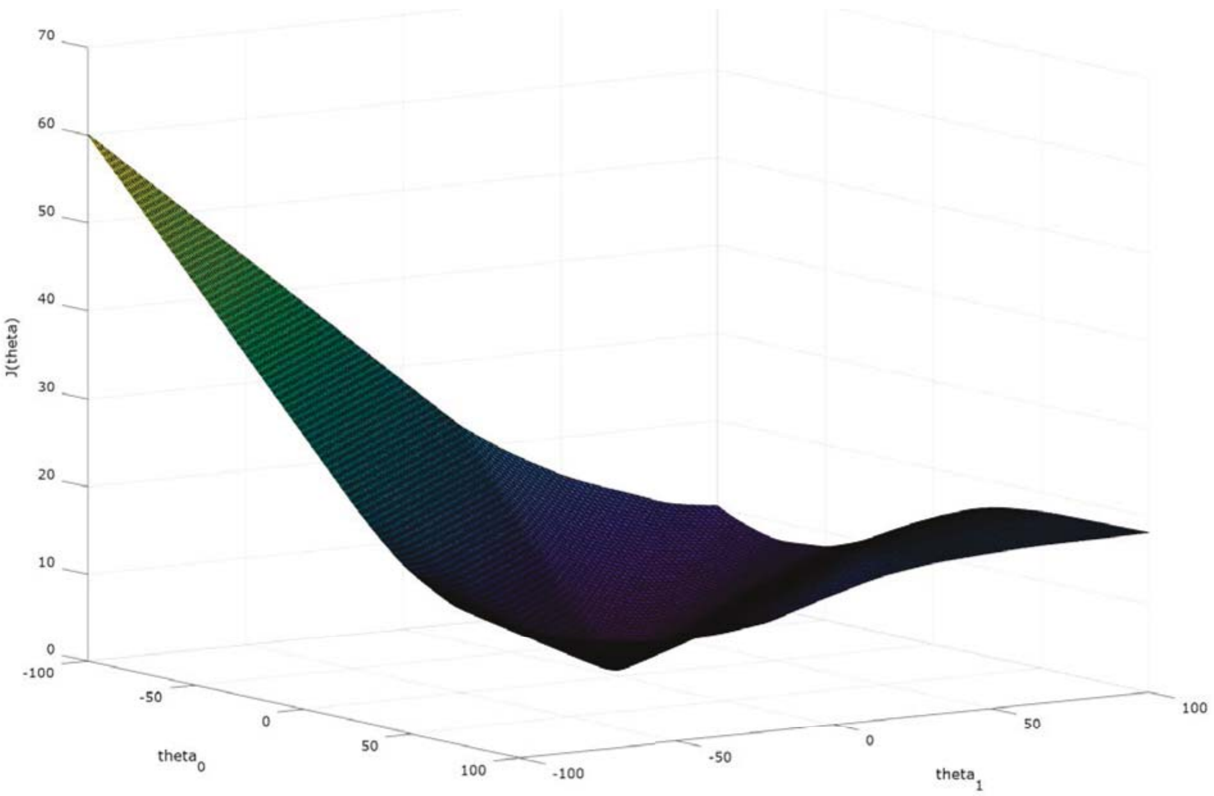
- Ova funkcija gubitka se naziva logističkim/log gubitkom (engl. *logistic /log loss*) ili gubitkom unakrsne entropije (engl. *cross-entropy loss*)
- Unakrsna entropija između dve raspodele p i q se definiše kao:

$$H(p, q) = -\sum_i p_i \ln q_i$$

- Ako se raspodele p i q definišu kao:

$$p \in \{y, 1 - y\} \qquad q \in \{h(x), 1 - h(x)\}$$

$$H(p, q) = -y \ln\big(h(x)\big) - (1 - y) \ln(1 - h(x))$$



Primer izgleda logističke funkcije gubitka u logističkoj regresiji

Slika preuzeta sa: http://stats.stackexchange.com/questions/267400/logistic-regression-cost-surface-not-convex

Funkcija greške i gradijentni spust

- Funkcija greške ima oblik:

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \ln h\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \ln(1 - h(x^{(i)}))$$

- Cilj obučavanja se opet svodi na traženje minimuma funkcije greške $J(w)$ u zavisnosti od w putem gradijentnog spusta
- Izraz za ažuriranje težinskih vrednosti w pri gradijentnom spustu:

$$w_r := w_r - \alpha \frac{\partial J(w)}{\partial w_r}$$

- Da bi se pronašao parcijalni izvod funkcije $J(w)$ potrebno je prvo pronaći parcijalne izvode logističke funkcije g i hipoteze h

Parcijalni izvod logističke funkcije

$$g(z) = \frac{1}{1+e^{-z}}$$

$$\frac{\partial g(z)}{\partial z} = -\frac{1}{(1+e^{-z})^2}\frac{\partial}{\partial z}(1+e^{-z}) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}}\frac{e^{-z}}{1+e^{-z}}$$

$$= g(z)\frac{(1+e^{-z})-1}{1+e^{-z}} = g(z)\left(1-\frac{1}{1+e^{-z}}\right) = g(z)(1-g(z))$$

Parcijalni izvod hipoteze $h(x)$

$$h(x) = g\left(\sum_{i=0}^{n} w_i x_i\right) = \frac{1}{1 + e^{-\sum_{i=0}^{n} w_i x_i}}$$

$$\frac{\partial h(x)}{\partial w_r} = \frac{\partial g\left(\sum_{i=0}^{n} w_i x_i\right)}{\partial w_r}$$

$$= g\left(\sum_{i=0}^{n} w_i x_i\right)\left(1 - g\left(\sum_{i=0}^{n} w_i x_i\right)\right)\frac{\partial}{\partial w_r}\sum_{i=0}^{n} w_i x_i$$

$$= h(x)(1 - h(x))x_r$$

Parcijalni izvod funkcije greške

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \ln h\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \ln\left(1 - h\left(x^{(i)}\right)\right)$$

$$\frac{\partial J(w)}{\partial w_r} = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \frac{1}{h\left(x^{(i)}\right)} \frac{\partial h\left(x^{(i)}\right)}{\partial w_r} + \left(1 - y^{(i)}\right) \frac{1}{1 - h\left(x^{(i)}\right)} \frac{\partial}{\partial w_r}\left(1 - h\left(x^{(i)}\right)\right)$$

$$= -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \left(1 - h\left(x^{(i)}\right)\right) x_r^{(i)} - \left(1 - y^{(i)}\right) h\left(x^{(i)}\right) x_r^{(i)}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \left(h\left(x^{(i)}\right) - y^{(i)}\right) x_r^{(i)}$$

Obučavanje putem gradijentnog spusta

- Postupak izgleda identično kao kod linearne regresije
- Simultano ažurirati sve w_k i ponavljati do konvergencije
- Algoritam grupnog gradijentnog spusta:

$$w_r := w_r - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h\left(x^{(i)}\right) - y^{(i)}\right) x_r^{(i)}$$

- Algoritam stohastičkog gradijentnog spusta:

$randomly\ shuffle\ the\ data$

$for\ i = 1\ to\ m:$

$$w_r := w_r - \alpha \frac{1}{m} \left(h\left(x^{(i)}\right) - y^{(i)}\right) x_r^{(i)}$$

Obučavanje putem gradijentnog spusta

- Radi ubrzanja gradijentnog spusta u praksi se često koriste naprednije metode optimizacije
  - Konjugovani spust
  - Kvazi-Njutnove metode
    - BFGS (*Broyden-Fletcher-Goldfarb-Shanno*) algoritam
    - L-BFGS (*Limited memory BFGS*) algoritam
- Dodatna prednost ovih metoda – nije potrebno ručno zadati brzinu učenja α (metode same biraju optimalnu vrednost α)
- Mana ovih metoda – velika kompleksnost
  - Nisu pogodne za samostalnu implementaciju već treba koristiti gotove biblioteke

Preterana prilagođenost modela podacima

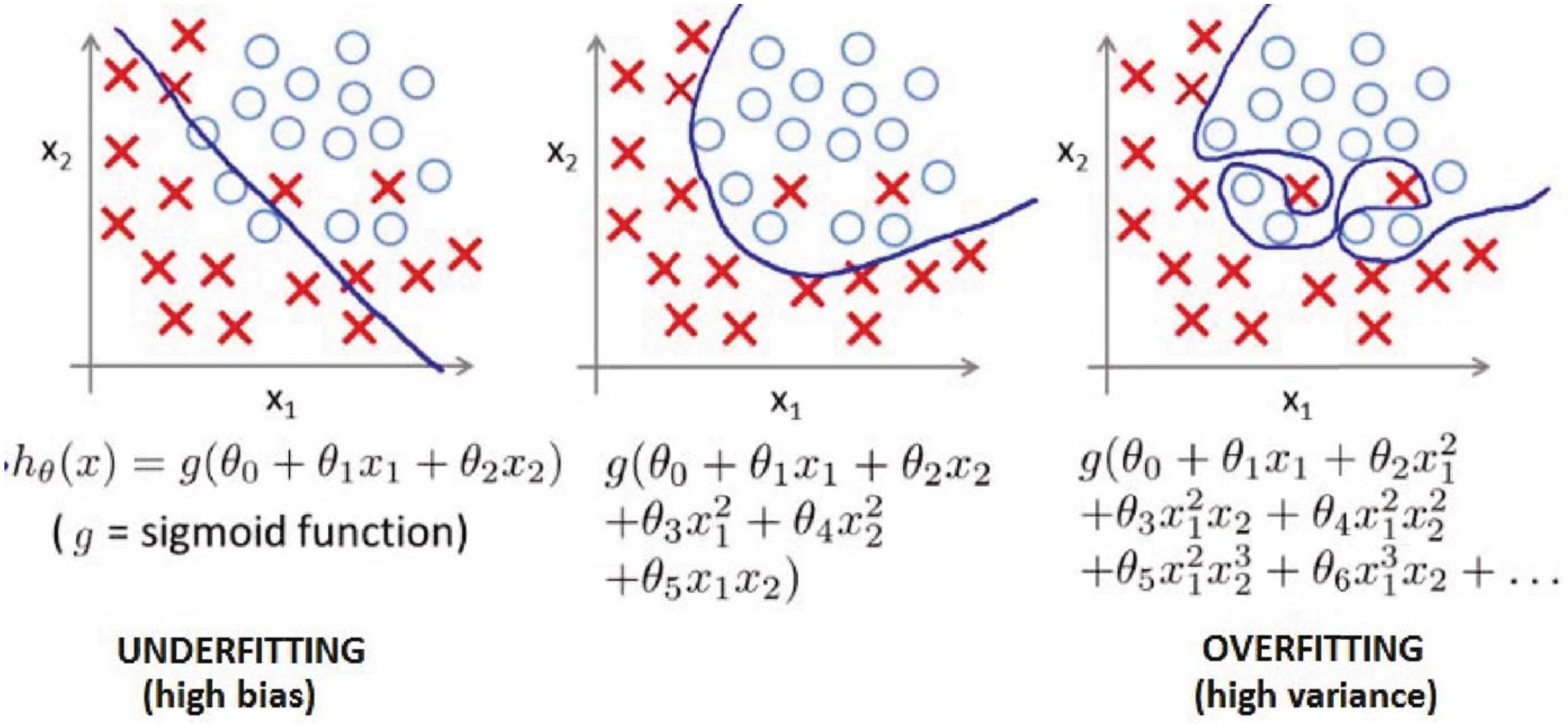
- Kao i u regresiji, i u klasifikaciji preterana prilagođenost modela podacima (*overfitting*) dovodi do toga da apsolutne vrednosti nekih težinskih parametara w budu jako velike (u krajnjoj instanci teže $\rightarrow \infty$)
- Primer – klasifikacija dokumenata po tematici između sporta $(y = 1)$ i informatike $(y = 0)$
  - Pretpostavimo *bag-of-words* pristup – svaka reč je jedna odlika
  - U skupu za obučavanje reč *fudbal* se javlja samo u dokumentima čija je tema sport
  - Algoritam stoga teži da postavi $P(y = 1 | x; x_{fudbal} \geq 1) = 1$
  - Pošto se reč *fudbal* javlja samo u dokumentima jedne klase, obučavanje dovodi do visoke vrednosti za w_{fudbal}

Preterana prilagođenost modela podacima

$$w_{fudbal} \to +\infty \quad \Longrightarrow \quad P\left(y = 1 | x; x_{fudbal} \geq 1\right) = \frac{1}{1 + e^{-\sum_{i=0}^{n} w_i x_i}} \to 1$$

- Visoka vrednost parametra w_{fudbal} efektivno dovodi do toga da se svaki dokument u kome se javlja reč *fudbal* automatski klasifikuje u sportsku tematiku – sve ostale odlike se ignorišu
- Ovo ponašanje nije poželjno, jer je moguće da se reč *fudbal* javi u informatičkom dokumentu čija je tema opis nekog fudbalskog simulatora
- Do neželjenog ponašanja dolazi zbog proređenosti podataka u skupu za obučavanje

- Kao i kod regresije, regularizacija sprečava preteranu prilagođenost modela podacima ograničavanjem magnituda težinskih parametara
- Moguće je koristiti bilo L_1 bilo L_2 regularizaciju



Ilustracija efekta nedovoljne i preterane prilagođenosti modela podacima u klasifikaciji podataka

Slika preuzeta sa: Andrew Ng, Machine Learning, Coursera

L_2-regularizovana logistička regresija

- Funkcija greške, njen parcijalni izvod i pravilo ažuriranja težinskih vrednosti za grupni spust u L_2-regularizovanoj logističkoj regresiji:

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \ln h\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \ln\left(1 - h\left(x^{(i)}\right)\right) + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$$

$$\frac{\partial J(w)}{\partial w_r} = \frac{1}{m}\sum_{i=1}^{m}\left(h\left(x^{(i)}\right) - y^{(i)}\right) x_r^{(i)} + \frac{\lambda}{m} w_r$$

$$w_r := w_r\left(1 - \alpha\frac{\lambda}{m}\right) - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h\left(x^{(i)}\right) - y^{(i)}\right) x_r^{(i)}$$

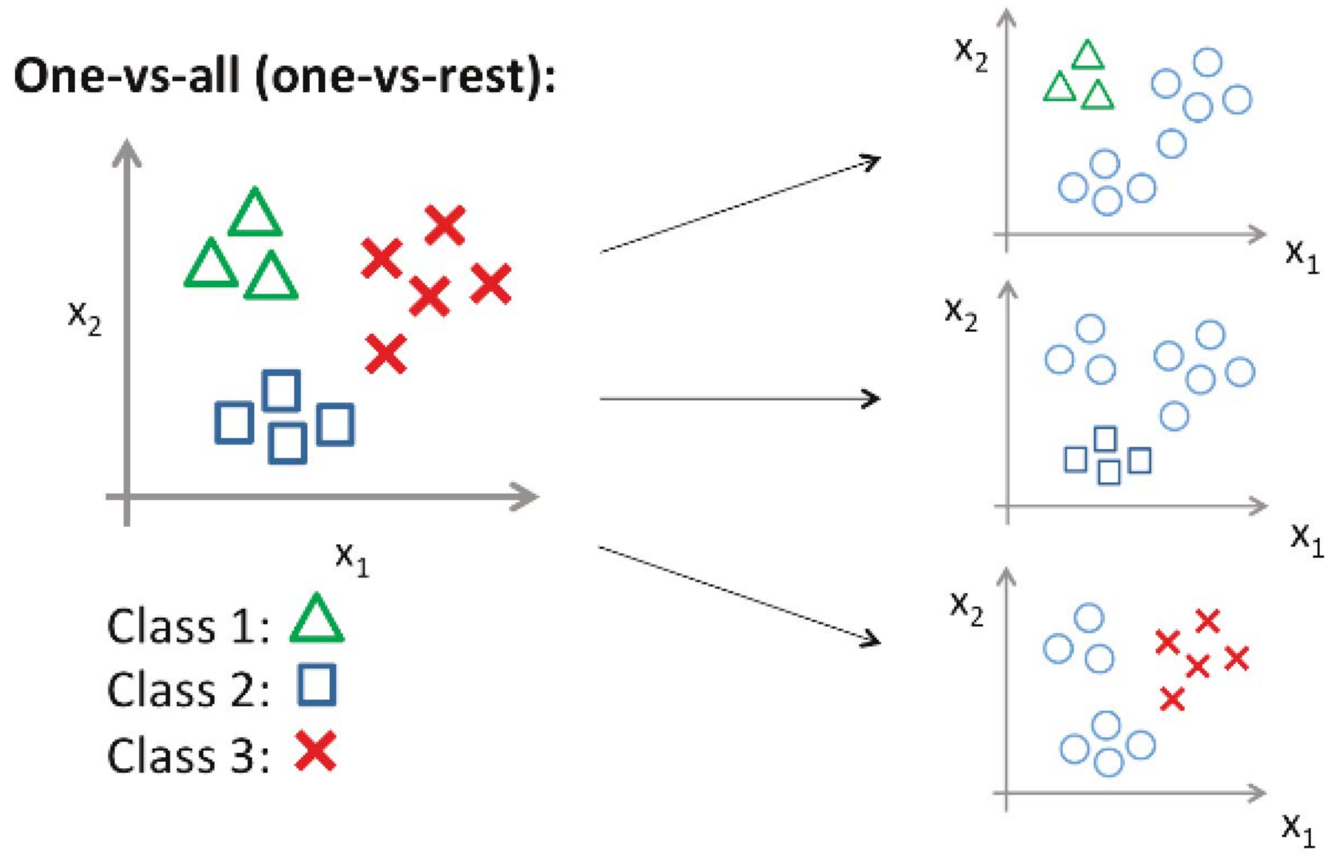
- Kao i kod linearne regresije, težinski parametar w_0 obično ne podleže regularizaciji

Višeklasna klasifikacija

- Logistička regresija se može primeniti i u slučaju prisustva većeg broja klasa, i to na dva načina:
  - Kombinovanje rezultata većeg broja binarnih klasifikatora tj. binarnih logističkih regresija
    - Princip „jedan nasuprot svima“ (engl. *one-versus-all*) / „jedan nasuprot ostalima“ (engl. *one-versus-rest*)
    - Princip „jedan nasuprot jednom“ (engl. *one-versus-one*)
    - Ove metode su primenjive i na ostale binarne klasifikatore (na primer – metodu potpornih vektora)
  - Multinomijalna logistička regresija

Princip „jedan nasuprot svima“

- Ako je k broj klasa, konstruiše se k binarnih klasifikatora
- Svaki binarni klasifikator je dodeljen jednoj klasi
  - Tretira dodeljenu klasu kao jednu klasu, a sve ostale klase zajedno kao drugu klasu
  - Problem neuravnoteženosti broja primera po klasama – primera druge klase tipično ima daleko više nego primera prve
- Nov podatak se svrstava u onu klasu čiji binarni klasifikator proizvodi najveću verovatnoću pripadnosti tog podatka posmatranoj klasi



Ilustracija pristupa „jedan nasuprot svima“ u višeklasnoj klasifikaciji

Slika preuzeta sa: Andrew Ng, Machine Learning, Coursera

Princip „jedan nasuprot jednom“

- Ako je k broj klasa, konstruiše se $k(k-1)/2$ binarnih klasifikatora, po jedan za svaki par klasa
- Nov podatak se svrstava u onu klasu koju je odabralo najviše binarnih klasifikatora – mehanizam glasanja
- Broj klasifikatora je u ovoj varijanti znatno veći nego u pristupu „jedan nasuprot svima“
- Ipak, u zavisnosti od broja klasa, vreme rada može da bude i kraće nego u pristupu „jedan nasuprot svima“
    - Skup podataka za obučavanje svakog binarnog klasifikatora je znatno manji u pristupu „jedan nasuprot jednom“

Multinomijalna logistička regresija

- „Prirodno“ proširenje logističke regresije na rad sa više klasa
- Naziva se i *softmax* regresijom jer se verovatnoća pripadnosti klasi t dobija pomoću tzv. softmax funkcije čiji je oblik:

$$P(y = t|x) = \frac{e^{\sum_{i=0}^{n} w_i^{\{t\}} x_i}}{\sum_{j=1}^{k} e^{\sum_{i=0}^{n} w_i^{\{j\}} x_i}} = \frac{e^{\left(W^{\{t\}}\right)^T X}}{\sum_{j=1}^{k} e^{\left(W^{\{j\}}\right)^T X}}$$

- gde je k broj klasa, n broj odlika, $w_i^{\{t\}}$ težinski parametar i-te odlike za t-tu klasu, $\left(W^{\{t\}}\right)^T$ (transponovani) vektor težinskih parametara za t-tu klasu, a X vektor vrednosti odlika
- Imenilac razlomka služi da normalizuje raspodelu, tako da suma svih verovatnoća bude jednaka jedinici

Multinomijalna logistička regresija

- Izlaz softmax funkcije se koristi da reprezentuje multinomijalnu raspodelu (raspodelu sa k mogućih ishoda)
- Hipoteza multinomijalne logističke regresije je k-dimenzionalni vektor koji izražava verovatnoće pripadnosti podatka svakoj od k klasa:

$$h(x) = \begin{bmatrix} P(y = 1|x) \\ P(y = 2|x) \\ \vdots \\ P(y = k|x) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\left(W^{\{j\}}\right)^T X}} \begin{bmatrix} e^{\left(W^{\{1\}}\right)^T X} \\ e^{\left(W^{\{2\}}\right)^T X} \\ \vdots \\ e^{\left(W^{\{k\}}\right)^T X} \end{bmatrix}$$

Multinomijalna logistička regresija

- Iako svaka klasa ima svoj vektor težinskih parametara $W^{\{t\}}$, broj slobodnih vektora je zapravo $k-1$, pošto zbir verovatnoća svih klasa mora biti jednak jedinici
- Model je „preterano parametrizovan“ (engl. *overparameterized*), što znači da za svaki skup podataka za obučavanje postoji veći broj vrednosti parametara koje proizvode identičnu hipotezu $h(x)$
- Iz ovoga se može izvesti i dokaz da je binarna logistička regresija samo jednostavniji slučaj multinomijalne

Ekvivalentnost binarne i multinomijalne logističke regresije za $k = 2$

- Ako postoje samo dve klase – $y = 0$ i $y = 1$ – može se proizvoljno usvojiti da su vrednosti svih težinskih parametara klase 0 jednaki nuli
- Sledi:

$$P(y = 1|x) = \frac{e^{\left(W^{\{1\}}\right)^T X}}{e^{\left(W^{\{0\}}\right)^T X} + e^{\left(W^{\{1\}}\right)^T X}} = \frac{e^{\left(W^{\{1\}}\right)^T X}}{e^{0} + e^{\left(W^{\{1\}}\right)^T X}} = \frac{e^{\left(W^{\{1\}}\right)^T X}}{1 + e^{\left(W^{\{1\}}\right)^T X}}$$

- što je jednako već viđenom izrazu za hipotezu binarne logističke regresije:

$$P(y = 1|x) = h(x) = \frac{e^{W^T X}}{1 + e^{W^T X}}$$

Funkcija gubitka multinomijalne logističke regresije

- Kao i kod binarne logističke regresije, funkcija gubitka klase t bi u slučaju pripadnosti podatka klasi t trebalo da bude velika kada je vrednost $h(x)^{\{t\}} = P(y = t|x)$ blizu nuli, a mala kada je vrednost $h(x)^{\{t\}}$ blizu jedinici:

$$L(h(x), y)^{\{t\}} = -\ln h(x)^{\{t\}} = -\ln P(y = t|x) = -\ln \frac{e^{\left(W^{\{t\}}\right)^T X}}{\sum_{j=1}^{k} e^{\left(W^{\{j\}}\right)^T X}}$$

$$= -\left(\left(W^{\{t\}}\right)^T X - \ln \sum_{j=1}^{k} e^{\left(W^{\{j\}}\right)^T X}\right)$$

- (Multinomijalna) Logistička regresija spada u tzv. log-linearne modele – logaritam hipoteze sadrži linearnu kombinaciju parametara : $\left(W^{\{t\}}\right)^T X$

Funkcija greške multinomijalne logističke regresije

- Funkcija greške ima sledeći oblik:

$$J(w) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{k} 1\{y^{(i)} = t\} L\left(h\left(x^{(i)}\right), y^{(i)}\right)^{\{t\}}$$

- gde je $1\{y^{(i)} = t\}$ indikatorska funkcija čija je vrednost 1 kada je $y^{(i)} = t$, odnosno 0 u suprotnom (Kronekerova delta funkcija)

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{k} 1\left\{y^{(i)} = t\right\} \ln \frac{e^{\left(W^{\{t\}}\right)^T X^{(i)}}}{\sum_{j=1}^{k} e^{\left(W^{\{j\}}\right)^T X^{(i)}}}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{k} 1\left\{y^{(i)} = t\right\} \ln P\left(y^{(i)} = t | x^{(i)}\right)$$

Ekvivalentnost funkcije greške binarne i multinomijalne logističke regresije za $k = 2$

- Funkcija greške binarne logističke regresije se može transformisati na sledeći način:

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \ln h\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \ln\left(1 - h\left(x^{(i)}\right)\right)$$

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \ln P\left(y^{(i)} = 1 \middle| x^{(i)}\right) + \left(1 - y^{(i)}\right) \ln P\left(y^{(i)} = 0 \middle| x^{(i)}\right)$$

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{1} 1\{y^{(i)} = t\} \ln P(y^{(i)} = t|x^{(i)})$$

- čime se dobija oblik funkcije greške ekvivalentan onom za multinomijalnu logističku regresiju

Parcijalni izvod funkcije gubitka multinomijalne logističke regresije

$$\frac{\partial}{\partial w_r^{\{t\}}} L(h(x), y)^{\{t\}} = -\frac{\partial}{\partial w_r^{\{t\}}}\left(\left(W^{\{t\}}\right)^T X - \ln \sum_{j=1}^{k} e^{\left(W^{\{j\}}\right)^T X}\right)$$

$$= -\frac{\partial}{\partial w_r^{\{t\}}}\left(\sum\nolimits_{i=0}^{n} w_i^{\{t\}} x_i - \ln \sum_{j=1}^{k} e^{\sum_{i=0}^{n} w_i^{\{j\}} x_i}\right)$$

$$= -x_r + \frac{e^{\sum_{i=0}^{n} w_i^{\{t\}} x_i}}{\sum_{j=1}^{k} e^{\sum_{i=0}^{n} w_i^{\{j\}} x_i}} x_r = -x_r\left(1 - P(y = t|x)\right)$$

Parcijalni izvod funkcije gubitka multinomijalne logističke regresije

$$\frac{\partial}{\partial w_r^{\{t\}}} L(h(x), y)^{\{q|q\neq t\}} = -\frac{\partial}{\partial w_r^{\{t\}}}\left(\left(W^{\{q\}}\right)^T X - \ln\sum_{j=1}^{k} e^{\left(W^{\{j\}}\right)^T X}\right)$$

$$= -\frac{\partial}{\partial w_r^{\{t\}}}\left(\sum\nolimits_{i=0}^{n} w_i^{\{q\}} x_i - \ln\sum_{j=1}^{k} e^{\sum_{i=0}^{n} w_i^{\{j\}} x_i}\right)$$

$$= \frac{e^{\sum_{i=0}^{n} w_i^{\{t\}} x_i}}{\sum_{j=1}^{k} e^{\sum_{i=0}^{n} w_i^{\{j\}} x_i}} x_r = P(y = t|x) x_r$$

Parcijalni izvod funkcije greške multinomijalne logističke regresije

$$J(w) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{k} 1\{y^{(i)} = t\} L\left(h\left(x^{(i)}\right), y^{(i)}\right)^{\{t\}}$$

$$\frac{\partial}{\partial w_r^{\{t\}}} L(h(x), y)^{\{t\}} = -x_r\left(1 - P(y = t|x)\right)$$

$$\frac{\partial}{\partial w_r^{\{t\}}} L(h(x), y)^{\{q|q \neq t\}} = P(y = t|x) x_r$$

$$\frac{\partial J(w)}{\partial w_r^{\{t\}}} = -\frac{1}{m}\sum_{i=1}^{m} x_r^{(i)}\left(1\{y^{(i)} = t\} - P\left(y^{(i)} = t \middle| x^{(i)}\right)\right)$$

Ekvivalentnost parcijalnog izvoda funkcije greške binarne i multinomijalne logističke regresije za $k = 2$

$$\frac{\partial J(w)}{\partial w_r^{\{t\}}} = -\frac{1}{m}\sum_{i=1}^{m} x_r^{(i)}\left(1\{y^{(i)} = t\} - P\left(y^{(i)} = t \middle| x^{(i)}\right)\right)$$

- Ako postoje samo dve klase – $y = 0$ i $y = 1$ – može se proizvoljno usvojiti da su vrednosti svih težinskih parametara klase 0 jednaki nuli

$$\frac{\partial J(w)}{\partial w_r} = \frac{1}{m}\sum_{i=1}^{m} x_r^{(i)}\left(P\left(y^{(i)} = 1 \middle| x^{(i)}\right) - 1\{y^{(i)} = 1\}\right)$$

$$\frac{\partial J(w)}{\partial w_r} = \frac{1}{m}\sum_{i=1}^{m}\left(P\left(y^{(i)} = 1 | x^{(i)}\right) - y^{(i)}\right) x_r^{(i)}$$

$$\frac{\partial J(w)}{\partial w_r} = \frac{1}{m}\sum_{i=1}^{m}\left(h\left(x^{(i)}\right) - y^{(i)}\right) x_r^{(i)}$$

Regularizacija multinomijalne logističke regresije

- Izgled funkcije greške sa L_2 regularizacijom:

$$J(w) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{k} 1\{y^{(i)} = t\} L\left(h\left(x^{(i)}\right), y^{(i)}\right)^{\{t\}} + \frac{\lambda}{2m}\sum_{j=1}^{n}\sum_{s=1}^{k}\left(w_j^{\{s\}}\right)^2$$

- Parcijalni izvod funkcije greške sa L_2 regularizacijom:

$$\frac{\partial J(w)}{\partial w_r^{\{t\}}} = -\frac{1}{m}\sum_{i=1}^{m} x_r^{(i)}\left(1\left\{y^{(i)} = t\right\} - P\left(y^{(i)} = t \middle| x^{(i)}\right)\right) + \frac{\lambda}{m} w_r^{\{t\}}$$

Multinomijalna logistička regresija – model maksimalne entropije

- Multinomijalna logistička regresija se javlja i pod imenom modela maksimalne entropije (engl. *Maximum Entropy Models*)
- Entropija raspodele slučajne promenljive x se definiše kao:

$$H(x) = -\sum_{x} P(x) \log P(x)$$

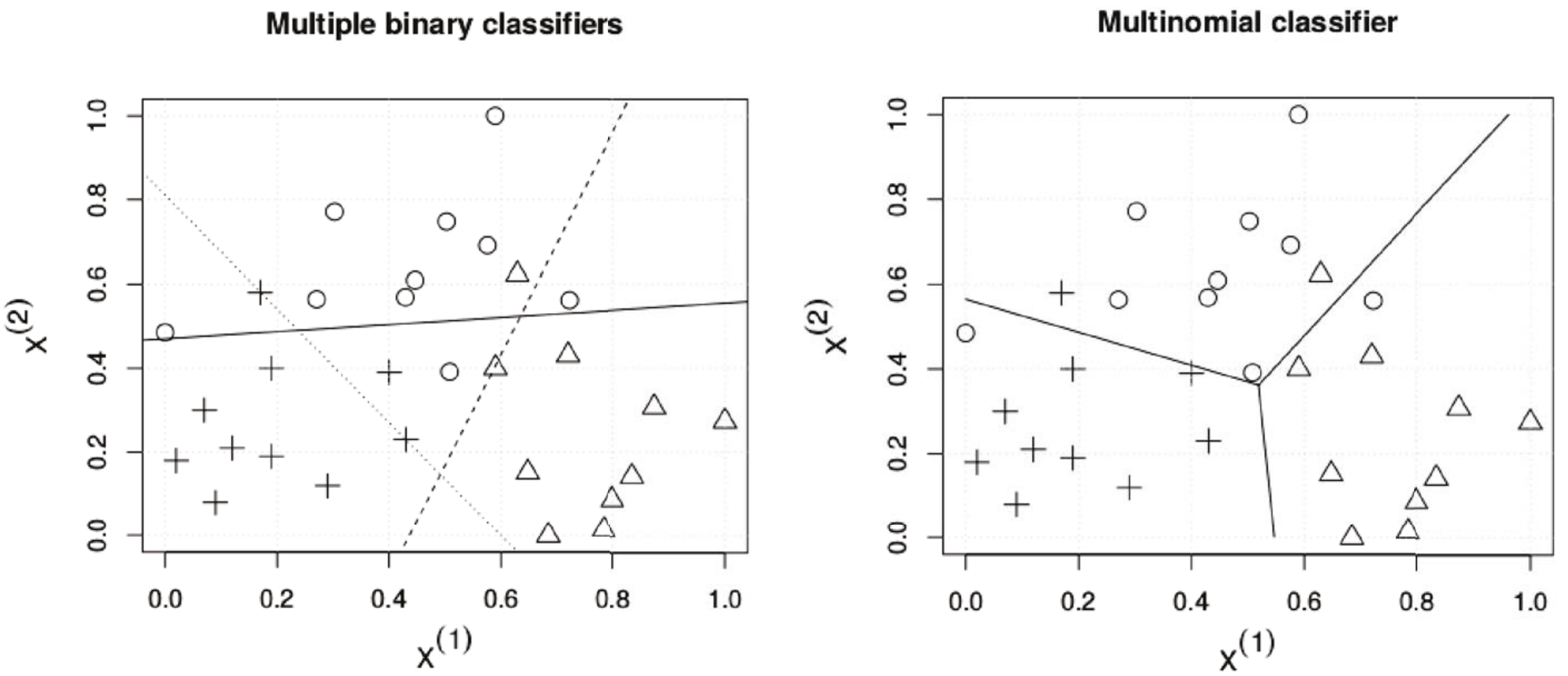
- Intuicija modela maksimalne entropije jeste da probabilistički model koji se izgrađuje treba da poštuje sva ograničenja koja mu se zadaju, ali da sem toga sledi princip Okamove oštrice
  - Treba da donosi zaključke na osnovu što manje dodatnih pretpostavki o izgledu raspodele verovatnoće

Multinomijalna logistička regresija – model maksimalne entropije

- Ako bi se zanemarila sva ograničenja, model sa maksimalnom entropijom bi bio onaj koji svim klasama daje podjednaku verovatnoću, a model sa minimalnom onaj koji svu verovatnoću daje samo jednoj klasi
- Moguće je pokazati da model koji ima maksimalnu entropiju jeste upravo model multinomijalne logističke regresije čiji težinski faktori w maksimizuju uslovne verovatnoće $P(y^{(i)}|x^{(i)})$ nad podacima za obučavanje

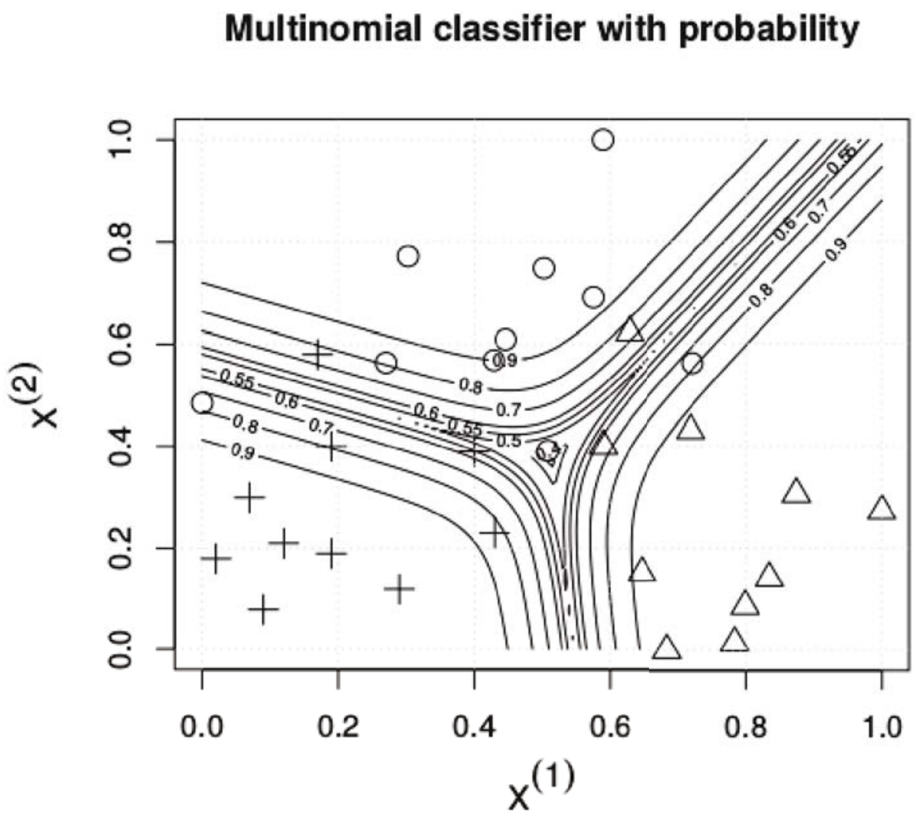
Multinomijalna logistička regresija vs kombinovanje binarnih klasifikatora

- Binarni klasifikatori nisu u stanju da modeluju celu složenost problema kada je broj klasa veći od dve
  - Ignorišu se interakcije između klasa pri optimizaciji parametara w
- Kombinovanje binarnih klasifikatora je osetljivije na *outlier*-e nego multinomijalni pristup
- Multinomijalni pristup omogućava globalni probabilistički izlaz modela
  - Ovo je primereno kada podaci mogu da pripadaju samo jednoj od k klasa
  - Ako podaci mogu da pripadaju većem broju klasa u različitoj meri, onda je primerenije koristiti kombinaciju binarnih klasifikatora



Poređenje kombinacije binarnih klasifikatora i jednog multinomijalnog klasifikatora

Slike preuzete sa: https://www.quora.com/In-multi-class-classification-what-are-pros-and-cons-of-One-to-Rest-and-One-to-One

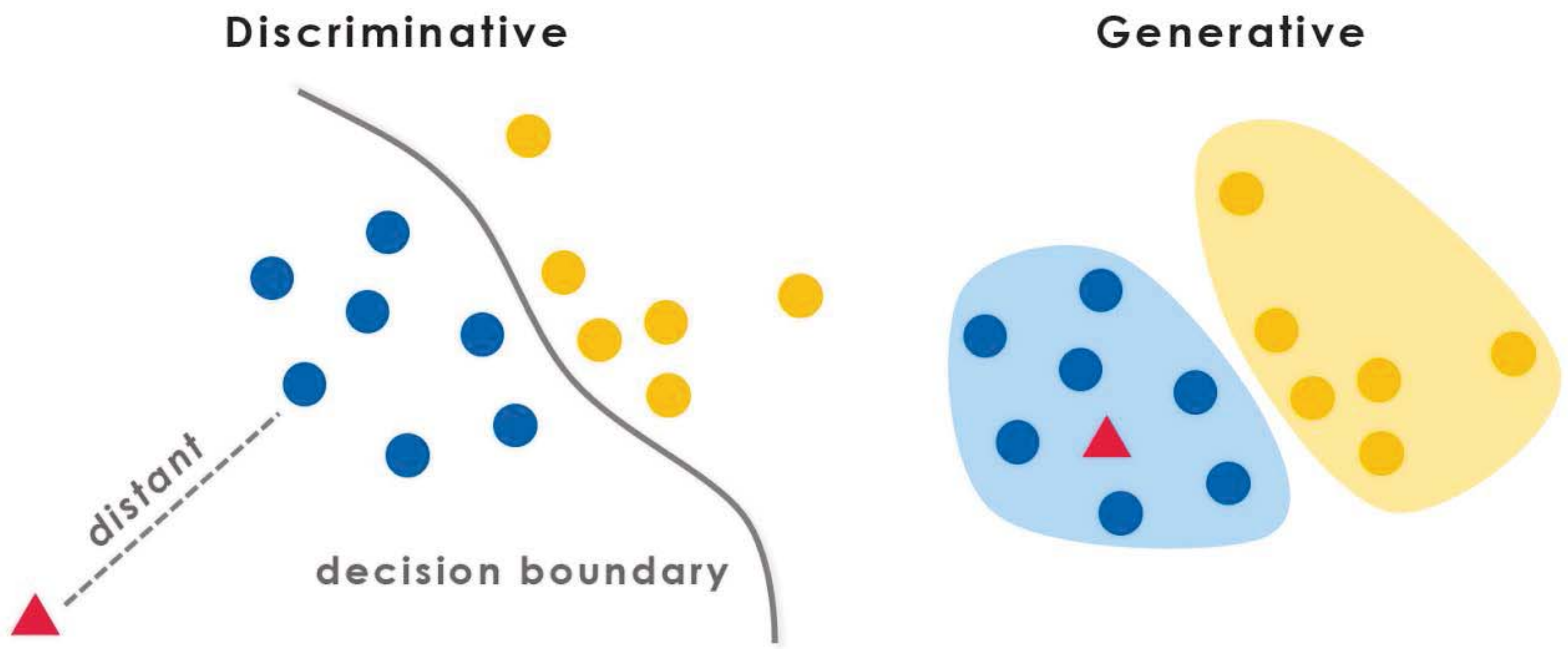


Prikaz probabilističkog izlaza multinomijalnog klasifikatora pri radu sa tri klase

Slika preuzeta sa: https://www.quora.com/In-multi-class-classification-what-are-pros-and-cons-of-One-to-Rest-and-One-to-One

Diskriminativni modeli

- Direktno modeluju verovatnoću $P(y|x)$
- Šire govoreći, diskriminativni modeli se fokusiraju samo na što pravilnije razlikovanje članova različitih klasa tj. na modelovanje granice između klasa (ne nužno i verovatnoće $P(y|x)$)
- Ne zanima ih proces generisanja parova (x, y), za razliku od generativnih modela
  - Model ne uči ono što nije potrebno za rešavanje konkretnog zadatka
- Pored logističke regresije, u diskriminativne modele spadaju:
  - Stabla odlučivanja (engl. *Decision Trees*)
  - Metoda potpornih vektora (engl. *Support Vector Machines*)
  - Uslovna nasumična polja (engl. *Conditional Random Fields*)



Ilustracija razlike između diskriminativnih i generativnih modela

Slika preuzeta sa: http://www.evolvingai.org/fooling

Diskriminativni vs. generativni modeli

- Modelovanje klasifikacione granice
  - Diskriminativni – eksplicitno
  - Generativni – implicitno, tamo gde jedna klasa postaje verovatnija od druge
- Količina podataka za obučavanje
  - Srednja/velika – diskriminativni modeli obično daju bolje performanse
  - Mala – generativni modeli obično daju bolje performanse
- Korišćenje neobeleženih podataka u obučavanju
  - Diskriminativni modeli – teže – modeli po prirodi zahtevaju obeležene podatke za obučavanje
  - Generativni modeli mogu lakše da iskoriste i neobeležene podatke

Diskriminativni vs. generativni pristup - primer

- Zadatak – određivanje na kom jeziku neka osoba govori
- Generativni pristup
  - Naučiti svaki jezik
  - Odabrati onaj naučeni jezik kome je jezik govornika najsličniji
- Diskriminativni pristup
  - Naučiti lingvističke razlike između jezika bez učenja samih jezika
  - Koristeći naučene razlike odrediti jezik govornika
  - Mnogo lakši pristup

Prednosti i mane logističke regresije

- Prednosti
  - Ne pretpostavlja statističku nezavisnost odlika
  - Dobre performanse uz dovoljnu količinu podataka za obučavanje
  - Izlaz je probabilističkog tipa
  - Lako je primenjiva na višeklasnu klasifikaciju
- Mane
  - Lako dolazi do *overfitting*-a kada je količina podataka za obučavanje mala
    - Neophodna je regularizacija
  - Efekti interakcija između odlika se moraju eksplicitno modelovati