Strojno učenje
4
Linearni modeli

Tomislav Šmuc

PMF, Zagreb, 2014

Regresija

Osnovni pojmovi

- Ulazni vektor varijabli (engl. *attributes, features*): $\mathbf{x} = (x_1, x_2, \dots, x_d)$
  - Broj ulaznih varijabli: d
- Izlazna ili ciljna varijabla (engl. *target variable*): y
- Primjer za učenje (engl. *training example*): $(\mathbf{x}, y)$
- Skup primjera za učenje (engl. *training examples*): $\mathrm{D} = \{(\mathbf{x_i}, y_i); i = 1 \dots N\}$ (= poznati podaci)
  - Broj primjera za učenje: N
- Nepoznata ciljna (idealna) funkcija (koncept): $f\colon \mathcal{X} \longrightarrow \mathcal{Y}\ ,\ y = f(\mathbf{x})$
- **Regresija**: y – kontinuirana varijabla

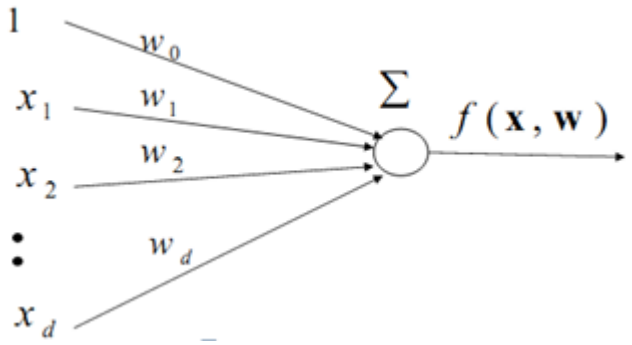
Linearna regresija

- $f: X \rightarrow Y$ je linearna kombinacija ulaznih varijabli

- $f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d = w_0 + \sum_{j=1}^{d} w_j x_j$
  - $w_0, w_1, \ldots, wd$ – parametri modela (težine)

- Radi pojednostavljenja tretmana možemo dodati još jednu (konstantu) varijablu (x_0=1)
-

$$f(\mathbf{x}) = w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d = \boldsymbol{w}^T \boldsymbol{x}$$



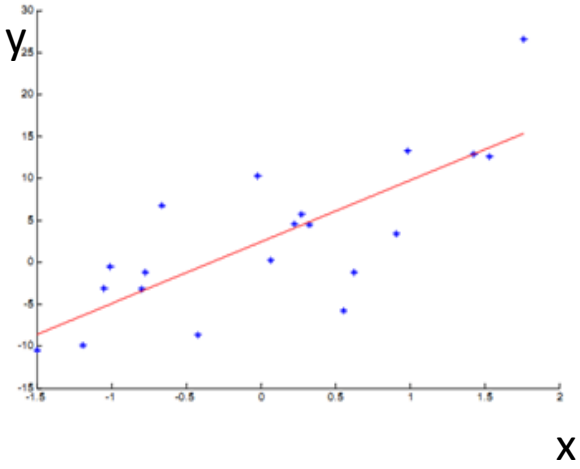
Funkcija greške

- Mjeri koliko predikcije odstupaju od željenih vrijednosti y

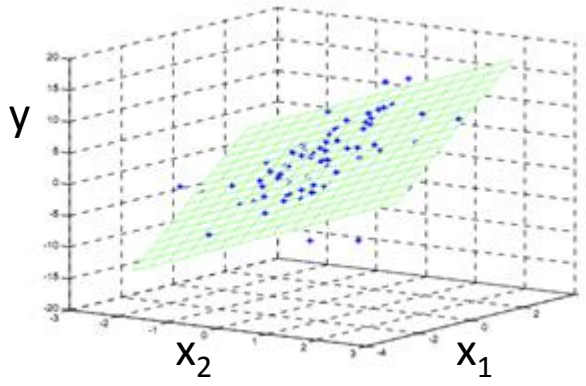
- **Srednja kvadratna pogreška (MSE – Mean Squared Error)**:

$$J = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2$$

d=1



d=2



Minimum funkcije greške => optimizacija

- $f: X \rightarrow Y$ je linearna kombinacija ulaznih varijabli

$$J(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(yi - \boldsymbol{w}^T\boldsymbol{x}_i)^2$$

- Za optimum (optimalne težine **w**) vrijedi da je derivacija $J(w^*) = 0$

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w_j}} = -\frac{2}{n}\sum_{i=1}^{n}(yi - w_0 x_{i,0} - w_1 x_{i,1} - \cdots - wdxi,_d)xi,_j = 0$$

- Ustvari – vektor derivacija = **0** !

$$\boldsymbol{\nabla_w} J(w) =) = -\frac{2}{n}\sum_{i=1}^{n}(yi - \boldsymbol{w}^T\boldsymbol{x}_i)\boldsymbol{x_i}\boldsymbol{=0}$$

Rješenje linearnog regresijskog problema

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w_j}} = -\frac{2}{n}\sum_{i=1}^{n}(yi - w_0 x_{i,0} - w_1 x_{i,1} \ldots - wdxi,_d)xi,_j = 0$$

$$\boldsymbol{\nabla_w} J(\boldsymbol{w}) = -\frac{2}{n}\sum_{i=1}^{n}(yi - \boldsymbol{w}^T \boldsymbol{x}_i)\boldsymbol{x_i} = \mathbf{0}$$

Sistem linearnih jednadžbi sa (d+1) nepoznanicom**:**

$$\mathbf{Aw}=\mathbf{b}$$

Sistem linearnih jednadžbi – *j*-ta komponenta:

$$w_0 \sum_{i=1}^{n} x_{i,0} x_{i,j} + w_1 \sum_{i=1}^{n} x_{i,0} x_{i,j} + \cdots + w_j \sum_{i=1}^{n} x_{i,j} x_{i,j} + w_d \sum_{i=1}^{n} x_{i,d} x_{i,j} = \sum_{i=1}^{n} y_i x_{i,j}$$

Rješenje – inverzija matrica

$$\mathbf{w}=\mathbf{A}^{-1}\mathbf{b}$$

Rješenje linearnog regresijskog problema

$$J(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

$$\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) = -2\boldsymbol{X}^T\,(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = \boldsymbol{0}$$

Rješenje**:**

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Što ako je **X –** singularna (determinanta =0, kolone matrice linearno ovisne)

Rješenje – izbaciti redundantne (linearno ovisne) kolone

Alternativno rješenje

Gradijentno spuštanje !

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \boldsymbol{\beta}\, \nabla_{\boldsymbol{w}} J(\boldsymbol{w}(t))$$

Standardno učenje

- Funkcija greške – sumacija preko grešaka na svim primjerima iz skupa za učenje

$$J(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(yi - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$

Online učenje

- Umjesto ove sume - koristimo grešku za slučajno odabrani primjer $\boldsymbol{x}_i$:

$$J_{online} = \frac{1}{2}\left(y_i - f(\boldsymbol{x}_i)\right)^2$$

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \beta(t+1)\, \nabla_{\boldsymbol{w}} J_{online}(\boldsymbol{w}(t))$$

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \beta(t+1)\left(y_i - f(\boldsymbol{x}_i)\right)\boldsymbol{x}_i$$

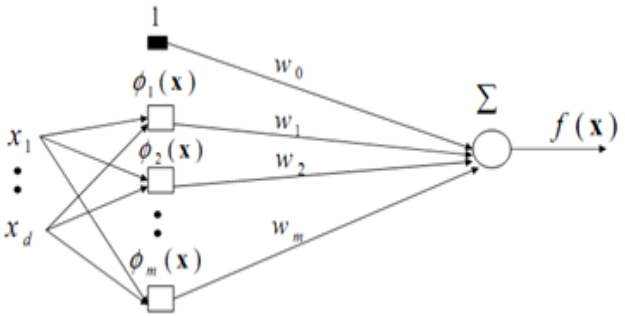
$\beta(t)>0$

Proširenja jednostavnog linearnog modela

Umjesto direktnog korištenja ulaznih varijabli – bazne funkcije (basis functions) – nelinearni modeli:

$$f(\mathbf{x}) = w_0 + \sum_{j=1}^{m} w_j \varphi_j(\boldsymbol{x})$$

gdje su $\varphi_j(\boldsymbol{x})$ – arbitrarne funkcije $\boldsymbol{x}$



Primjeri baznih funkcija $\varphi_j(\boldsymbol{x})$:

$\boldsymbol{x}=(x)$
$\varphi_1(x) = x;\ \varphi_1(x) = x^2;\ \varphi_1(x) = x^3$

$\boldsymbol{x}=(x_1,x_2)$
$\varphi_1(\boldsymbol{x}) = x_1;\ \varphi_2(\boldsymbol{x}) = (x_1)^2;\ \varphi_3(\boldsymbol{x}) = x_2;$
$\varphi_4(\boldsymbol{x}) = (x_2)^2;\ \varphi_5(\boldsymbol{x}) = x_1 x_2$

- Iste tehnike učenja mogu se koristiti za $\boldsymbol{w}$, kao i za obične linearne modele !

Složenost linearnih modela

- Metoda najmanjih kvadrata – tipično mala pristranost (bias) velika varijanca modela

- Prediktivna točnost ili bolje rečeno generalizacija modela može se poboljšati tako da se neki parametri (težine) izjednače s nulom!
  - tako se povećava pristranost nauštrb varijance modela

Rješenja

- Regularizacija
  - Ridge regression
  - Lasso algoritam

- Selekcija (najinformativnijih) varijabli (u jednom od slijedećih predavanja)
- Regresija sa glavnim komponentama (Principal Component Regression)

Ridge regresija (ridge = greben)

$$J(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(yi - \boldsymbol{w}^T\boldsymbol{x}_i)^2 + \lambda||\boldsymbol{w}||^2$$

- Gdje je

$$||\boldsymbol{w}||^2 = \sum_{i=0}^{d}(w_i)^2$$

$$\text{i} \quad \lambda \geq 0$$

- $||\boldsymbol{w}||^2 = \sum_{i=0}^{d}(w_i)^2$ penalizira težine različite od nule sa λ

- Isti efekt je kao da imamo običnu grešku najmanjih kvadrata uz ograničenje na ukupnu sumu kvadrata ***w***:

$$\sum_{i=0}^{d}(w_i)^2 \leq r$$

- Kada imamo nezavisne varijable koje su jako međusobno korelirane njihove težine *w* imaju veliku varijancu - RR uz ograničenje na ukupnu sumu kvadrata težina, rješava efektivno ovaj problem.
- λ – regularizacijski koeficijent (još i shrinkage coefficient)

Lasso regresija / algoritam

$$J(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(yi - \boldsymbol{w}^T\boldsymbol{x}_i)^2 + \lambda||\boldsymbol{w}||$$

- Gdje je

$$||\boldsymbol{w}|| = \sum_{i=0}^{d}||w_i||$$

$$\text{i} \quad \lambda \geq 0$$

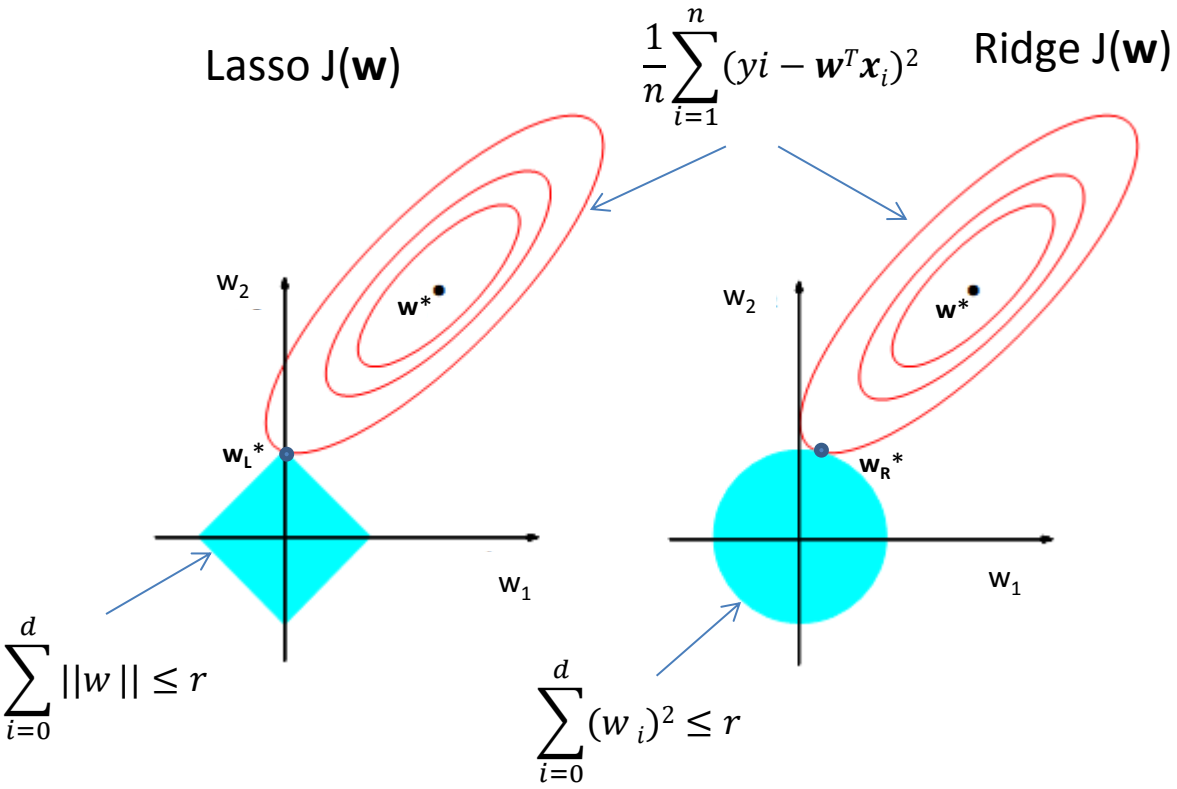
- Isti efekt je kao da imamo običnu grešku najmanjih kvadrata uz ograničenje na ukupnu sumu apsolutnih vrijednosti *w*:

$$\sum_{i=0}^{d}||w_i|| \leq r$$

- Slično kao i kod ridge regresije – no efektivno Lasso sa smanjenjem λ inkrementalno radi selekciju varijabli (težine *w* manje važnih varijabli postaju nula ako se λ postepeno povećava)

- Kako izgledaju $J(\boldsymbol{w})$ za ridge i lasso regresiju ? Kako se ponašaju koeficijenti za različite vrijednosti λ ?

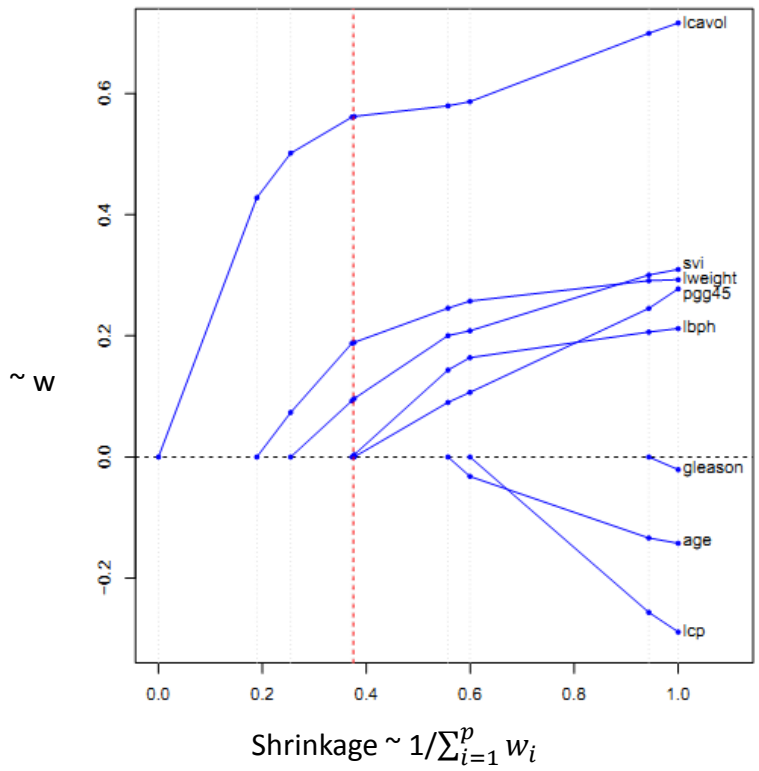
Linearna regresija



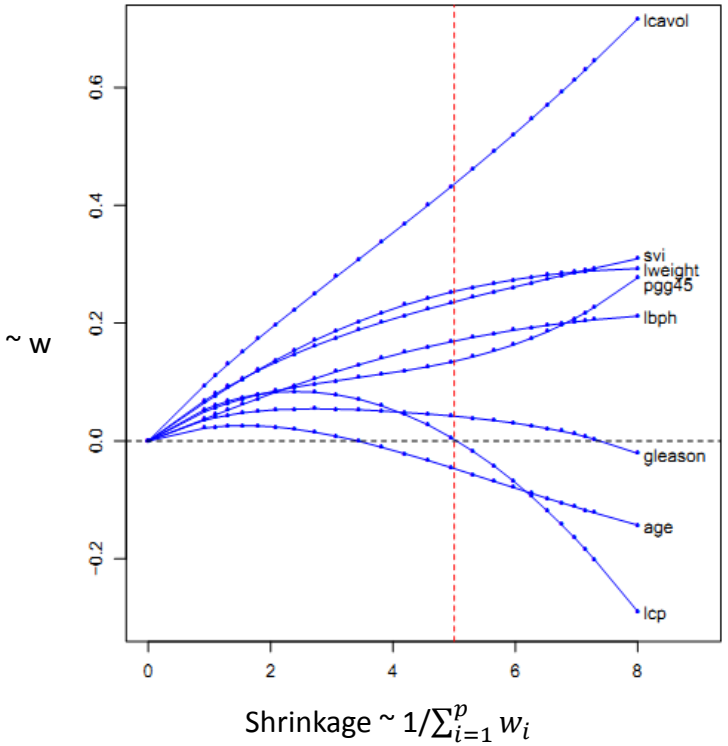
w* - Optimalne težine/parametri ***w*** = (w_1, w_2) bez ograničenja na ***w***
$\mathbf{w}_L$* - Optimalne težine/parametri ***w*** = (w_1, w_2) uz Lasso ograničenje
$\mathbf{w}_R$* - Optimalne težine/parametri ***w*** = (w_1, w_2) uz Ridge ograničenje

Ponašanje težina varijabli **w** za različite vrijednosti λ (odnosno ograničenje *r*)

Lasso **w**

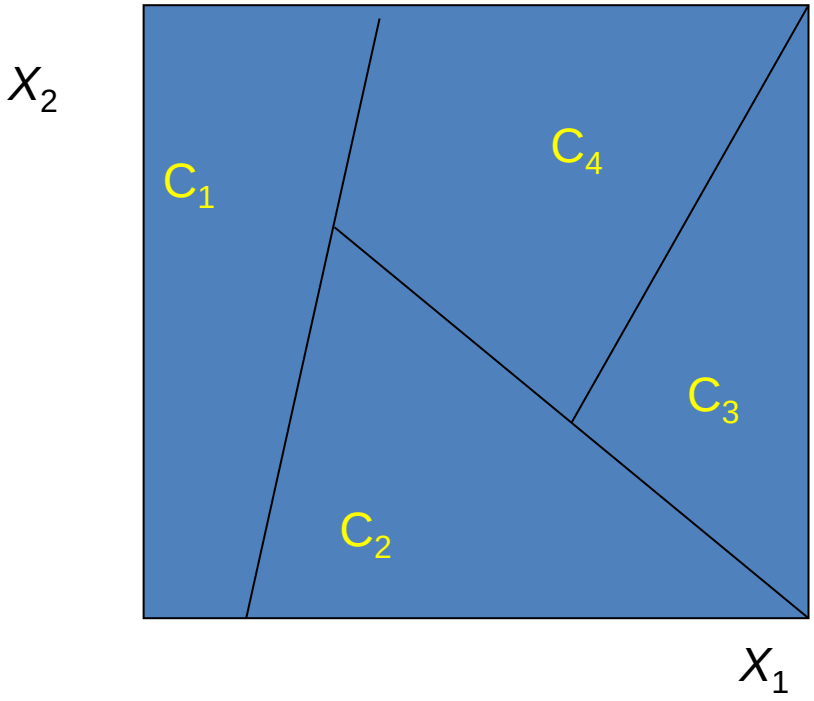


Ridge **w**



Linearni klasifikacijski modeli

- Što je linearna klasifikacija
  - Plohe između primjera različitih klasa su linearne (po dijelovima !)



Linearna metode

- Što su Linearne metode klasifikacije?

  Metode koje daju linearne granice između različitih klasa

  $\{x: \beta_0 + \beta_1^T x = 0\}$

- Dva pristupa kako definirati granice između klasa
  - Modeliranje diskriminantne funkcije $\delta_k(x)$ za svaku od klasa kao linearne
    - Linearna regresija indikatorske matrice klasa
    - Logistička regresija (LOGREG)
    - Linearna diskriminantna analiza (LDA)
  - Modeliranje granice između klasa kao linearne funkcije
    - Perceptron
    - Metoda potpornih vektora (Support Vector Machines)

Modeliranje diskriminantne funkcije $\delta_k(x)$

- Model – diskriminantne funkcije $\delta_k(x)$

  Različit za linearnu regresiju, LogReg i LDA
- Na granici između klasa *j* i *k*

  $\{x: \delta_j(x) = \delta_k(x)\}$
- Klasa je određena kao *k* za koju je funkcija $\delta_k(x)$ najveća

$$C(x) = \arg\max_{k \in g} \delta_k(x)$$

Linearna Regresija ciljne (klasne) - indikatorske varijable

- Imamo *K* ciljnih varijabli (indikatorske)
  K = broj klasa
- Linearni model za *k*-tu indikatorsku varijable → $f_k(\mathbf{x}) = w_{k0} + \mathbf{w}_k^T \mathbf{x}$
- Linearna diskriminantna funkcija za klasu *k:* → $\delta_k(\mathbf{x}) = f_k(\mathbf{x})$
- Granica između klasa je skup točaka za koje je: → $\{\mathbf{x} : f_k(\mathbf{x}) = f_l(\mathbf{x})\} = \{x : (w_{k0} - w_{l0}) + (\mathbf{w}_k - \mathbf{w}_l)^T \mathbf{x} = 0\}$
- Klasifikacija novog primjera – u klasu sa najvećom $\delta_k(x)$ → $C(\mathbf{x}) = \arg\max_{k \in C} \delta_k(\mathbf{x})$

Linearna Regresija ciljne (klasne) - indikatorske varijable

- Određivanje parametara
  - Funkcija greške - cilj optimizacije => suma najmanjih kvadrata (RSS)

$$\mathbf{W} = \arg\min_{\mathbf{w}} \mathrm{RSS}(\mathbf{W}) = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \| y_i - [(1, \mathbf{x}_i)\mathbf{W}]^T \|^2$$

  - Određivanje koeficijenata - težina

$$\mathbf{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} w_{10} & \cdot & \cdot & \cdot & w_{K0} \\ w_d & \cdot & \cdot & \cdot & w_{Kd} \end{bmatrix}_{(d+1)\times K}$$

Linearna Regresija ciljne (klasne) - indikatorske varijable

- Ako naš klasifikacijski problem *C* ima *K* klasa – imamo *K* klasnih indikatorski varijabli y_k, *k=1,K:*

C	y_1	y_2	y_3	…	y_K
1	1	0	0	..	0
3	0	0	1	..	0
4	0	0	0	..	0
K	0	0	0	..	1
2	0	1	0	..	0

- Odredimo regresijski model za svaku y_k :

$$\hat{\mathbf{y}}_k(\mathbf{x}) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_k$$

Linearna Regresija ciljne (klasne) - indikatorske varijable

- Definiramo matricu procjene za sve indikatorske varijable:

$$\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K)$$

Klasifikacijska procedura

- Definiramo matricu **W:** ⟹ $\mathbf{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

- Za neki novi primjer **x,** izračunamo: ⟹ $\mathbf{f}(\mathbf{x}) = [(1, \mathbf{x})\mathbf{W}]^T = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \ldots \\ \ldots \\ f_K(\mathbf{x}) \end{pmatrix}$

- Na kraju – klasa **x** određuje se prema najvećoj komponenti **f(x)**: ⟹

$$C(\mathbf{x}) = \arg\max_{k \in C} f_k(\mathbf{x})$$

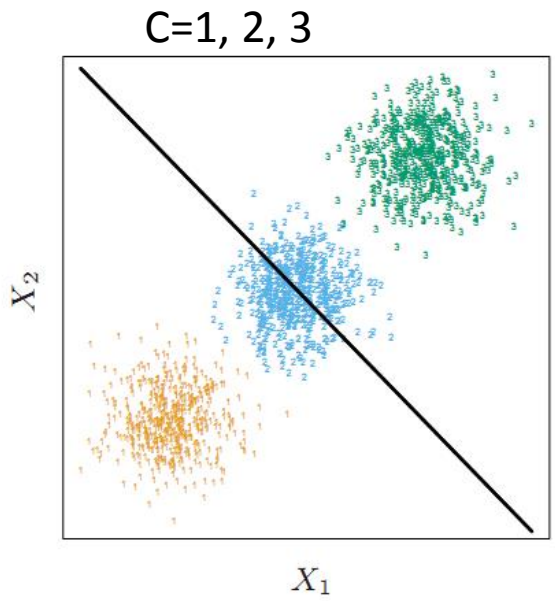
- Pojašnjenje
  - Linearna Regresija indikatorske varijable (klase) Y_k linearna je aproksimacija očekivanja $E(Y_k|X)$

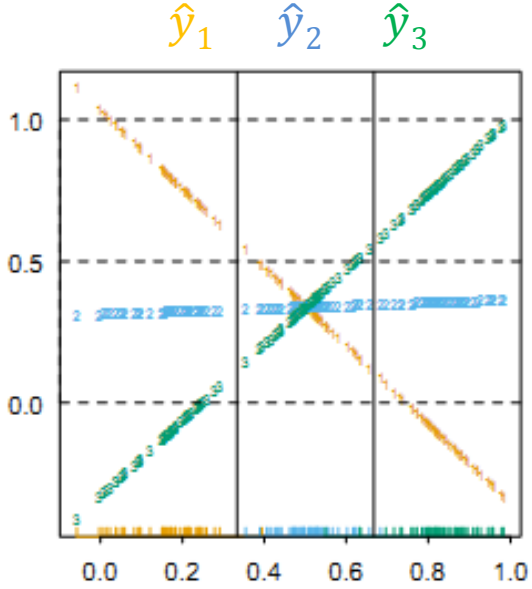
$$f_k(\mathbf{x}) = E(Y_k \mid X = \mathbf{x}) = P(C = k \mid X = \mathbf{x})$$

  - Odnosno aposteriorne vjerojatnosti indeksa klase ciljne varijable

$$\begin{aligned} E(Y_k \mid X = \mathbf{x}) &= P(Y_k = 1 \mid X = \mathbf{x}) \cdot 1 + P(Y_k = 0 \mid X = \mathbf{x}) \cdot 0 \\ &= P(Y_k = 1 \mid X = \mathbf{x}) \\ &= P(C = k \mid X = \mathbf{x}) \end{aligned}$$

- Problemi s linearnom regresijom – „maskiranje” klasa





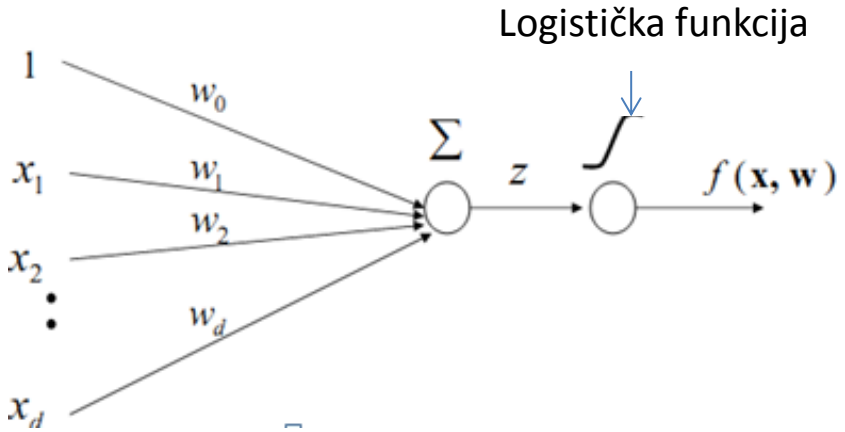
- $\hat{y}_2$ – nikada ne dominira nad $\hat{y}_1$ i $\hat{y}_3$
- Svi primjeri klase C=2 se klasificiraju kao 1 ili 3 !?
- „maskiranje” klasa - za velike K (>3)

Logistička regresija

- Definira linearne granice između klasa – diskriminativni algoritam
- Diskriminativne funkcije

  $g_1(\boldsymbol{x}) = g(\boldsymbol{w^T x})$; $g_0(\boldsymbol{x}) = 1 - g(\boldsymbol{w^T x})$;

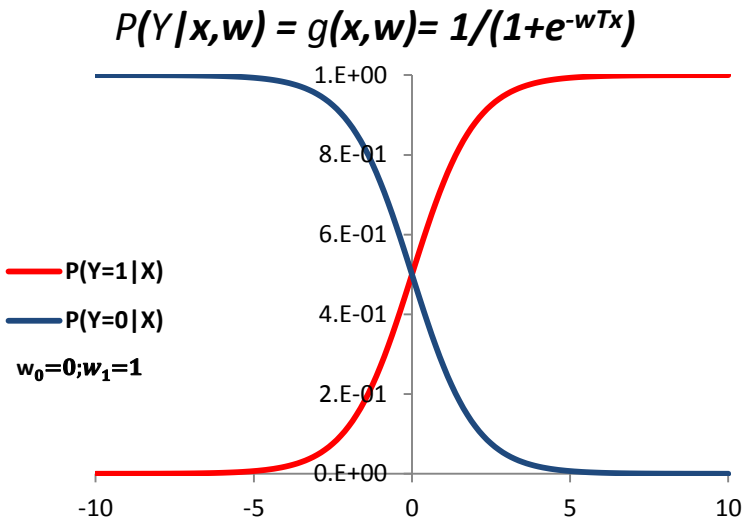
- Gdje je $g(\boldsymbol{w^T x}) = \frac{1}{1+e^{-\boldsymbol{w^T x}}} = f(\boldsymbol{x}, \boldsymbol{w})$ - logistička funkcija (sigmoidalna funkcija)



- Vrijednosti logističke funkcije $\in [0,1]$!

Logistička regresija

Probabilistička interpretacija !



$$p(y = 1|\boldsymbol{x}, \boldsymbol{w}) = f(\boldsymbol{x}, \boldsymbol{w}) = g(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}}$$
$$p(y = 0|\boldsymbol{x}, \boldsymbol{w}) = 1 - p(y = 1|\boldsymbol{x}, \boldsymbol{w})$$

Klasifikacija:

Ako $p(y = 1|\boldsymbol{x}, \boldsymbol{w}) \geq 0.5$ tada y=1

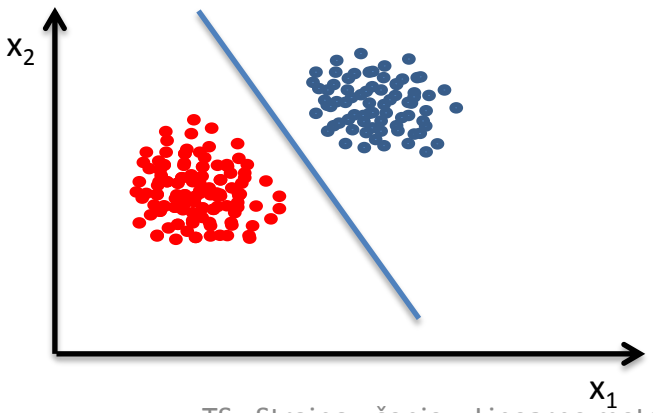
Inače $y = 0$

Logistička regresija

- Definira linearnu diskriminativnu plohu između klasa
- Zašto?
  - Na plohi vrijedi da su diskriminativne funkcije jednake: $g_1(\boldsymbol{x}) = g_0(\boldsymbol{x})$, dakle:

$$log\left(\frac{g_0(\boldsymbol{x})}{g_1(\boldsymbol{x})}\right) = \log\left(\frac{1 - g(\boldsymbol{w}^T\boldsymbol{x})}{g(\boldsymbol{w}^T\boldsymbol{x})}\right) = 0$$

$$log\left(\frac{g_0(\boldsymbol{x})}{g_1(\boldsymbol{x})}\right) = log\frac{\frac{exp^{-\boldsymbol{w}^T\boldsymbol{x}}}{1 + exp^{-\boldsymbol{w}^T\boldsymbol{x}}}}{\frac{1}{1 + exp^{-\boldsymbol{w}^T\boldsymbol{x}}}} = \log(exp^{-\boldsymbol{w}^T\boldsymbol{x}}) = -\boldsymbol{w}^T\boldsymbol{x} = 0$$



Logistička regresija

- Učenje parametara $\boldsymbol{w}$
- Vjerojatnost podataka $L(D, \boldsymbol{w})$ uz $\boldsymbol{w}$
  - $D_i = \langle \boldsymbol{x_i}, y_i \rangle$; $\mu_i = p(y_i = \mathbf{1} | \boldsymbol{x_i}, \boldsymbol{w}) = g(\boldsymbol{w^T x})$
  - $L(D, \boldsymbol{w}) = \prod_{i=1}^{n} P(y = yi | \boldsymbol{x}_i, \boldsymbol{w}) = \prod_{i=1}^{n} {\mu_i}^{y_i} \ (1 - \mu_i)^{(1-yi)}$

- Odrediti težine $\boldsymbol{w}$ koje maksimiziraju vjerojatnost podataka
- Trik: Logaritam vjerojatnosti (Log-Likelihood)
  - Optimalne težine jednake su i za $L(D, \boldsymbol{w})$ i za $log(L(D, \boldsymbol{w}))$!

$$log\big(L(D, \boldsymbol{w})\big) = log \prod_{i=1}^{n} {\mu_i}^{y_i} \ (1 - \mu_i)^{(1-yi)} = \sum_{i=1}^{n} log \, {\mu_i}^{y_i} (1 - \mu_i)^{(1-yi)} =$$

$$= \sum_{i=1}^{n} (y_i log \, \mu_i + (1 - yi) log(1 - \mu_i))$$

Logistička regresija - Učenje parametara *w*

- Log-likelihood

$$log\big(L(D,\boldsymbol{w})\big) = \sum_{i=1}^{n} (y_i log\, \mu_i + (1 - yi) log(1 - \mu_i))$$

- Derivacija $log\big(L(D,\boldsymbol{w})\big)$ => negativni gradijent

$$-\frac{\partial}{\partial w_j} log\big(L(D,\boldsymbol{w})\big) = \sum_{i=1}^{n} -xi,_j(y_i - g(\boldsymbol{w^T x}_i))$$

$$\nabla_{\boldsymbol{w}}[-log\big(L(D,\boldsymbol{w})\big)] = \sum_{i=1}^{n} -\boldsymbol{x}_i(y_i - g(\boldsymbol{w^T x}_i))$$

- Gradijentno spuštanje

$$\boldsymbol{w}^{(k)} = \boldsymbol{w}^{(k)} - \beta(k)\nabla_{\boldsymbol{w}}[-log(L(D,\boldsymbol{w}))]_{\boldsymbol{w}^{(k\_1)}}$$

$$\boldsymbol{w}^{(k)} = \boldsymbol{w}^{(k)} - \beta(k) \sum_{i=1}^{n} -\boldsymbol{x}_i(y_i - g(\boldsymbol{w^T x}_i))$$

- Parametri LogReg se također mogu učiti korištenjem online metode !

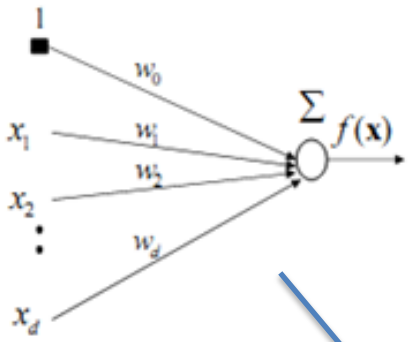
Logistička regresija - Algoritam za online učenje parametara *w*

Online LogReg (D, broj_iteracija)

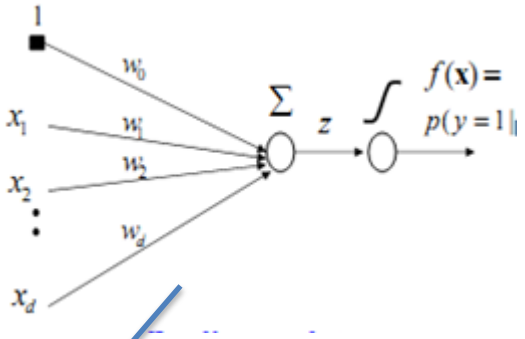
Incijaliziraj težine $w^{(0)}=(w_0, w_1, w_2,\ldots w_d)$
For *i=1:broj_iteracija*
    do
        izaberi primjer iz $D=<\boldsymbol{x}_k, y_k>$
        postavi $\beta_i = 1/i$
        odredi *nove težine*
        $\boldsymbol{w}=\boldsymbol{w}+ \beta_i [\, y_i - g(\boldsymbol{w}^T\boldsymbol{x}) \,] \, \boldsymbol{x}_i$
end for
Vrati težine ***w***

Usporedba Linearna regresija i Logistička regresija

$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$



$f(\boldsymbol{x}) = p(y = 1 \mid \boldsymbol{x}, \boldsymbol{w}) = g(\boldsymbol{w}^T\boldsymbol{x})$



$$\boldsymbol{w} = \boldsymbol{w} + \beta \sum_{i=1}^{n} (y - f(\boldsymbol{x}))\boldsymbol{x}$$

Učenje modela je isto !

Nelinearna ekstenzija LogReg i Linearne regresije

Korištenje nelinearnih baznih funkcija

Linearna regresije

$$f(\boldsymbol{x}) = w_0 + \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} w_j \varphi_j(\boldsymbol{x})$$

Logistička regresija

$$f(\boldsymbol{x}) = g(w_0 + \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} w_j \varphi_j(\boldsymbol{x}))$$

$\varphi j(\boldsymbol{x})$ **-** arbitrarne funkcije **x**

LDA

Pretpostavke:
- (više-dim.) Gaussovu distribuciju kao model gustoće uvjetne vjerojatnosti po klasama
- Istu kovarijancu po varijablama za sve klase

Klasifikacija se bazira na određivanju aposteriorne vjerojatnosti za klasu:

$$P(C = k \mid \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{l=1}^{K} f_l(\mathbf{x})\pi_l}$$

← Bayesovo pravilo

π_k – apriorna vjerojatnost pojave klase *k*

Klasa se modelira preko $f_k(\boldsymbol{x})$ – gustoća uvjetne vjerojatnosti ($p(\boldsymbol{x} \mid C=k)$)

$$f(\mathbf{x}) = p(\mathbf{x} \mid c_k) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma}_k \mid^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}-\boldsymbol{\mu}_k)}$$

Parametri:

Na bazi skupa primjera za učenje:

apriorne vjerojatnosti: $\hat{\pi}_k = N_k / N$

srednje vrijednosti: $\hat{\mu}_k = \sum_{g_i=k} x_i / N_k$

kovarijacijska matrica: $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)$

(Parametri su određeni po principu ML (max likelihood) – uz prethodne pretpostavke).

uz korištenje log-odds…

$$\log \frac{P(C=k \mid X=\mathbf{x})}{P(C=l \mid X=\mathbf{x})} = \log \frac{\pi_k}{\pi_l} + \log \frac{f_k(\mathbf{x})}{f_l(\mathbf{x})}$$

$$= \underbrace{(\log \pi_k + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \boldsymbol{\Sigma}^{-1} \mu_k)}_{\delta_k(\mathbf{x})} - \underbrace{(\log \pi_l + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mu_l - \frac{1}{2} \mu_l^T \boldsymbol{\Sigma}^{-1} \mu_l)}_{\delta_l(\mathbf{x})}$$

Diskriminativne funkcije za klasu *k* i *l*:

Klasifikacija: $C(\mathbf{x}) = \arg\max_k \delta_k(\mathbf{x}) \equiv \arg\max_k P(C=k \mid \mathbf{x})$

Granica između klasa je definirana gdje vrijedi ($\delta_k(\boldsymbol{x}) = \delta_l(\boldsymbol{x})$):

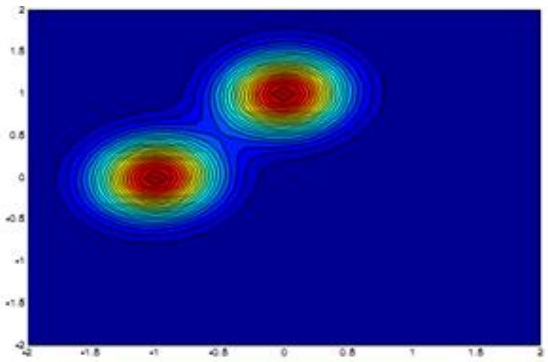
$$\log \pi_k + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \boldsymbol{\Sigma}^{-1} \mu_k = \log \pi_l + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mu_l - \frac{1}{2} \mu_l^T \boldsymbol{\Sigma}^{-1} \mu_l$$

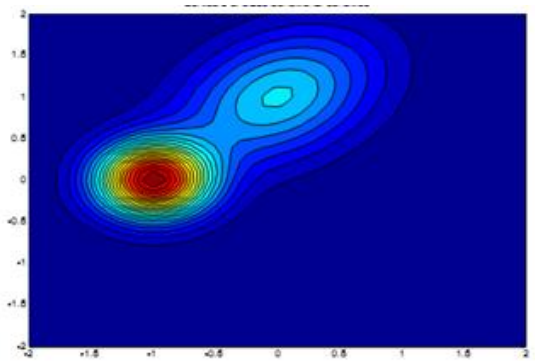
Kvadratna diskriminativna analiza - QDA

- Relaksira pretpostavku-uvjet iste kovarijacijske matrice – gustoće vjerojatnosti za klase (multivarijantne Gaussove distribucije) mogu imati različite kovarijacijske matrice
- Posljedica – granice između klasa nisu linearne, nego kvadratne !

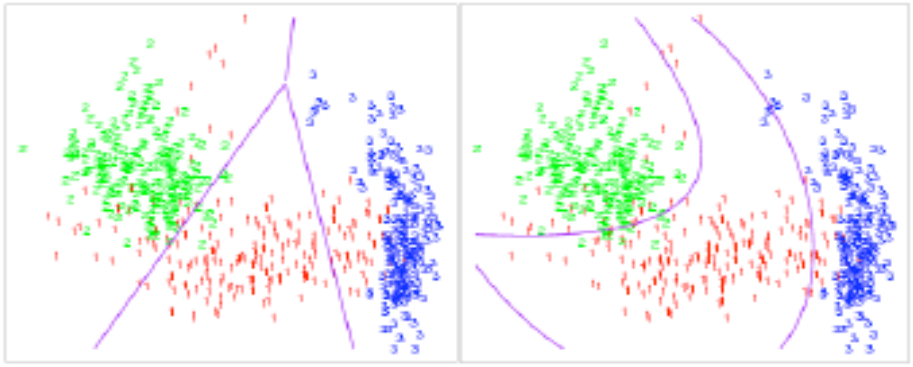
$$\underbrace{\log \pi_k + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \boldsymbol{\Sigma}^{-1} \mu_k}_{\delta_k(x) - \mathbf{LDA}}$$

$$\underbrace{\log \pi_k - \frac{1}{2}(\mathbf{x} - \mu_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \mu_k) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k|}_{\delta_k(x) - \mathbf{QDA}}$$





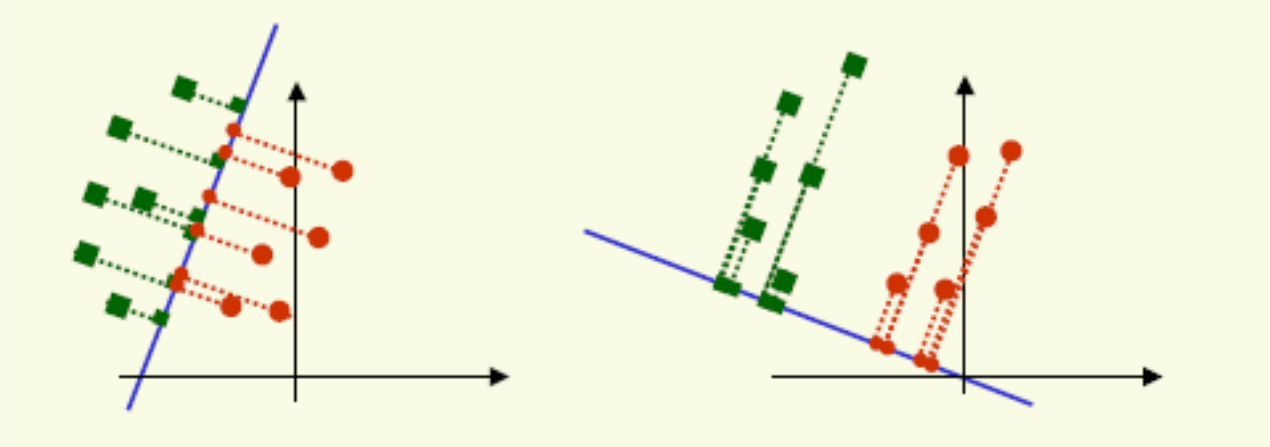
LDA vs QDA - Granice između klasa



LDA QDA

Fisherova linearna diskriminativna metoda - FDA

Osnovna ideja – naći projekciju na liniju u d- dimenzionalnom prostoru tako da se primjerci različitih klasa mogu na njoj lako odvojiti



loša projekcija

dobra projekcija

Fisherova linearna diskriminativna metoda

Neke veličine:

$$\boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

Srednja vrijednost u d-dimenzionalnom prostoru – za klasu *i*

$$\overline{\mu}_i = \frac{1}{N_i} \sum_{y \in Y_i} y = \frac{1}{N_i} \sum_{x \in D_i} \mathbf{w}^T \mathbf{x} = \mathbf{w}^T \boldsymbol{\mu}_i$$

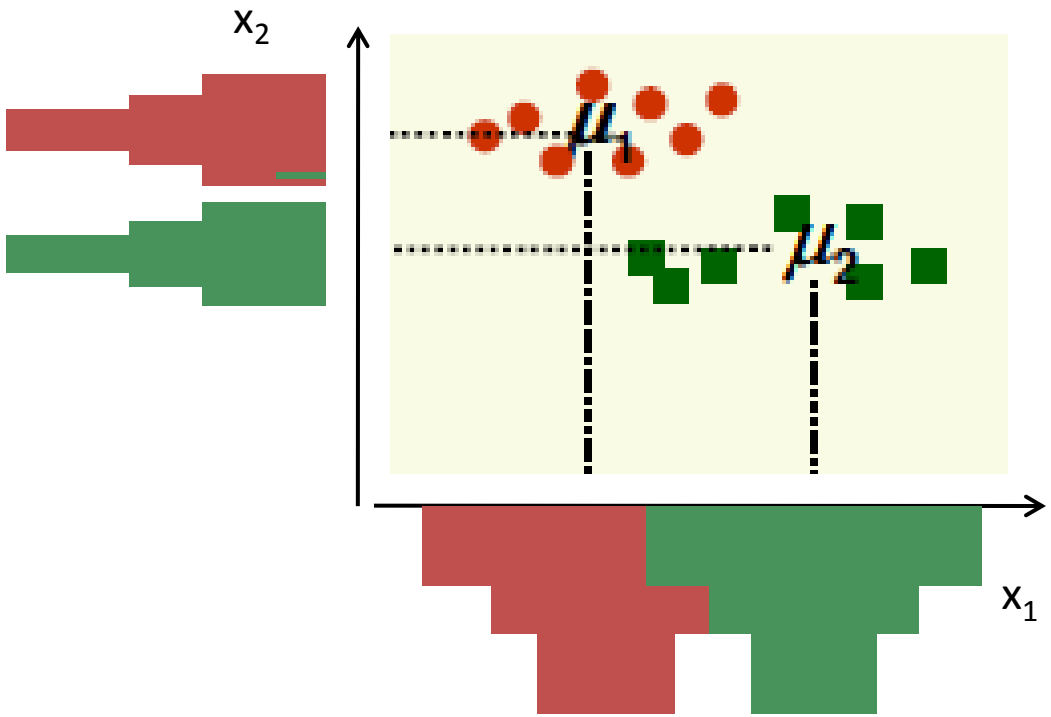
Srednja vrijednost za točke klase *i* projicirane na **w**

$$\left| \overline{\mu}_1 - \overline{\mu}_2 \right| = \left| \mathbf{w}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \right|$$

Udaljenost između projiciranih srednjih vrijednosti za dvije klase

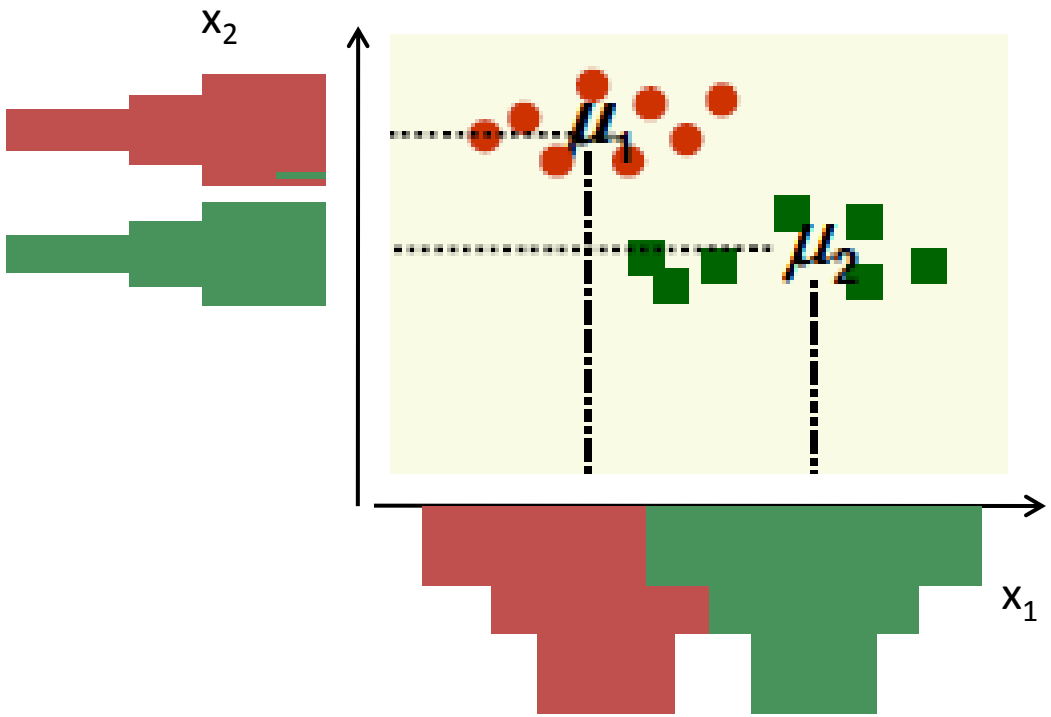
Fisherova linearna diskriminativna metoda

Koliko je dobra mjera separacije $\left|\overline{\mu}_1 - \overline{\mu}_2\right|$?



Koja od osi je bolja za razdvajanje klasa, x_1 ili x_2 ?

Koliko je dobra mjera separacije $|\overline{\mu}_1 - \overline{\mu}_2|$?



x_2 je bolja, no:

$$|\overline{\mu}_1 - \overline{\mu}_2|_{x1} > |\overline{\mu}_1 - \overline{\mu}_2|_{x2}$$

problem je što $|\overline{\mu}_1 - \overline{\mu}_2|$

ne uzima u obzir varijancu

distribucije primjera.

Fisherova linearna diskriminantna metoda

Ako definiramo:

$$y_i = \mathbf{w}^T \mathbf{x}_i$$ projicirani primjeri

$$\tilde{s}_1^2 = \sum_{y_i \in Cl_1} (|y_i - \bar{\mu}_1|)$$ "raspršenje" primjera klase1

$$s_2^2 = \sum_{y_i \in Cl_2} (|y_i - \bar{\mu}_2|)$$ "raspršenje" primjera klase2

Možemo raspršenje koristiti za normalizaciju udaljenosti projiciranih "centara" između klasa !

Fisherova linearna diskriminantivna metoda

- Moramo normalizirati koristeći raspršenje klase 1 i klase 2 !
- FDA dakle svodi se na pronalaženje projekcije na liniju **w** koja maksimizira J(**w**):

$$J(w) = \frac{\left|\overline{\mu}_1 - \overline{\mu}_2\right|}{\left|\widetilde{S}_1^{\,2} + \widetilde{S}_2^{\,2}\right|}$$

želimo da brojnik bude što veći

a nazivnik što manji!

$$J(w) = \frac{\left|\bar{\mu}_1 - \bar{\mu}_2\right|}{\left|\tilde{s}_1^2 + \tilde{s}_2^2\right|}$$

- Kako izraziti J(**w**) kao funkciju – radi se zapravo o optimizaciji J(**w**),
  - u ovisnosti o **w !**

- Treba eksplicitno prikazati J u ovisnosti o **w !**

- Definiramo matrice raspršenja za svaku klasu – prije projekcije - za originalne primjere

$$\mathbf{S}_1 = \sum_{\mathbf{x}_i \in Cl_1} (\mathbf{x}_i - \mu_1)(\mathbf{x}_i - \mu_1)^T$$

$$\mathbf{S}_2 = \sum_{\mathbf{x}_i \in Cl_2} (\mathbf{x}_i - \mu_2)(\mathbf{x}_i - \mu_2)^T$$

Fisherova linearna diskriminantivna metoda

- Definiramo matricu raspršenja *unutar* klasa za svaku klasu

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

- Uz prethodnu definiciju

$$\tilde{s}_1^2 = \sum_{y_i \in Cl_1} (y_i - \bar{\mu}_1)^2$$

- I ako koristimo:

$$y_i = \mathbf{w}^T \mathbf{x}_i \quad \text{i} \quad \bar{\mu}_1 = \mathbf{w}^T \boldsymbol{\mu}_1$$

- Dobivamo:

$$\begin{aligned}
\tilde{s}_1^2 &= \sum_{y_i \in Cl} (\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \boldsymbol{\mu}_1)^2 \\
&= \sum_{y_i \in Cl} (\mathbf{w}^T (\mathbf{x}_i - \boldsymbol{\mu}_1))^T (\mathbf{w}^T (\mathbf{x}_i - \boldsymbol{\mu}_1)) \\
&= \sum_{y_i \in Cl} ((\mathbf{x}_i - \boldsymbol{\mu}_1)^T \mathbf{w})^T ((\mathbf{x}_i - \boldsymbol{\mu}_1)^T \mathbf{w}) \\
&= \sum_{y_i \in Cl} \mathbf{w}^T (\mathbf{x}_i - \boldsymbol{\mu}_1)(\mathbf{x}_i - \boldsymbol{\mu}_1)^T \mathbf{w} = \mathbf{w}^T \mathbf{S}_1 \mathbf{w}
\end{aligned}$$

- Slično kao i za klasu 1

$$\tilde{s}_2^2 = \mathbf{w}^T \mathbf{S}_2 \mathbf{w}$$

- Dakle

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^T \mathbf{S}_1 \mathbf{w} + \mathbf{w}^T \mathbf{S}_2 \mathbf{w} = \mathbf{w}^T \mathbf{S}_W \mathbf{w}$$

- Ako definiramo matricu raspršenja između klasa $\mathbf{S}_B$ kao mjeru separacije između srednjih vrijednosti između klasa prije projekcije

$$\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$$

- A razlika između projiciranih srednjih vrijednosti je:

$$\begin{aligned}(\bar{\mu}_1 - \bar{\mu}_2)^2 &= (\mathbf{w}^T \boldsymbol{\mu}_1 - \mathbf{w}^T \boldsymbol{\mu}_2)^2 \\ &= \mathbf{w}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{w} \\ &= \mathbf{w}^T \mathbf{S}_B \mathbf{w}\end{aligned}$$

Fisherova linearna diskriminantivna metoda

- Na kraju je naša funkcija cilja

$$J(\mathbf{w}) = \frac{(\overline{\mu}_1 - \overline{\mu}_2)^2}{\widetilde{s}_1^2 + \widetilde{s}_2^2} = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}}$$

  - Da bi je optimirali, našli maksimum – prva derivacija po **w** = 0

$$\frac{d}{d\mathbf{w}} J(\mathbf{w}) = 0$$

    - Na koncu se to svede na problem određivanja svojstvenih vrijednosti

$$\Rightarrow \mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

$$\Rightarrow \mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}$$

- Ako postoji inverzna matrica - nakon sređivanja:

$$\mathbf{w} = \mathbf{S}_w^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

- Za potrebe klasifikacije - još je potrebno odrediti graničnu vrijednost *t,* kojom se konačno određuje output diskriminativne funkcije:

$$y_i = \mathbf{w}^T\mathbf{x}_i < \mathrm{t} \Rightarrow \mathrm{y}_1$$

$$y_i = \mathbf{w}^T\mathbf{x}_i \geq \mathrm{t} \Rightarrow \mathrm{y}_2$$

Literatura

Linearne metode

The Elements of Statistical Learning

Hastie, Tibshirani, Friedman (chapter 4)