

Strojno učenje 6 Linearni modeli

Tomislav Šmuc

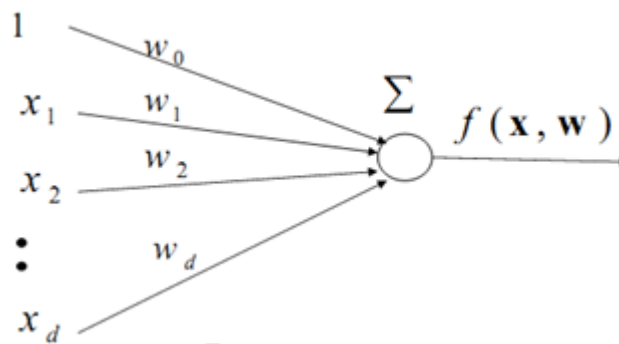
Osnovni pojmovi

- 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*): $D = \{(\mathbf{x}_i, y_i); i = 1 \dots N\}$ (= poznati podaci)
 - Broj primjera za učenje: N
- Nepoznata ciljna (idealna) funkcija (koncept): $f: \mathcal{X} \rightarrow \mathcal{Y}, y = f(\mathbf{x})$
- **Regresija:** y – kontinuirana varijabla

Linearna regresija

- $f: X \rightarrow Y$ je linearna kombinacija ulaznih varijabli
- $f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + \dots + w_dx_d = w_0 + \sum_{j=1}^d w_jx_j$
 - w_0, w_1, \dots, w_d – parametri modela (težine)
- Radi pojednostavljenja (vektorski zapis) možemo dodati još jednu (konstantu) varijablu ($x_0=1$)

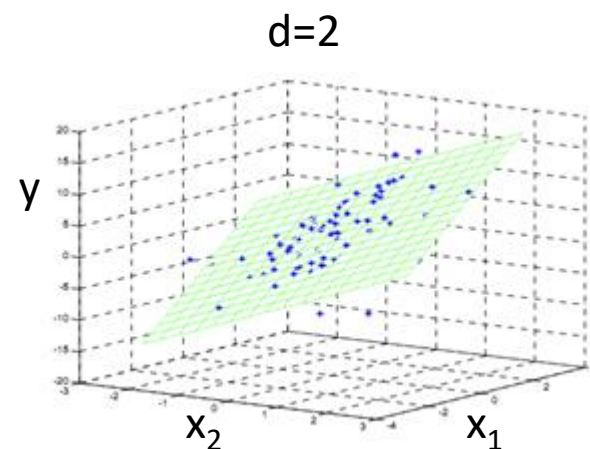
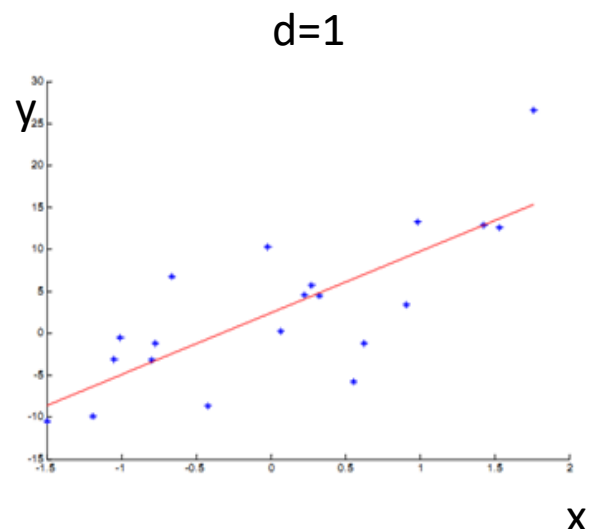
$$f(\mathbf{x}) = w_0x_0 + w_1x_1 + w_2x_2 + \dots + w_dx_d = \mathbf{w}^T \mathbf{x}$$



Funkcija greške

- Mjeri koliko predikcije odstupaju od željenih vrijednosti y
- **Srednja kvadratna pogreška (MSE – Mean Squared Error):**

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



Minimum funkcije greške => optimizacija

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

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

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

$$\frac{\partial J(\mathbf{w})}{\partial w_j} = -\frac{2}{n} \sum_{i=1}^n (y_i - w_0 x_{i,0} - w_1 x_{i,1} - \dots - w_d x_{i,d}) x_{i,j} = 0$$

Rješenje linearnog regresijskog problema

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}_j} = -\frac{2}{n} \sum_{i=1}^n (y_i - w_0 x_{i,0} - w_1 x_{i,1} \dots - w_d x_{i,d}) x_{i,j} = 0$$

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = -\frac{2}{n} \sum_{i=1}^n (y_i - \mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i = 0$$

Sistem linearnih jednadžbi sa (d+1) nepoznanicom:

$$\mathbf{A}\mathbf{w}=\mathbf{b}$$

Sistem linearnih jednadžbi – j -ta komponenta:

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

Rješenje linearnog regresijskog problema

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

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

Rješenje – inverzija matrica

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

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

Što ako je \mathbf{X} – singularna (determinanta =0, kolone matrice linearno ovisne)

Rješenje – izbaciti redundantne (linearno ovisne) kolone

Alternativno rješenje

Gradijentno spuštanje !

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

Standardno učenje

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

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

Online učenje

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

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

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

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

$$\beta(t) > 0$$

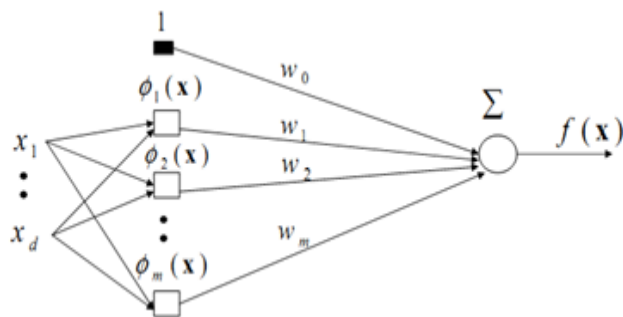
Proširenja jednostavnog linearnog modela

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

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

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

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



$$\mathbf{x} = (x)$$

$$\varphi_1(x) = x; \varphi_2(x) = x^2; \varphi_3(x) = x^3$$

$$\mathbf{x} = (x_1, x_2)$$

$$\varphi_1(\mathbf{x}) = x_1; \varphi_2(\mathbf{x}) = (x_1)^2; \varphi_3(\mathbf{x}) = x_2;$$

$$\varphi_4(\mathbf{x}) = (x_2)^2; \varphi_5(\mathbf{x}) = x_1 x_2$$

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

Složenost linearnih modela

- Metoda najmanjih kvadrata – tipično velika varijanca modela (naročito osjetljiva na outlier primjere!)
- 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)

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

- Gdje je

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

i $\lambda \geq 0$

- $\|\mathbf{w}\|^2 = \sum_{i=0}^d (w_i)^2$ penalizira težine različite od nule sa λ
- Isti efekt je kao da imamo običnu grešku najmanjih kvadrata uz ograničenje na ukupnu sumu kvadrata \mathbf{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 (en. „shrinkage coefficient”)

Lasso regresija / algoritam

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

- Gdje je

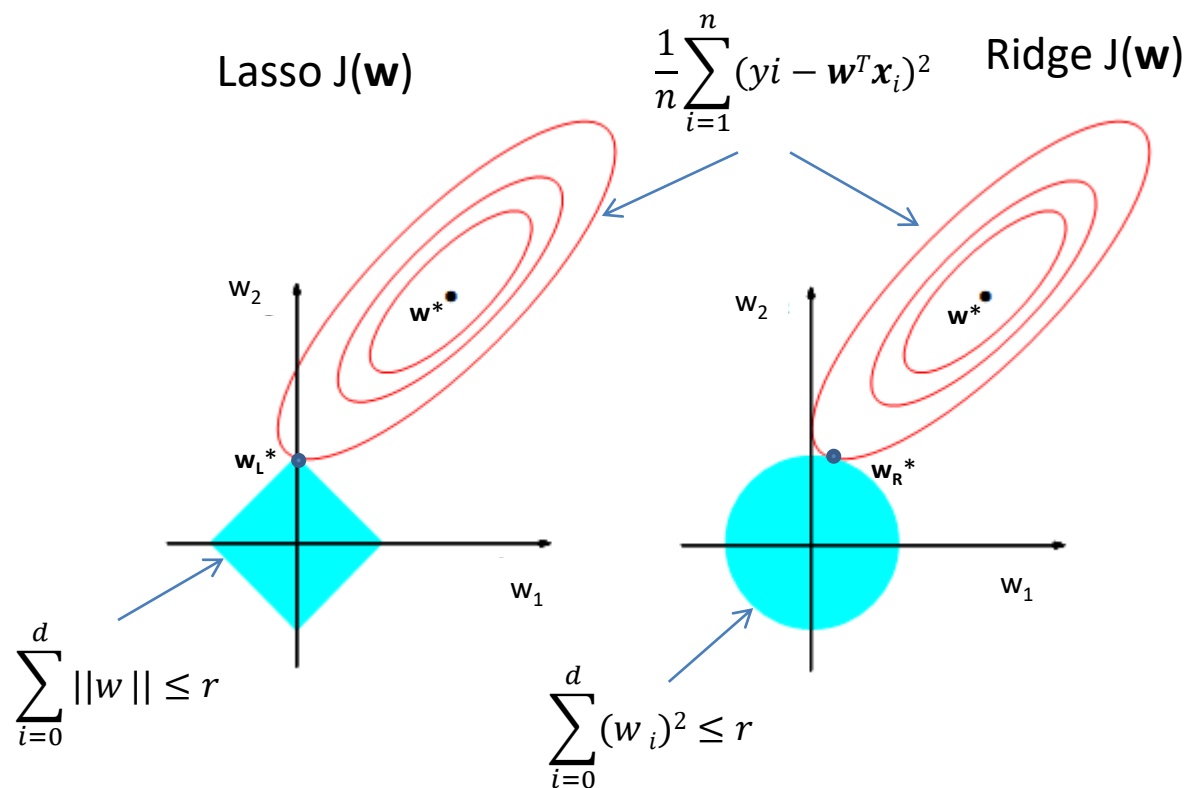
$$\|\mathbf{w}\| = \sum_{i=0}^d \|w_i\|$$

i $\lambda \geq 0$

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



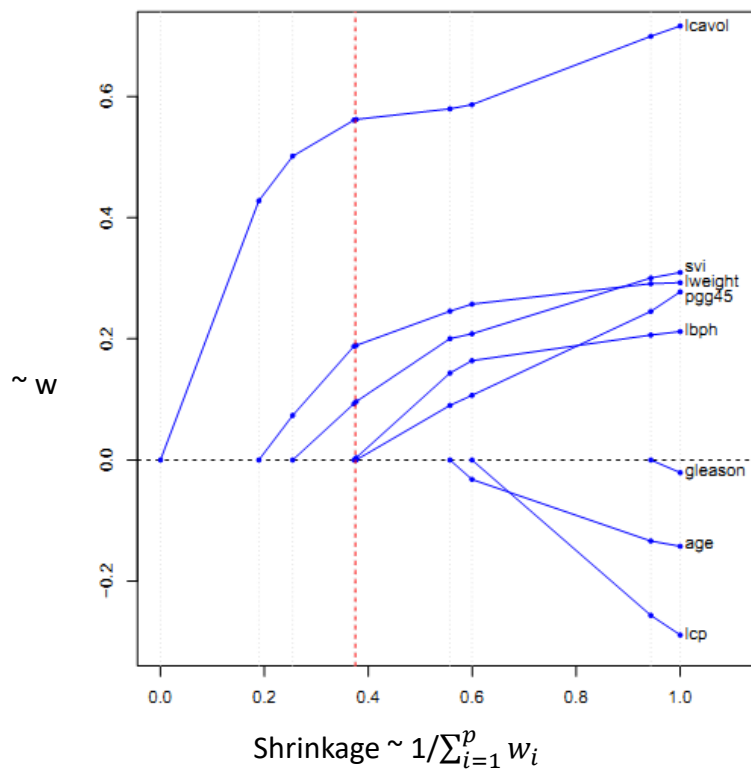
\mathbf{w}^* - Optimalne težine/parametri $\mathbf{w} = (w_1, w_2)$ bez ograničenja na \mathbf{w}

\mathbf{w}_L^* - Optimalne težine/parametri $\mathbf{w} = (w_1, w_2)$ uz Lasso ograničenje

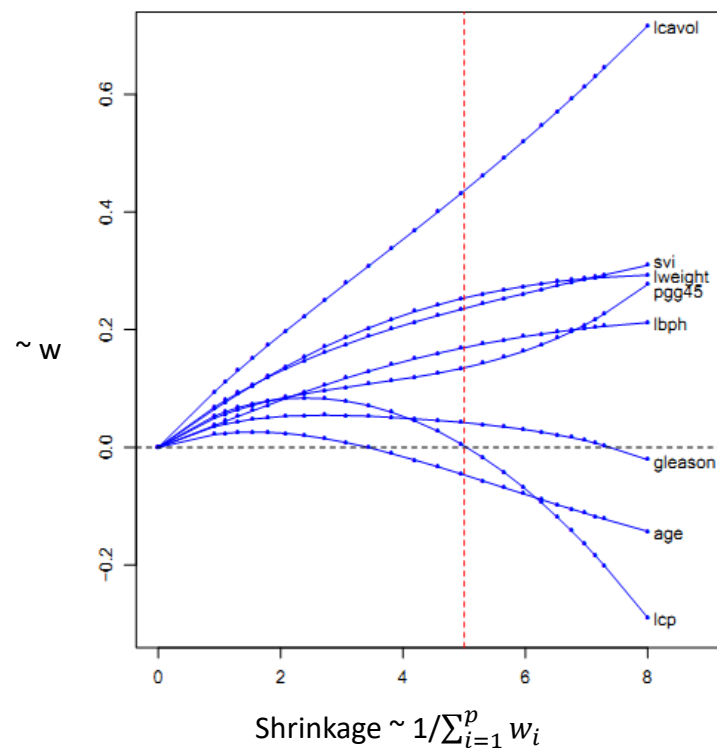
\mathbf{w}_R^* - Optimalne težine/parametri $\mathbf{w} = (w_1, w_2)$ uz Ridge ograničenje

Ponašanje težina varijabli \mathbf{w} za različite vrijednosti λ (odnosno ograničenje r)

Lasso \mathbf{w}

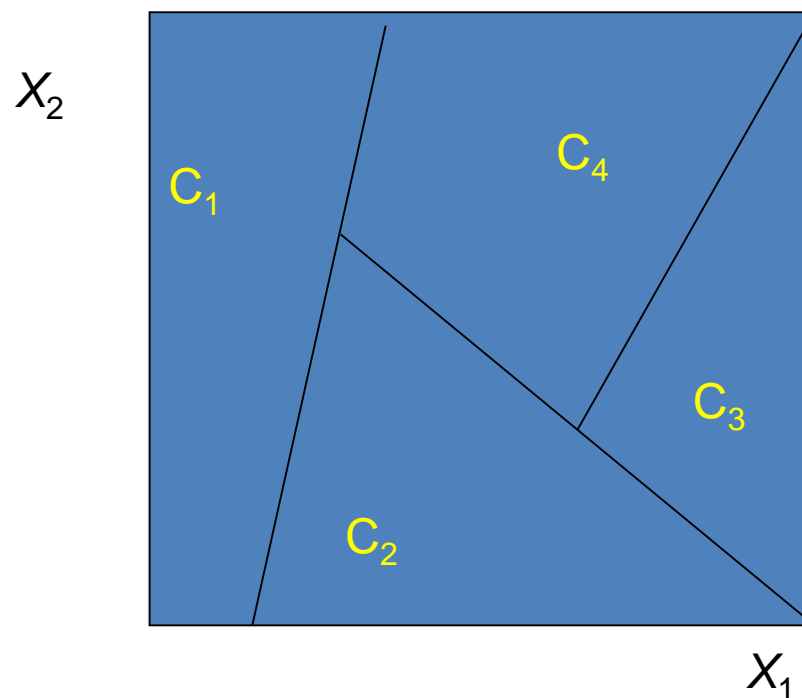


Ridge \mathbf{w}



Linearni klasifikacijski modeli

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



- 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
 - Metoda potpornih vektora (Support Vector Machines)
 - Perceptron

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 $\longrightarrow f_k(\mathbf{x}) = w_{k0} + \mathbf{w}_k^T \mathbf{x}$
- Linearna diskriminantna funkcija za klasu k : $\longrightarrow \delta_k(\mathbf{x}) = f_k(\mathbf{x})$
- Granica između klasa je skup točaka za koje je:
 $\longrightarrow \{\mathbf{x} : f_k(\mathbf{x}) = f_l(\mathbf{x})\} = \{\mathbf{x} : (w_{k0} - w_{l0}) + (\mathbf{w}_k - \mathbf{w}_l)^T \mathbf{x} = 0\}$
- Klasifikacija novog primjera – u klasu sa najvećom $\delta_k(\mathbf{x})$
 $\longrightarrow C(\mathbf{x}) = \arg \max_{k \in C} \delta_k(\mathbf{x})$

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

Klasifikacijska procedura

- Definiramo matricu \mathbf{W} :

$$\longrightarrow \mathbf{W} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- Za neki novi primjer \mathbf{x} , izračunamo:

$$\longrightarrow \mathbf{f}(\mathbf{x}) = [(1, \mathbf{x}) \mathbf{W}]^T = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \dots \\ \dots \\ f_K(\mathbf{x}) \end{pmatrix}$$

- Na kraju – klasa \mathbf{x} određuje se prema najvećoj komponenti $\mathbf{f}(\mathbf{x})$:

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

- Pojašnjenje

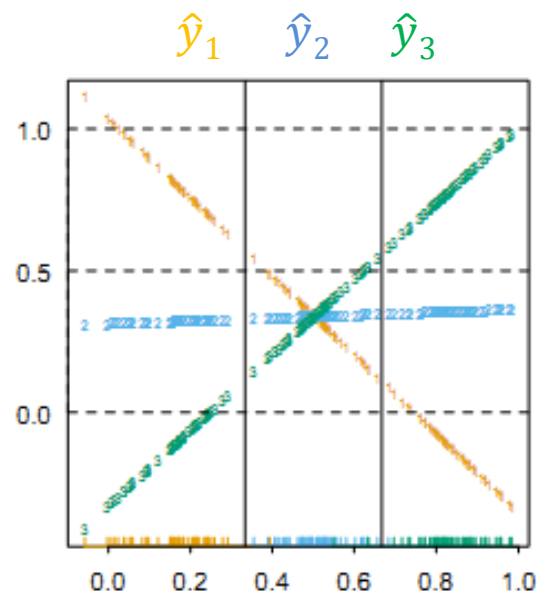
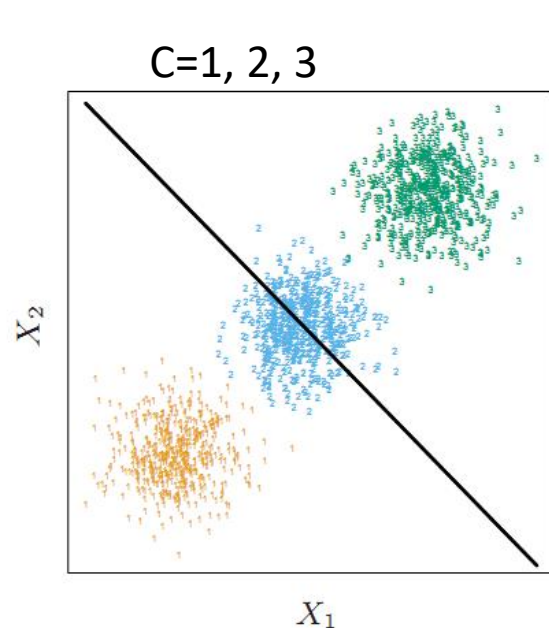
- Linearna Regresija indikatorske varijable (klase) Y_k linearna je aproksimacija očekivanja $E(Y_k/X)$

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

- Odnosno aposteriorne vjerojatnosti indeksa klase ciljne varijable

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

- Problemi s linearnom regresijom – „maskiranje” klasa



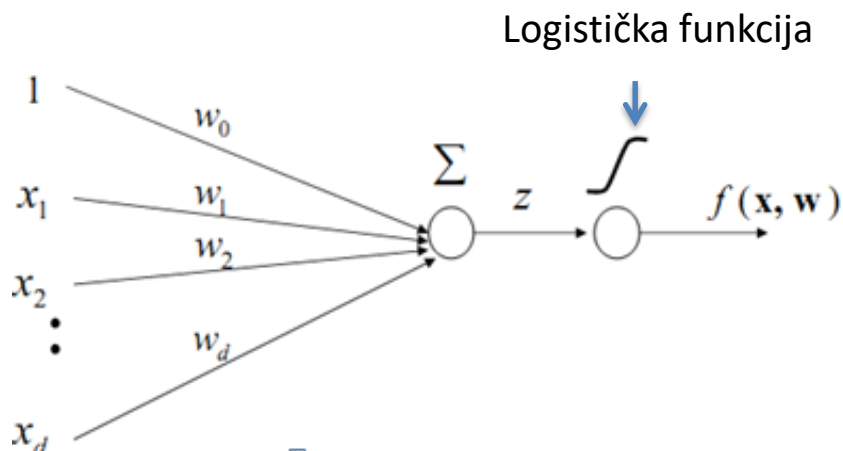
- \hat{y}_2 – nikada ne dominira nad \hat{y}_1 i \hat{y}_3
- Svi primjeri klase C=2 se klasificiraju kao 1 ili 3 !?
- „maskiranje” klasa - za velike K (>3)

Logistička regresija

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

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

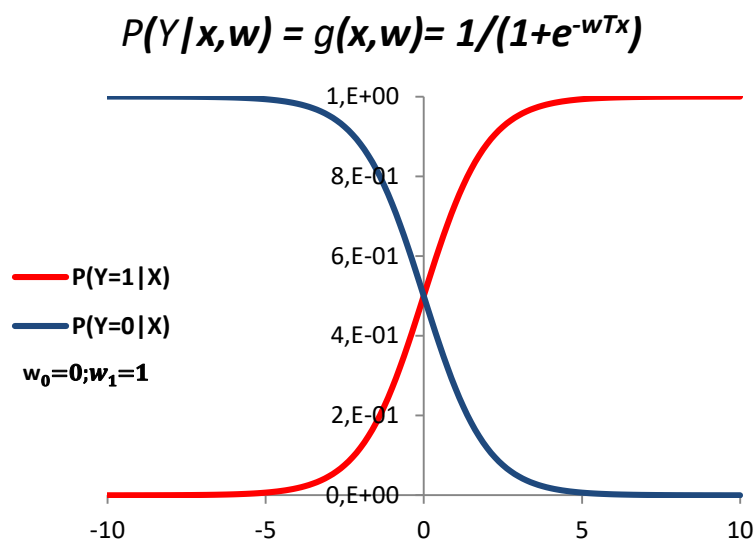
- Gdje je $g(\mathbf{w}^T \mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T \mathbf{x}}} = f(\mathbf{x}, \mathbf{w})$ - logistička funkcija (sigmoidalna funkcija)



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

Logistička regresija

Probabilistička interpretacija !



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

Klasifikacija:

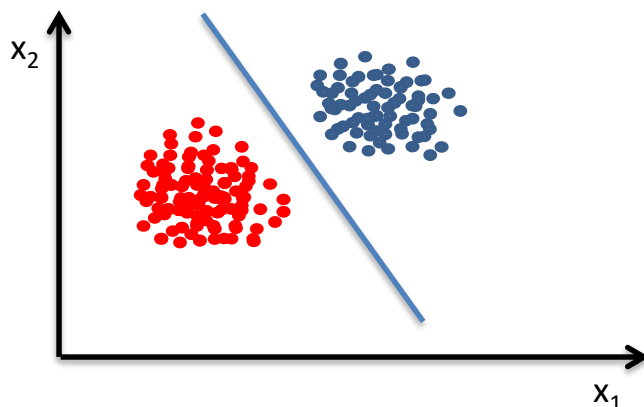
Ako $p(y = 1|x, \mathbf{w}) \geq 0.5$ tada $y=1$
Inače $y = 0$

Logistička regresija

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

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

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



Logistička regresija

Učenje parametara \mathbf{w}

- Vjerojatnost podataka $L(D, \mathbf{w})$ uz \mathbf{w}
 - $D_i = \langle \mathbf{x}_i, y_i \rangle$; $\mu_i = p(y_i = \mathbf{1} | \mathbf{x}_i, \mathbf{w}) = g(\mathbf{w}^T \mathbf{x}_i)$
 - $L(D, \mathbf{w}) = \prod_{i=1}^n P(y = y_i | \mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^n \mu_i^{y_i} (1 - \mu_i)^{(1-y_i)}$
- Odrediti težine \mathbf{w} koje maksimiziraju izglednost (likelihood) podataka
- Trik: Logaritam izglednosti (Max Log-Likelihood)
 - Optimalne težine jednake su i za $L(D, \mathbf{w})$ i za $\log(L(D, \mathbf{w}))$!

$$\begin{aligned} \log(L(D, \mathbf{w})) &= \log \prod_{i=1}^n \mu_i^{y_i} (1 - \mu_i)^{(1-y_i)} = \sum_{i=1}^n \log \mu_i^{y_i} (1 - \mu_i)^{(1-y_i)} = \\ &= \sum_{i=1}^n (y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)) \end{aligned}$$

Logistička regresija - Učenje parametara \mathbf{w}

- Log-likelihood

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

- Derivacija $\log(L(D, \mathbf{w})) \Rightarrow$ negativni gradijent

$$-\frac{\partial}{\partial w_j} \log(L(D, \mathbf{w})) = \sum_{i=1}^n -x_{i,j} (y_i - g(\mathbf{w}^T \mathbf{x}_i))$$

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

- Gradijentno spuštanje

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \beta(k) \nabla_{\mathbf{w}}[-\log(L(D, \mathbf{w}))]_{\mathbf{w}^{(k-1)}}$$

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

- Parametri LogReg se također mogu učiti korištenjem online metode !

Logistička regresija - Algoritam za online učenje parametara w

Online LogReg (D , broj_iteracija)

Inicijaliziraj težine $w^{(0)} = (w_0, w_1, w_2, \dots, w_d)$

For $i=1:\text{broj_iteracija}$

do

izaberi primjer iz $D = \langle \mathbf{x}_k, y_k \rangle$

postavi $\beta_i = 1/i$

odredi nove težine

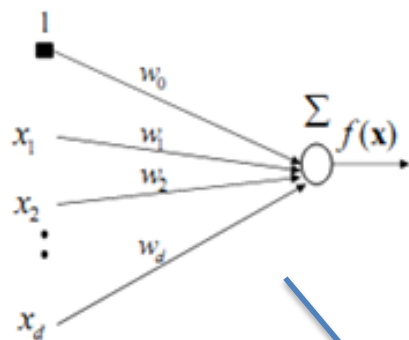
$w = w + \beta_i [y_i - g(w^T \mathbf{x})] \mathbf{x}_i$

end for

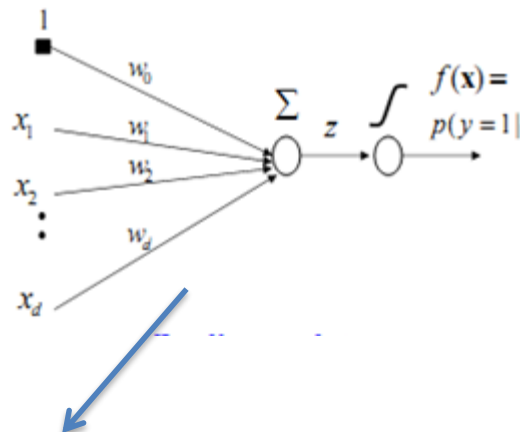
Vrati težine w

Usporedba Linearna regresija i Logistička regresija

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



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



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

Učenje modela ima isti oblik !

Nelinearna ekstenzija LogReg i Linearne regresije

Korišćenje nelinearnih baznih funkcija

Linearna regresija

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

Logistička regresija

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

$\varphi_j(\mathbf{x})$ - arbitrarne funkcije \mathbf{x}

LDA

Pretpostavke:

- (više-dimenzionalna) Gaussova distribuciju kao model gustoće uvjetne vjerojatnosti po klasama
- Ista kovarijacija po varijablama za sve klase

Klasifikacija se bazira na određivanju aposteriorne vjerojatnosti za klasu:

$$P(C = k | \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{l=1}^K f_l(\mathbf{x})\pi_l} \quad \longleftarrow \quad \text{Bayesovo pravilo}$$

π_k – apriorna vjerojatnost pojave klase k

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

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

$\Sigma_k = \Sigma$ - kovarijacijska matrica
kod LDA metode - $\Sigma_k = \Sigma$ ista za sve klase !

Parametri:

Na bazi skupa primjera za učenje:

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

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

kovarijacijska matrica: $\Rightarrow \hat{\Sigma}_k = \sum_i r_i^k (x_i - \hat{\mu}_k) * (x_i - \hat{\mu}_k)^T$
 $r_i^k = 1, \text{ ako } x_i \in C_k, \text{ inače } r_i^k = 0$

(Parametri su određeni po principu maksimalne izglednosti („max. likelihood“).

uz korištenje log-odds...

$$\begin{aligned} \log \frac{P(C = k \mid X = \mathbf{x})}{P(C = l \mid X = \mathbf{x})} &= \log \frac{\pi_k}{\pi_l} + \log \frac{f_k(\mathbf{x})}{f_l(\mathbf{x})} \\ &= \underbrace{\left(\log \pi_k + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k \right)}_{\delta_k(\mathbf{x})} - \underbrace{\left(\log \pi_l + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l - \frac{1}{2} \boldsymbol{\mu}_l^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l \right)}_{\delta_l(\mathbf{x})} \end{aligned}$$

Diskriminativne funkcije
za klasu k i l :

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

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

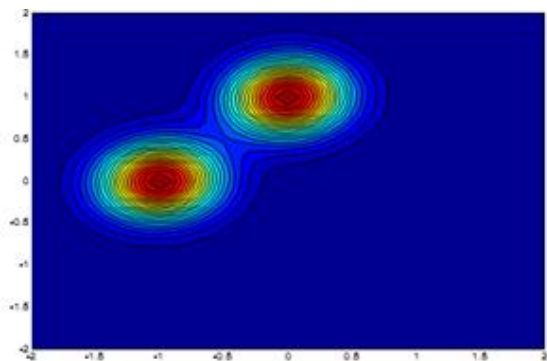
$$\log \pi_k + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k = \log \pi_l + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l - \frac{1}{2} \boldsymbol{\mu}_l^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l$$

Kvadratna diskriminativna analiza - QDA

- **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 !

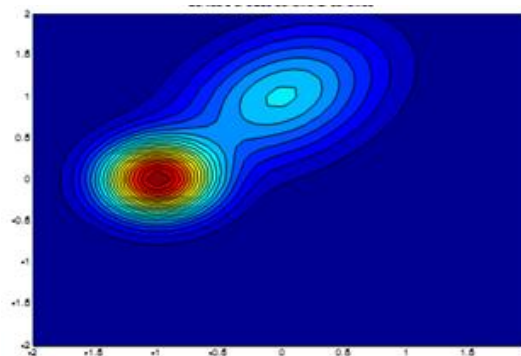
$$\log \pi_k + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k$$

$\delta_k(x)$ – **LDA**

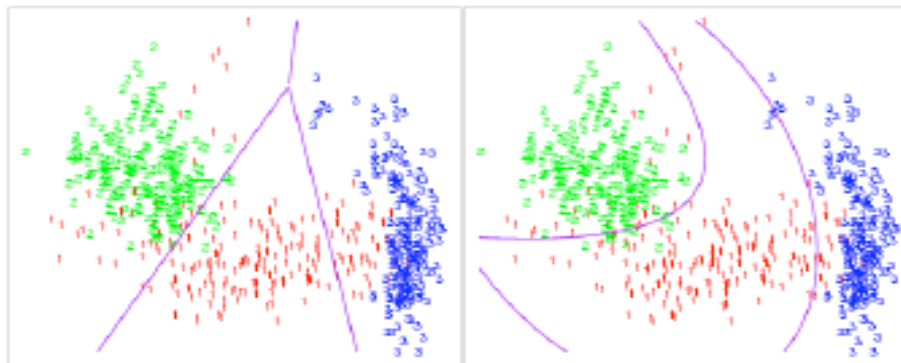


$$\log \pi_k - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) - \frac{1}{2} \log |\boldsymbol{\Sigma}_k|$$

$\delta_k(x)$ – **QDA**



LDA vs QDA - Granice između klasa

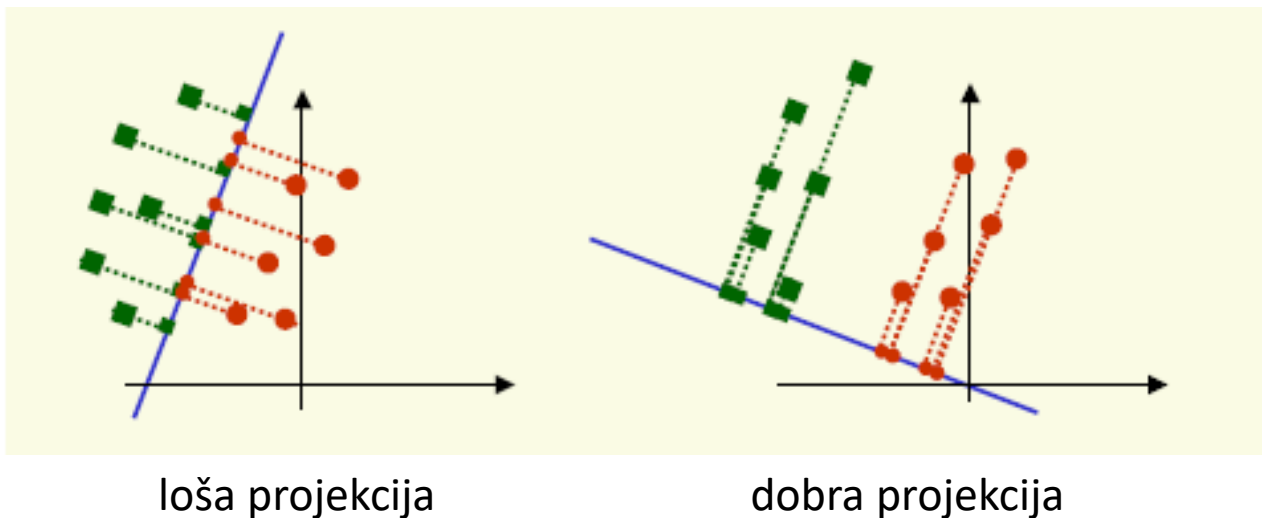


LDA

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



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

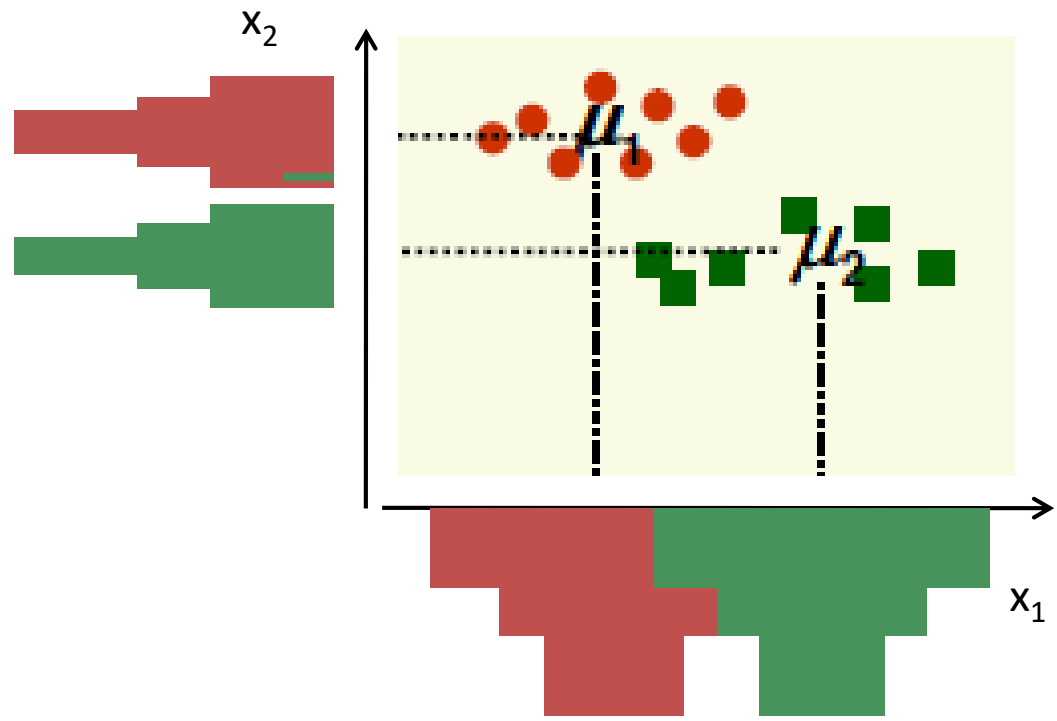
$$\bar{\mu}_i = \frac{1}{N_i} \sum_{y \in Y_i} y = \frac{1}{N_i} \sum_{x \in D_i} \mathbf{w}^T \mathbf{x} = \mathbf{w}^T \boldsymbol{\mu}_i$$

Srednja vrijednost za točke klase i projicirane na \mathbf{w}

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

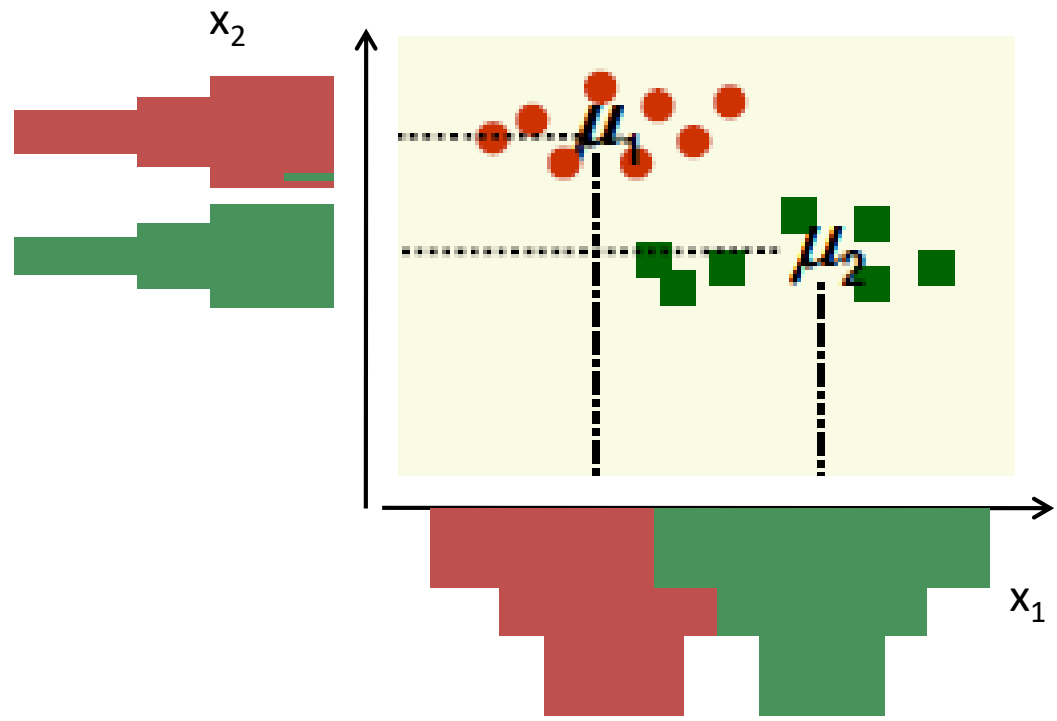
Udaljenost između projiciranih srednjih vrijednosti za dvije klase

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



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

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



x_2 je bolja, no:

$$|\bar{\mu}_1 - \bar{\mu}_2|_{x_1} > |\bar{\mu}_1 - \bar{\mu}_2|_{x_2}$$

problem je što $|\bar{\mu}_1 - \bar{\mu}_2|$
ne uzima u obzir varijancu
distribucije primjera!

Ako definiramo:

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

$$\tilde{s}_1^2 = \sum_{y_i \in Cl_1} (|y_i - \bar{\mu}_1|) \quad \text{“raspršenje” primjera klase1 (na projekciji)}$$

$$s_2^2 = \sum_{y_i \in Cl_2} (|y_i - \bar{\mu}_2|) \quad \text{“raspršenje” primjera klase2 (na projekciji)}$$

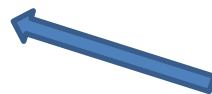
Možemo raspršenje koristiti za normalizaciju udaljenosti projiciranih “centara” između klasa !

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

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



želimo da brojnik bude što veći



a nazivnik što manji!

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

- Treba eksplicitno prikazati J u ovisnosti o **w** !
- Radi se zapravo o optimizaciji J(**w**) – odnosno određivanju optimalnog **w**
- Definirajmo matrice raspršenja za svaku klasu – prije projekcije - dakle za originalne primjere

$$\mathbf{S}_1 = \sum_{\mathbf{x}_i \in Cl_1} (\mathbf{x}_i - \mu_1)(\mathbf{x}_i - \mu_1)^T$$
$$\mathbf{S}_2 = \sum_{\mathbf{x}_i \in Cl_2} (\mathbf{x}_i - \mu_2)(\mathbf{x}_i - \mu_2)^T$$

- 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:

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

- Na kraju je naša funkcija cilja

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

- Da bi je optimirali, našli maksimum – prva derivacija po $\mathbf{w} = 0$

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

- Na koncu se to svede na problem određivanja svojstvenih vrijednosti

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

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

- Odnosno, nakon sređivanja – ako postoji inverzna matrica:

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

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

$$y_i = \mathbf{w}^T \mathbf{x}_i < t \Rightarrow y_1$$

$$y_i = \mathbf{w}^T \mathbf{x}_i \geq t \Rightarrow y_2$$