INTELLIGENTNI SISTEMI

LINEARNI MODELI IN KLASIFIKACIJA

Osnove Linearnih klasifikacijskih modelov

Prof. Jurij F. Tasič
Emil Plesnik

Plan

- Regresija vs Razvrščanju
- Linearni Modeli za razvrščanje
- Pretvarjanje verjetnostne regresije v razvrščanje
- Trije razredi klasifikacijskih modelov

Regresija vs Klasifikaciji

- Regresija:
  - dodeli vhodni vektor x eni ali več zveznim ciljnim spremenljivkam t
- Klasifikacija:
  - dodeli vhodni vektor x enemu od K diskretnih razredov C_k, $k = 1, \ldots, K$.
- Regresija razvrščanja:
  - Diskretni razredi so urejeni

Linearni modeli razvrščanja

- Skupni scenarij: razredi predstavljajo razčlenitev
- Vhodni prostor je razdeljen na odločitvene regije
- Linearni model ravnin odločanja je linearna funkcija vhoda x
  - *D-1* dim. Hiperravnina znotraj *D* dim. vhodnega prostora
- Podatkovni nizi, katerih razredi so ločeni z linernimi ločevalnimi ravninami so – Linearno ločljivi podatki

Verjetnostni modeli: Pretvarjanje regresije na izhod razreda

- Ciljne spremenljivke *t* predstavljajo razrede oznak
- Binarni razred predstavitev je najbolj primeren za verjetnostne modele
  - Dva razreda: $t \in \{0, 1\}$, $t = 1$ predstavlja c_1, $t = 0$ predstavlja razred c_2
  - Tolmačenje vrednosti t kot verjetnosti, da je razred c_1
  - Verjetnost zavzema ekstremni vrednosti 0 in 1
  - Za $K > 2$ - uporabimo *1-od-K* kodirno shemo.

- Npr. $K = 5$, vzorec razreda 2 ima ciljni vektor $t = (0, 1, 0, 0, 0)^T$
- Vrednost t_k razlagamo kot verjetnost razreda C_k

Dva pristopa k klasifikaciji

1. Diskriminantna funkcija
   - Direktno dodeli x določenemu razredu
     - npr. Perceptron,
2. Verjetnostni Modeli
   - Model $p(C_k / x)$ v fazi sklepanja (neposredno ali $p(x | C_k)$)
   - Uporabimo ga za optimalno odločitev

Ločevanje s sklepanjem je boljše:

- zmanjšuje tveganja (izgubo funkcijo lahko spremeni v finančnno app)
- Zavrača opcije (zmanjšanje pričakovane izgube)
- Kompenzira neuravnoteženost podatkov
  - Uporaba spremenjenih in uravnoteženih podatkov & skaliranje po razredih - frakcijah
- Združevanje modelov

Dva pristopa k klasifikaciji

- Modeliramo $p(C_k/x)$ v fazi *sklepanja* in ga uporabimo za optimalno odločanje
- Poznamo dva pristopa računanja $p(C_k/x)$
  - Generativni - razmnoževalen
    - Modelira razred pogojne gostote $p(x|C_k)$ skupaj s priorno (predhodno) verjetnostjo $p(C_k)$
    - Nato uporabi Bayesovo pravilo za izračun posteriorne verjetnosti

      $p(C_k|x) = p(x|C_k)p(C_k)/p(x)$
  - Diskriminativni – razločevalen
    - Neposreden model pogojne verjetnosti $p(C_k|X)$

Pretvorba linearnega regresijskega modela v linearni klasifikator

- Linearni regresijski model $y(x,w)$ je linearna funkcija parametrov w
- V enostavnejšem primeru je tudi linearna funkcija spremenljivke x
  - torej je oblike $y(x)=w^T x+w_0$
- Za klasifikacijo želimo pridobiti diskretne izhod ali zadnji verjetnosti v območju $(0,1)$
  - Uporabimo posplošitev modela

$$y(x)=f(w^T x+w_0)$$

Posplošen linearni model

- $f(.)$ je znana aktivacijska funkcija
- Površina odločanja zadošča izrazoma $y(x)=constant$ ali $w^T x+w_0=constant$
- Meje odločanja so linearne, tudi če je npr. $f(.)$ nelinearna funkcija
- Zato ga imenujemo splošen linearni model
- Vendar ni več dolgo linearen po parametru w zaradi prisotnosti $f(.)$
  - Omogoča kompleksnejše modele razvrščanja kot pa regresija

Pregled linearnih razvrščevalnikov - klasifikatorjev

1. Discriminantne funkcije

- dvo in več vrstne (Linear discriminant analysis-LDA)
- razvrščanje z metodo najmanših kvadratov
- Fisherjeva linearna discriminantna analiza
- Algoritem Perceptron

2. Verjetnostni generativni- razmnoževalni modeli

- Zveznost vhodov in maksimalna verjetnost
- Diskretni vhodi, eksponentna rast

3. Verjetnostni diskriminativni modeli

- – Logistična regresija za enojne in večkratne razrede
- – Laplaceova aproksimacija
- – Bayesova logistična regresija

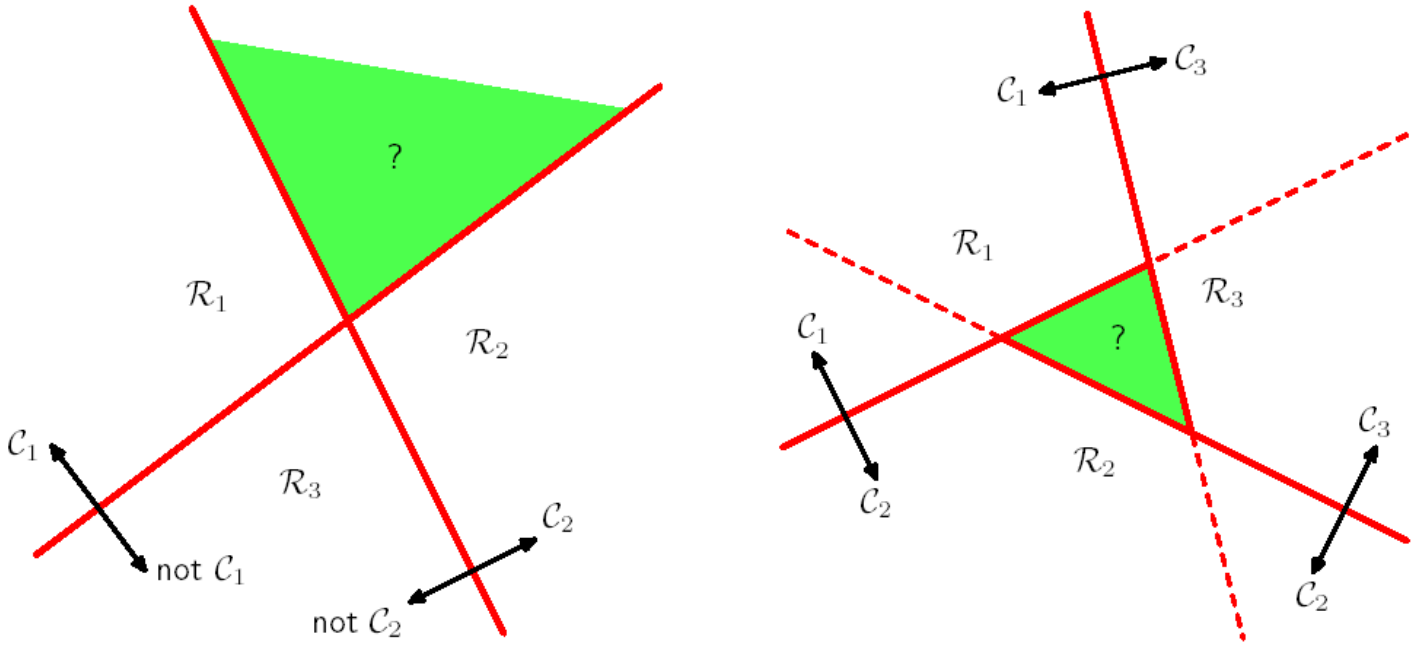
Nadaljevanje

- Linearne Diskriminantne Funkcije
  - Definicije (2-razreda), Geometrija
  - Generalizacija za razrede $K > 2$
- Metode učenja parametrov
  1. Klasifikacija z metodo najmanjših kvadratov
  2. Fisher jeva Linearna metoda diskriminacije
  3. Perceptroni

Diskriminantna Funkcija

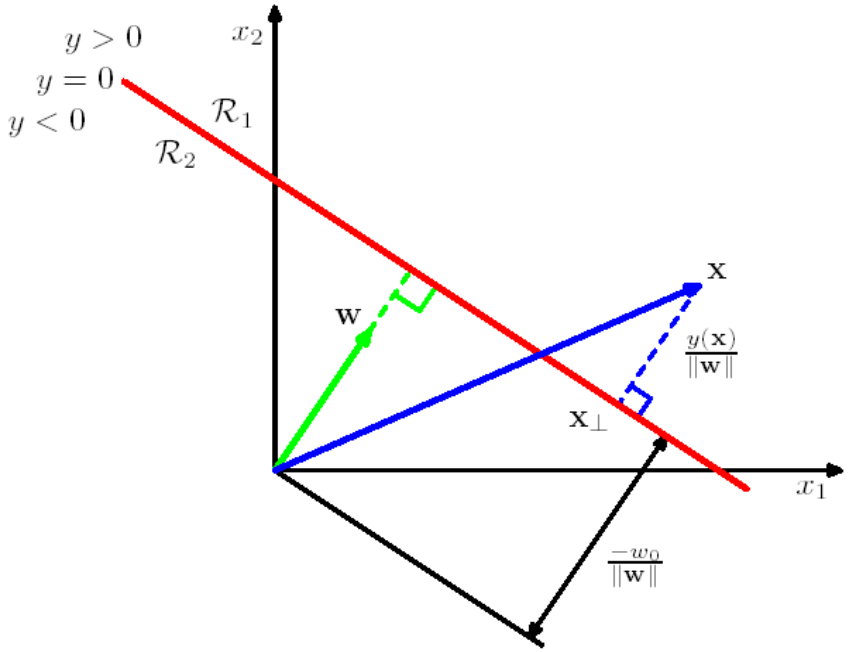
- Vhodnemu vektorju x dodelimo enega od razredov – grozdov K
- Označimo ga s Ck
- Omejimo se le na linearno diskriminacijo
  - – odločitvene površine so hiperravnine
- Naj prej vzemimo $K = 2$, kasneje pa razširimo število razredov- grozdov na $K > 2$

Več razredov z dvema diskriminantnima funkcijama



Poskus razvrščanja v *K* razredov; Več diskriminantnih premic vodi v nejasne regije (prikazana v zeleni barvi). Na levi je primer, ki ponazarja diskriminantni funkciji, ki omogočata razlikovanje dveh točk v razredu *Ck* od točk, ki niso v razredu Ck. Na desni je primer, ki vključuje tri diskriminantne funkcije vsaka posamezno ločuje par razredov Cj in Ck.

Geometrija Linearnih Diskriminantnih Funkcij



- Dva razreda linearnih diskriminantnih funkcij:

$$y(x) = w^T x + w_0$$

w je w^T vektor in w_0 je odmik - bias

Naj bo x element $C1$ če $y(x) \geq 0$

drugače $C2$

- Mejo določa $y(x) = 0$
- w določa orientacijo površine

Ker je $w^T(x_A - x_B) = 0$, je w pravokoten na vsak

- *Ko je* $y(x)=0$ tedaj w_0 določa razdaljo od izhodišča do površine

$w^T x/||w|| = -w_0/||w||$

- $y(x)$ podaja vrednost pravokotne razdalje r točke x do decezijske ravnine,

$r = y(x)/||w||$

- Če vpeljemo kompaktnejšo predstavitev, uporabimo vrednost $x_0 = 1$ in

$\omega = (w_0, w)$ tedeaj je $y(x) = \omega^T x$

– V tem primeru prehaja D dimenzionalna hyperravnina skozi izhodišče $D+1$ dimenzionalnega razširjenega prostora

Večkratni razredi s K diskriminantami

- Vzemimo enostavno K razredno diskriminanto oblike

  $$y_k(x) = w^T{}_k\, x + w_{k0}$$

- Dodelimo točko x razredu Ck če $y_k(x) > yj\ (x)$ for all $j \neq k$

- Odločitvena meja med razredoma Ck in Cj je podana z

  $$y_k(x) = y_j\ (x)$$

  – ki pripada $D - 1$ dimensionalni hiperravnini določeni z

  – $(w_k - w_j\ )^T\, x + (w_{k0} - w_{j0}\ ) = 0$

  – Vidimo isto obliko kot v primeru 2 razredov

  $w^T x + w_0 = 0$

- Ravnine decezije so v takih primerih enostavno povezane in so konveksne

  – dokaz sledi

Konveksnost odločitvenih regij

Vzemimi točki x_A in x_B , obe ležita na površini regije R_k

Vsako točko na premici, ki povezuje x_A z x_B lahko predstavimo kot

$$\hat{x} = \lambda x_A + (1-\lambda) x_B \text{ kjer je } 0 < \lambda < 1.$$

Linearne diskriminantne funkcije

$$y_k(x) = w^T{}_k\, x + w_{k0}$$

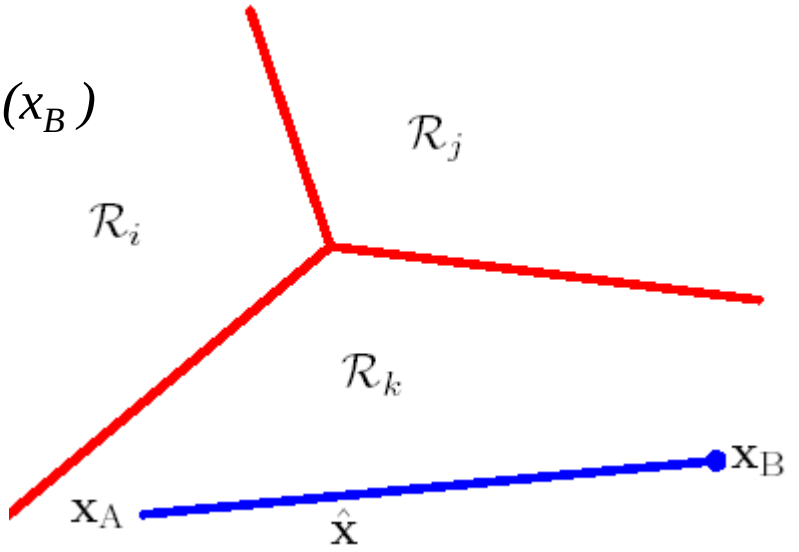
Če kombiniramo dve dobimo

$$y_k(\hat{x}) = \lambda y_k(x_A) + (1-\lambda) y_k(x_B)$$

Ker x_A in x_B ležita v regiji R_k sledi

$y_k(x_A) > y_j(x_A)$ in $y_k(x_B) > y_j(x_B)$ za vse $j \neq k$.

Torej $\hat{x}$ tudi leži znotraj regije R_k
in R_k je enojno povezana in konveksna



Pomen parametrov
Linearne Diskriminantne Funkcije

- Tri metode
  - metoda najmanjših kvadratov
  - Fisherjeva Linearna Diskriminantna m.
  - Perceptroni
- Vse so enostavne, vska ima svoje slabosti

Klasifikacija po metodi najmanjših kvadratov

- Podobna je regresiji: obstoja enostavna zaprta rešitev za dane parametre
- Vsak C_k , $k=1,..K$ je podan s svojim linearnim modelom

$$y_k(x) = w^T{}_k\, x + w_{k0}$$

- Vzemimo pomožen vektor $x=(1,x^T)$ in $w_k=(w_{k0},w_k{}^T)$
- Če zapis posplošimo $y(x) = W^T x$

  W je parameter matrike katere k^{ti} stolpec je $D+1$ dimenzionalni vektor (vsebuje odmik)
- Novi vhodni vektor x pripada razredu, katerega izhod $y_k = w^T{}_k\, x$ je največji
- W določimo z minimizacijo kvadrata napake

Parameti optimiranja

- Vzemimo niz učnih podatkov $\{x_n, t_n\}, n = 1, \ldots, N$ in določimo matriko T katere n_{ta} vrstica je vektor t^T_n
- Določi matrike
- $T \equiv n^{th}$ vrstica je vektor t^T_n
- $X \equiv n^{th}$ vrstica ki je x^T_n
- Vsota kvadratov funkcije napake

$$E_D(W) = 1/2 \,\mathrm{Tr}\,\{(XW - T)^T (XW - T)\}$$

Opomnik:

$(XW\text{-}T)$ je vektor napake, katere kvadrat je diagonalna matrika

Trace – sled je vsota diagonalnih elementov matrike

Minimizacija vsote kvadratov

- Vsota kvadratov funkcije napake

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^{\mathrm{T}}(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\}$$

- Odvajamo po W in postavimo na nič, kar da rešitev

$$\widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{T} = \widetilde{\mathbf{X}}^{\dagger}\mathbf{T}$$

  kjer je $\boldsymbol{X}^{\dagger}$ pseudo-inverz matrike $\boldsymbol{X}$

- Po preureditvi, je diskriminantna funkcija

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}} = \mathbf{T}^{\mathrm{T}}\left(\widetilde{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}\widetilde{\mathbf{x}}$$

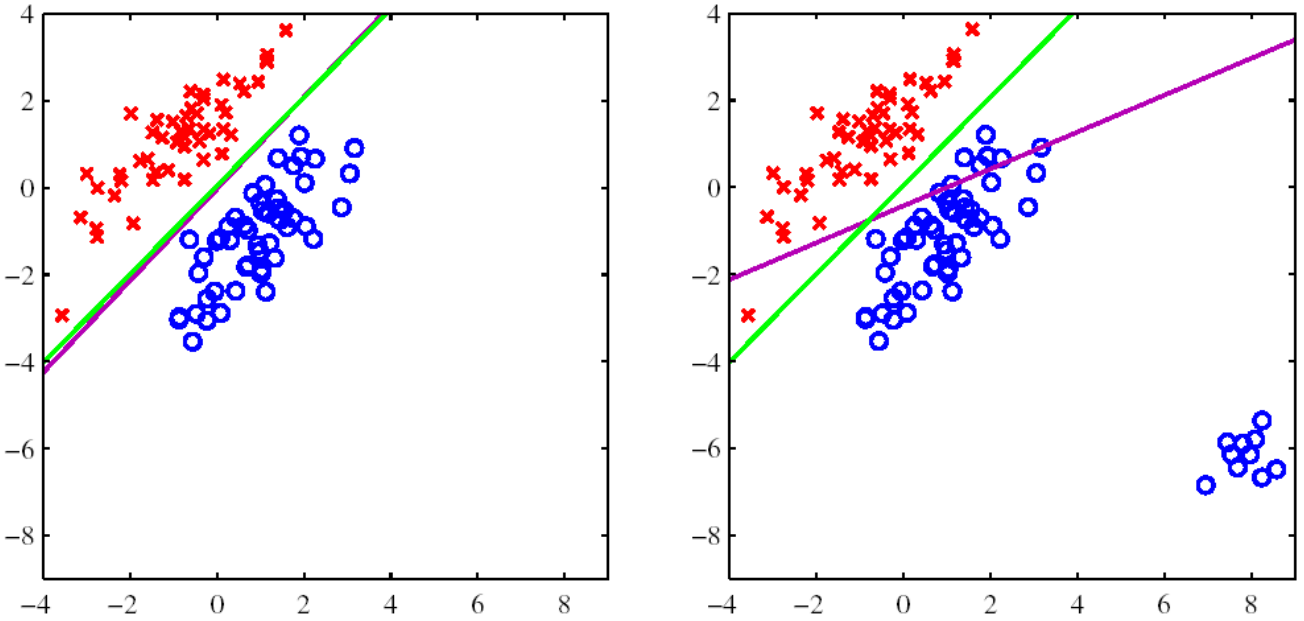
- Zanimiva lastnost LMS z več spremenljivkami je, da v primeru učenja z učnim nizom, ta zadošča omejitvam

$$\mathbf{a}^{\mathrm{T}}\mathbf{t}_n + b = 0$$

  Za konstanti a in b, mora predikcijski model za vsak vhod x zadostiti istim omejitvam

$$\mathbf{a}^{\mathrm{T}}\mathbf{y}(\mathbf{x}) + b = 0.$$

Metoda najmanjših kvadratov je občutljiva na ubežnike

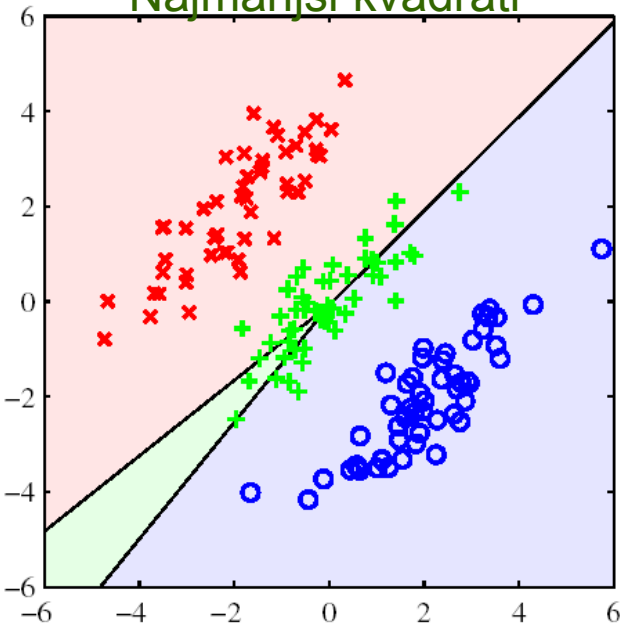


škrlatna: Najmanjši kvadrati
zelena: Regresija (robustnejša)

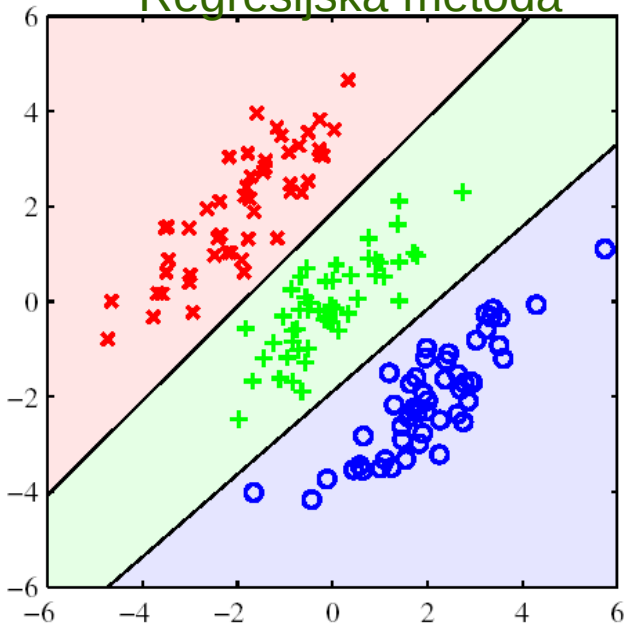
Levo slika prikazuje podatke dveh razredov, označena z rdečimi križi in modrimi krogi, skupaj mejama metode najmanjših kvadratov (magenta krivulja) in logistične regresije (zelena krivulja). Desna slika prikazuje dosežene rezultate ko smo uvedli dodatne točke - podatke na spodnji levi strani diagrama; Vidimo, da je metoda najmanjših kvadratov zelo občutljiva na ubežnike (outlayers), regresija pa ne.

Slabosti metode najmanjših kvadratov

Najmanjši kvadrati



Regresijska metoda



Regija zelenih vzorcev je premajhna, zato jo večkrat zgrešimo
Vidimo, da logična regresijska metoda daje odlične rezultate z linearnimi klasifikatorji

- Pomanjkanje robustnosti na ubežnike
- Določenim podatkovni nizi niso primerni za metodo najmanjših kvadratov
- Meje odločanja pomenijo rešitve strojnega učenja ML

  Gaussova porazdelitev
- Binarne ciljne vrednosti imajo porazdelitev vrednosti daleč od Gaussove

Fisherjeva Linearna Diskriminacijska funkcija

- Klasifikacijo gledamo v smislu redukcije dimenzionalnosti
  - –D-dimenzionalni vhodni vektor x projeciramo v eno dimenzijo z $y = w^T x$
- Za klasifikacijo postavimo prag spremenljivke y tako, da je $y \geq -w_0$ za C_1 drugače za C_2

  Tako dobimo standarden linearni klasifikator
- Razredi so v *D*-prostoru dobro ločeni, a se lahko močno prekrivajo v*1*-dimenzijskem prostoru
  - – Prilagajamo komponenete utežnega vektorja w tako ta maksimiziramo ločljivost razredov
  - – Izbiramo tako projekcijo, da minimiziramo ločevanje med razredi

Fisher: Maksimizacija razdalj ločevalnimi funkcijami

- Vzemimo problem dveh razredov:

– N_1 točk razreda C_1 in N_2 točk razreda C_2

- Tako sta srednji vrednosti enega $\boldsymbol{m_1}$ in drugega razreda $\boldsymbol{m_2}$ podani z $\boldsymbol{m_1} = (\frac{1}{N_1}) \sum_{\boldsymbol{n} \in \boldsymbol{C_1}} x_1$ in $\boldsymbol{m_2} = (\frac{1}{N_2}) \sum_{\boldsymbol{n} \in \boldsymbol{C_2}} x_2$
- Izbira uteži $\boldsymbol{w}$, ki maksimira razdaljo med srednjima vrednostima obeh razredov
- Maksimiramo $m_2 - m_1 = \boldsymbol{w}^T (\boldsymbol{m_2} - \boldsymbol{m_1})$,

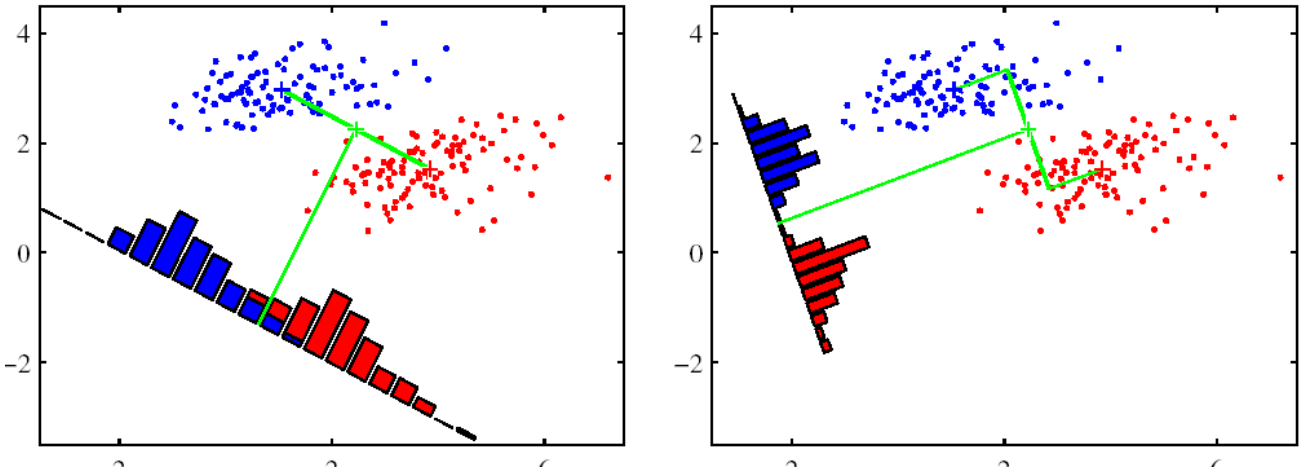
  kjer je $m_k = \boldsymbol{w}^T m_k$

  že omenjena srednja vrednost razreda $\boldsymbol{C_k}$
- Razdalja je lahko poljubna z večanjem vrednosti utežnega vektorja $\boldsymbol{w}$.

  Da rešimo ta problem, se omejimo $\boldsymbol{w}$ na enotno dolžino $\sum_i w_i^2 = 1$
- Za problem maksimiranja z omejitvami uporabimo Lagrangeove multiplikatorje, kjer so $\boldsymbol{w} \propto (\boldsymbol{m_2} - \boldsymbol{m_1})$.

Fisher: Minimizacija variance

Sredinni sta dovolj narazen, razredi pa
Se še prekrivajo



Znotraj nediagonalne kovariance maksimiranje razdalje običajno ni zadostno

- Fisherjeva formulacija

1. Maksimaj funkcijo tako, da maksimiraš razdaljo med srednjo vrednostjo projeciranih podatkov
2. Znotraj vsakega razreda naj bo varianca najmanjša,
   tako minimizaramo prekrivanje razredov

Fisher: Odvajanje

Projekcijska formula $y = \boldsymbol{w}^T \boldsymbol{x}$ preslika niz podatkovnih točk x v niz podatkovnih točk enodimenzionalnega prostora y.
Varianca preslikanega razreda podatkovnih točk $\boldsymbol{C_k}$ je

$$s_k^2, \quad \text{kjer je } y_n = w^T x_n$$

- Varianca celotnega podatkovnega niza pa je enostavno

$$s_1^2 + s_2 = \sum_{n \varepsilon C_k} (yn - mk)2$$

- Fisherjev kriterij je $J(w) = (m_2 - m_1)^2/(s_1^2+s_2^2)$

odvisnost od uteži w je podana v drugi obliki Fisherjevega kriterija

$$J(w) = \boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w} \,/\, \boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}$$

kjer je medrazredna kovar. matrika $\boldsymbol{S}_B = (m2 - m1)(m2 - m1)^T$ in

S_W je kovariančna matrika razreda

$$\boldsymbol{S}_W = \sum_{n \varepsilon C_k} (\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^T + \sum_{n \varepsilon C_k} (\boldsymbol{x}_n - \boldsymbol{m}_2)(\boldsymbol{x}_n - \boldsymbol{m}_2)^T$$

- Če odvajamo izraz $J(\boldsymbol{w})$ po $\boldsymbol{w}$ in ga maksimiziramo ko je

$$(\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w})\boldsymbol{S}_W \boldsymbol{w} = (\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w})\boldsymbol{S}_B \boldsymbol{w}$$

Iz enačbe $\boldsymbol{S}_B$ vidimo, da ta kaže v smeri $(m_2 - m_1)$. Zato lahko izpustimo skalarna faktorja (v oklepajih) & ga pomnožimo s $\boldsymbol{S}_W^{-1}$ kar da $\boldsymbol{w} \propto \boldsymbol{S}_W^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1)$ Fisherjeva linearna diskriminanta

Primerjava z metodo najmanjših kvadratov

- Metoda najmanjših kvadratov: Model predikcije je blizu ciljnim vrednostim
- Fisher: Maksimira ločljivost razredov
- Za problem dveh razredov (grup) je Fisherjeva metoda poseben primer metode najmanjših kvadratov
  – Dokaz pričnemo z vsoto kvadrata napake in dokažemo, da utežni vektor v tem primeru sovpada s Fisherjevim kriterijem

Fisherjevo razločevanje več podatkovni razredov

- Metodo lahko posplošimo na več podatkovnih razredov
- Odvod je korektno upoštevan [Fukunaga 1990]

Rosenblattov Algoritem Perceptron

- Ustreza modelu dveh razredov

  – Vhodni vektor x transformiramo z uporabo stalne nonlinear funkcje, katere rezultat je vektor $\Phi(x)$

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

kjer je nelinearna aktivacijska funkcija $f(.)$ funkcija stopnice

$$f(a) = \begin{cases} +1, & a \geq 0 \\ -1 & a < 0 \end{cases}$$

- Uporabimo ciljno kodno shemo

$t = +1$, za razred $C1$ in $t = -1$ za $C2$ ujemanja z aktivacijsko funkcijo

Funkcija napake Perceptrona

- Funkcijo napake predstavlja število zgrešenih klasifikacij
- Ta funkcija napake je po delih konstantna funkcija w z nezveznostmi (drugače kot regresijska metoda)
- Zato ne moremo računati odvodov, kar pomeni da gradieentne metode niso uporabne
- Funkcija napake ni zvezna

Kriterij Perceptrona

- Iščemo w ki je za vzorce $xn \in C_1$ in za katerega velja
$$\boldsymbol{w}^T \Phi(\boldsymbol{x}_n) > 0$$
- ter za vzorce $x_n \in C_2$ ima $\boldsymbol{w}^T \Phi(\boldsymbol{x}_n) < 0$
- Z upoštevanjem ciljne kodne sheme, kjer je $t \in \{+1, -1\}$, tedaj morajo vsi vzorci zadostiti enačbi
$$w^T \Phi(x_n) t_n > 0$$
- Za vse pravilno razvrščene vzorce je vrednost kriterija perceptrona nič, za vse zgrešeno razvrščene vzorce $\boldsymbol{x}_n$ skušamo minimizirati $-w^T \Phi(x_n) t_n$ ali *kriterij perceptrona*
$$E_P(\boldsymbol{w}) = -\sum_{n \in M} w^T \Phi_n t_n$$
kjer *M* pomeni niz zgrešenih vzorcev in $\Phi_n = \Phi(x_n)$

Algoritem perceptrona

- Funkcija napake $E_P(\boldsymbol{w}) = -\sum_{n \in M} w^T \Phi_n t_n$
- Stohastična Gradientna Metoda Spusta
  - sprememba uteži je podana z

$$w^{\tau+1} = w^{\tau} - \eta \nabla E_P(w) = w^{\tau} + \eta \Phi_n t_n$$

η predstavlja hitrost učenja, τ poa indeks koraka

- Algoritem učenja je enostaven

Krožimo po vzorcih učne množice in za vsak vzorec ocenjujemo funkcijo $y(\boldsymbol{x}) = f(\boldsymbol{w}^T \Phi(x))$

- Če vzorec korektno razvrstimo, tedaj se vrednost uteži ne spremeni
- Če vzorec napačno razvrstimo potem za razred C_1 prištejemo vektor $\Phi(x_n)$ k vektorju uteži $\boldsymbol{w}$ če ga napačno razvrstimo za razred C_2 potem vektor $\Phi(x_n)$ odštejemo od vektorja uteži $\boldsymbol{w}$.

Algoritem učenja pri Perceptronu

Če upoštevamo učinek enega koraka posodobitve v algoritmu učenja perceptrona vidimo, da bo prispevek napake zaradi zgrešene klasifikacije zmanjšan, zaradi

$$-\mathbf{w}^{(\tau+1)\mathrm{T}}\boldsymbol{\phi}_n t_n = -\mathbf{w}^{(\tau)\mathrm{T}}\boldsymbol{\phi}_n t_n - (\boldsymbol{\phi}_n t_n)^{\mathrm{T}}\boldsymbol{\phi}_n t_n < -\mathbf{w}^{(\tau)\mathrm{T}}\boldsymbol{\phi}_n t_n$$

kjer smo postavili vrednost parametra hitrosti učenja $\eta=1$

in uporabili izraz $abs(\Phi_n t_n)^2 > 0$

Sprememba vektorja uteži lahko povzroči, da predhodno pravilno razvrščeni vzorci bod sedaj napačno razvrščeni.

To pomeni, da algoritem učenja parceptrona ne zagotavlja zmanjšanja vrednosti funkcije napake v vsakem koraku.

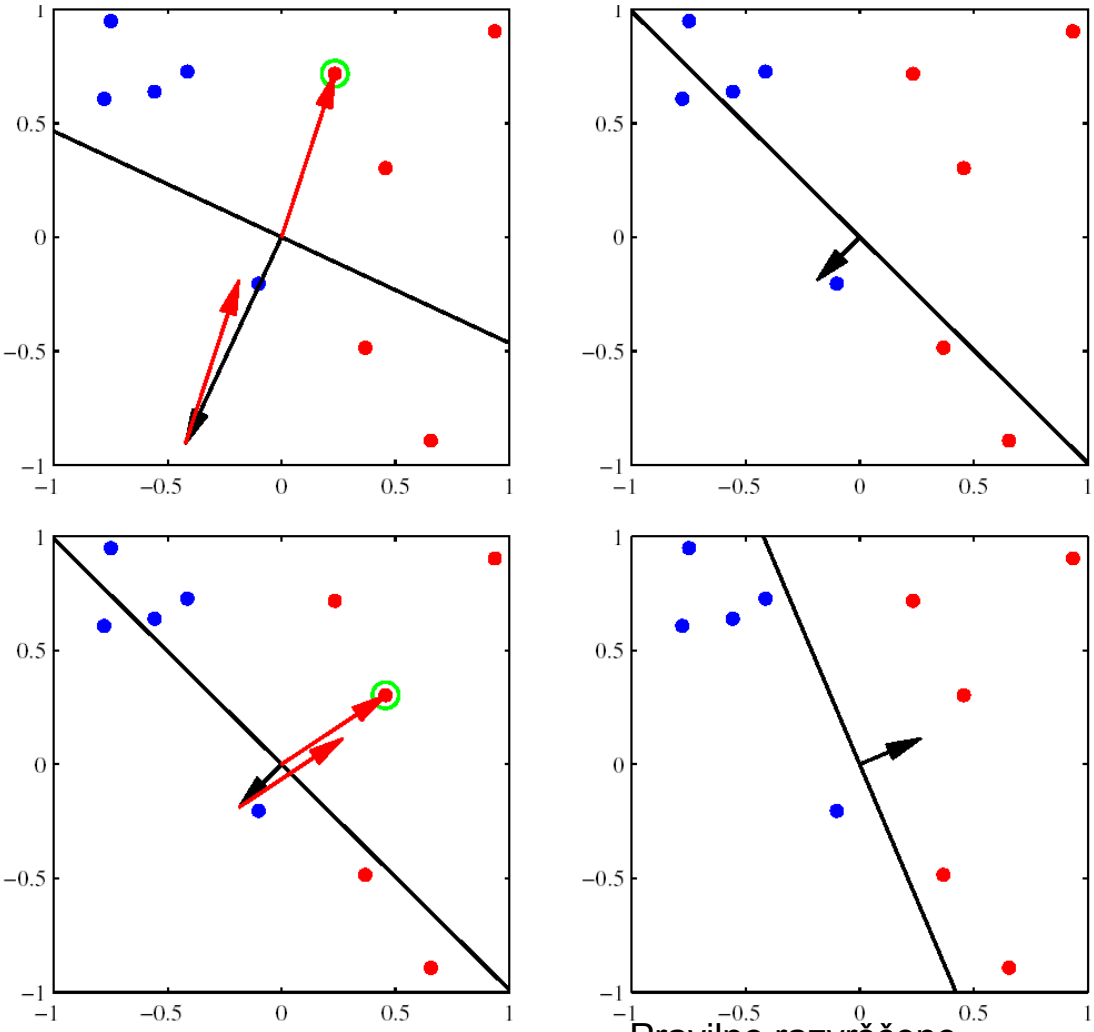
Vendar izrek konvergence učenja pravi, da če obstaja natančna rešitev,

(drugimi besedami, če so podatki linearno ločljivi), potem algoritem učenja perceptrona zagotavlja točno rešitev v končnem številu korakov.

Primer učenja perceptrona

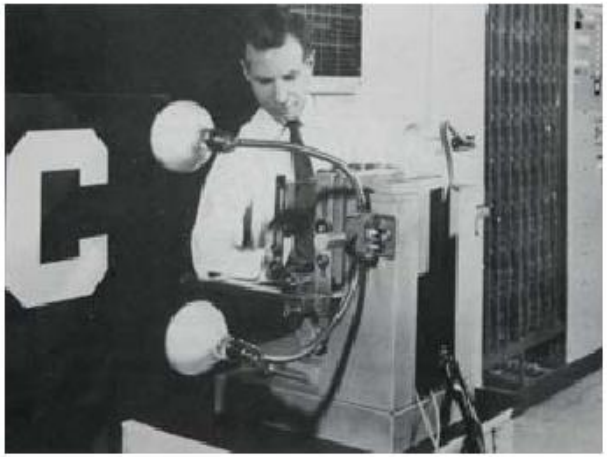
Vzemimo dvodimenzionalni prostor značilk Φ_1 , Φ_2 in dva razreda vektorja uteži v črnem

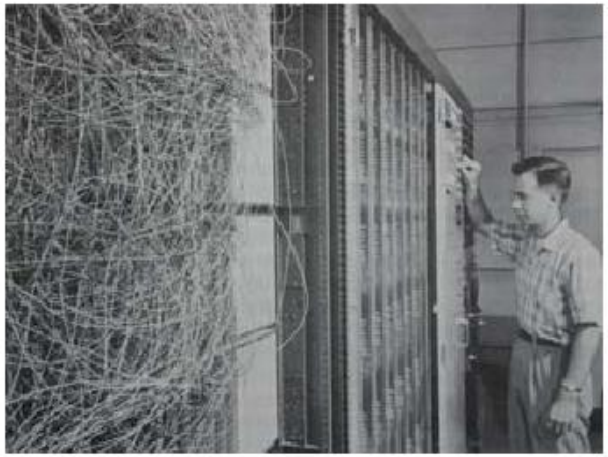
Zelene točke pomenijo napačno razvrstitev, kar prištejemo vektorju uteži

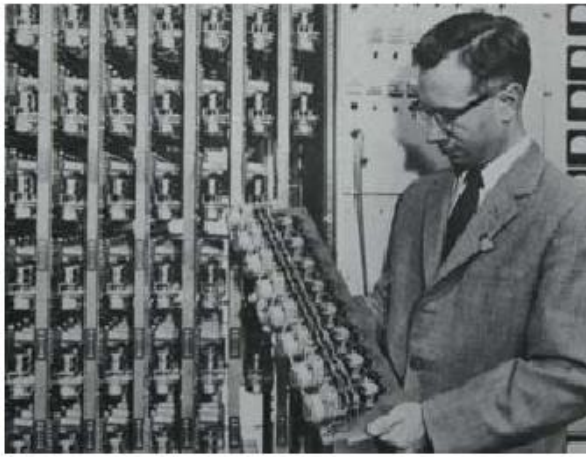


Pravilno razvrščene podatkovne točke

Primer prvega sistema zasnovanega na perceptronu







- Učenje za ločevanje oblike črk
- Celica 20x20

Povezovalna plošča

Ohišja z adaptivnimi utežmi (potenciometri)

Slabosti perceptrona

- Algoritme ne konvergira če posamezni razredi niso linearno ločljivi.
- Ne dopšuščajo verjetnega odločanja
- Niso primerni za večje število razredov $K > 2$

Povzetek

- Linearne diskrimin. funkcije imajo enostavno geometrijo
- Razširljive na več razredov
- Parametre lahko učimo z
  - metodo najmanjših kvadratov
    - Ni odporna na ubežnike, model close to target values
  - Fisherjevo linearno diskriminantno funkcijo
    - Dva razreda sta poseben primer LMS
    - Težko razširljiva na več razredov kot 2
  - Perceptrons
    - ne konvergirajo, če razredi niso linearno ločljivi
    - Ne dajejo verjetnega izhoda
    - niso uporabni za K >2

Linearna klasifikacija: verjetnostni generativni modeli

Linearna klasifikacija z generativnimi modeli

- Teme
  1. Pregled (Generativni vs Diskriminativni)
  2. Bayesov Klasifikator
     - using Logistic Sigmoid and Softmax
  3. zvezni vhodi
     - Gausov porazdeljeni razredi-pogoji
       – Ocena parametrov
  4. Diskretne značilke (features)
  5. Družina eksponentov

Pregled metod klasifikacije

- Generativni Modeli (dva koraka)
  - Sklepamo na osnovi gostote porazdelitve pogojev razredov $p(x|Ck)$ in verjetnosti $p(Ck)$
  - Uporabimo Bayesov teorem za določitev posteriorne verjetnosti

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

- Diskriminativni Modeli (En korak)
  - Direktno sklepamo o posteriorni vejetnosti $p(Ck|x)$
- Teorija odločanja
  - V obeh primerih uporabljamo teorijo odločanja za določitev razreda vsakemu novemu vhodnemu elementu x

Bayesov klasifikator in sigmoidna funkcija

- Modeliramo pogoje razredov $p(x|Ck)$ in verjetje $p(Ck)$
- Izračunamo posteriorno verjetje $p(Ck|x)$ z uporabo Bayesovega teorema
- Predpostavimo dva razreda
  Posteriorna verjetnost za $C1$ je

$$p(x) = \Sigma i p(x,Ci)$$
$$= \Sigma i p(x/Ci)p(Ci)$$

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

$$= \frac{1}{1+\exp(-a)} = \sigma(a)$$

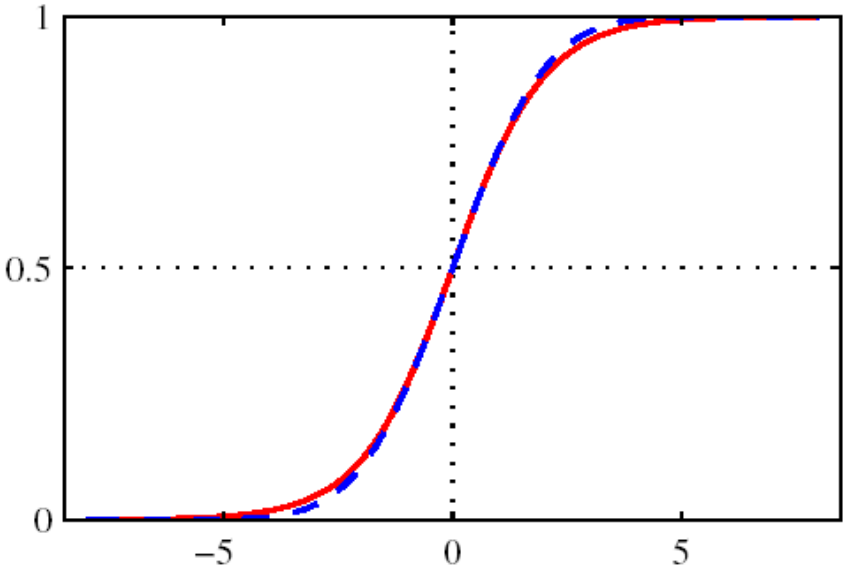
kjer je

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

  - Vedi da sta σ in $1\text{-}\sigma$ verjetnosti
  - vpeljemo σ saj bomo na koncu uporabili nelinearno aktivacijsko funkcijo

Logistična sigmoidna funkcija

Graf logistične sigmoidne funkcije, prikazane z rdečo, skupaj s skalirano funkcijo $\Phi(\lambda a)$, za $\lambda^2 = \pi/8$, Prikazana je z modro črtkano linijo, kjer je $\Phi(a) = \int_{-\infty}^{a} N(\theta|0, 1)\, d\theta$. Skalirni faktor $\pi/8$ izberemo tako da so odvodi obeh krivulj enaki v $a = 0$.



Sigmoida: "S"- oblikovana funkcija preslika $a \in (-\infty, +\infty)$ na končnem intervalu $(0,1)$

Črtkana linija predstavlja $\sigma(a)$ je sigmoidna funkcija definirana z

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

S simetričnimi lastnostmi

$$\sigma(-a) = 1 - \sigma(a)$$

Če $\sigma(a) = P(C_1|x)$ tedaj je inverz $ln[p(C1|x)/p(C2|x)$

Softmax: generalizacija sigmoidne funkcije

- Za $K=2$ imamo sigmoido
- Za $K > 2$, imamo

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

Če $K=2$ se izraz reducira

$p(C_1/x)=exp(a_1)/[exp(a_1)+exp(a_2)]$
$=1/[1+exp(a_2-a_1)]$
$=1/[1+exp(lnp(x/C_2)p(C_2)-ln(x/C_1)p(C_1)]$
$=1/[1+p(x/C_2)p(C_2)/p(x/C_1)p(C_1)]$
$=1/[1+exp(-a)]$ *where*

$$a = lnx \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$$

  - Veličine a_k so določene z

$$a_k = ln\ p(x|C_k)p(C_k)$$

- Znana pod imenom *soft-max* funkcija
  - to je zvezna max funkcija
    - Če $a_k >> a_j\ for\ all\ j \neq k$ potem $p(C_k|x) = 1$ in 0 za ostalo
    - Generalni pristop iskanja a_k

Zvezni vhodi: Sigmoidna funkcija

- Gaussova pogojna verjetnost z isto kovariančno matriko

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

- Primer dveh razredov

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$$

Kjer je

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

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

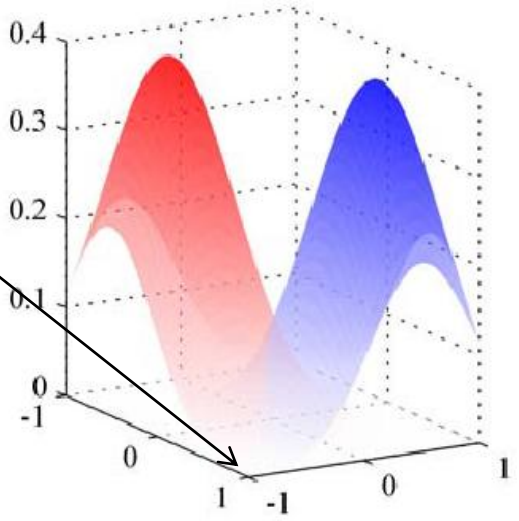
– kvadratični izraz po x

– A linear function of x in argument of logistic sigmoid

Dva Gaussova razreda

Dvo-dimenzionalni vhodni prostor $x=(x_1,x_2)$

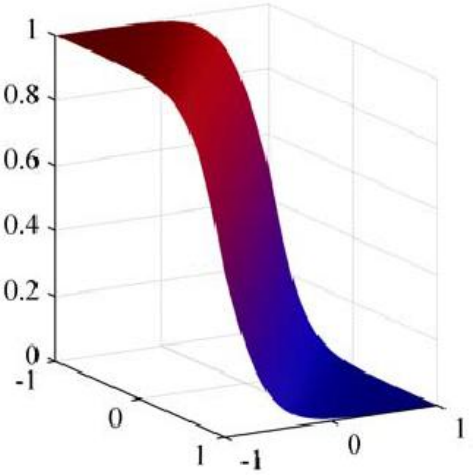
Pogojne gostote razredov $p(x|C_k)$



Linearna odločitvena meja

Vrednosti so pozitivne (seštevanje do 1 ni potrebno)

Posteriorna verjetnost $p(C_1|x)$



Logistična sigmoida linearne funkcije od x

Rdeče je sorazmerno z $p(C_1/x)$

Modro pa z $p(C_2/x)=1-p(C_1/x)$

Vrednost 1 ali 0

Zvezni primer za $K > 2$

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

S pogoji Gaussovih razredov

$$a_k(x) = w_k^T x + w_{k0}$$

Kvadratni členi se izničijo, kar vodi v linearnost

Kjer je

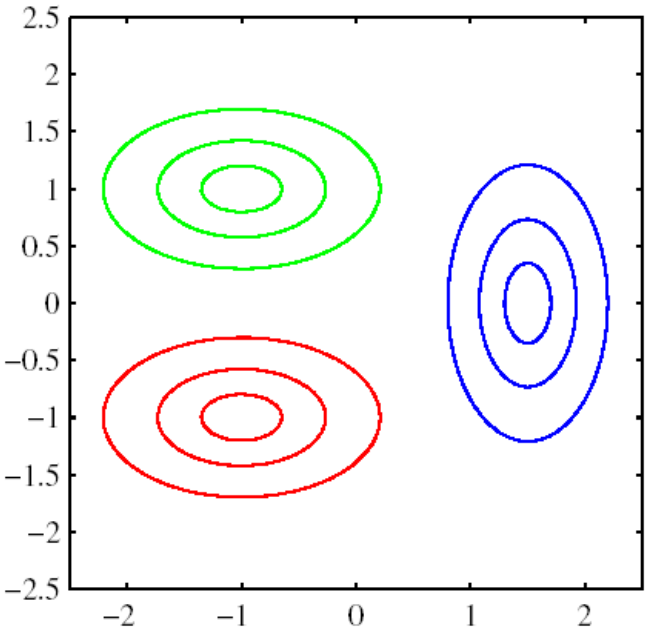
$$w_k = \Sigma^{-1}\mu$$

$$w_{k0} = -1/2\,\mu_k^T \Sigma^{-1}\mu_k + \ln p(C_k)$$

– Če nismo predvideli deljene kovariančne matrike, dobimo kvadratično diskriminanto

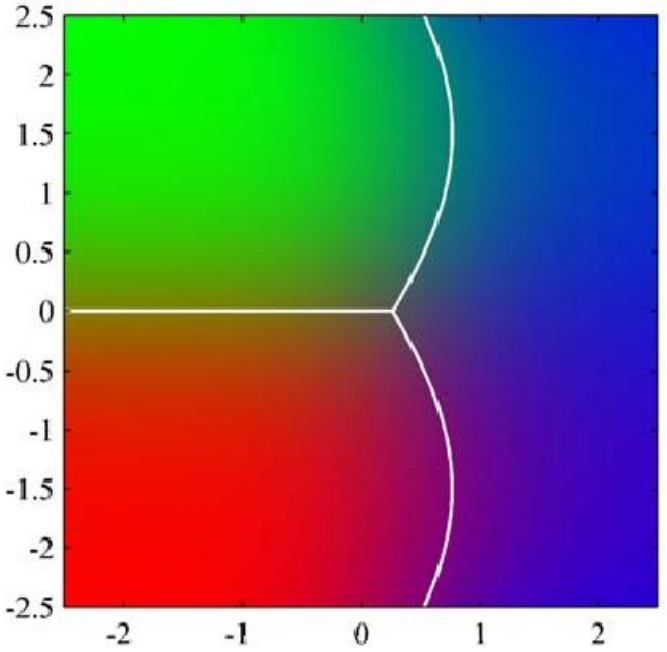
Tro-razredni primer z Gaussovimi modeli

Tako linearne kot tudi kvadratične odločitve



Pogojni gostoti razredov

C1 in *C2* imata enako kovariančno matriko



Posteriorne verjetnosti

Med *C1* in *C2* je linearna meja, drugi dve sta kvadratični

RGB vrednosti odgovarjajo posteriornim verjetnostim

Ocena največje verjetnosti za Gaussove parametre

- z upoštevanjem parametrične oblike zapisa za $p(x|C_k)$ lahko določimo vrednosti parametrov in $p(C_k)$ z uporabo maksimalne verjetnosti

Podani so podatki $\{x_n, t_n\}, n = 1, \ldots, N$, kjer t_n=*1* določa razred $\mathcal{C}_1$ in t_n=*0* določa razred $\mathcal{C}_2$

Naj bodo priorne verjetnosti: $p(\mathcal{C}_1) = \pi$ in $p(\mathcal{C}_2) = 1 - \pi$

$$p(x_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(x_n|\mathcal{C}_1) = \pi\, \mathcal{N}(x_n|\mu_1, \Sigma)$$

$$p(x_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(x_n|\mathcal{C}_2) = (1 - \pi)\, \mathcal{N}(x_n|\mu_2, \Sigma)$$

Verjetnost je podana z

$$p(t|\pi, \mu_1, \mu_2, \Sigma) == \prod_{n=1}^{N} [\pi \mathcal{N}(x_n|\mu_1, \Sigma)]^{t_n} [(1 - \pi)\mathcal{N}(x_n|\mu_2, \Sigma)]^{1-t_n}$$

Log verjetnost odvisna od π je $\sum_{n=1}^{N} \{t_n \ln \pi + (1 - t_n) \ln(1 - \pi)\}$

Če postavimo odvod na nič in uredimo, dobimo $\pi = \frac{1}{N} \sum_{n=1}^{N} t_n = \frac{N_1}{N_1 + N_2}$, kjer je N_1 število točk v razredu $\mathcal{C}_1$ in N_1 v razredu $\mathcal{C}_2$.

Optimiziramo glede na μ_1. Izberemo funkcijo log verjetnosti odvisno le od μ_1

$$\sum_{n=1}^{N} t_n \ln \mathcal{N}(x_n|\mu_1, \Sigma) = -\frac{1}{2} \sum_{n=1}^{N} t_n (x - \mu_1)^T \Sigma^{-1} (x - \mu_1) + const$$

Kjer $t = (t_1, .., t_N)^T$

Primerno za maksimizacijo log verjetnosti

- Literatura:
  - Sargur N. Srihari, Lecture notes, University at Buffalo, State University of New York

HVALA ZA POZORNOST