5. Linearni diskriminativni modeli

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.9

1 Linearni diskriminativni modeli

- Linearni diskriminativni modeli – granica je linearna ⇒ **hiperravnina**:

$$h(\mathbf{x}; \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

- Granica između klasa je na $h(\mathbf{x}) = 0$ (ponekad: $h(\mathbf{x}) = 0.5$)
- **Diskriminativan model** – izravno modelira granicu između klasa
- Generativni vs. diskriminativni modeli

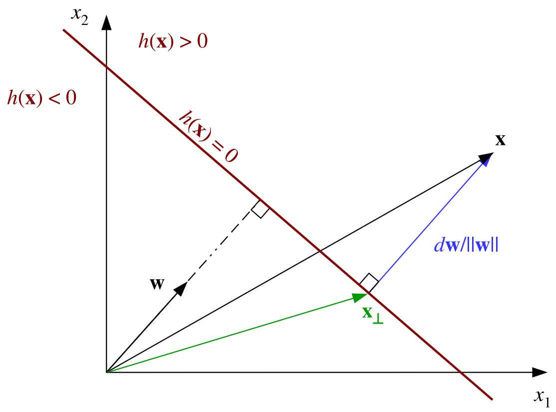
2 Geometrija linearnog modela

- BSO, razmatramo sljedeći model:

$$h(\mathbf{x}; \mathbf{w}) = w_1x_1 + w_2x_2 + w_0$$

- Granica je pravac:

$$w_1x_1 + w_2x_2 + w_0 = 0$$



- $\mathbf{w}$ je **normala** hiperravnine:

$$
\begin{aligned}
h(\mathbf{x}_1) &= h(\mathbf{x}_2)\\
\mathbf{w}^\mathrm{T}\mathbf{x}_1 + w_0 &= \mathbf{w}^\mathrm{T}\mathbf{x}_2 + w_0\\
\mathbf{w}^\mathrm{T}(\mathbf{x}_1 - \mathbf{x}_2) + w_0 - w_0 &= 0\\
\mathbf{w}^\mathrm{T}(\mathbf{x}_1 - \mathbf{x}_2) &= 0
\end{aligned}
$$

- Udaljenost primjera $\mathbf{x}$ od hiperravnine:

$$
\begin{aligned}
\mathbf{x} &= \mathbf{x}_\perp + d\frac{\mathbf{w}}{\|\mathbf{w}\|}\\
\underbrace{= \mathbf{w}^\mathrm{T}\mathbf{x} + w_0}_{h(\mathbf{x})} &= \underbrace{\mathbf{w}^\mathrm{T}\mathbf{x}_\perp + w_0}_{=h(\mathbf{x}_\perp)=0} + d\frac{\mathbf{w}^\mathrm{T}\mathbf{w}}{\|\mathbf{w}\|}\\
h(\mathbf{x}) = d\|\mathbf{w}\| &\quad \Rightarrow d = \frac{h(\mathbf{x})}{\|\mathbf{w}\|}
\end{aligned}
$$

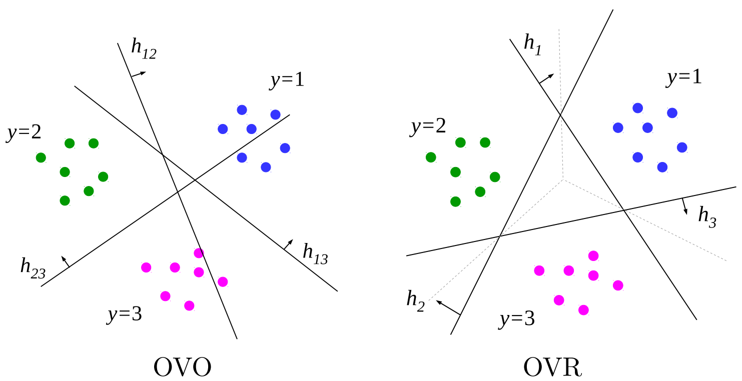
3 Višeklasna klasifikacija

- Shema **jedan-naspram-jedan** (**one-vs-one, OVO**) – $\binom{K}{2}$ binarnih modela:

$$
h(\mathbf{x}) = \underset{i}{\mathrm{argmax}} \sum_{i \neq j} \mathrm{sgn}\big(h_{ij}(\mathbf{x})\big), \quad h_{ji}(\mathbf{x}) = -h_{ij}(\mathbf{x})
$$

- Shema **jedan-naspram-ostali** (**one-vs-rest, OVR**) – K binarnih modela:

$$
h(\mathbf{x}) = \underset{j}{\mathrm{argmax}}\ h_j(\mathbf{x})
$$



- OVR ima manje modela od OVO, ali potencira neuravnotežnost klasa

4 Klasifikacija regresijom

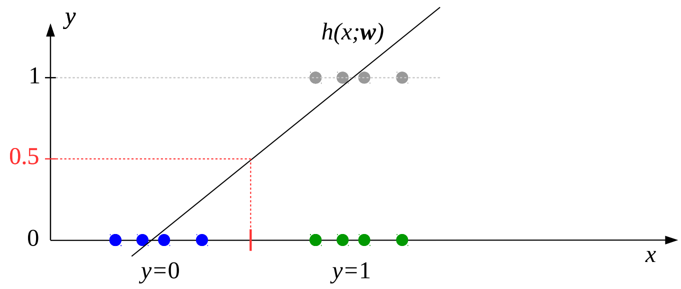
- Funkcija pogreške:

$$
E(\mathbf{w}|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}\big(\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}^{(i)}) - y^{(i)}\big)^2 = \frac{1}{2}(\boldsymbol{\Phi}\mathbf{w} - \mathbf{y})^\intercal(\boldsymbol{\Phi}\mathbf{w} - \mathbf{y})
$$

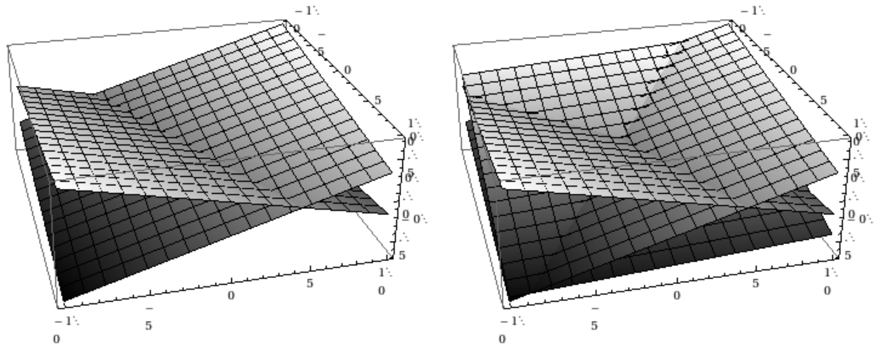
- Minimizator:

$$\mathbf{w}^* = (\mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal\mathbf{y} = \mathbf{\Phi}^+\mathbf{y}$$

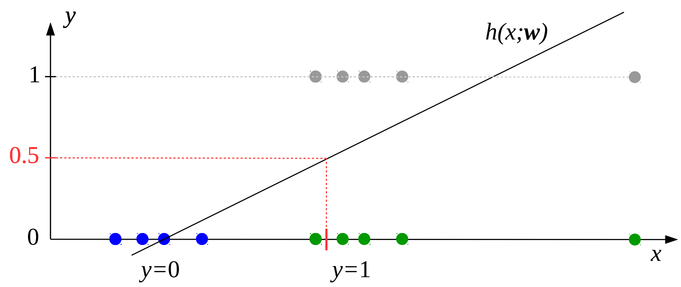
- Ideja: hipoteza koja predviđa $y = 1$ i $y = 0$ za primjere prve odnosno druge klase
- Model: $h(\mathbf{x};\mathbf{w}) = \mathbf{1}\{\mathbf{w}^\mathrm{T}\phi(\mathbf{x}) \geqslant 0.5\}$
- Skica za $n = 1$:



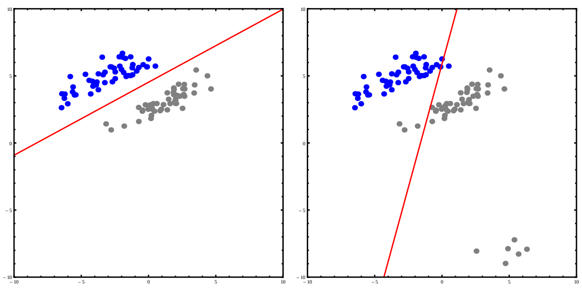
- Proširivo na $K > 2$ klasa shemom OVR ili OVO



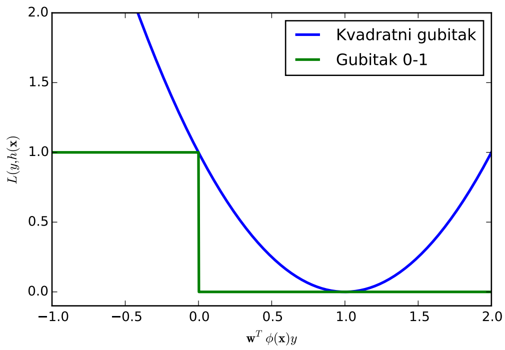
- Nedostatci: izlazi nisu vjerojatnosti, nerobusnost na vrijednosti koje odskaču
- Skica: nerobusnost na vrijednosti koje odskaču ($n = 1$):



- Nerobusnost na vrijednosti koje odskaču ($n = 2$):



- Uzrok: funkcija gubitka L kažnjava i dobro klasificirane primjere
- Za $y \in \{-1, +1\}$: ispravna klasifikacija $\Leftrightarrow \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})y > 0$
- Skica: L kao funkcija od $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})y$



- Idealan gubitak je gubitak 0-1, ali nije konveksan, pa nije pogodan za optimizaciju

5 Perceptron

- Model:

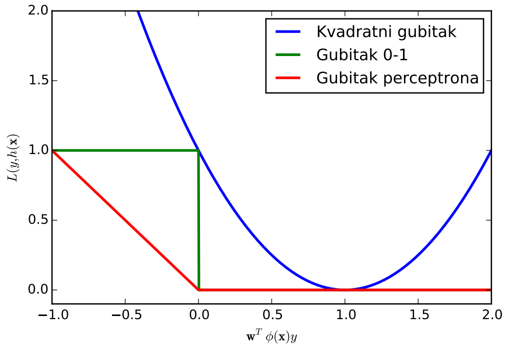
$$h(\mathbf{x}) = f\big(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\big)$$

- **Funkcija praga** kao aktivacijska funkcija:

$$f(\alpha) = \begin{cases} +1 & \text{ako } \alpha \geqslant 0 \\ -1 & \text{inače} \end{cases}$$

- Perceptron – umjetni neuron (McCulloch & Pitts, 1943.)
- Funkcija gubitka:

$$L(y, h(\mathbf{x})) = \max(0, -\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})y)$$

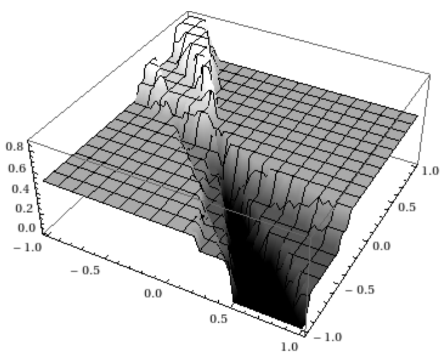


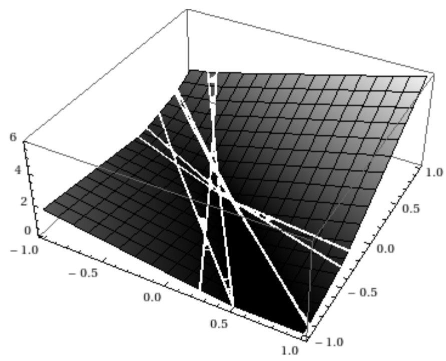
- Funkcija pogreške:

$$E(\mathbf{w}|\mathcal{D}) = -\sum_{i\,:\,f(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^{(i)}))\neq y^{(i)}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^{(i)})y^{(i)} = \sum_{i=1}^{N} \max\left(0, -\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^{(i)})y^{(i)}\right)$$

  $\Rightarrow$ kažnjava samo netočno klasificirane primjere (za razliku od regresije)

- Površina pogreške u prostoru parametara:


udio pogrešnih klasifikacija


pogreška perceptrona

- Ne postoji minimizator u zatvorenoj formi $\Rightarrow$ primjenjujemo **gradijentni spust**
- Gradijentni spust: težine ažuriramo u smjeru suprotnome od gradijenta
- Gradijent funkcije gubitka za netočno klasificirane primjere:

$$\nabla_{\mathbf{w}} L = \nabla_{\mathbf{w}}(-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})y) = -\boldsymbol{\phi}(\mathbf{x})y$$

- Ažuriranja težina – **Widrow-Hoffovo pravilo**:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\nabla L(\mathbf{w}|\mathcal{D})$$

  tj.

$$\mathbf{w} \leftarrow \mathbf{w} + \eta\boldsymbol{\phi}(\mathbf{x})y$$

  gdje je η stopa učenja

Algoritam perceptrona

1: inicijaliziraj $\mathbf{w} \leftarrow (0, \dots, 0)$
2: **ponavljaj** do konvergencije
3:     **za** $i = 1, \dots, N$
4:         **ako** $f\big(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^{(i)})\big) \neq y^{(i)}$ **onda** $\mathbf{w} \leftarrow \mathbf{w} + \eta\boldsymbol{\phi}(\mathbf{x}^{(i)})y^{(i)}$

- Nedostatci:
  - Izlazi modela nisu vjerojatnosti
  - Konvergira samo ako su primjeri linearno odvojivi (Rosenblatt, 1962)
  - Rezultat ovisi o početnim težinama