6. Logistička regresija

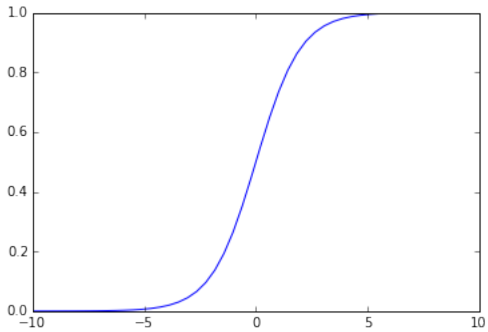
Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.5

1 Model logističke regresije

- **Logistička (sigmoidalna) funkcija**:

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



- Funkcija je derivabilna:

$$\frac{\partial\sigma(\alpha)}{\partial\alpha} = \frac{\partial}{\partial\alpha}\big(1+\exp(-\alpha)\big)^{-1} = \sigma(\alpha)\big(1-\sigma(\alpha)\big)$$

- Model logističke regresije:

$$h(\mathbf{x};\mathbf{w}) = \sigma\big(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\big) = \frac{1}{1+\exp(-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}))} = P(y=1|\mathbf{x})$$

  $\Rightarrow$ izlaz modela možemo tumačiti kao vjerojatnost da primjer pripada klasi $y=1$

- Ovo je primjer **poopćenog linearnog modela** (*generalized linear model, GLM*)
- GLM – linearni modeli s (nelinarnom) **aktivacijskom funkcijom** f:

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

  gdje je $f:\mathbb{R}\to[0,1]$ ili $f:\mathbb{R}\to(0,1)$ ili ($f:\mathbb{R}\to[-1,+1]$ ili $f:\mathbb{R}\to(-1,+1)$)

2 Pogreška unakrsne entropije

- Izlaz modela je **Bernoullijeva varijabla**:

$$P(y|\mu) = \begin{cases} \mu & \text{ako } y = 1 \\ 1-\mu & \text{inače} \end{cases} = \mu^y (1-\mu)^{1-y}$$

- U našem slučaju, y je oznaka primjera, a μ je izlaz modela, tj. $\mu = h(\mathbf{x};\mathbf{w})$, pa:

$$P(y^{(i)}|\mathbf{x}^{(i)}) = h(\mathbf{x};\mathbf{w})^y \big(1 - h(\mathbf{x};\mathbf{w})\big)^{1-y}$$

- Log-izglednost oznaka iz skupa označenih primjera:

$$\begin{aligned} \ln P(\mathbf{y}|\mathbf{X},\mathbf{w}) &= \ln \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)}) = \\ &= \sum_{i=1}^{N} \Big(y^{(i)} \ln h(\mathbf{x}^{(i)};\mathbf{w}) + (1-y^{(i)}) \ln \big(1 - h(\mathbf{x}^{(i)};\mathbf{w})\big)\Big) \end{aligned}$$

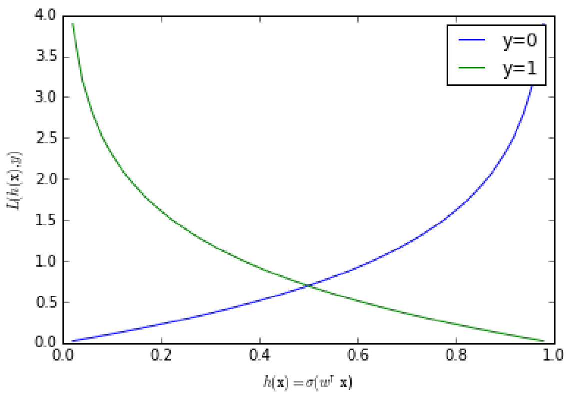
- Empirijska pogreška je negativna log-izglednost:

$$E(\mathbf{w}|\mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \Big(- y^{(i)} \ln h(\mathbf{x}^{(i)};\mathbf{w}) - (1-y^{(i)}) \ln \big(1 - h(\mathbf{x}^{(i)};\mathbf{w})\big)\Big)$$

  $\Rightarrow$ **pogreška unakrsne entropije** (**cross-entropy error**)

- **Gubitak unakrsne entropije** (**cross-entropy loss**):

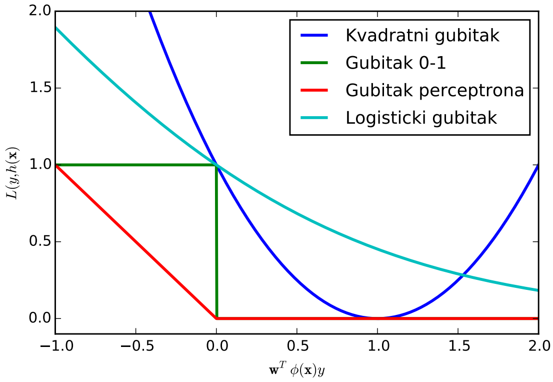
$$L(y, h(\mathbf{x})) = -y \ln h(\mathbf{x}) - (1-y) \ln \big(1 - h(\mathbf{x})\big)$$



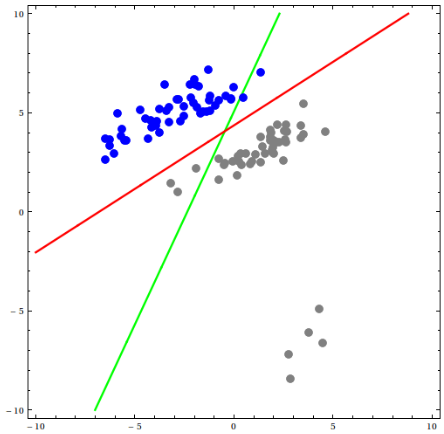
- Reformulacija $y \in \{0,1\} \to y \in \{-1,+1\}$ i skaliranje sa $1/\ln 2$:

$$L(y, h(\mathbf{x})) = \frac{1}{\ln 2} \ln \big(1 + \exp(-y\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}))\big)$$

- Usporedba funkcija gubitaka:



- Logistička regresija robusnija je od modela linearne regresije:



- Minimizacija u zatvorenoj formi nije moguća $\Rightarrow$ iterativna optimizacija

3 Gradijentni spust

- **Gradijentni spust** – minimum nalazimo krećući se u smjeru suprotnom od gradijenta:

$$\mathbf{x} \leftarrow \mathbf{x} - \eta \nabla f(\mathbf{x})$$

- η je **stopa učenja**: prevelika $\eta \Rightarrow$ divergencija; premalena $\eta \Rightarrow$ spora konvergencija
- Želimo **globalnu konvergenciju** (konvergencija uvijek i svugdje)
- Ostvarivo **linijskim pretraživanjem** – η koji minimizira $f(\mathbf{x})$ u smjeru spusta $\Delta\mathbf{x}$:

$$g(\eta) = f(\mathbf{x} + \eta \Delta \mathbf{x})$$

- Pronađeni optimum bit će globalni optimum ako je $f(\mathbf{x})$ **konveksna**
- Funkcija $f : \mathbb{R}^n \rightarrow \mathbb{R}$ je **konveksna** akko

(1) Njezina domena $\mathrm{dom}(f)$ je **konveksni skup**:
Za svaki $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathrm{dom}(f)$ i za svaki $\alpha_1, \ldots \alpha_n$ takav da $\sum_i \alpha_1 = 1$ vrijedi:

$$\sum_{i=1}^{n} \alpha_i \mathbf{x}_i \in \mathrm{dom}(f)$$

(2) Za svaki $\mathbf{x}_1, \mathbf{x}_2 \in \mathrm{dom}(f)$ i svaki $\alpha \in [0,1]$ vrijedi:

$$f(\mathbf{x}) = f\big(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2\big) \leqslant \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2)$$

- Empirijska pogreška je konveksna $\Leftrightarrow$ funkcija gubitka L je konveksna
- Dvije varijante gradijentnog spusta:
  - **Batch** (grupni): $\mathbf{w} \leftarrow \mathbf{w} - \eta \sum_{i=1}^{N} \nabla L\big(y^{(i)}, h(\mathbf{x}^{(i)}; \mathbf{w})\big)$
  - **Stohastički** (**SGD**): $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla L\big(y^{(i)}, h(\mathbf{x}^{(i)}; \mathbf{w})\big)$
- SGD je pogodan za on-line učenje (big data, data streams)

4 Gradijentni spust za logističku regresiju

- Gradijent funkcije gubitka i funkcije pogreške:

$$\begin{aligned}
E(\mathbf{w}|\mathcal{D}) &= \frac{1}{N} \sum_{i=1}^{N} \Big(- y^{(i)} \ln h(\mathbf{x}^{(i)}; \mathbf{w}) - (1 - y^{(i)}) \ln \big(1 - h(\mathbf{x}^{(i)}; \mathbf{w})\big)\Big) \\
\nabla_{\mathbf{w}} E(\mathbf{w}|\mathcal{D}) &= \frac{1}{N} \sum_{i=1}^{N} \nabla L\big(y^{(i)}, h(\mathbf{x}^{(i)}; \mathbf{w})\big) \\
\nabla L\big(y, h(\mathbf{x})\big) &= \Big(-\frac{y}{h(\mathbf{x})} + \frac{1-y}{1-h(\mathbf{x})}\Big) h(\mathbf{x})\big(1 - h(\mathbf{x})\big)\boldsymbol{\phi}(\mathbf{x}) = \big(h(\mathbf{x}) - y\big)\boldsymbol{\phi}(\mathbf{x}) \\
\nabla E(\mathbf{w}|\mathcal{D}) &= \sum_{i=1}^{N} \big(h(\mathbf{x}^{(i)}) - y^{(i)}\big)\boldsymbol{\phi}(\mathbf{x}^{(i)})
\end{aligned}$$

(faktor $1/N$ može se apsorbirati u stopu učenja η)

Logistička regresija (grupni gradijentni spust)

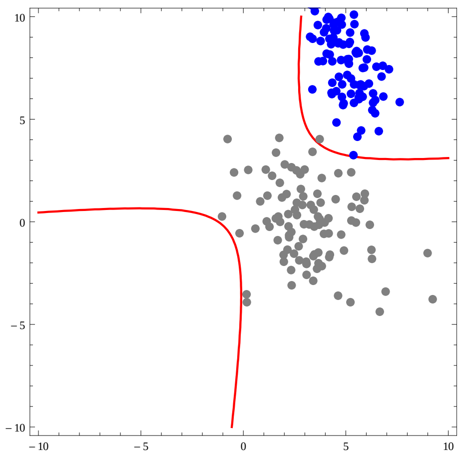
$\mathbf{w} \leftarrow (0, 0, \ldots, 0)$
ponavljaj do konvergencije
  $\Delta\mathbf{w} \leftarrow (0, 0, \ldots, 0)$
  za $i = 1, \ldots, N$
    $h \leftarrow \sigma(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}^{(i)}))$
    $\Delta\mathbf{w} \leftarrow \Delta\mathbf{w} - (h - y^{(i)})\, \boldsymbol{\phi}(\mathbf{x}^{(i)})$
  $\eta \leftarrow$ optimum linijskim pretraživanjem u smjeru spusta $\Delta\mathbf{w}$
  $\mathbf{w} \leftarrow \mathbf{w} + \eta \Delta\mathbf{w}$

Logistička regresija (stohastički gradijentni spust)

$\mathbf{w} \leftarrow (0, 0, \ldots, 0)$
ponavljaj do konvergencije
  slučajno permutiraj primjere u $\mathcal{D}$
  za $i = 1, \ldots, N$
    $h \leftarrow \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^{(i)}))$
    $\Delta\mathbf{w} \leftarrow -(h - y^{(i)})\boldsymbol{\phi}(\mathbf{x}^{(i)})$
    $\eta \leftarrow$ optimum linijskim pretraživanjem u smjeru spusta $\Delta\mathbf{w}$
    $\mathbf{w} \leftarrow \mathbf{w} + \eta\Delta\mathbf{w}$

5 Regularizirana regresija

- Prednosti regularizacije:
  - Sprječavanje pretjerane nelinearnosti
  - Suzbijanje nepotrebnih značajki
  - Sprječavanje otvrdnjavanja sigmoide kod linearno odvojivih problema
- Primjer prenaučenosti ($n = 2, \boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$):



- L2-regularizirana pogreška:

$$E_R(\mathbf{w}|\mathcal{D}) = \sum_{i=1}^{N}\Big(-y^{(i)}\ln h(\mathbf{x}^{(i)}) - (1 - y^{(i)})\ln\big(1 - h(\mathbf{x}^{i})\big)\Big) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

- Ažuriranje težina:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\Big(\sum_{i=1}^{N}\big(h(\mathbf{x}^{(i)}) - y^{(i)}\big)\boldsymbol{\phi}(\mathbf{x}^{(i)}) + \lambda\mathbf{w}\Big)$$

  ekvivalentno:

$$\mathbf{w} \leftarrow \mathbf{w}(1 - \eta\lambda) - \eta\sum_{i=1}^{N}\big(h(\mathbf{x}^{(i)}) - y^{(i)}\big)\boldsymbol{\phi}(\mathbf{x}^{(i)})$$

- Napomena: Težina w_0 se ne regularizira

L2-regularizirana logistička regresija (grupni gradijentni spust)

$\tilde{\mathbf{w}} \leftarrow (0, 0, \ldots, 0)$ // $\tilde{\mathbf{w}}$ je prošireni vektor $(w_0, \mathbf{w})$
ponavljaj do konvergencije
  $\Delta w_0 \leftarrow 0$
  $\Delta \mathbf{w} \leftarrow (0, 0, \ldots, 0)$
  za $i = 1, \ldots, N$
    $h \leftarrow \sigma(\tilde{\mathbf{w}}^{\mathrm{T}} \boldsymbol{\phi}(\tilde{\mathbf{x}}^{(i)}))$
    $\Delta w_0 \leftarrow \Delta \mathbf{w}_0 - (h - y^{(i)})$
    $\Delta \mathbf{w} \leftarrow \Delta \mathbf{w} - (h - y^{(i)}) \boldsymbol{\phi}(\mathbf{x}^{(i)})$
  $\eta \leftarrow$ optimum linijskim pretraživanjem u smjeru spusta $\Delta \tilde{\mathbf{w}}$
  $w_0 \leftarrow w_0 + \eta \Delta w_0$
  $\mathbf{w} \leftarrow \mathbf{w}(1 - \eta\lambda) + \eta \Delta \mathbf{w}$

L2-regularizirana logistička regresija (stohastički gradijentni spust)

$\tilde{\mathbf{w}} \leftarrow (0, 0, \ldots, 0)$ // $\tilde{\mathbf{w}}$ je prošireni vektor $(w_0, \mathbf{w})$
ponavljaj do konvergencije:
  slučajno permutiraj primjere u $\mathcal{D}$
  za $i = 1, \ldots, N$
    $h \leftarrow \sigma(\tilde{\mathbf{w}}^{\mathrm{T}} \boldsymbol{\phi}(\tilde{\mathbf{x}}^{(i)}))$
    $\Delta w_0 \leftarrow -(h - y^{(i)})$
    $\Delta \mathbf{w} \leftarrow -(h - y^{(i)}) \boldsymbol{\phi}(\mathbf{x}^{(i)})$
    $\eta \leftarrow$ optimum linijskim pretraživanjem u smjeru spusta $\Delta \tilde{\mathbf{w}}$
    $w_0 \leftarrow w_0 + \eta \Delta w_0$
    $\mathbf{w} \leftarrow \mathbf{w}(1 - \eta\lambda) + \eta \Delta \mathbf{w}$