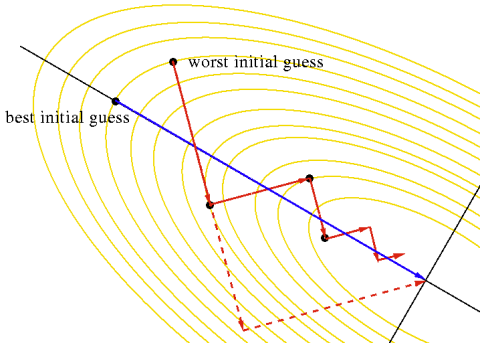
7. Logistička regresija II

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

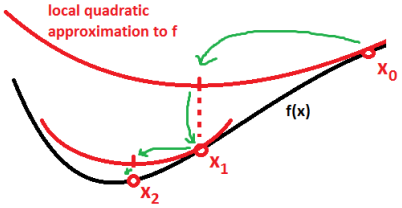
Jan Šnajder, natuknice s predavanja, v1.5

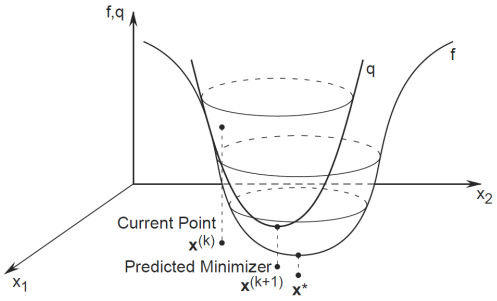
1 Alternative gradijentnom spustu

- Gradijentni spust s linijskim pretraživanjem ima cik-cak trajektoriju ⇒ sporo



- Alternativa: **optimizacija drugog reda**, npr. **Newtonov postupak**
- Ideja: skok iz trenutačnog minimuma do minimuma kvadratne aproks. funkcije





- Kvadratna aproksimacija $f(\mathbf{x})$ u točki $\mathbf{x}_t$ razvojem u **Taylorov red** drugog reda:

$$f(\mathbf{x}) \approx f_{\text{quad}}(\mathbf{x}) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_t) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^{\mathrm{T}}\mathbf{H}_t(\mathbf{x} - \mathbf{x}_t)$$

gdje je $\mathbf{H}_t$ **Hesseova matrica** funkcije $f(\mathbf{x})$ u točki $\mathbf{x}_t$

$$\mathbf{H} = \nabla\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

- $f(\mathbf{x})$ je konveksna $\Leftrightarrow$ $\mathbf{H}$ je pozitivno semi-definitna (ali ne nužno pozitivno definitna!)
- Ažuriranje parametara:
$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{H}_t^{-1} \nabla f(\mathbf{x}_t)$$
- Ne radi ako $\mathbf{H}$ nije invertibilna $\Rightarrow$ multikolinearnost $\Rightarrow$ treba regularizirati
- Specifično, za logističku regresiju:
$$\mathbf{H} = \mathbf{\Phi}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi}$$
gdje $\mathbf{S} = \mathrm{diag}\big(h(\mathbf{x}^{(i)})(1 - h(\mathbf{x}^{(i)}))\big)$
- Pravilo ažuriranja:
$$\mathbf{w} \leftarrow \mathbf{w} - \mathbf{H}^{-1} \nabla E(\mathbf{w}|\mathcal{D}) \qquad (\eta = 1)$$
$\Rightarrow$ algoritam **iteratively reweighted least squares (IRLS)**
- Izračun $\mathbf{H}_t$ u svakom koraku je potencijalno skup
- Alternativa: **kvazi-Newtonovi postupci** (BFSG, L-BFSG) – aproksimiraju $\mathbf{H}_t$
- Uključivanje L2-regularizacije je jednostavno:
$$\nabla E_R(\mathbf{w}|\mathcal{D}) = \nabla E(\mathbf{w}|\mathcal{D}) + \lambda \mathbf{w}$$
$$\mathbf{H_R} = \mathbf{H} + \lambda I$$
- L1-regularizacija: **podgradijentne metode** (koordinatni spust, proksimalne metode)

2 Višeklasna logistička regresija

- OVO/OVR ne daje vjerojatnosnu distribuciju po klasama
- **Funkcija softmax**: $\mathrm{softmax} : \mathbb{R}^n \to \mathbb{R}^n$, gdje za komponentu k vrijedi:
$$\mathrm{softmax}_k(x_1, \ldots, x_n) = \frac{\exp(x_k)}{\sum_j \exp(x_j)}$$
$\Rightarrow$ normalizira tako da $\sum x_k = 1$ te smanjuje male i povećava velike vrijednosti
- **Multinomijalna logistička regresija** (**MNR**, maximum entropy classifier):
$$h_k(\mathbf{x}; \mathbf{W}) = \frac{\exp(\mathbf{w}_k^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}))}{\sum_j \exp(\mathbf{w}_j^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}))} = P(y = k|\mathbf{x}, \mathbf{W})$$
gdje $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$

- Izlaz je **multinulijeva** (kategorička) varijabla $\mathbf{y} = (y_1, y_2, \ldots, y_K)^{\mathrm{T}}$, s distribucijom:

$$P(\mathbf{y}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{y_k}$$

- Log-izglednost označenih primjera:

$$\begin{aligned}\ln P(\mathbf{y}|\mathbf{X}, \mathbf{W}) &= \ln \prod_{i=1}^{N} P(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) = \ln \prod_{i=1}^{N} \prod_{k=1}^{K} \mu_k^{y_k^{(i)}} = \ln \prod_{i=1}^{N} \prod_{k=1}^{K} h_k(\mathbf{x}^{(i)}; \mathbf{W})^{y_k^{(i)}} \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} y_k^{(i)} \ln h_k(\mathbf{x}^{(i)}; \mathbf{W})\end{aligned}$$

  $\Rightarrow$ poopćena **pogreška unakrsne entropije**:

$$E(\mathbf{W}|\mathcal{D}) = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_k^{(i)} \ln h_k(\mathbf{x}^{(i)}; \mathbf{W})$$

- Funkcija gubitka:

$$L(\mathbf{y}, h_k(\mathbf{x})) = -\sum_{k=1}^{K} y_k \ln h_k(\mathbf{x}; \mathbf{W})$$

- Gradijent za klasu k:

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}|\mathcal{D}) = \sum_{i=1}^{N} \big(h_k(\mathbf{x}^{(i)}; \mathbf{W}) - y_k^{(i)}\big)\boldsymbol{\phi}(\mathbf{x}^{(i)})$$

  $\Rightarrow$ gradijent je isti kao i za binarnu logističku funkciju

- On-line ažuriranje:

$$\mathbf{w}_k \leftarrow \mathbf{w} - \eta\big(h(\mathbf{x}^{(i)}; \mathbf{w}) - y^{(i)}\big)\boldsymbol{\phi}(\mathbf{x}^{(i)})$$

  $\Rightarrow$ **algoritam least-mean-squares (LMS)** ili **Widrow-Hoffovo pravilo**

- Isto dobivamo za on-line optimizaciju linearne regresije

3 Poopćeni linearni modeli i eksponencijalna familija

- Unificirani pogled na tri poopćena linearna modela koja smo razmatrali
- Linearna regresija:

$$\begin{aligned}h(\mathbf{x}; \mathbf{w}) &= \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) \\ P(y|\mathbf{x}, \mathbf{w}) &= \mathcal{N}(\mu, \sigma^2) = \mathcal{N}(h(\mathbf{x}), \sigma^2) \\ L(y, h(\mathbf{x})) &= (h(\mathbf{x}) - y)^2 \\ \nabla_{\mathbf{w}} L\big(y, h(\mathbf{x})\big) &= \big(h(\mathbf{x}) - y\big)\boldsymbol{\phi}(\mathbf{x})\end{aligned}$$

- Logistička regresija:

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

- Multinomijalna logistička regresija:

$$h_k(\mathbf{x};\mathbf{W}) = \mathrm{softmax}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})) = \frac{\exp(\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}))}{\sum_j \exp(\mathbf{w}_j^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}))} = P(y=k|\mathbf{x},\mathbf{w})$$
$$P(\mathbf{y}|\mathbf{x},\mathbf{w}) = \prod_{k=1}^{K}\mu_k^{y_k} = \prod_{k=1}^{K} h_k(\mathbf{x})^{y_k}$$
$$L(\mathbf{y},h_k(\mathbf{x})) = -\sum_{k=1}^{K} y_k \ln h_k(\mathbf{x};\mathbf{W})$$
$$\nabla_{\mathbf{w}_k} L\big(y_k,h_k(\mathbf{x})\big) = \big(h_k(\mathbf{x})-y_k\big)\boldsymbol{\phi}(\mathbf{x})$$

- Sve tri korištene distribucije pripadaju **eksponencijalnoj familiji distribucija**:

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})\exp\big(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})\big)$$

- Ključno za poopćene linearne modele – distribucija određuje aktivacijsku funkciju:
  - Gauss ↔ funkcija identiteta, Bernoulli ↔ logistička, Multinoulli ↔ softmax

4 Adaptivne bazne funkcije

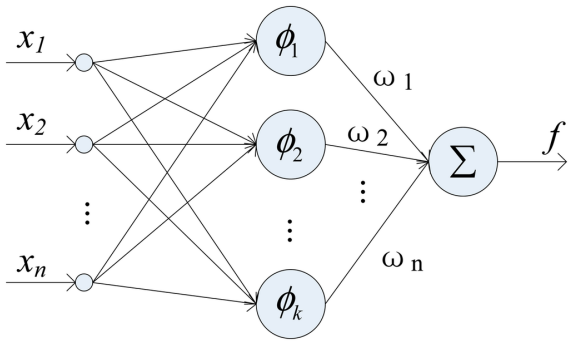
- Model s baznim funkcijama:

$$h(\mathbf{x};\mathbf{w}) = f\big(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\big) = f\Big(\sum_{j=0}^{m} w_j\phi_j(\mathbf{x})\Big)$$

- Fiksne (u obliku i broju) adaptivne funkcije mogu biti ograničavajuće
- **Parametrizirane bazne funkcije** – svaka bazna funkcija je poopćeni linearan model:

$$h(\mathbf{x};\mathbf{w}) = f\Big(\sum_{j=0}^{m} w_j^{(2)} \underbrace{f\big(\sum_{i=0}^{n} w_{ji}^{(1)} x_i\big)}_{=\phi_j(\mathbf{x})}\Big) = f\big(\mathbf{w}^{(2)\mathrm{T}} f(\mathbf{W}^{(1)}\mathbf{x})\big)$$

- Dobili smo dvoslojnu **neuronsku mrežu**



- Složeniji model, ali ga je lakše pretrenirati te optimizacija nije konveksna