Maschinelles Lernen

Johannes Eifler und co
Matrikel: 144915
Stand: 19. April 2016 19:56

lecture 1 (04.04.2016)

Introduction

- supervised learning: learn relaionships between variables
- unsupervised learning: learn some structure of measured variables

Dependend variables are measured at independant variables (covariates). Variables are measured on some **scale**:

- nominal (gender, color)
- ordinal (ranking of soccerteams)
- interval (temperature in degree celsius)
- rational (temperature in kelvin, weight, height), has meaningful zero in comparison to interval

$\Rightarrow$ quotients make sense on ratio scale; quotiens of differences make sense on interval scale
metric scale: interval- and ratio scale

problems in machine learning

:

1. **regression**: one dependent variable on metric scale
   one or more independent variables on metric scale
2. **variance analysis**: one dependent variable on metric scale
   one or more independent variables on nominal scale
3. **classification**: one dependent variable on nominal scale
   one or more independent variables on metric scale
4. **contingency analysis**: one dependent variable on nominal scale
   one or more independent variables on nominal scale
5. **scaling problems**: independent variables on arbitrary scale but measurements on ordinal scale
   dependent variables on metric scale

linear regression

data/measurements: $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$

$x^{(i)}$ independent/covariates $\in \mathbb{R}^n$ (n - variables)
$y^{(i)}$ dependent/variates $\in \mathbb{R}^n$
plot suggests a linear dependence between x and y
$y = \Theta_1 x + \Theta_0$

in the multivate case: $y = \Theta_0 + \Theta_1 X_1 + \cdots + \Theta_n X_n$
$= \Theta^T X, X = (1, X_1, \dots, X_n) \in \mathbb{R}^{n+1}$

problem: estimate the parameter vector $\Theta in \mathbb{R}^{n+1}$ from the measurements $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$
loss function: $L(\Theta) = \frac{1}{2}\Sigma_{i=1}^m (\Theta^T X^{(i)} - y^{(i)})^2$
model loss $\hat{=}$ loss for parameter vector Θ

goal: choose $\Theta \in \mathbb{R}^{n+1}$ that minimizes the loss function
reformulation:

data matrix:

$$X = \begin{pmatrix} x^{(1)^T} \\ \vdots \\ x^{(n)^T} \end{pmatrix} \in \mathbb{R}^{m \times (n+1)}$$

response vector:

$$Y = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{pmatrix} \in \mathbb{R}^n$$

parameter vector:

$$\Theta = \begin{pmatrix} \Theta_0 \\ \vdots \\ \Theta_n \end{pmatrix} \in \mathbb{R}^{n+1}$$

loss function in vectorized form:

$$L(\Theta) = \frac{1}{2}\Sigma_{i=1}^m (\Theta^T * X^{(i)} - Y^{(i)})^2 = \frac{1}{2}\|X * \Theta - Y\|_2^2$$

(vector of predictions vector of observation response)

$$= \frac{1}{2}(X * \Theta - Y)^T * (X * \Theta - Y)$$

(definition of the euclidian norm)

$$= \frac{1}{2}(\Theta^T X^T \times \Theta - \Theta^T X^T Y - Y^T X * \Theta + Y^T Y)$$

$$= -2\Theta^T X^T Y \text{ since the dot product is symmetric } (X^T Y = Y^T X)$$

$$= \frac{1}{2}\Theta^T X^T X\Theta - \Theta^T X^T Y + \frac{1}{2}Y^T Y$$

remember from calculus: A neccessary condition for an optimum of the (loss-) function is that the gradient vanishes.

$\bigtriangledown_\Theta L(\Theta) \stackrel{!}{=} 0$ $\qquad$ $t(x) = \frac{1}{2}x^2 + ax + b$

$\bigtriangledown_\Theta L(\Theta) = X^T X\Theta * X^T Y \stackrel{!}{=} 0$ $\qquad$ $\bigtriangledown_x t(x) = x + a$

here we have used that $X^T X$ is symmetric

$\Rightarrow X^T X\Theta = X^T Y$ $\qquad$ $t(\Theta) = \Theta^T X\Theta$

$\Rightarrow \Theta = (X^T X)^{-1} X^T Y$ $\qquad$ $\bigtriangledown_\Theta t(\Theta) = (X + X^T)\Theta$

privided that $(X^T X)^{-1}$ exists

$(X^T X)_{ij} = X^{(i)^T} X^{(j)}$

operation matrix

dot product of i-th dataa point and j-th data point

hence, the last square solution of the linear regression problem is $\Theta = (X^T X)^{-1} X^T Y$

more robust solution:

$\Theta = (X^T X + \gamma \mathbb{1})^{-1} X^T Y, \qquad \gamma > 0$ regularization parameter

ridge regression solution is not only more robust numerically, but also statistically (it is not so sensitive to small measurement errors in X).

Natural question: which loss function gives us the ridge regression solution?

answer: $\qquad L_{ridge}(\Theta) = \frac{1}{2}\|X * \Theta - Y\|_2^2 + \gamma\|\Theta\|_2^2$

loss term $\quad$ reularisation term

probabilistic interpretation of least squares

$$Y = \Theta^T * X + \epsilon$$

deterministic part $\quad$ random/noise part

model of the noise: gaussian noise $\quad p(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{\epsilon^2}{2\sigma^2}) \quad$ (probability density function)

$P[a \leq \epsilon \leq b] = \int_a^b p(\epsilon) \quad d\epsilon$

Y is a function of the random noise term ϵ als a random variable. The probability density function of Y is:

$$p(Y) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\|Y - \Theta^T X\|_2^2}{2\sigma^2})$$

lecture 2 (06.04.2016)

linear regression

data:

$$(x^{(1)}, y^{(1)}), \dots, (x^n, y^n)$$
$$x^{(i)} \in \mathbb{R}^n \qquad covariates$$
$$y^{(i)} \in \mathbb{R} \qquad variates/response$$

assumption:

(1) $y = t(x)$ $\qquad$ y is function of x

linear regression $\underbrace{\boxed{y = \Theta^T x}}_{\sum_{i=0}^n \Theta_i x_i}$ $\Theta \in \mathbb{R}^{n+1}$ (parameter vector)

$x = (1, x_1, \dots, x_n) \in \mathbb{R}^{n+1} \qquad x_0 = 1$

(2) data are obscured by random noise:

$y = \Theta^T x + \epsilon, \qquad \epsilon =$ random noise term

$p(\epsilon) = \frac{1}{\sqrt{2\pi}} r \exp(-\frac{\epsilon^2}{2\sigma^2})$

since ϵ is random, also y is random with density $p(y|x, \Theta) - \frac{1}{\sqrt{2\pi r}} \exp(-\frac{\|y - \Theta^T x\|^2}{2r^2})$

To specify the model we have to estimate $\Theta \in \mathbb{R}^{n+1}$ from the data

idea: choose Θ that maximizes the likelihood

likelihood function:

$$L(\Theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}; \Theta)$$

the product form means: the observation $(x^{(1)}, y^{(1)}), \dots, (x^{(i)}, y^{(i)})$ are independent of each other

estimate:

$$\Theta_{ML} = \overset{argmax}{\Theta \in \mathbb{R}^{n+1}} L(\Theta)$$
$$= \overset{argmax}{\Theta \in \mathbb{R}^{n+1}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \Theta^T x^{(i)})^2}{2\sigma^2}\right)$$

since we are only interested in the positoin where the maximum is attained we can apply a monoton transformation to $L(\Theta)$ with changing this position

$\Rightarrow$ log-likelihood function: $l(\Theta) = \log L(\Theta)$

$$= \Theta_{ML} = \overset{argmax}{\Theta \in \mathbb{R}^{n+1}}\ l(\Theta)$$

$$= \overset{argmax}{\Theta \in \mathbb{R}^{n+1}}\ \Sigma_{i=1}^{m} - \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma}(y^{(i)} - \Theta^T x^{(i)})^2$$

$$\overset{argmax}{\Theta \in \mathbb{R}^{n+1}}\ -m\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\Sigma_{i=1}^{m}(y^{(i)} - \Theta^T x^{(i)})^2$$

does not depend on Θ scaling vector does not influence the optimal Θ

$$\overset{argmax}{\Theta \in \mathbb{R}^{n+1}}\ -\frac{1}{2}\Sigma_{i=1}^{m}(y^{(i)} - \Theta x^{(i)})^2$$

X: data matrix
Θ: parameter vector
Y: response vector

$$\overset{argmax}{\Theta \in \mathbb{R}^{n+1}}\ -\frac{1}{2}\|X * \Theta - Y\|_2^2$$

$$\overset{argmax}{\Theta \in \mathbb{R}^{n+1}}\ \frac{1}{2}\|X * \Theta - Y\|_2^2$$

$L(\Theta)$ loss function
Minimizing the loss function that we discussed already

remark

going non-linear $x \in \mathbb{R}, y \in \mathbb{R}\quad y = t(x)$

observations: $(x^{(1)}, y^{(1)}), \ldots, (x^{(i)}, y^{(i)}) \in \mathbb{R} \times \mathbb{R}$ but $f(.)$ not neccessarily linear function

$((x^{(1)}, x^{(1)^2}, x^{(1)^3}), y^{(1)}), \ldots, ((x^{(m)}, x^{(m)^2}, x^{(m)^3}), y^{(m)})$
apply linear regression to argumented data points:

$\Rightarrow y = \Theta_0 + \Theta_1 x + \Theta_2 x^2 + \Theta_3 x^3$

linear regression gives „good estimates“ for $\Theta_0, \Theta_1, \Theta_2, \Theta_3$

overfitting problem!

logistic regression for binary classification

data/observations: $(x^{(1)}, y^{(1)}), \ldots, (x^{(i)}, y^{(i)})$

$$x^{(i)} \in \mathbb{R}^n \quad \text{covariates}$$
$$y^{(i)} \in 0,1 \quad \text{variates/response}$$

probabilistic model of logistic regression

$$P[y=1|x;\Theta] = h_\Theta(x) \in (0,1)$$

$$P[y=0|x;\Theta] = 1 - h_\Theta(x)$$

$h_\Theta(x) = g(\Theta^T x)$, where $g(.)$ is the logistic function $g(z) = \frac{1}{1+\exp(-z)}$

goal(as in linear regression): estimate $\Theta \in \mathbb{R}^n$ (Parameter vector) from data.
likelihood function for parameter vector Θ:

$$L(\Theta) = \prod_{i=1}^{m} P[y^{(i)}|x^{(i)};\Theta]$$

again, assumption of independent observation

$$= \prod_{i=1}^{m} h_\Theta(x^{(i)})^{y^{(i)}} (1 - h_\Theta(x^{(i)}))^{1-y^{(i)}}$$

$$= \begin{cases} 1 & \text{if } y^{(i)} = 0 \\ h_\Theta(x^{(i)}) & \text{if } y^{(i)} = 1 \end{cases} = \begin{cases} 1 & \text{if } y^{(i)} = 1 \\ 1 - h_\Theta(x^{(i)}) & \text{if } y^{(i)} = 0 \end{cases}$$
$$h_\Theta := P[y^{(i)} = 1|x^{(i)};\Theta] \quad 1 - h_\Theta(x^{(i)}) := P[y^{(i)} = 0|x^{(i)};\Theta]$$

instead of working with the likelihood function it is easier to work with the log-likelihood function:

$$\Theta_{ML} = \overset{argmax}{\Theta \in \mathbb{R}^n}\ L(\Theta) = \overset{argmax}{\Theta \in \mathbb{R}^n} \log \underbrace{L(\Theta)}_{l(\Theta)}$$

$$= \overset{argmax}{\Theta \in \mathbb{R}^n}\ \Sigma_{i=1}^{m} y^{(i)} \log h_\Theta(x^{(i)}) + \underbrace{(1 - y^{(i)}) \log(1 - h_\Theta(x^{(i)}))}_{\text{log likelihood function}}$$

neccessary for optimum is a vanishing gradient
$\bigtriangledown_\Theta l(\Theta) \overset{!}{=} 0$
for computing the gradient:

$$\frac{d}{dz} g(z) = \frac{d}{dz} \frac{1}{1+\exp(-z)}$$

$$= \frac{\exp(-z)}{(1+\exp(-z))^2}$$

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

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

$$= \boxed{\text{g(z)(1-g(z))}}$$

$$\nabla_\Theta l(\Theta) = (\frac{\delta}{\delta\Theta_1}l(\Theta), \ldots, \frac{\delta}{\delta\Theta_1}l(\Theta))$$

$$\frac{\delta}{\delta\Theta_j}l(\Theta) = \frac{\delta}{\delta\Theta_j}\Sigma_{i=1}^m y^{(i)}\log h_\Theta(x^{(i)}) + (1-y^{(i)})\log(1-h_\Theta(x^{(i)}))$$

$$= \Sigma_{i=1}^m \left(\frac{y^{(i)}}{h_\Theta(x^{(i)})} - \frac{1-y^{(i)}}{1-h_\Theta(x^{(i)})}\right)\frac{\delta}{\delta\Theta_j}\underbrace{h_\Theta(x^{(i)})}_{g(\Theta^T x^{(i)})}$$

$$= \Sigma_{i=1}^m \left(\frac{y^{(i)}}{h_\Theta(x^{(i)})} - \frac{1-y^{(i)}}{1-h_\Theta(x^{(i)})}\right)$$

$$h_\Theta(x^{(i)})(1-h_\Theta(x^{(i)})) \qquad \underbrace{x_j^{(i)}}_{\text{j-th component i-th data vetor}}$$

$$= \Sigma_{i=1}^m y^{(i)}(1-h_\Theta(x^{(i)})) - (1-y^{(i)})h_\Theta(x^{(i)})x_j^{(i)}$$

$$= \Sigma_{i=1}^m (y^{(i)} - h_\Theta(x^{(i)}))x_j^{(i)}$$

Unfortunetaly the system of equations $\frac{\delta}{\delta\Theta_j}l(\Theta) \stackrel{!}{=} 0$ is highly non-linear and thus difficult to solve!

$$= \Sigma_{i=1}^m (y^{(i)} - h_\Theta(x^{(i)}))x_j^{(i)} = 0$$

turn to a numerical scheme(gradient ascend):

initialize: $\Theta^{(0)}$ arbitrary with some vector in $\mathbb{R}^n$
repeat

for $i = 1$ to m

$$\Theta_j^{(k)} = \Theta_j^{(k-1)} + \underbrace{\alpha}_{learningrate} \Sigma_{i=1}^m (y^{(i)} - h_{\Theta^{(k)}}x^{(i)})x_j^{(i)}$$

end for

until convergence

exercise 1 (08.04.2016)

(1)

$$L(\Theta) = \frac{1}{2}\|X\Theta - Y\|_2^2 + \gamma\|\Theta\|_2^2 = \frac{1}{2}\Sigma_{i=1}^m(\underbrace{x^{(i)}\Theta}_{=\Sigma_{j=0}^n x_j^{(i)}\Theta_j} - y^{(i)})^2 + \Sigma_{j=0}^n \Theta_j^2$$

$$= \frac{1}{2}\Sigma_{i=1}^m((\Sigma_{j=0}^n x_j^{(i)}\Theta_j) - y^{(i)})^2 + \gamma\Sigma_{j=0}^n\Theta_j^2$$

$$\frac{\delta}{\delta\Theta_l}(\dots) = \Sigma_{i=1}^m((\Sigma_{j=0}^n x_j^{(i)}\Theta_j) - y^{(i)})x_l^{(i)} + 2\gamma\Theta_l$$

$$\nabla_\Theta L(\Theta) = \underbrace{\frac{1}{2}\nabla_\Theta\|X\Theta - Y\|_2^2}_{=X^TX\Theta - X^TY} + \underbrace{\gamma\nabla_\Theta\|\Theta\|_2^2}_{=2\gamma\Theta}$$

$$= X^TX\Theta - X^TY + 2\gamma\Theta \overset{!}{=} \Theta$$

$$\Rightarrow (X^TX + 2\gamma\mathbb{1})\Theta = X^TY$$

$$\Rightarrow \Theta_{ridge} = (X^TX + 2\gamma\mathbb{1})^{-1}X^TY$$

$$\|X\Theta - Y\|_2^2 = (X\Theta - Y)^T(X\Theta - Y)$$

$$= \Theta^TX^TX\Theta - \underbrace{\Theta^TX^TY}_{(X\Theta)^TY} - \underbrace{Y^TX\Theta}_{Y^T(X\Theta)\Rightarrow -2(X\Theta)^TY = 2\Theta^TX^TY = (X\Theta)^TY} + Y^TY$$

(2) Bayes:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

here: generative $\underbrace{\text{model}}_{\text{has parameters: } \phi, \Sigma, \mu_0, \mu_1}$ that specifies how y abd x are generated

goal: estimate parameters from data $\rightarrow$ user likelihood function to do so

$$L(\phi, \Sigma, \mu_0, \mu_1) = \prod_{i=1}^m p(x^{(i)}, y^{(i)}; \phi, \Sigma, \mu_0, \mu_1)$$

$$\prod_{i=1}^m p(y^{(i)}, \phi)p(x^{(i)}; \Sigma, \mu_0)^{1-y^{(i)}}p(x^{(i)}; \Sigma, \mu_1)^{y^{(i)}}$$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \quad \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

density of 1-dim normal distribution $x \in \mathbb{R}$

$$p(x) = \frac{1}{\underbrace{(2\pi)^{\frac{n}{2}}}_{\sqrt{\det(\epsilon)}}} \underbrace{|\epsilon|^{\frac{n}{2}}}_{\text{multi. var. extension of normal dist.}} \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$$

density of multi-variate distribution $x \in \mathbb{R}^n$

parameters: $\underbrace{\mu \in \mathbb{R}^n}_{mem.vector}, \Sigma \in \mathbb{R}^{n\times n}$
symmetric and positive definite (covariance matrix)

a matrix $\Sigma \in \mathbb{R}^{n\times n}$ is symmetric, if $\Sigma_{ij} = \Sigma_{ji} \forall i < j$
example:

$$\begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix}$$

a symmetric matrix is called positive definite, if $x^T\Sigma x > 0$ for $x \neq 0$

example: data matrix $X \in \mathbb{R}^{m\times(n+1)}$

$$X^TX \in \mathbb{R}^{(n+1)\times(n+1)} (X^TX)_{ij} = x^{(i)^T}x^{(j)}$$
$$= x^{(j)^T}x^{(i)}$$
$$v^T(X^TX) = (Xv)^T(Xv) = \|Xv\|_2^2 \underbrace{\geq}_{\text{equality also allowed}} 0 \quad \text{positive semi-definite}$$

loss function, eg. for logistic regression $L(\Theta)$

$$\Theta \in \mathbb{R}^{n+1} \to L(\Theta) \in \mathbb{R}, \text{that is, } L : \mathbb{R}^{n+1} \to \mathbb{R}$$

to get optimal Θ_i neccessary
$\nabla_\Theta L(\Theta) \stackrel{!}{=} 0$
vector with $(n+1)$ entries, we require that every entry is zero!

lecture 3 (11.04.2016)

differentiability and convexiability

Differentiability: function $t : \mathbb{R}^n \to \mathbb{R}, x \mapsto t(x)$
t is differentiable at $x \in \mathbb{R}^n$ if t can approximated well in x by a linear function.

$$\exists t'(x) : t(y) = t(x) + t'(x)^T(y-x) + o(\|y-x\|)$$

$$\in \mathbb{R}^n \quad \text{little o-notation } r \xrightarrow{lim} 0 \ \tfrac{1}{r}o(r) = 0$$

not always possible:

the gradient $t'(x)$ in coordinates

using the definition of differentiability we can plug in special values for y.

$$y^{(i)}(t) = x + te_i$$

i-th standard basis vector $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T, o(0) = 0$ (the 1 is position i)

by differentiability we have:

$$\begin{aligned}
t(y^{(i)}(t)) &= t(x) + t'(x)^T y^{(i)}(t) + o(\|y^{(i)}(t) - x\|) \\
&= t(x) + tt'(x)^T e_i + o(t) \\
\Rightarrow t(y^{(i)}(t)) &- t(x) - tt'(x)^T e_i = o(t) \\
\Rightarrow \frac{t(y^{(i)}(t))}{} &- t'(x)^T e_i = \frac{1}{t} o(t) \\
\Rightarrow t \overset{lim}{\to} 0 \ \frac{t(y^{(i)}(t))}{t} &= t'(x)^T e_i
\end{aligned}$$

$= \frac{\delta t}{\delta x_i}(x)$ partial derivative i-th component of the vector $t'(x)$ (gradient)

$$\Rightarrow t'(x) = (\frac{\delta t}{\delta x_1}, \ldots, \frac{\delta t}{\delta < n}(x))$$

generalization to functions $t : \mathbb{R}^n \to \mathbb{R}^m$
t is called differentiable in $x \in \mathbb{R}^n$ if $\exists t'(x) \in \mathbb{R}^{m \times n} sit \forall y \in \mathbb{R} : t(y) :$

$$\underbrace{t(x) + t'(x)(y - x)}_{\text{linear approx.}} + \underbrace{(|y - x|)}_{\text{vector little o-notation}}$$

here $t'(x)$ is called Jacobi-Matrix

in coordinates:

$$t'(x) = \begin{pmatrix} \frac{\delta t_1}{\delta x_1} \cdots \frac{\delta t_1}{\delta x_n} \\ \vdots \qquad \vdots \\ \frac{\delta t_m}{\delta x_1} \cdots \frac{\delta t_m}{\delta x_n} \end{pmatrix}$$

$$\lim_{v \to 0} \frac{1}{\|v\|} \quad o(v) = o \in \mathbb{R}^m$$

$$o(\underbrace{0}_{\in \mathbb{R}^m}) = 0$$

special case: assume, that $t : \mathbb{R}^n \to \mathbb{R}$ is differentiable in every $x \in \mathbb{R}^n$. That means $t'(x)$ exists for every $x \in \mathbb{R}^n$! Use this to define a new function:

$$t' : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto t'(x)$$

in coordiantes:

$$t'(x) = \begin{pmatrix} \frac{\delta^2 t}{\delta x_1 \delta_{x_1}}(x) \dots \frac{\delta^2 t}{\delta x_n \delta_{x_1}}(x) \\ \vdots \quad\quad \vdots \\ \frac{\delta^2 t}{\delta x_1 \delta_{x_n}}(x) \dots \frac{\delta^2 t}{\delta x_n \delta_{x_n}}(x) \end{pmatrix}$$

convex functions

a subset $K \subset \mathbb{R}^n$ is called convex, if every $p, q \in K$ also $\lambda p + (1-\lambda)q \in K$ for $\lambda \in [0,1]$

a function $t : \mathbb{R}^n \to \mathbb{R}$ is called convex, if the epi-graph of the function

$$epi(t) = \{(x, y) \in \mathbb{R}^{n+1} | y \geq f(x)\}$$

is a convex set

alternative characterization of convexity of functions:
$t : \mathbb{R}^n \to \mathbb{R}$ is convex, if for all $x, y \in \mathbb{R}^n$:

$$t(\lambda x + (1-\lambda)y) \leq \lambda t(x) + (1-\lambda) + t(x)$$

lemma: a differentiable function $t : \mathbb{R} \to \mathbb{R}$ is convex if and only if $t' : \mathbb{R} \to \mathbb{R}$ $(x \mapsto t'(x))$ is increasing

that is, it is enough to look at point $p, q \in \mathbb{R}^{n+1}$ that we on the boundary of the epi-graph of t

corrollary: a twice differentiable function is convex if its second derivative is always non-negative.
follows from: a differential function $t : \mathbb{R} \to \mathbb{R}$ is increasing if its derivative is non-negative

theorem: a twice differential function $t : \mathbb{R}^n \to \mathbb{R}$ is convex, if and only if the hessian os positive semi-definite.

exercise 2 (15.04.2016)

(1) likelihood Funktion: $L(\Theta) = \prod_{i=1}^m h_\Theta(x^{(i)})(1 - h_\Theta(x^{(i)}))^{1-y^{(i)}}$
$h_\Theta(x) = \frac{1}{1+\exp(-\Theta^T x)}$
gesucht: $\Theta^k = \overset{argmax}{\Theta \in} \mathbb{R}^n \; l(\Theta) = \overset{argmax}{\Theta \in} \mathbb{R}^n \log L(\Theta) = \overset{argmax}{\Theta \in} \mathbb{R}^n \; \Sigma_{i=1}^m y^{(i)} \log h_\Theta(x^{(i)}) + (1 -$

$y^{(i)}) \log(1 - h_\Theta(x^{(i)}))$
Ziel heute: Zeige, dass $l(\Theta)$ konkave Funktion ist, d. h.

$$l(\alpha\Theta_1 + (1-\alpha)\Theta_2) \geq \alpha l(\Theta_1) + (1-\alpha) l(\Theta_2)$$

Es reicht zu zeigen, dass die Hesse-Matrix von $l(\Theta)$ negativ semidefinit ist.

$$g(\epsilon)\frac{1}{1+\exp(-\epsilon)} \in (0,1) \qquad \text{logistische Funktion}$$

$$\frac{d_g(t)}{dt} = \frac{1}{(1+\exp(-t))}^2 - \exp(-t) = \frac{1}{1+\exp(-t)}\left(1 - \frac{1}{1+\exp(t)}\right)$$

$$g(t)(1-g(t))$$

$$\Rightarrow \frac{\delta h_\Theta(x)}{\delta\Theta_j} = h_\Theta(x)(1-h_\Theta(x))\frac{\delta}{\delta\Theta_j}\Theta^T x$$

$$= h_\Theta(x)(1 - H_\Theta(x))x_j$$

Kettenregel

$$h_\Theta(x) = g(\Theta^T x)$$

$$\underbrace{x}_{\in\mathbb{R}^n} \underbrace{\mapsto}_{\text{innere Fkt.}} \underbrace{\Theta^T}_{\in\mathbb{R}^n x} \underbrace{\mapsto}_{\text{äußere Fkt.}} \underbrace{g(\Theta^T x)}_{\in\mathbb{R}^n}$$

$$\frac{\delta}{\delta\Theta_j}\Theta^T x = \frac{\delta}{\delta\Theta_j}(\Sigma_{i=0}^m \Theta_i x_i) = \frac{\delta}{\delta\Theta_j}(\Theta_1 x_1 + \Theta_1 x_1 + \cdots + \Theta_n x_n) = x_j$$

Wir wissen schon:

$$\frac{\delta l(\Theta)}{\delta\Theta_j} = \Sigma_{i=1}^m(\underbrace{y^{(i)}}_{\in\{0,1\}} - \underbrace{\Theta(x^{(i)}))}_{\in(0,1)} x_j^{(i)}$$

$$\frac{\delta^2 l(\Theta)}{\delta\Theta_k\delta\Theta_j} = \frac{\delta}{\delta\Theta_k}\Sigma_{i=1}^m(y^{(i)} - h_\Theta(x^{(i)}))x_j^{(i)}$$

$$= \underbrace{\frac{\delta}{\delta\Theta_k}(\Sigma_{i=1}^m y^{(i)} x_j^{(i)})}_{=0,\text{hängt nicht von }\Theta\text{ ab}} - \frac{\delta}{\delta\Theta_k}(\Sigma_{i=1}^m h_\Theta(x^{(i)})x_j^{(i)})$$

$$= \frac{\delta}{\delta\Theta_k}\Sigma_{i=1}^m h_\Theta(x^{(i)})x_j^{(i)}$$

$$= -\Sigma_{i=1}^m \frac{\delta}{\delta\Theta_k}(h_\Theta(x^{(i)})x_j^{(i)})$$

$$= -\Sigma_{i=1}^{m} x_j^{(i)} \frac{\delta}{\delta\Theta_k} h_\Theta(x^{(i)})$$

$$= -\Sigma_{i=1}^{m} \underbrace{h_\Theta(x^{(i)})}_{>0} \underbrace{(1 - h_\Theta(x^{(i)}))}_{>0} x_j^{(i)} x_k^{(i)}$$

Wir müssen zeigen, dass $v^T \quad H(\Theta)v \geq 0$ für alle $v \in \mathbb{R}^n$

$$\underbrace{H(\Theta)}_{\in\mathbb{R}^{n\times n}} = (H_{kj}(\Theta)) = \left(\frac{\delta^2 l(\Theta)}{\delta\Theta_k \delta\Theta_j}\right)$$