# Maschinelles Lernen

Johannes Eifler und co
Matrikel: 144915
Stand: 18. April 2016 19:14

## 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: $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$

$x^{(i)}$ 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: $y = \Theta_0 + \Theta_1 X_1 + \cdots + \Theta_n X_n$
$= \Theta^T X, X = (1, X_1, \ldots, X_n) \in \mathbb{R}^{n+1}$

problem: estimate the parameter vector $\Theta in \mathbb{R}^{n+1}$ from the measurements $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$
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 $\Theta$

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)

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

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

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

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

$$\nabla_\Theta L(\Theta) \overset{!}{=} 0 \qquad\qquad t(x) = \frac{1}{2}x^2 + ax + b$$
$$\nabla_\Theta L(\Theta) = X^T X \Theta * X^T Y \overset{!}{=} 0 \qquad \nabla_x t(x) = x + a$$

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

$$\Rightarrow X^T X \Theta = X^T Y \qquad\qquad t(\Theta) = \Theta^T X \Theta$$
$$\Rightarrow \Theta = (X^T X)^{-1} X^T Y \qquad\qquad \nabla_\Theta t(\Theta) = (X + X^T)\Theta$$
privided that $(X^T X)^{-1}$ exists

${\color{red}(X^T X)_{ij} = X^{(i)^T} X^{(j)}}$
${\color{red}\text{operation matrix}}$
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:

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

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

$\qquad\qquad\qquad\qquad {\color{red}\text{loss term}} \quad {\color{blue}\text{reularisation term}}$

probabilistic interpretation of least squares

$$Y = {\color{red}\Theta^T * X} + {\color{blue}\epsilon}$$

$\qquad\qquad {\color{red}\text{deterministic part}} \quad {\color{blue}\text{random/noise part}}$

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

assumption:

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

linear regression $\boxed{y = \Theta^T x}$ $\Theta \in \mathbb{R}^{n+1}$ (parameter vector)
$$\Sigma_{i=0}^n \Theta_i x_i$$
$x = (1, x_1, \ldots, x_n) \in \mathbb{R}^{n+1} \qquad x_0 = 1$

(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 $\epsilon$ 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 $\Theta$ that maximizes the likelihood

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

the product form means: the observation $(x^{(1)}, y^{(1)}), \ldots, (x^{(i)}, y^{(i)})$ 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: $l(\Theta) = \log L(\Theta)$

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

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

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

<span style="color:red">does not depend on $\Theta$</span>   <span style="color:blue">scaling vector does not influence the optimal $\Theta$</span>

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

$X$: data matrix
$\Theta$: parameter vector
$Y$: response vector

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

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

<span style="color:red">$L(\Theta)$ loss function</span>
Minimizing the loss function that we discussed already

## remark

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

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 $\Theta$:

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

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:

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

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

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

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

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

$$\bigtriangledown_\Theta l(\Theta) = \left( \frac{\delta}{\delta\Theta_1} l(\Theta), \ldots, \frac{\delta}{\delta\Theta_1} l(\Theta) \right)$$

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

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

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

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

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

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

Unfortunetaly the system of equations $\frac{\delta}{\delta\Theta_j} l(\Theta) \overset{!}{=} 0$ 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

# lecture 3

## 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.

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

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

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:

$$t(y^{(i)}(t)) = t(x) + t'(x)^T y^{(i)}(t) + o(\|y^{(i)}(t) - x\|)$$

$$= t(x) + tt'(x)^T e_i + o(t)$$

$$\Rightarrow t(y^{(i)}(t)) - t(x) - tt'(x)^T e_i = o(t)$$

$$\Rightarrow \frac{t(y^{(i)}(t))}{\rule{2cm}{0.4pt}} - t'(x)^T e_i = \frac{1}{t}o(t)$$

$$\Rightarrow t \overset{lim}{\to} 0 \; \frac{t(y^{(i)}(t))}{t} = t'(x)^T e_i$$

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

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

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:

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