

Maschinelles Lernen

Johannes Eifler und co
Matrikel: 144915
Stand: 15. April 2016 17:23

lecture 1 (04.04.2016)

Introduction

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

Dependent variables are measured at independent variables (covariates). Variables are measured on some **scale**:

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

⇒ quotients make sense on ratio scale; quotients 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 multivariate case: $y = \Theta_0 + \Theta_1 X_1 + \dots + \Theta_n X_n$
 $= \Theta^T X, X = (1, X_1, \dots, X_n) \in \mathbb{R}^{n+1}$

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

loss function: $L(\Theta) = \frac{1}{2} \sum_{i=1}^m (\Theta^T X^{(i)} - y^{(i)})^2$

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:

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

(vector of predictions vector of observation response)

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

(definition of the euclidian norm)

$$\begin{aligned}
 &= \frac{1}{2}(\Theta^T X^T \times \Theta - \Theta^T X^T Y - Y^T X * \Theta + Y^T Y) \\
 &= -2\Theta^T X^T Y \text{ since the dot product is symmetric } (X^T Y = Y^T X) \\
 &= \frac{1}{2}\Theta^T X^T X \Theta - \Theta^T X^T Y + \frac{1}{2}Y^T Y
 \end{aligned}$$

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

$$\begin{aligned}
 \nabla_{\Theta} L(\Theta) &\stackrel{!}{=} 0 & t(x) &= \frac{1}{2}x^2 + ax + b \\
 \nabla_{\Theta} L(\Theta) &= X^T X \Theta * X^T Y \stackrel{!}{=} 0 & \nabla_x t(x) &= x + a
 \end{aligned}$$

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

$$\begin{aligned}
 \Rightarrow X^T X \Theta &= X^T Y & t(\Theta) &= \Theta^T X \Theta \\
 \Rightarrow \Theta &= (X^T X)^{-1} X^T Y & \nabla_{\Theta} t(\Theta) &= (X + X^T) \Theta
 \end{aligned}$$

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

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

operation matrix

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

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

more robust solution:

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

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?

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

loss term reularisation term

probabilistic interpretation of least squares

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

deterministic part random/noise part

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

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

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:

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

assumption:

- (1) $y = t(x)$ y is function of x
 linear regression $\boxed{y = \Theta^T x}$ $\Theta \in \mathbb{R}^{n+1}$ (parameter vector)
 $\sum_{i=0}^n \Theta_i x_i$
 $x = (1, x_1, \dots, x_n) \in \mathbb{R}^{n+1}$ $x_0 = 1$

- (2) data are obscured by random noise:

$$y = \Theta^T x + \epsilon, \quad \epsilon = \text{random noise term}$$

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

since ϵ is random, also y is random with density $p(y|x, \Theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\|y - \Theta^T x\|_2^2}{2\sigma^2})$

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:

$$\Theta_{ML} = \underset{\Theta \in \mathbb{R}^{n+1}}{\operatorname{argmax}} L(\Theta)$$

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

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

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

$$\begin{aligned} &= \underset{\Theta \in \mathbb{R}^{n+1}}{\operatorname{argmax}} l(\Theta) \\ &= \underset{\Theta \in \mathbb{R}^{n+1}}{\operatorname{argmax}} \sum_{i=1}^m -\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} (y^{(i)} - \Theta^T x^{(i)})^2 \\ &= \underset{\Theta \in \mathbb{R}^{n+1}}{\operatorname{argmax}} -m \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^m (y^{(i)} - \Theta^T x^{(i)})^2 \end{aligned}$$

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

$$\underset{\Theta \in \mathbb{R}^{n+1}}{\operatorname{argmax}} -\frac{1}{2} \sum_{i=1}^m (y^{(i)} - \Theta^T x^{(i)})^2$$

X : data matrix

Θ : parameter vector

Y : response vector

$$\begin{aligned} &\underset{\Theta \in \mathbb{R}^{n+1}}{\operatorname{argmax}} -\frac{1}{2} \|X * \Theta - Y\|_2^2 \\ &\underset{\Theta \in \mathbb{R}^{n+1}}{\operatorname{argmax}} \frac{1}{2} \|X * \Theta - Y\|_2^2 \end{aligned}$$

$L(\Theta)$ 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: $(x^{(1)}, y^{(1)}), \dots, (x^{(i)}, y^{(i)}) \in \mathbb{R} \times \mathbb{R}$ but $f(\cdot)$ not necessarily linear function

$$((x^{(1)}, x^{(1)^2}, x^{(1)^3}), y^{(1)}), \dots, ((x^{(m)}, x^{(m)^2}, x^{(m)^3}), y^{(m)})$$

apply linear regression to augmented 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: $(x^{(1)}, y^{(1)}), \dots, (x^{(i)}, y^{(i)})$

$$\begin{aligned} x^{(i)} &\in \mathbb{R}^n \quad \text{covariates} \\ y^{(i)} &\in \{0, 1\} \quad \text{variables/response} \end{aligned}$$

probabilistic model of logistic regression

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

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

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

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

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

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

$$\begin{aligned} \Theta_{ML} &= \underset{\Theta \in \mathbb{R}^n}{\operatorname{argmax}} L(\Theta) = \underset{\Theta \in \mathbb{R}^n}{\operatorname{argmax}} \underbrace{\log L(\Theta)}_{l(\Theta)} \\ &= \underset{\Theta \in \mathbb{R}^n}{\operatorname{argmax}} \sum_{i=1}^m y^{(i)} \log h_{\Theta}(x^{(i)}) + \underbrace{(1 - y^{(i)}) \log(1 - h_{\Theta}(x^{(i)}))}_{\text{log likelihood function}} \end{aligned}$$

necessary for optimum is a vanishing gradient

$$\nabla_{\Theta} l(\Theta) \stackrel{!}{=} 0$$

for computing the gradient:

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

$$\begin{aligned}
&= \frac{\exp(-z)}{(1 + \exp(-z))^2} \\
&= \frac{1}{1 + \exp(-z)} \left(\frac{1 + \exp(-z) - 1}{1 + \exp(-z)} \right) \\
&= \frac{1}{1 + \exp(-z)} \left(1 - \frac{1}{1 + \exp(-z)} \right) \\
&= \boxed{g(z)(1-g(z))}
\end{aligned}$$

$$\begin{aligned}
\nabla_{\Theta} l(\Theta) &= \left(\frac{\delta}{\delta \Theta_1} l(\Theta), \dots, \frac{\delta}{\delta \Theta_1} l(\Theta) \right) \\
\frac{\delta}{\delta \Theta_j} l(\Theta) &= \frac{\delta}{\delta \Theta_j} \sum_{i=1}^m y^{(i)} \log h_{\Theta}(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_{\Theta}(x^{(i)})) \\
&= \sum_{i=1}^m \left(\frac{y^{(i)}}{h_{\Theta}(x^{(i)})} - \frac{1 - y^{(i)}}{1 - h_{\Theta}(x^{(i)})} \right) \frac{\delta}{\delta \Theta_j} \underbrace{h_{\Theta}(x^{(i)})}_{g(\Theta^T x^{(i)})} \\
&= \sum_{i=1}^m \left(\frac{y^{(i)}}{h_{\Theta}(x^{(i)})} - \frac{1 - y^{(i)}}{1 - h_{\Theta}(x^{(i)})} \right) \\
&\quad h_{\Theta}(x^{(i)})(1 - h_{\Theta}(x^{(i)})) \underbrace{x_j^{(i)}}_{\text{j-th component i-th data vector}} \\
&= \sum_{i=1}^m y^{(i)}(1 - h_{\Theta}(x^{(i)})) - (1 - y^{(i)})h_{\Theta}(x^{(i)})x_j^{(i)} \\
&= \sum_{i=1}^m (y^{(i)} - h_{\Theta}(x^{(i)}))x_j^{(i)}
\end{aligned}$$

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

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

turn to a numerical scheme (gradient ascend):

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

repeat

$$\begin{aligned}
&\text{for } i = 1 \text{ to } m \\
&\quad \Theta_j^{(k)} = \Theta_j^{(k-1)} + \underbrace{\alpha}_{\text{learningrate}} \sum_{i=1}^m (y^{(i)} - h_{\Theta^{(k)}}(x^{(i)}))x_j^{(i)}
\end{aligned}$$

end for

until convergence

lecture 3

differentiability and convexiability

Differentiability: function $t : \mathbb{R}^n \rightarrow \mathbb{R}, x \mapsto t(x)$

t is differentiable at $x \in \mathbb{R}^n$ if t can be approximated well in x by a linear function.

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

$$\in \mathbb{R}^n \quad \text{little o-notation} \quad \lim_{r \rightarrow 0} \frac{1}{r} o(r) = 0$$

not always possible:

the gradient $t'(x)$ in coordinates

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

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

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

by differentiability we have:

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

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

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

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

$$\Rightarrow \lim_{t \rightarrow 0} \frac{t(y^{(i)}(t))}{t} = t'(x)^T e_i$$

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

$$\Rightarrow t'(x) = \left(\frac{\delta t}{\delta x_1}, \dots, \frac{\delta t}{\delta x_n}(x) \right)$$

generalization to functions $t : \mathbb{R}^n \rightarrow \mathbb{R}^m$

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

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

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

in coordinates:

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