Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Machine Learning 1

Winter semester 2016/17

Group APXNLE

Exercise 2

Members

Jing Li[387272] jing.li.1@campus.tu-berlin.de

Kumar Awanish[386697] k.awanish@campus.tu-berlin.de

Manjiao Xu[386498] manjiao.xu@campus.tu-berlin.de

Rudresha Gulaganjihalli Parameshappa[386642]

Gulaganjihalliparameshappa@ campus.tu-berlin.de

Sonali Nayak[386995] sonali.nayak@campus.tu-berlin.de

Maximilian Ernst[364862] maximilian.ernst@campus.tu-berlin.de

Exercise 1: Maximum-Likelihood Estimation

1. (a). $p(x) = \int_{-\infty}^{+\infty} p(x,y)\,dy$ (supported on $\mathbb{R}_+^2$)

$= \int_0^{+\infty} \lambda\eta e^{-\lambda x-\eta y}\,dy$

$= \lambda\eta \int_0^{+\infty} e^{-\lambda x} * e^{-\eta y}\,dy$

$= \lambda\eta e^{-\lambda x} \int_0^{+\infty} e^{-\eta y}\,dy$

$= \lambda\eta e^{-\lambda x} * (-\frac{1}{\eta} e^{-\eta y}\Big|_0^{+\infty})$

$= \lambda\eta e^{-\lambda x} * [-\frac{1}{\eta}(0-1)]$

$= \lambda\eta e^{-\lambda x} \cdot \frac{1}{\eta}$

$= \lambda e^{-\lambda x}$

$p(y) = \int_{-\infty}^{+\infty} p(x,y)\,dx$

$= \int_0^{+\infty} \lambda\eta e^{-\lambda x-\eta y}\,dx$

$= \lambda\eta \int_0^{+\infty} e^{-\lambda x} * e^{-\eta y}\,dx$

$= \lambda\eta e^{-\eta y} \int_0^{+\infty} e^{-\lambda x}\,dx$

$= \lambda\eta e^{-\eta y} * (-\frac{1}{\lambda} e^{-\lambda x}\Big|_0^{+\infty})$

$= \lambda\eta e^{-\eta y} * [-\frac{1}{\lambda}(0-1)]$

$= \lambda\eta e^{-\eta y} \cdot \frac{1}{\lambda}$

$= \eta e^{-\eta y}$

$\therefore p(x) * p(y) = \lambda e^{-\lambda x} * \eta e^{-\eta y}$

$= \lambda\eta e^{-\lambda x-\eta y}$

$= p(x,y)$

$\therefore$ x and y are independent.

(b) $L(\lambda) = \prod_{i=1}^{n} p(x_i, y_i)$

$= \prod_{i=1}^{n} \lambda\eta e^{-\lambda x_i - \eta y_i}$

$= \lambda^n \eta^n e^{-\lambda \sum_{i=1}^{n} x_i - \eta \sum_{i=1}^{n} y_i}$

$\ln L(\lambda) = n\ln\lambda + n\ln\eta - \lambda\sum_{i=1}^{n} x_i - \eta\sum_{i=1}^{n} y_i$

$\frac{d\ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} + 0 - \sum_{i=1}^{n} x_i - 0$

$= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$

$\frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i = 0$. according to $\sum_{i=1}^{n} x_i = n\bar{x}$

$\frac{n}{\hat{\lambda}} - n\bar{x} = 0$

$\hat{\lambda} = \frac{1}{\bar{x}}$

(c) $\eta = \frac{1}{\lambda}$

$L(\lambda) = \prod_{i=1}^{n} p(x_i, y_i)$

$= \prod_{i=1}^{n} \lambda \cdot \frac{1}{\lambda} \cdot e^{-\lambda x_i - \frac{1}{\lambda} y_i}$

$= e^{-\lambda\sum_{i=1}^{n} x_i - \frac{1}{\lambda}\sum_{i=1}^{n} y_i}$

$\ln L(\lambda) = -\lambda\sum_{i=1}^{n} x_i - \frac{1}{\lambda}\sum_{i=1}^{n} y_i$

$\frac{d\ln L(\lambda)}{d\lambda} = -\sum_{i=1}^{n} x_i - (-\frac{1}{\lambda^2}\sum_{i=1}^{n} y_i)$

$= \frac{1}{\lambda^2}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i$

$\frac{1}{\hat{\lambda}^2}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i = 0$

$\frac{n}{\hat{\lambda}^2}\bar{y} - n\bar{x} = 0$

$\frac{\bar{y}}{\hat{\lambda}^2} = \bar{x}$

$\hat{\lambda}^2 = \bar{y}/\bar{x}$

$\hat{\lambda} = \pm\sqrt{\bar{y}/\bar{x}}$

we have $\lambda > 0$

so $\hat{\lambda} = \sqrt{\bar{y}/\bar{x}}$

(d). $\eta = 1-\lambda$

$$L(\lambda) = \prod_{i=1}^{n} p(x_i, y_i)$$

$$= \prod_{i=1}^{n} \lambda(1-\lambda)\, e^{-\lambda x_i - (1-\lambda) y_i}$$

$$= \lambda^n (1-\lambda)^n e^{-\lambda \sum_{i=1}^{n} x_i - (1-\lambda)\sum_{i=1}^{n} y_i}$$

$$\ln L(\lambda) = n\ln\lambda + n\ln(1-\lambda) + \left(-\lambda \sum_{i=1}^{n} x_i - (1-\lambda)\sum_{i=1}^{n} y_i\right)$$

$$= n\ln\lambda + n\ln(1-\lambda) - \lambda \sum_{i=1}^{n} x_i - (1-\lambda)\sum_{i=1}^{n} y_i$$

$$\frac{d\ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} + \left(-\frac{n}{1-\lambda}\right) - \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

$$\frac{n}{\hat{\lambda}} - \frac{n}{1-\hat{\lambda}} - n\bar{x} + n\bar{y} = 0$$

$$\frac{n}{\hat{\lambda}} - \frac{n}{1-\hat{\lambda}} - n\bar{x} + n\bar{y} = 0$$

$$1-\hat{\lambda} - \hat{\lambda} - \hat{\lambda}(1-\hat{\lambda})\bar{x} + \hat{\lambda}(1-\hat{\lambda})\bar{y} = 0$$

$$(\bar{y}-\bar{x})\hat{\lambda}^2 + (\bar{x}-\bar{y}+2)\hat{\lambda} - 1 = 0$$

$$\hat{\lambda} = \frac{-(\bar{x}-\bar{y}+2) \pm \sqrt{(\bar{x}-\bar{y}+2)^2 - 4(\bar{y}-\bar{x})(-1)}}{2(\bar{y}-\bar{x})}$$

$$= \frac{\bar{y}-\bar{x}-2 \pm \sqrt{(\bar{x}-\bar{y}+2)^2 + 4(\bar{y}-\bar{x})}}{2(\bar{y}-\bar{x})}$$

we have $\lambda > 0 \quad \eta > 0$

$\eta = 1-\lambda > 0 \quad \lambda < 1$

so $0 < \lambda < 1$

Exercise 2: Multiple Linear Regression

Given: Multiple linear regression problem $y = x^T\beta + \epsilon$, with:
$x \in \mathbb{R}^d$: the predictor variables,
$y \in \mathbb{R}$: the response variable, and
$\beta \in \mathbb{R}^d$: the linear regression coefficients.
Data: $\mathcal{D} = ((x_1, y_1), \ldots, (x_N, y_N))$ of N independent data pairs.
Note $Y := (y_1, \ldots, y_N)^T \in \mathbb{R}^N$, $\tilde{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T \in \mathbb{R}^N$ and
$X = (x_1, \ldots, x_N)^T \in \mathbb{R}^{N \times d}$.

(a) Show that $\hat{\beta} := (X^TX)^{-1}X^TY \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$

Proof. Since $y = x^T\beta + \epsilon$, there is

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} x_1^1 \ldots x_d^1 \\ \vdots \ddots \vdots \\ x_1^1 \ldots x_d^1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix} = X\beta + \tilde{\epsilon}.$$

We know that $\epsilon \sim \mathcal{N}(0, \sigma^2)$, thus $\tilde{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix}$ is multivariate Gaussian distributed and $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$ holds, where I is the identity matrix.
It follows that with the **affine transformation**, $Y = X\beta + \tilde{\epsilon}$ is also multivariate Gaussian distributed: $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$. This is well-defined since I is not singular.
Because of the linear property of the expected value, it follows that

$$\begin{aligned} \mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^TX)^{-1}X^TY] \\ &= (X^TX)^{-1}X^T\mathbb{E}[Y] \\ &= (X^TX)^{-1}X^T\mathbb{E}[X\beta + \tilde{\epsilon}] \\ &= (X^TX)^{-1}(X^TX)\beta + (X^TX)^{-1}X^T\mathbb{E}[\tilde{\epsilon}] \\ &= \beta + 0 \\ &= \beta \end{aligned}$$

and

$$\begin{aligned}
\Sigma[\hat{\beta}] &= \Sigma(X^TX)^{-1}X^TY] \\
&= ((X^TX)^{-1}X^T) \cdot \Sigma[Y] \cdot ((X^TX)^{-1}X^T)^T \\
&= ((X^TX)^{-1}X^T) \cdot \sigma^2 I \cdot X(X^TX)^{-T} \\
&= ((X^TX)^{-1}X^T) \cdot \sigma^2 \cdot X(X^TX)^{-T} \\
&= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-T}) \\
&= \sigma^2(X^TX)^{-T} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

As the affine transformed of the multivariate Gaussian distributed random variable Y, $\hat{\beta}$ is also Gaussian distributed. It has mean β and covariance matrix $\sigma^2(X^TX)^{-1}$. The covariance is well defined since $(X^TX)^{-1}$ is not singular.

□

(b) Knowing the full distribution of $\hat{\beta}$ makes it possible for us to choose the variable β from the whole real-axis $\mathbb{R}$, i.e. the estimated value is no longer one single value anymore, but rather a set of real numbers or an interval.

(c) Given new data point x_*, we want to predict the response for x_*.
We know that the response is $y_* := x_*^T\hat{\beta} = x_*^T(X^TX)^{-1}X^TY$, where $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$.
Thus, as the result of an affine transformation, y_* is 1-dim normally distributed and has expectation $x_*^T\beta$ and variance $\sigma^2x_*^T(X^TX)^{-1}x_*$

(d) Now that we have the distribution of the new predicted response, we are able to see with which values the response is inclined to appear. We can also deduce other information from it, e.g.constructing the confidence interval of the predicted response.