Systems analysis and decision making

Exercises – List No. 4

Authors: A. Gonczarek, J. Tomczak

Exemplary problem

A guest of a bar drinks x litres of a beer by his first visit, and y liters of a beer by the second visit. The following quantities were observed for three random customers of the bar:

x	1	2	3
y	1	1	3

Tabela 1: Data for three clients.

We aim at predicting how many liters y a client will drink for the second visit if he drank x liters for the first time.

Linear regression: maximum likelihood estimation

Feature extraction

Feature extraction is as in least squares problem.

Model

Let us assume the following model:

$$y = \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{w} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(\varepsilon|0, \sigma^2)$ is a random variable (noise)

Task 1: Learning

For given data $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$:

- Task 1A: Write negative log-likelihood function $\ell_{ML}(\mathbf{w})$
- Task 1B: Derive maximum likelihood estimator for model parameters $\mathbf{w}$
- Task 1C: Derive maximum likelihood estimator for σ^2 given model parameters $\hat{\mathbf{w}}_{ML}$ from Task 1B. What kind of information does this parameter provide?

Task 2: Prediction

Use formulas derived in Task 1 to calculate (numerically) values of $\mathbf{w}$ (model parameters), σ (standard deviation) and $\overline{\mathbf{y}}$ (model prediction). Use input data that is provided in Table 1.

Linear regression: maximum *a posteriori* estimation

Feature extraction

Feature extraction is as in least squares problem.

Model

Let us assume the following model:

$$p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(y|\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}, \sigma^2)$$

with the *a priori* distribution on parameters $\mathbf{w}$:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0, \alpha^2\mathbf{I}),$$

where α^2 i σ^2 are known.

Task 3: Learning

For given data $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$:

- Task 3A: Write negative log-likelihood function $\ell_{ML}(\mathbf{w})$
- Task 3B: Derive maximum a posteriori estimator for model parameters $\mathbf{w}$
- Task 3C: Derive maximum a posteriori estimator for σ^2 given model parameters $\hat{\mathbf{w}}_{MAP}$ from Task 3B. What kind of information does this parameter provide?

Task 4: Prediction

Use formulas derived in Task 3 to calculate (numerically) values of $\mathbf{w}$ (model parameters), σ (standard deviation) and $\overline{\mathbf{y}}$ (model prediction). Use input data that is provided in Table 1 and $\lambda = 10^{-3}$.

HANDOUT FOR LINEAR REGRESSION (LISTS 3 & 4)

Bernoulli distribution:

$\mathrm{B}(x|\theta) = \theta^x(1-\theta)^{1-x}$, where $x \in \{0,1\}$ i $\theta \in [0,1]$

$\mathbb{E}[x] = \theta$

$\mathrm{Var}[x] = \theta(1-\theta)$

Categorical (Multinoulli) distribution:

$\mathrm{M}(\mathbf{x}|\boldsymbol{\theta}) = \prod_{d=1}^{D} \theta_d^{x_d}$, where $x_d \in \{0,1\}$ i $\theta_d \in [0,1]$ for all $d = 1, 2, \ldots, D$, $\sum_{d=1}^{D} \theta_d = 1$

$\mathbb{E}[x_d] = \theta_d$

$\mathrm{Var}[x_d] = \theta_d(1-\theta_d)$

Normal distribution:

$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$

$\mathbb{E}[x] = \mu$

$\mathrm{Var}[x] = \sigma^2$

Multivariate normal distribution:

$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right\}$,

where $\mathbf{x}$ is D-dimensional vector, $\boldsymbol{\mu}$ – D-dimensional vector of means, $\boldsymbol{\Sigma}$ – $D \times D$ covariance matrix

$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$

$\mathrm{Cov}[\mathbf{x}] = \boldsymbol{\Sigma}$

Beta distribution:

$\mathrm{Beta}(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$,

where $x \in [0,1]$ and $a > 0$ i $b > 0$, $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$

$\mathbb{E}[x] = \frac{a}{a+b}$

$\mathrm{Var}[x] = \frac{ab}{(a+b)^2(a+b+1)}$

Marginal distribution:

In the continuous case:

$$p(x) = \int p(x,y) \mathrm{d}y$$

and in the discrete case:

$$p(x) = \sum_y p(x,y)$$

Conditional distribution:

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

Marginal distribution and conditional distribution for multivariate normal distribution:
Assume $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_c \\ \boldsymbol{\Sigma}_c^T & \boldsymbol{\Sigma}_b \end{bmatrix},$$

then we get the following dependencies:
$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$,
$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\hat{\boldsymbol{\mu}}_a, \hat{\boldsymbol{\Sigma}}_a)$, where
$\hat{\boldsymbol{\mu}}_a = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_b^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$,
$\hat{\boldsymbol{\Sigma}}_a = \boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_c\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\Sigma}_c^T$.

Bayes' rule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Maximum likelihood estimator:
There are given N independent examples of $\mathbf{x}$ from the identical distribution $p(\mathbf{x}|\theta)$, $\mathcal{D} = \{\mathbf{x}_1 \dots \mathbf{x}_N\}$.
The likelihood function is the following function:

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(\mathbf{x}_n|\theta).$$

The logarithm of the likelihood function $p(\mathcal{D}|\theta)$ is given by the following expression:

$$\log p(\mathcal{D}|\theta) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta).$$

Maximum likelihood estimator of the parameters θ_{ML} minimizes the likelihood function:

$$p(\mathcal{D}|\theta_{ML}) = \max_{\theta} p(\mathcal{D}|\theta).$$

Maximum *a posteriori* estimator:
There are given N independent examples of $\mathbf{x}$ from the identical distribution $p(\mathbf{x}|\theta)$, $\mathcal{D} = \{\mathbf{x}_1 \dots \mathbf{x}_N\}$.

Maximum *a posteriori* (MAP) estimator of the parameters θ_{MAP} minimizes the *a posteriori* distribution:

$$p(\theta_{MAP}|\mathcal{D}) = \max_{\theta} p(\theta|\mathcal{D}).$$

Risk in the decision making:
Risk (expected loss) is defined as follows:

$$\mathcal{R}[\overline{y}] = \iint L(y, \overline{y}(\mathbf{x}))\ p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y,$$

where $L(\cdot, \cdot)$ is the loss function.

Chosen properties of matrix calculus:
For given vectors $\mathbf{x}$, $\mathbf{y}$ and a matrix $\mathbf{A}$, which is symmetric and positive definite, the following equations hold true:

- $\dfrac{\partial}{\partial \mathbf{y}}(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{A}(\mathbf{x}-\mathbf{y}) = -2\mathbf{A}(\mathbf{x}-\mathbf{y})$
- $\dfrac{\partial(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{A}^{-1}(\mathbf{x}-\mathbf{y})}{\partial \mathbf{A}} = -\mathbf{A}^{-1}(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{A}^{-1}$
- $\dfrac{\partial \ln\det(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{A}^{-1}$