Systems analysis and decision making

Exercises – List No. 2

Authors: A. Gonczarek, J. Tomczak

Maximum likelihood (ML) estimation

Task 1

Occurrence of a spam message is described by a Bernoulli random variable x with a parameter $\theta \in [0, 1]$, where the random variable x takes the value 1, if the message is spam. Some user tagged N messages. Calculate the maximum likelihood estimator of the parameter θ.

Task 2

Population of the students at Wroclaw University of Technology has been divided into the following three groups:

1. Students with the mean of grades below 3.5.
2. Students with the mean of grades between 3.5 and 4.5.
3. Students with the mean of grades above 4.5.

Each student in the population is described by a vector of random variables $\mathbf{x} = (x^1\ x^2\ x^3)^{\mathrm{T}}$, taking one of three possible states: $(1\ 0\ 0)^{\mathrm{T}}$ if the student belongs to the first group, $(0\ 1\ 0)^{\mathrm{T}}$ if the student belongs to the second group, and $(0\ 0\ 1)^{\mathrm{T}}$ if the student belongs to the third group. The distribution of $\mathbf{x}$ is categorical distribution (also known as generalized Bernoulli distribution or Multinoulli distribution) with parameters $\boldsymbol{\theta} = (\theta_1\ \theta_2\ \theta_3)^{\mathrm{T}}$. From the population of the studets N examples were drawn. Calculate the maximum likelihood estimator of $\boldsymbol{\theta}$.

Task 3

Car alarm switches on if the ultrasonic sensor placed in the cabin exceeds some threshold. However, the sensor needs to be calibrated before installing it in the alarm. It is assumed that the measurements of the sensor are realizations of the normally distributed random variable $x \sim \mathcal{N}(x|\mu, \sigma^2)$. N measurements were done when there was no movement in the cabin. Calculate the maximum likelihood estimators of μ and σ^2.

Task 4

Some word spoken by a man is characterized by a vector normally distributed random variables $\mathbf{x} = (x^1 \ \dots \ x^D)^{\mathrm{T}}$, $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Several different people were asked to say this word several times and thus N examples were collected. Calculate the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Maximum *a posteriori* (MAP) estimation

Task 5

Let the random variable $x \in \{0, 1\}$ denote loss and wind of Śląsk Wrocław, respectively. The random variable x is described by the Bernoulli distribution $\mathrm{B}(x|\theta)$. N results in the season were collected. Assuming beta *a priori* distribution $\mathrm{Beta}(\theta|a, b)$, calculate *MAP* estimator (maximum *a posteriori*) for θ. How the hyperparamters a i b can be interpreted?

HANDOUT

Bernoulli distribution:

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

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

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

Categorical (Multinoulli) distribution:

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

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

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

Normal distribution:

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

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

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

Multivariate normal distribution:

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

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:

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

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:

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