“Machine Learning and Computational Statistics”

5th Homework

Exercise 1 (multiple choices question): Which of the following statements are true?

1. The Least Squares estimation method takes into account the statistical nature of the training data.

2. Ridge Regression does not take into account any statistical information related to the available data set.

3. The Maximum Likelihood method assumes that the training data stem from some probability density distribution (pdf).

4. The Maximum Likelihood method treats the parameters involved in the assumed pdf as random variables.

Exercise 2 (multiple choices question): Assume that for the data set $X = \{-1.1, -0.5,\ 0.1,\ 0.6,\ 1.0\}$, it is known that its elements have been drawn independently from a unit variance normal distribution of unknown mean μ. The likelihood function of μ, with respect to X, $p(X;\mu) \equiv p(-1.1, -0.5,\ 0.1,\ 0.6,\ 1.0;\mu)$, is:

1. $\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(-1.1-\mu)^2}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(-0.5-\mu)^2}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(0.1-\mu)^2}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(0.6-\mu)^2}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(1.0-\mu)^2}{2}\right)$

2. $\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(-1.1-\mu)^2}{2}\right)+\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(-0.5-\mu)^2}{2}\right)+\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(0.1-\mu)^2}{2}\right)+\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(0.6-\mu)^2}{2}\right)+\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(1.0-\mu)^2}{2}\right)$

3. $\frac{1}{\sqrt{2\pi}}\ln\left(-\frac{(-1.1-\mu)^2}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}\ln\left(-\frac{(-0.5-\mu)^2}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}\ln\left(-\frac{(0.1-\mu)^2}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}\ln\left(-\frac{(0.6-\mu)^2}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}\ln\left(-\frac{(1.0-\mu)^2}{2}\right)$

4. $-5\cdot\ln\left(\sqrt{2\pi}\right)+\left(-\frac{(-1.1-\mu)^2}{2}\right)+\left(-\frac{(-0.5-\mu)^2}{2}\right)+\left(-\frac{(0.1-\mu)^2}{2}\right)+\left(-\frac{(0.6-\mu)^2}{2}\right)+\left(-\frac{(1.0-\mu)^2}{2}\right)$

Exercise 3 (multiple choices question): Assume that for the data set $X = \{0.1,\ 0.5,\ 0.7,\ 1.1,\ 2.0\}$, it is known that its elements have been drawn independently from the exponential distribution $f(x;\lambda) = \lambda e^{-\lambda x}$ $(x > 0)$, parameterized by the (unknown) parameter λ. The log-likelihood function of λ, with respect to X, $p(X;\lambda) \equiv p(0.1,\ 0.5,\ 0.7,\ 1.1,\ 2.0;\lambda)$, is:

1. $\lambda^5 e^{-\lambda(0.1+\ 0.5+\ 0.7+\ 1.1+2.0)}$

2. $5 \cdot \ln\lambda - \lambda(0.1 \cdot 0.5 \cdot 0.7 \cdot 1.1 \cdot 2.0)$

3. $5 \cdot \ln\lambda - \lambda(0.1 + 0.5 + 0.7 + 1.1 + 2.0)$

4. $\lambda^5 + e^{-\lambda(0.1+\ 0.5+\ 0.7+\ 1.1+2.0)}$

Exercise 4 (multiple choices question): Consider the two-dimensional Gaussian distribution $p(\boldsymbol{x}) := N(\boldsymbol{\mu}, \Sigma)$ with mean $\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and covariance matrix $\Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$. Recall that the inverse A^{-1} of a two-dimensional matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ (if it exists) is $A^{-1} = \frac{1}{D}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$, where $D = ad - bc$. Then, the value of $p(\boldsymbol{x})$ for $\boldsymbol{x} = [2,\ 1]^T$is (in four decimals accuracy)

1. 0.4523

2. 0.0294

3. 0.9612

4. 1.0102

Exercise 5 (multiple choices question): Consider a data set X, whose elements are drawn independently from a pdf of a known form, parameterized by an (unknown) parameter vector $\boldsymbol{\theta}$. Which one of the following optimization problems is not equivalent to the other three ones (in the sense that it does not return the same solution):

1. $\arg max_{\boldsymbol{\theta}} p(X; \boldsymbol{\theta})$

2. $\arg max_{\boldsymbol{\theta}} \ln(p(X; \boldsymbol{\theta}))$

3. $\arg min_{\boldsymbol{\theta}} (-p(X; \boldsymbol{\theta}))$

4. $\arg max_{\boldsymbol{\theta}} \frac{1}{p(X;\boldsymbol{\theta})}$

Exercise 6 (multiple choices question): Consider the data sets $, X_j,\ j = 1, 2, \ldots,$ of finite cardinality N, whose elements are drawn independently from an l-dimensional normal pdf of known covariance matrix Σ and of an unknown mean $\boldsymbol{\mu}_o$. Let $\widehat{\boldsymbol{\mu}}_{ML}{}^{(j)}$ be the maximum likelihood estimation of $\boldsymbol{\mu}_o$, associated with X_j, and let $\widehat{\boldsymbol{\mu}}_{ML}$ be the maximum likelihood estimator of $\boldsymbol{\mu}_o$, whose instances are the $\widehat{\boldsymbol{\mu}}_{ML}{}^{(j)}$'s. Which of the following statements are true?

1. $\mathrm{E}[\widehat{\boldsymbol{\mu}}_{ML}] = \boldsymbol{\mu}_o$

2. $\mathrm{Prob}\{|\,\widehat{\boldsymbol{\mu}}_{ML} - \boldsymbol{\mu}_o| > \epsilon\} = 0$, for any $\epsilon > 0$.

3. $\widehat{\boldsymbol{\mu}}_{ML}$ is an efficient estimator of $\boldsymbol{\mu}_o$

4. $\mathrm{E}[\widehat{\boldsymbol{\mu}}_{ML}]$ is expected to be "close" to $\boldsymbol{\mu}_o$, for large enough values of N.

Exercise 7 (multiple choices question): Consider the data sets $X_j,\ j = 1, 2, \dots$, of cardinality N, whose elements are drawn independently from an l-dimensional normal pdf of known covariance matrix Σ and unknown mean $\boldsymbol{\mu}_o$. Let $\widehat{\boldsymbol{\mu}}_{ML}{}^{(j)}$ be the maximum likelihood estimation of $\boldsymbol{\mu}_o$, associated with X_j, and let $\widehat{\boldsymbol{\mu}}_{ML}$ be the maximum likelihood estimator of $\boldsymbol{\mu}_o$, whose instances are the $\widehat{\boldsymbol{\mu}}_{ML}{}^{(j)}$'s. Which of the following statements are true?

1. As $N \to \infty$, although $\mathrm{E}[\widehat{\boldsymbol{\mu}}_{ML}] = \boldsymbol{\mu}_o$, the variance of $\widehat{\boldsymbol{\mu}}_{ML}{}^{(j)}$'s around $\boldsymbol{\mu}_o$ may be large.

2. As $N \to \infty$, $\widehat{\boldsymbol{\mu}}_{ML}$ is an efficient estimator of $\boldsymbol{\mu}_o$.

3. As $N \to \infty$, $\widehat{\boldsymbol{\mu}}_{ML}$ is a biased estimator of $\boldsymbol{\mu}_o$.

4. As $N \to \infty$ and for any $\epsilon > 0$, it is always probable to have an $\widehat{\boldsymbol{\mu}}_{ML}{}^{(j)}$ at distance greater than ϵ, from $\boldsymbol{\mu}_o$.

Exercise 8 (multiple choices question): Consider a data set $X = \{x_1, x_2, \dots, x_N\}$, whose elements are drawn from an one-dimensional normal distribution, with mean μ and variance σ^2. The values of μ and σ^2 are assumed to be unknown. Let $\hat{\mu}$ and $\hat{\sigma}^2$ be the Maximum Likelihood estimates of these parameters that are based on X. In order to derive the expressions for $\hat{\mu}$ and $\hat{\sigma}^2$, one should equate to zero the derivatives (since both parameters are scalars in the present case) of the log-likelihood function with respect to μ and σ^2, and solve the resulting system of the two equations and the two unknowns. The resulting values for $\hat{\mu}$ and $\hat{\sigma}^2$ are:

1. $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$

2. $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$

3. $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} {x_n}^2$ and $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$

4. $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})$

Exercise 9 (multiple choices question): Consider a data set $X = \{x_1, x_2, \dots, x_N\}$, whose elements are drawn independently from a mixture $p(x)$ comprising two one-dimensional unit variance normal distributions, $p_1(x)$ and $p_2(x)$, with mean values -2 και 2, respectively. The data points of X are drawn with equal probability from the two distributions. In mathematical terms, this is expressed as $p(x) = \frac{1}{2}p_1(x) + \frac{1}{2}p_2(x)$. Assume that it is erroneously assumed that the elements of X stem from a single normal distribution with mean value μ and variance σ^2. Let $\hat{\mu}$ and $\hat{\sigma}^2$ be the maximum likelihood estimates of these parameters that are based on X. Which of the following statements are true?

1. $\hat{\mu}$ and $\hat{\sigma}^2$ are expected to be close to 2 and 1, respectively.

2. $\hat{\mu}$ and $\hat{\sigma}^2$ are expected to be close to -2 and 1, respectively.

3. $\hat{\mu}$ is expected to be close to 0 and $\hat{\sigma}^2$ to be greater than 4.

4. $\hat{\mu}$ and $\hat{\sigma}^2$ are expected to be close to 0 and 1, respectively.

Exercise 10 (multiple choices question): Consider the linear regression task $\mathrm{y} = \boldsymbol{\theta}^T\mathbf{x} + \eta$. Assume that we are given a data set X consisting of N data points, $(y_n, \boldsymbol{x}_n), n = 1, \dots, N$, where the noise samples $\eta_n, n = 1, \dots, N$, originate from a jointly Gaussian N-dimensional distribution with zero mean and covariance matrix (of size $N \times N$) equal to Σ_η. Let $\widehat{\boldsymbol{\theta}}_{ML}$ and $\widehat{\boldsymbol{\theta}}_{LS}$ be the maximum likelihood and the least squares estimates of $\boldsymbol{\theta}$, based on X. Which of the following statements are true?

1. If Σ_η equals to the N-dimensional identity matrix, then $\widehat{\boldsymbol{\theta}}_{ML} \neq \widehat{\boldsymbol{\theta}}_{LS}$

2. If Σ_η has non-zero off-diagonal entries, then $\widehat{\boldsymbol{\theta}}_{ML} \neq \widehat{\boldsymbol{\theta}}_{LS}$

3. If Σ_η is diagonal and has a single diagonal entry equal to zero, then $\widehat{\boldsymbol{\theta}}_{ML} = \widehat{\boldsymbol{\theta}}_{LS}$.

4. If $N = 1$, it always holds $\widehat{\boldsymbol{\theta}}_{ML} = \widehat{\boldsymbol{\theta}}_{LS}$

Exercise 11 (multiple choices question): Consider the linear regression task $\mathrm{y} = \boldsymbol{\theta}^T\mathbf{x} + \eta$. Assume that we are given a finite data set X consisting of N data points, $(y_n, \boldsymbol{x}_n), n = 1, \dots, N$. Assume also that the noise samples $\eta_n,\ n = 1, \dots, N$, originate from a jointly Gaussian N-dimensional distribution with zero mean and covariance matrix (of size $N \times N$) equal to (the non-diagonal matrix) Σ_η. However, let us pretend that the latter information is not available to us; that is, we have at our disposal only the data set X and no additional information. Let $\widehat{\boldsymbol{\theta}}_{ML}$ and $\widehat{\boldsymbol{\theta}}_{LS}$ be the Maximum Likelihood and the Least Squares estimates of $\boldsymbol{\theta}$, based on X. Which of the following statements are true?

1. $\widehat{\boldsymbol{\theta}}_{LS}$ is a biased estimate of $\boldsymbol{\theta}$.

2. $\widehat{\boldsymbol{\theta}}_{ML}$ is an unbiased estimate of $\boldsymbol{\theta}$.

3. $\widehat{\boldsymbol{\theta}}_{ML}$ is an efficient estimate of $\boldsymbol{\theta}$.

4. $\widehat{\boldsymbol{\theta}}_{LS}$ is an efficient estimate of $\boldsymbol{\theta}$.

<u>Exercise 12:</u>

Consider the Erlang distribution $p(x) = \theta^2 \ x \, exp(-\theta \, x) u(x)$, (where $u(x) = 1(0)$, if $x \geq 0$ (< 0)).

(a) Given a set of N measurements $x_1, \ldots, x_N$, for the random variable x that follows the Erlang distribution, prove that the ML estimate of θ is

$$\theta_{ML} = \frac{2N}{\sum_{i=1}^{N} x_i}$$

(b) For $N = 5$ and $x_1 = 2$, $x_2 = 2.2$, $x_3 = 2.7$, $x_4 = 2.4$, $x_5 = 2.6$, estimate the θ_{ML}. Utilizing this estimate, determine $\hat{p}(x)$, for $x = 2.3$ and $x = 2.9$.