"Machine Learning and Computational Statistics"

3rd Homework

Exercise 1:

Let $\mathbf{x} = [\mathrm{x}_1, \dots, \mathrm{x}_l]^T$ be an l-dimensional random vector with mean vector $\boldsymbol{\mu} = [\mu_1, \dots, \mu_l]^T$ and let $cov(\mathbf{x}) = E[(\mathbf{x} - \boldsymbol{\mu}) \cdot (\mathbf{x} - \boldsymbol{\mu})^T]$ and $R_{\mathbf{x}} = E[\mathbf{x} \cdot \mathbf{x}^T]$ be the corresponding covariance and correlation matrices, respectively. Prove that

$$R_{\mathbf{x}} = cov(\mathbf{x}) + \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

Exercise 2:

(a) Prove that the **mean** and the **variance** of a random variable x that follows the Bernoulli distribution $Bern(\mathrm{x}|p)$ $(0 < p < 1)$ are $E[\mathrm{x}] = p$ and $\sigma_{\mathrm{x}}{}^2 = p(1-p)$, respectively.

(b) Prove that the **mean** of the random variable x that follows the binomial distribution $Bin(\mathrm{x}|n,p)$ $(0 < p < 1)$ is $E[\mathrm{x}] = np$.

Hint: For (b) use the binomial expansion equation

$$(x+y)^n = \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \dots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n$$

Exercise 3:

Let $\mathbf{x} = [\mathrm{x}_1, \dots, \mathrm{x}_l]^T$ be an l-dimensional random vector that follows the ($l-$ dim.) normal distribution

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{l/2} |\Sigma|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)$$

where $\boldsymbol{\mu} = [\mu_1, \dots, \mu_l]^T$ and $\Sigma = \begin{bmatrix} \sigma_1{}^2 & \sigma_{12} & \cdots & \sigma_{1l} \\ \sigma_{21} & \sigma_2{}^2 & \cdots & \sigma_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{l1} & \sigma_{l2} & \cdots & \sigma_l{}^2 \end{bmatrix}$. Prove that **if** Σ is diagonal (that is, $\sigma_{ij} = 0, i = 1, \dots, l, j = 1, \dots, l, i \neq j$), the coordinates (random variables) $\mathrm{x}_i, i = 1, \dots, l$, of $\mathbf{x}$ are statistically independent.

Hint: Prove that $p(\boldsymbol{x}) = \prod_{i=1}^{l} p_i(x_i) = \prod_{i=1}^{l} \frac{1}{\sqrt{2\pi\sigma_i{}^2}} exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i{}^2}\right)$.

Exercise 4:

Consider a regression problem where both the independent and dependent quantities are scalars and are related via the following linear model

$$y = \theta \cdot x + \eta$$

where η follows the zero mean normal distribution with variance σ^2 (Note that only a **single parameter** is involved here). Consider also the data set

$$X = \{(y_1, x_1), \dots, (y_N, x_N)\}.$$

Derive the **least squares estimate** for the **scalar** θ, based on the above data set.

Hint: Work either **(a)** by considering the general least squares solution where $\boldsymbol{\theta}$ is a vector and deriving the solution for the specific case where θ is a scalar, or **(b)** Formulating explicitly the optimization problem for this specific case and deriving the estimate (i.e., writing explicitly the cost function for this case, taking the derivative wrt θ and setting it equal to zero…)

Exercise 5 (python code + text):

Consider a regression problem where both the independent and dependent quantities are scalars and are related via the following linear model

$$y = \theta_o \cdot x + \eta$$

where η follows the zero mean normal distribution with variance σ^2 and $\theta_o = 2$ (thus, the actual model is $y = 2 \cdot x + \eta$).

(a) Generate $d = 50$ data set as follows:

- Generate a set D_1 of $N = 30$ data pairs (y_i', x_i), where $y' = 2 \cdot x$.
- Add zero mean and $\sigma^2 = 64$ variance Gaussian noise to the y_i' 's, resulting to y_i's.
- The **observed** data pairs are (y_i, x_i), $i = 1, \dots, 30$, which constitute the data set D_1.

Repeat the above procedure d=50 times in order to generate 50 different data sets.

(b) Compute the LS linear **estimates** of θ_o based on $D_1, D_2, \dots, D_d$ (thus, $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_d$ numbers/estimates will result).

(c) Consider now the **estimator** $\hat{\theta}$ that models $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_d$ and

(c1) compute the $MSE = E\left[\left(\hat{\theta} - \theta_o\right)^2\right]$ and

(c2) depict graphically the values $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_d$ and comment on how they are spread around θ_o.

<u>Hint:</u> For (c) approximate MSE as $MSE = \frac{1}{d}\sum_{i=1}^{d}\left(\hat{\theta}_i - \theta_o\right)^2$.