"Machine Learning and Computational Statistics"

2nd Homework

Exercise 1:

Consider the following nonlinear model:

$$y = 3{x_1}^2 + 4{x_2}^2 + 5{x_3}^2 + 7x_1x_2 + x_1x_3 + 4x_2x_3 - 2x_1 - 3x_2 - 5x_3 + \eta$$

Define a suitable function φ that transforms the problem to a space where the problem of estimating the model becomes linear. What is the dimension of the original and the transformed spaces?

Exercise 2:

Consider the following two-class nonlinear classification task:

$$\boldsymbol{x} = [x_1, x_2, x_3]^T : {x_1}^2 + 3{x_2}^2 + 6{x_3}^2 + x_1x_2 + x_2x_3 > (<) 3 \longrightarrow \boldsymbol{x} \in \omega_1(\omega_2)$$

Define a suitable function φ that transforms the problem to a space where the problem of estimating the border of the two classes becomes linear. What is the dimension of the original and the transformed spaces?

Exercise 3:

Consider the exclusive OR (XOR) classification problem, where the points $(0,0)$ and $(1,1)$ are assigned to class -1 and the points $(0,1)$ and $(1,0)$ are assigned to class $+1$.

(a) Draw the points on the paper and **prove** that the classification problem is not linearly separable.

(b) Propose a transformation $\varphi(\cdot)$ that maps the above points to a new space, where the XOR classification problem becomes linearly separable.

Hint: For (a), assume that the problem is linearly separable, that is, there exists a hyperplane (H) $\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$, that leaves the points from class +1 (resp. -1) on its positive (resp. negative) side (thus, if a point $({x_1}', {x_2}')$ lies on the positive (resp. negative) side of (H), it holds $\theta_0 + \theta_1 {x_1}' + \theta_2 {x_2}' > (<) 0$). Write down the relative four inequalities (one for each data point) and combine them, so that to reach a contradiction.

Exercise 4:

Consider a two-class one-dimensional classification problem where the points $-3, -1, 0, 2$ belong to class -1 and the points $-8, -5, 4, 7$ belong to class $+1$. Propose a transformation $\varphi(\cdot)$ that maps the above points to a **new one-dimensional space**, where the classification problem becomes linearly separable.

Exercise 5:

Consider a two-class, two-dimensional classification problem, where the data points $[1,1]^T, [1,2]^T, [2,1]^T$ belong to class $+1$, while the data points $[-1,-1]^T, [-1,-2]^T, [-2,-1]^T$ belong to class -1.

(a) Determine the hyperplane (line) that separates the data from the two classes, utilizing the Least Squares criterion.
(b) Draw on the paper the data points from the classes along with the line determined in (a).
(c) Is this the unique line that separates the data from the two classes? Is this line the only one that results from the minimization of the sum of squared error criterion?

Perform all the necessary operations on the paper.

Hint: Consult the pdf file "Detailed_derivation_of_LS_estimation", which has been uploaded to e-class.

Exercise 6:

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 7:

(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

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

Exercise 8 (python code + text)**:**

(a) **Generate** a set $X=\{(y_i,\boldsymbol{x}_i),\boldsymbol{x}_i=[x_{i1},x_{i2}]^T\in R^2,y_i\in R,i=1,\dots,200\}$ from the model

$$y=\theta_0+\theta_1x_1+\theta_2x_2+\theta_3x_1x_2+\eta$$

where η is an i.i.d. normal zero mean noise, with variance 0.05. Use $\theta_0=3,\theta_1=2,\theta_2=1,\theta_3=1$ (adopt the strategy given in the example of the 2nd slide of the 2nd lecture). In the sequel, pretend that you do not know the model that generates the data. All you have at your disposal is the data set X.

(b) **Adopting** the linear model assumption in the **original space** (that is, assuming that $y=\theta_0+\theta_1x_1+\theta_2x_2$), estimate the parameters of the model $(\theta_0,\theta_1,\theta_2)$ that minimize the sum of error squares criterion.

(c) For each one of the 200 data points $\boldsymbol{x}_i$ of X, determine the associated estimate $\hat{y}_i$ provided from the model estimated in **(b)** and compute the $MSE=\frac{1}{200}\sum_{i=1}^{200}(y_i-\hat{y}_i)^2$.

(d) **Apply** the transformation $\varphi(\boldsymbol{x})=\begin{bmatrix}\varphi_1(\boldsymbol{x})\\ \varphi_2(\boldsymbol{x})\\ \varphi_3(\boldsymbol{x})\end{bmatrix}=\begin{bmatrix}x_1\\ x_2\\ x_1\cdot x_2\end{bmatrix}$, on all $\boldsymbol{x}_i$’s of X. Denoting by $\boldsymbol{x}_i'(\in R^3)$ the image of $\boldsymbol{x}_i$ (that is, $\boldsymbol{x}_i'=\varphi(\boldsymbol{x}_i)$), form a new data set $X'=\{(y_i,\boldsymbol{x}_i'),\boldsymbol{x}_i'\in R^3,y_i\in R,i=1,\dots,200\}$.

(e) **Adopting** the linear model assumption in the **transformed space** and the sum of error squares criterion, estimate the parameters of the model that minimize this criterion.

(f) For each one of the 200 data points $\boldsymbol{x}_i$ of X, determine the associated estimate $\hat{y}_i$ provided from the model estimated in **(e)** and compute the $MSE=\frac{1}{200}\sum_{i=1}^{200}(y_i-\hat{y}_i)^2$.

(g) Comment on the results obtained in (c) and (f).

Exercise 9 (python code + text)**:**

(a) **Generate** a set $X = \{(y_i, \boldsymbol{x}_i), \boldsymbol{x}_i \in R^2, y_i \in \{-1, +1\}, i = 1, \ldots, 2000\}$, as follows: Select 2000 points in the squared area $[-2,2]x[-2,2]$ of the R^2 space, using the uniform distribution. All points that lie on the positive side of the curve ${x_2}^2 - {x_1}^2 = 0$, are assigned to the class “+1”, while all the others are assigned to class “-1”. Plot the data using different colors for points from different classes. In the sequel, pretend that you do not know how the data were generated. All you have at your disposal is the data set X.

(b) **Apply** the transformation $\varphi(\boldsymbol{x}) = \begin{bmatrix} \varphi_1(\boldsymbol{x}) \\ \varphi_2(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} x_1^2 \\ x_2^2 \end{bmatrix}$, on all $\boldsymbol{x}_i$’s of X. Denoting by $\boldsymbol{x}_i{}'$ the image of $\boldsymbol{x}_i$ (that is, $\boldsymbol{x}_i' = \varphi(\boldsymbol{x}_i)$), we form a new data set $X' = \{(y_i, \boldsymbol{x}_i{}'), i = 1, \ldots, 2000\}$

(c) **Plot** the $\boldsymbol{x}_i$’s using again different colors for points from different classes and compare the resulting plot with that of (a). Comment on them.

(d) **Adopting** the linear model assumption in the transformed space and the sum of error squares criterion, estimate the parameters of the model that minimize this criterion.