Statistical Machine Learning for Data Science Assignment 1

Due Date: January 10, 2021 - By 11:59pm

All your answers must be written on a separate sheet, properly typeset and submitted in the form of a report, in pdf format. No MS Word report will be accepted. Make sure your report in pdf format is uploaded to the designated folder before the deadline.

Exercise 1

Let $\mathscr{D}_n = \{(\mathrm{x}_i, \mathrm{y}_i) \overset{iid}{\sim} p_{\mathrm{xy}}(\mathrm{x}, \mathrm{y}),\ \mathrm{x}_i \in \mathbb{R}, \mathrm{y}_i \in \mathbb{R}, i = 1, \cdots, n\}$. Consider using the data $\mathscr{D}_n$, to build mappings $f : \mathscr{X} \longrightarrow \mathscr{Y}$, such that $f \in \mathscr{H}$, where

$$\mathscr{H} := \left\{ \mathrm{x} \mapsto f(\mathrm{x}) = \theta \mathrm{x},\ \theta \in \mathbb{R}^\star \right\} \tag{1}$$

Further suppose that $\forall i \in [n]$, we have

$$p(\mathrm{y}_i | \mathrm{x}_i, \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mathrm{y}_i - f(\mathbf{x}_i))^2}, \tag{2}$$

where $f \in \mathscr{H}$. Finally, let $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_n)^\top$, and $\mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_n)^\top$ and define

$$\mathtt{SSE}_n(a) = \sum_{i=1}^{n} (\mathrm{y}_i - a\mathrm{x}_i)^2 = (\mathbf{y} - a\mathbf{x})^\top (\mathbf{y} - a\mathbf{x}) = \|\mathbf{y} - a\mathbf{x}\|_2^2. \tag{3}$$

1. Specify the input space in this problem, and clearly indicate its dimensionality.
2. Specify the output space in this problem, and clearly indicate its dimensionality.
3. Specify the dimensionality of $\mathbf{x}$.
4. Specify the dimensionality of $\mathbf{y}$.
5. Determine in this case the assumed conditional distribution of Y_i given x_i and deduce the distribution $\mathbf{y}$ given $\mathbf{x}$.
6. Rewrite the model defined by Equation (2) in its additive form featuring the deterministic component (signal) and the stochastic component (noise or error term). Be sure to reflect the fact that $f \in \mathscr{H}$ as defined in Equation (1), but also the clear probability distribution used.
7. Indicate which type statistical machine learning task is being solved here. Justify your answer.

8. Use the appropriate tools to find $\frac{\partial \texttt{SSE}_n(a)}{\partial a}$
9. Show that
$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^\star}{\texttt{argmin}}\{\texttt{SSE}_n(\theta)\} = \frac{\mathbf{x}^\top \mathbf{y}}{\mathbf{x}^\top \mathbf{x}}.$$
Now, let $\widehat{Y}_i = \widehat{\theta}\mathrm{x}_i$, for $i \in [n]$, and define $\widehat{\mathbf{y}} = (\widehat{\mathrm{y}}_1, \cdots, \widehat{\mathrm{y}}_n)^\top$.
10. Find $\texttt{mean}[\widehat{\theta}] = \mathbb{E}[\widehat{\theta}]$.
11. Find $\texttt{variance}[\widehat{\theta}] = \mathbb{V}[\widehat{\theta}]$.
12. Find $\texttt{mean}(Y_i|\mathrm{x}_i) = \mathbb{E}[Y_i|\mathrm{x}_i]$ for $i \in [n]$ and deduce $\texttt{mean}(\mathbf{y}|\mathbf{x}) = \mathbb{E}(\mathbf{y}|\mathbf{x})$.
13. Write down $\mathbf{y}$ as a function of $\mathbf{x}$, θ and all other necessary parts of the assumed model in keeping with Equation (2).
14. Write down $\widehat{\mathbf{y}}$ as function of $\mathbf{x}$ and $\mathbf{y}$.
15. Find $\texttt{variance}[Y_i|\mathrm{x}_i] = \mathbb{V}[Y_i|\mathrm{x}_i]$ for $i \in [n]$ and deduce $\texttt{variance}(\mathbf{y}|\mathbf{x})$.
16. Find $\texttt{mean}(\widehat{Y}_i|\mathrm{x}_i) = \mathbb{E}[\widehat{Y}_i|\mathrm{x}_i]$ for $i \in [n]$ and deduce $\texttt{mean}(\widehat{\mathbf{y}}|\mathbf{x}) = \mathbb{E}(\widehat{\mathbf{y}}|\mathbf{x})$.
17. Find $\texttt{variance}(\widehat{Y}_i|\mathrm{x}_i) = \mathbb{V}[\widehat{Y}_i|\mathrm{x}_i]$ for $i \in [n]$ and deduce $\texttt{variance}(\widehat{\mathbf{y}}|\mathbf{x}) = \mathbb{V}(\widehat{\mathbf{y}}|\mathbf{x})$.
18. Based on all the above, determine the distribution $\widehat{\theta}$.
19. Based on all the above, determine the distribution $\widehat{Y}_i$ given x_i and deduce the distribution of $\widehat{\mathbf{y}}$ given $\mathbf{x}$.
20. Find the estimator $\widehat{\sigma^2}$ of σ^2.

Exercise 2

```
prostate <- read.csv('prostate-cancer-1.csv')  # DNA MicroArray Gene Expression
```

1. Find by all means possible the history and description of this data and comment on it.
2. Plot the distribution of the response for this dataset and comment.
3. Comment on the shape of this dataset in terms of the sample size and the dimensionality of the input space
4. Comment succinctly from the statistical perspective on the type of data in the input space. It is absolutely crucial here to provide as many details as possible including distributional aspects via boxplots and others on randomly selected subsets of variables.

Exercise 3

Consider a linear regression analysis in $\mathbb{R}^2$ with $Y_i = \theta_1 \mathrm{x}_{i1} + \theta_2 \mathrm{x}_{i2} + \sigma Z_i$ where $Z_i \sim \mathcal{N}(0,1)$ for $i = 1, \cdots, n$, and $\sigma \in \mathbb{R}_+^\star$. You are given the data matrix $\mathbf{X}$ of Equation (4).

$$\mathbf{X} = \begin{bmatrix} 1 & -2 \\ -2 & 1 \\ 4 & 1 \end{bmatrix} \tag{4}$$

You are also given $\mathbf{Y} = (-5, 4, -3)^\top$ and the vector of observed response values.

1. Compute the important $\mathbf{X}^\top \mathbf{X}$
2. Comment on the shape of $\mathbf{X}^\top \mathbf{X}$
3. Find $(\mathbf{X}^\top \mathbf{X})^{-1}$ in the most straightforward way
4. Compute $\widehat{\boldsymbol{\theta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$
5. Compute the vector $\widehat{\mathbf{Y}}$ of estimated responses
6. Compute the vector $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ of residual values
7. Find the value of $\texttt{SSE}(\widehat{\theta})$
8. Find the estimate $\widehat{\sigma^2} = \frac{\texttt{SSE}(\widehat{\theta})}{n-2}$
9. Find and write down $\texttt{Variance}(\widehat{\theta}) = \widehat{\sigma^2}(\mathbf{X}^\top \mathbf{X})^{-1}$
10. Write the matrix $\mathbf{X}^\top \mathbf{X}$ as a function of the identity matrix when the data matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}. \tag{5}$$

Exercise 4

Consider the `gifted children` data set. For this dataset, the response variable is `score`, which contains the score of the child on a battery of tests. The explanatory variables in this case are self explanatory by their names. The goal here is to attempt to build the best linear model that captures the relationship between the response `score` and the other variables

1. Generate an upper triangular pairwise scatterplot for this data, and comment based on the scatterplots regarding which of the explanatory variables are more strongly related to the response. Can you tell from the plot the strongest of all the predictor variables?
2. Generate the correlation matrix for this data (please do not include the p-values in the matrix for now). Which variable does the correlation matrix appear to indicate as the strongest?
3. Plot a histogram of the response variable and also perform a test of normality on it
4. Perform a simple linear regression (SLR) model fitting featuring the response and the variable you singled out as the most important.
5. Generate the 4 residual analysis plots and comment on the suitability of your built model. Are there observations that might have badly influenced the estimation of your parameters?
6. Let's assume for a little while that you are to use the above SLR model. Then give an interpretation of your estimated slope in layman's term.
7. Generate both confidence bands and the prediction bands for this model and provide intelligent comments on what you get.
8. Build a multiple linear regression (MLR) model for this data comprising of all the provided explanatory variables