HW1 : Statistical Learning – Solutions

Problem 1. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide $n$ and $p$.

(a) Some disease is related to abnormal quantities of a given protein. Researchers want to identify the genes that are substantially involved in the production of that protein. A preliminary work allows to focus on 50 genes. Medical professionals measure the expression levels of the 50 genes for 500 healthy people and 500 sick people.

Answer. Classification (response is whether a person is sick or healthy), inference, $n = 1000$, $p = 50$

(b) A data set records the amounts of atmospheric concentrations for five specific pollutants in Seattle: carbon monoxide (CO), methane (CH4), particulate matter (PM), chlorofluorocarbons (CFCs), and ammonia (NH3). Data points correspond to quarterly measurements from January 1, 1997 to December 31, 2016. We can find mortality rates over the same period of time. We want to know if air pollution is related to mortality in in Seattle.

Answer. Regression (response is mortality rate), inference, $n = 20 \times 4 = 80$, $p = 5$

(c) Researchers consider launching a new vaccine and wish to know whether it will be safe and effective. For this purpose, they conduct a trial in 3,500 patients. For each patient, we measure 10 characteristics of their blood sample.

Answer. Classification (response is whether the vaccine is safe and effective or not), prediction, $n = 3500$, $p = 10$

Problem 2. A given phenomenon involves three predictors $X_1, X_2, X_3$ and one quantitative response $Y$. The relationship between $Y$ and $X = (X_1, X_2, X_3)$ is modeled as follows

$$Y = f(X) + \varepsilon,$$

where the function $f$ is given by

$$f(X) = \frac{5 \ln(1 + X_2^2)}{2 + X_1X_3},$$

and $\varepsilon$ follows a normal distribution with mean 0 and standard deviation 0.5.
(a) Assume that a statistical learning method is applied to a training set and results in the following estimate \( \hat{f} \) of \( f \)

\[
\hat{f}(X) = 1 - X_1 + 2X_2 + 3X_3
\]

Is \( \hat{f} \) a parametric or a non-parametric model?

**Answer.** Parametric, \( \hat{f} \) has an explicit functional form.

(b) Compute \( \hat{Y} \) the predicted value for an observation \( X = (0, 1, 2)^T \) using \( \hat{f} \).

**Answer.** \( \hat{Y} = \hat{f}(X) = 1 - 0 + 2 \cdot 1 + 3 \cdot 2 = 9 \)

(c) Compute the expected prediction error \( E[(Y - \hat{Y})^2] \) for the observation \( X = (0, 1, 2)^T \) using the reducible-irreducible error decomposition.

**Answer.**

\[
E[(Y - \hat{Y})^2] = [f(X) - \hat{f}(X)]^2 + \text{Var}(\varepsilon) \\
= \left[5 \ln(1 + 1^2) - 9\right]^2 + 0.5^2 \\
\approx 53.06.
\]

(d) It turns out that our estimate \( \hat{f} \) provides poor predictions in practice. What type of statistical method would you use to improve the performance?

**Answer.** We need a more flexible method: either a parametric model with a larger number of parameters, or a non-parametric method if there are enough points in the training set.

**Problem 3.** A company wants to develop a personalized recommendation system based on a movie’s running time (in minutes), production budget (in millions of dollars), and ticket sales (in millions of dollars). The data set below records the features of six movies that a given individual watched, and whether she liked them or not.

<table>
<thead>
<tr>
<th>Movie</th>
<th>Time</th>
<th>Budget</th>
<th>Sales</th>
<th>Like</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>129</td>
<td>160</td>
<td>750</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>91</td>
<td>11</td>
<td>30</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>133</td>
<td>290</td>
<td>1000</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>139</td>
<td>258</td>
<td>890</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>35</td>
<td>470</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>110</td>
<td>75</td>
<td>340</td>
<td>No</td>
</tr>
</tbody>
</table>

We wish to predict whether that person will like a movie whose running time is 121 minutes, production budget $12 million, and ticket sales $81 million using \( K \)-nearest neighbors.
(a) Compute the Euclidean distance between each movie in the training set and the test movie.

**Answer.** The Euclidean distances to the test movie are given in the table below:

<table>
<thead>
<tr>
<th>Movie</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>685.22</td>
<td>59.18</td>
<td>938.11</td>
<td>845.77</td>
<td>390.24</td>
<td>266.78</td>
</tr>
</tbody>
</table>

where for example the distance between Movie 1 and the test movie is equal to $\sqrt{(129 - 121)^2 + (160 - 12)^2 + (750 - 81)^2}$.

(b) What is your prediction with $K = 1$?

**Answer.** The closest movie to the test movie is Movie 2. We thus predict that the person will like the test movie.

(c) What is your prediction with $K = 3$?

**Answer.** The three movies closest to the test movie are Movies 2, 5, and 6. In this case, we thus predict that the person will not like the test movie.

(d) What shape do you expect the Bayes decision boundary to be like? Justify your answer.

**Answer.** The Bayes decision boundary is likely to be wiggly since the relationship between the response (the individual’s opinion on a movie) and the predictors is highly nonlinear.

(e) Think of two other predictors that would improve the performance of the recommendation system.

**Answer.** Be creative ;-)!

**General note.** Actually, the study is not quite right! One should normalize the data set as a preprocessing step. Otherwise a variable with high variance would heavily skew the classifier.