HW1 : Statistical Learning

Directions. Show and explain all work to receive full credit. Homework is due on Thursday, April 6th at the beginning of class by 12:00 pm.

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.

(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.

(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.

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

\[
f(X) = \frac{5 \ln(1 + X_2^2)}{2 + X_1 X_3}
\]

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

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

(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?

**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>200</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.

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

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

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

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