MATH 373: Introduction to Machine Learning
Homework 3

Computational Problems

The following data set is coming from a Kaggle competition that came out on November 12, 2015. Here is the description from the competition:

Description:

What predicts an article popularity? In this competition, you'll be using data from Mashable (mashable.com), a digital media website, defined (by themselves) as a leading source for news, information and resources for the Connected Generation.

Time magazine noted Mashable as one of the 25 best blogs in 2009, and described it as a "one stop shop" for social media. As of November 2015, [Mashable] had over 6,000,000 Twitter followers and over 3,200,000 fans on Facebook. In this problem, you'll use data from thousands of articles published by Mashable in a period of two years to see which variables predict the popularity of an article.

Load and read more about the data:

1. Load the data *OnlineNewsPopularityTraining.csv* from the Canvas website, which contains a large portion of the data set from the above competition.

2. Read the variable descriptions for the variables at this website:
`https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity#`

3. I have added a binary label to the data set *popular*, which specifies whether or not each website is considered a popular website.

Prepare the data:

1. Remove the variables *shares*, *url* and *timedelta* from the dataset.

Questions: The aim of this computational exercise is to prepare a classifier to predict whether or not a new website will be popular.

1. Using this data, prepare a classifier using the following methods:

   (a) K-Nearest Neighbors

   (b) Linear Discriminant Analysis

   (c) Quadratic Discriminant Analysis

   (d) Empirical Naïve Bayes Classifier

   (e) Logistic Regression

For each of the methods, carefully describe how you choose any thresholds or tuning parameters. Use appropriate plots and documentation to describe your results. Feel free to remove variables or perform dimension reduction if you think it will help your predictions. I am being intentionally vague here because I want to see how you would handle such a data set in practice. All I ask is that for every step you perform, be principled. And try your best to get the best peformance.

2. Download the test data *OnlineNewsPopularityTest.csv* from Canvas. Apply your classifiers from (1) to the data set. Discuss the performance of each method using assessment measures such as MSPE, sensitivity, and specificity. Discuss which classifier you prefer and why.

Conceptual Problems

1. Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was 10, while the mean of those that didn't was 0. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was 7 last year.

2. Let $(X, Y) \in \mathbb{R} \times \{-1, +1\}$ be a random predictor-response pair. Suppose that Y has prior probabilities $\pi = \mathbb{P}(Y = +1)$ and $1 - \pi = \mathbb{P}(Y = -1)$, and that X is continuous with marginal density f and class conditional densities $f_1 = f(x \mid Y = +1)$ and $f_{-1} = f(x \mid Y = -1)$.

   a. Derive an expression for the Bayes rule $\phi^*(x)$ in terms of the logarithm of the ratio $\pi f_1(x)/(1-\pi)f_{-1}(x)$.

   Suppose that the distribution of X given $Y = +1$ is $\mathcal{N}(\mu_1, \sigma^2)$, and that the distribution of X given $Y = -1$ is $\mathcal{N}(\mu_{-1}, \sigma^2)$. Suppose that $\mu_1 > \mu_{-1}$.

   b. Using the result of part (a), find an expression for the Bayes rule $\phi^*(x)$ in terms of the parameters π, μ_{-1}, μ_1, and σ^2.

   c. What is the form of the rule in part (b) when $\pi = 1/2$? Explain why this makes intuitive sense.

   d. Suppose for simplicity that $\mu_{-1} = -\mu_1 = -u$ for $u > 0$. What form does the rule take when u increases (tends to infinity), and in particular, how does the rule depend on π versus $1 - \pi$? An informal but clear answer is fine.

3. Let $(X, Y) \in \mathbb{R}^2 \times \{-1, +1\}$ be a random predictor-response pair. Suppose that the predictor X is a pair (X_1, X_2) where $X_1, X_2 \in [0, 1]$ are independent, X_1 is uniform on $[0, 1]$, and X_2 has density $g(x_2) = 3x_2^2$ for $0 \le x_2 \le 1$. Suppose that $\eta(x_1, x_2) = (x_1 + x_2)/2$.

a. Find the Bayes rule ϕ^* for this problem and identify its decision boundary.

b. Find the unconditional density of X

c. Find the Bayes risk associated with (X, Y)

d. Find the prior probability that $Y = +1$.

e. Find the class-conditional density of X given $Y = +1$.

4. Consider a classification problem in which the conditional probability $\mathbb{P}(Y = 1 \mid X = x)$ is defined implicitly via the equation

$$\operatorname{logit}(\eta(x : \beta)) \;=\; \beta^t x \tag{1}$$

where $\operatorname{logit}(u) = \log[u/(1-u)]$ for $0 < u < 1$ is the logistic (or logit) function.

a. Show that, by inverting the relation (1) we have

$$\eta(x : \beta) \;=\; \frac{e^{\beta^t x}}{1 + e^{\beta^t x}} \;=\; \frac{1}{1 + e^{-\beta^t x}}$$

b. Consider the case that β and x are one-dimensional, and therefore real valued. Find the partial derivatives $\partial \log(\eta(x : \beta))/\partial\beta$ and $\partial^2 \log(\eta(x : \beta))/\partial^2\beta$, and show that the second partial is always negative.

5. Suppose that we collect data for a group of students in an MSAN course with variables X_1 = hours studied per week, X_2 = undergrad GPA, and Y = receives an A. We fit a logistic regression and produce estimated coefficients $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, and $\hat{\beta}_2 = 1$.

a. Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

b. How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?