Computational Statistics
Assignment 4

by James D. Wilson (University of San Francisco)

Directions: For all questions in this assignment, fully answer any question that is asked. Late assignments will automatically have 10 points deducted for each day that they are late.

Quantitative Questions

1. **Bayes Classifiers**: 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 $0 \leqslant x_2 \leqslant 1$. 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. (*Hint*: In the binary case, based on the Bayes decision rule in class, we will decide to classify a new label as $+1$ if $\mathbb{P}(Y = +1 \mid X = x) \geqslant 0.5$ and will classify as -1 otherwise)

   (b) Find the unconditional density of X (*Hint*: I mean find the joint density of (X_1, X_2))

   (c) Find the prior probability that $Y = +1$.

   (d) Find the class-conditional density of X given $Y = +1$.

2. **Logistic Regression**: Consider a classification problem in which the conditional probability $\mathbb{P}(Y = 1 \,|\, X = x)$ is defined implicitly via the equation

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

   where $\text{logit}(u) = \log[u/(1-u)]$ 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. What does this result suggest about estimation of a logistic regression model?

   (c) 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 $\hat{\beta}_2 = 1$.

      i. 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.

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

Computational Questions

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 compare and contrast the performance of logistic regression with Bayesian logistic regression for binary classification. Recall from our discussion in class, logistic regression can be viewed as treating the observed binary data $\mathbf{y}$ as counts (z_1, z_2) where z_j = number of data points in $\mathbf{y}$ that is of class j. As a result, logistic regression can be viewed as the following model:

$$z_j \sim Binomial(n, p_j)$$
$$\text{logit}(p_j) = X\beta$$

In the standard logistic regression setting (as seen in the above quantitative exercise), we treat β as fixed, unknown constants. However, we can alternatively treat β as a vector of random variables according to some well-defined prior distribution. From class, we showed that treating β_j as a Beta(α_j, γ_j) random variable provides a conjugate prior, and is known as the *Beta-Binomial* model.

1. Use `R` to fit the following methods on the *Mashable* training data:

   (a) Standard Logistic Regression

   (b) Bayesian Logistic Regression using the *Beta-Binomial* model described above. Importantly here, you will have to think through how to provide (or estimate) the hyperparameters α and γ by using the data that you observe X. This is the part where there is no correct answer, but up to your own creativity. *You don't have to calculate the posterior parameters analytically here, you can simply use the results for the Beta posterior talked over in class or from some other source!*

   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. So, you'll need some assessment of how well a model is performing. 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.