New York University Shanghai

Problem Set 3

BUSF-SHU 210: Business Analytics (Spring 2019)

1. Please briefly answer the following questions:

(a) Assume that a fitted logistic regression model is

$$\mathbb{P}(Y_i = 1|X_i) \approx \frac{\exp(1.5 + 0.7X_{i1} - 1.5X_{i2} + 0.9X_{i3})}{1 + \exp(1.5 + 0.7X_{i1} - 1.5X_{i2} + 0.9X_{i3})}$$

The model predicts that $Y_i = 1$ if $\mathbb{P}(Y_i = 1|X_i) \geq 0.5$ and predicts $Y_i = 0$ if $\mathbb{P}(Y_i = 1|X_i) < 0.5$. Under what condition of $X_i = (X_{i1}, X_{i2}, X_{i3})$ will the model predict $Y_i = 1$? In particular, if $X_i = (1, 2, 2)$, what is the prediction of Y_i with the model?

(b) Consider a classification problem with dependent variable $Y \in \{0, 1, 2, 3, 4\}$. We define $p_k = \mathbb{P}(Y = k)$ as the probability that $Y = k$ $(k = 0, 1, 2, 3, 4)$, where $\sum_{k=0}^{4} p_k = 1$. The probability distribution of Y is summarized in the following table:

k	0	1	2	3	4
p_k	0.3	0.15	0.1	0.25	0.2

We denote the 0-1 loss of predicting that $\hat{Y} = k$ as $l_k = \mathbb{P}(\hat{Y} \neq Y|\hat{Y} = k)$. What is the value of l_k for $k = 0, 1, 2, 3, 4$?

(c) Explain in your own words what is bias-variance trade-off. What will happen to the bias and variance of a model if the model complexity increases.

2. Click-Through Rate Prediction

In this problem, we will build a model to predict the click-through rate of an advertising demand-side platform (DSP). Your job is to predict whether an advertisement will be clicked by a mobile phone customer.

This dataset contains the following 8 variables:

- dc = Click-through. 1 means clicked, 0 means not clicked. This is the outcome variable the DSP cares about.
- $atype$ = the AdExchange platform where the advertisement slot is traded.
- $bidf$ = the lowest bid price of the advertisers on the AdExchange.
- $instl$ = full-screen advertisment. 1 means full-screen, 0 means half-screen.
- isp = code for the telecommunications company of the customer. 0 means unknown, 1 means China Mobile, 2 means China Unicom, 3 means China Telecome.
- nt = the broadband cellular network technology code, 0 means unknown, 1 means Wi-Fi, 2 means 2G, 3 means 3G, 4 means 4G, 5 means 5G
- mfr = the cellphone manufacturer.
- $period$ = time period of a day.

Use the data set "`RTB.csv`" to address the following questions.

(a) Build a $k-$NN model and use cross-validation to select the best number of neighbors and the covariates to include into the model.

(b) Build an L_2-regularized logistic regression model and use cross-validation to select the best regularization parameter and the covariates to include into the model?

(c) Build an L_1-regularized logistic regression model and use cross-validation to select the best regularization parameter and covariates?

(d) Consider the models you build in parts (a)-(c). Which model you would recommend to predict the click-throughs of this DSP? Report your model's generalization error (0-1 loss) and out-of-sample ROC-AUC.

3. Bias-Variance Trade-off

In this question, your task is to examine the bias-variance trade-off. Assume that, the data is generated according to the following model:

$$Y = f(X) + \epsilon = X^3 + 2X^2 + 3X + 1 + \epsilon, \tag{1}$$

where X is the independent variable normally distributed with mean 0 and standard deviation 1, ϵ is the noise term with mean 0 and standard deviation 10, Y is the dependent variable. The data set `BVTradeoff_train.csv` is the training set containing 2 variables:

- $X.train$: The independent variables in the training set, randomly sampled from the distribution $N(0, 1)$
- $Y.train$: The dependent variables in the training set, generated according to the model (1) with $X.train$ as the independent variable

Each row of the training set represents a data point in this data set. The data set `BVTradeoff_test.csv` is the testing set containing 2 variables:

- $X.test$: The independent variables in the testing set, randomly sampled from the distribution $N(0, 1)$
- $Y.test$: The dependent variables in the testing set, generated according to the model (1) with $X.test$ as the independent variable

Each row of the testing set represents a data point in this data set.

As a general hint for this problem, recall that, if we want to estimate the expectation of a random variable Z, one approach is to use the sample mean of an independent sample of this independent variable (this is called the law of large numbers), i.e.,

$$\mathbb{E}(Z) \approx \bar{Z} = \frac{1}{n}\sum_{j=1}^{n} Z_j, \text{ where } Z_j \text{ is independently drawn from the same distribution as } Z.$$

(a) Fit five models $\hat{f}_i(\cdot)$ $(i = 1, 2, 3, 4, 5)$, where $\hat{f}_i(X)$ is a polynomial of degree i, i.e.,

$$\begin{aligned}
\hat{f}_1(X) =& \hat{\beta}_0^1 + \hat{\beta}_1^1 X \\
\hat{f}_2(X) =& \hat{\beta}_0^2 + \hat{\beta}_1^2 X + \hat{\beta}_2^2 X^2 \\
\hat{f}_3(X) =& \hat{\beta}_0^3 + \hat{\beta}_1^3 X + \hat{\beta}_2^3 X^2 + \hat{\beta}_3^3 X^3 \\
\hat{f}_4(X) =& \hat{\beta}_0^4 + \hat{\beta}_1^4 X + \hat{\beta}_2^4 X^2 + \hat{\beta}_3^4 X^3 + \hat{\beta}_4^4 X^4 \\
\hat{f}_5(X) =& \hat{\beta}_0^5 + \hat{\beta}_1^5 X + \hat{\beta}_2^5 X^2 + \hat{\beta}_3^5 X^3 + \hat{\beta}_4^5 X^4 + \hat{\beta}_5^5 X^5
\end{aligned}$$

Take a screen shot of the summary of each model you fit with the training data.

(b) Make predictions under the testing data set. For each model you fit in part (a), draw the histogram for the squared error of each data point in the testing set. What observations do you have?

(c) Furthermore, estimate the expected errors $\mathbb{E}(Y - \hat{f}_i(X))^2$ $(i = 1, 2, 3, 4, 5)$ of each model built in part (a) using the testing data set. What observations do you have? Which model has the smallest squared error?

(d) The bias of a model $\hat{f}_i(\cdot)$ is defined as $\mathbb{E}(f(X) - \hat{f}_i(X))^2$ $(i = 1, 2, 3, 4, 5)$, i.e, the expected squared difference between the prediction under the "true" model and the prediction under the fitted model. Use the testing data set to estimate the bias of each model you built in part (a). What observations do you have? Which model has the smallest bias? Does this model has the smallest squared error as well? Why?

(e) The variance of a model $\hat{f}_i(\cdot)$ is defined as $\mathbb{E}(\hat{f}_i(X) - \mathbb{E}[\hat{f}_i(X)])^2$, i.e, the variance of the predicted outcome under the fitted model. Use the testing data set to estimate the variance of each model you built in part (a). What observations do you have? Which model has the smallest variance? Does this model has the smallest squared error as well? Why?

(f) Plot how the squared error, bias, and variance change when you increase the model complexity (i.e., the degree of the polynomial). What implications can you have from the plot your draw?