Machine Learning for Finance (FIN 570)
Midterm Exam

Instructor: Jaehyuk Choi

2018-19 Module 1 (2018. 10. 19)

1. (4 points) **(Bias vs variance trade-off)** For each of the following learning method, comment whether bias or variance (or both) is high.
   (a) Unpruned decision tree (i.e. decision tree with unlimited number of leaves)
   (b) Support vector machine with high allowance of violation
   (c) Linear regression for 10 samples in training dataset with 9 features and 1 response.
   (d) K-NN with large K.

   Solution:
   (a) High variance
   (b) High bias
   (c) High variance
   (d) High bias

2. (3 points) **(Probability)** Suppose that you are an equity analyst and you arrange two baskets of stocks for your clients. The clients pick a stock to invest in the following way: first, she randomly choose one basket and, then, randomly choose one stock in the basket. (The clients can not see the stocks in each basket.)

   Suppose that 101 stocks are currently traded in Shenzhen Stock Exchange (SZSE), and you know that only 51 of the stocks will bring positive return. (You know the exact list of those stocks.) How can you arrange the two baskets in order for your client to maximize the probability of selecting a stock with positive return? What is the maximized probability? The only restriction is that there should be at least one stock in each basket.

   Solution: You just put one winning stock in one basket and put the rest in the other basket. Then, the probability for picking a winning stock is

   $$\frac{1}{2} \times 1 + \frac{1}{2} \times \frac{50}{100} = 75\%.$$

   This value is maximum you can get because no matter how you arrange the stocks in the two baskets, (i) the basket with higher probability cannot have probability higher than 100%

and (ii) the basket with lower probability cannot have probability higher than 50%. (i) is obvious. (ii) is also true because the lower side probability has to be lower than or equal to the average 51/101. Considering that there is not common denominator of 51 and 101, it is impossible to arrange 51/101 probability for the two baskets exactly. The next available scenario is 50/100.

3. (2 points) **(Logistic Regression)** Suppose we collect data for a group of students in the MLF class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A grade. We fit a logistic regression and produce estimated coefficient, $w_0 = -6$, $w_1 = 0.05$, $w_2 = 1$.
   (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?

Solution: The probability is given as

$$P(X_1, X_2) = \frac{1}{1 + \exp(6 - 0.05X_1 - X_2)}.$$

(a)

$$P(X_1 = 40, X_2 = 3.5) = \frac{1}{1 + \exp(6 - 2 - 3.5)} = \frac{1}{1 + \sqrt{e}} \approx 38\%.$$

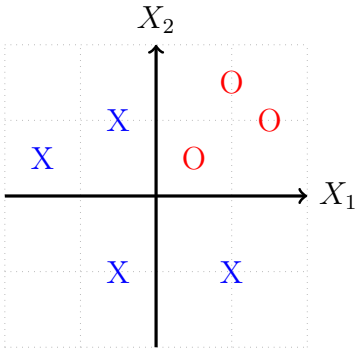
(b) $X_1 = 2.5/0.05 = 50$ hours because

$$P(X_1, X_2 = 3.5) = \frac{1}{1 + \exp(2.5 - 0.05X_1)} = \frac{1}{2}.$$

4. (2 points) **(Test accuracy)** Suppose that we take a dataset, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. $K = 1$) and get an average error rate (averaged over both test and training datasets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

Solution: The error rate for the training dataset for 1-NN is 0%, which implies that the error rate for the test dataset is 36%. Therefore, logistic regression should be preferred.

5. (3 points) **(Decision Tree)** Suppose that we want to apply decision tree to the following classification problem (O vs X):



(a) Compute the Gini impurity of the raw dataset before splitting?

(b) Let us consider the two criteria: (a) $X_1 > 0$ and (b) $X_2 > 0$. For the first split, which criteria should be selected? Compute the information gain for each.

Solution:

(a) Gini impurity:

$$I_G = 2\frac{3}{7}\frac{4}{7} = \frac{24}{49} \quad (\approx 0.49)$$

(b) After split (a), the 3-vs-1 branch ($X_1 > 0$) has Gini impurity as

$$I_G = 2\frac{3}{4}\frac{1}{4} = \frac{6}{16} \quad (\approx 0.38).$$

Therefore, the information gain is

$$\frac{24}{49} - \left(\frac{4}{7}\frac{6}{16} + \frac{3}{7}\,0\right) = \frac{27}{98} \quad (\approx 0.28)$$

For split (b), the 3-vs-2 branch ($X_2 > 0$) has Gini impurity as

$$I_G = 2\frac{3}{5}\frac{2}{5} = \frac{12}{25} \quad (\approx 0.48).$$

Therefore, the information gain is

$$\frac{24}{49} - \left(\frac{5}{7}\frac{12}{25} + \frac{2}{7}\,0\right) = \frac{36}{245} \quad (\approx 0.15).$$

The first split should be (a) because it has higher information gain. It is in line with intuition.

6. (3 points) **(PCA analysis)** Assume that the covariance matrix of 3-dimensional input (X_1, X_2, X_3) is given as

$$\mathbf{\Sigma} = \begin{pmatrix} 7 & 4 & 0 \\ 4 & 9 & 4 \\ 0 & 4 & 11 \end{pmatrix}.$$

(a) What is the total variance, i.e., $\sum_{k=1}^{3} \mathrm{Var}(X_k)$?

(b) Assume that we obtained the following three PCA component **without** correct order:

$$u_1 = \frac{1}{3}(2, -2, 1), \quad u_2 = \frac{1}{3}(1, 2, 2), \quad u_3 = \frac{1}{3}(2, 1, -2).$$

For example, the projection on the first PCA component is given as $P_1 = \frac{2}{3}X_1 - \frac{2}{3}X_2 + \frac{1}{3}X_3$. Calculate the variance of the projected features, $\mathrm{Var}(P_k)$ for $k = 1, 2, 3$. Order the PCA components correctly according to the computed variance.

(c) If we use the first two (correctly ordered) PCA components only, what is the ratio of explained variance by those two components over the total variance in (a)?

Solution:

(a) You can simply read off the diagonal values:

$$\sum_{k=1}^{3} \mathrm{Var}(X_k) = 7 + 9 + 11 = 27.$$

(b) The variance of the PCA components are same as the corresponding eigenvalues. Therefore, we check the eigenvalues first:

$$\boldsymbol{\Sigma}\, \boldsymbol{U} = \begin{pmatrix} 7 & 4 & 0 \\ 4 & 9 & 4 \\ 0 & 4 & 11 \end{pmatrix} \times \frac{1}{3} \begin{pmatrix} 2 & 1 & 2 \\ -2 & 2 & 1 \\ 1 & 2 & -2 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 6 & 15 & 18 \\ -6 & 30 & 9 \\ 3 & 30 & -18. \end{pmatrix}$$

Therefore, the eigenvalues for u_1, u_2, and u_3 are 3,15, and 9, respectively. The PCA component should be ordered as u_2, u_3, and u_1.

Alternatively, we directly compute the variance:

$$\begin{aligned} \mathrm{Var}(P_1) &= \frac{1}{9}(4\mathrm{Var}(X_1) + 4\mathrm{Var}(X_2) + \mathrm{Var}(X_3) \\ &\qquad - 8\mathrm{Cov}(X_1, X_2) + 4\mathrm{Cov}(X_1, X_3) - 4\mathrm{Cov}(X_2, X_3)) \\ &= \frac{1}{9}(4 \cdot 7 + 4 \cdot 9 + 11 - 8 \cdot 4 + 4 \cdot 0 - 4 \cdot 4) = 3. \end{aligned}$$

In the same way, we get $\mathrm{Var}(P_2) = 15$ and $\mathrm{Var}(P_3) = 9$.

(c) The ratio of the explained variance is $(15 + 9)/27 = 8/9 \approx 89\%$.

7. (3 points) **(XOR gate, K-NN, LOOCV)** The table below provides a training dataset containing four observations: two predictors and one classification response variable.

Obs.	X_1	X_2	Y
1	1	1	False
2	1	-1	True
3	-1	1	True
4	-1	-1	False

Suppose that we wish to use this dataset to make a prediction for Y using 1-NN (K-NN with $K = 1$). Here, we use the L-1 norm to compute the distance instead of the L-2 (Euclidean) norm. So, the distance between the two observations, (x_1, x_2) and (y_1, y_2), is given as

$$D = |x_1 - x_2| + |y_1 - y_2|.$$

(a) What is the prediction for Y when a new sample point is given as $(X_1, X_2) = (2, 1)$?

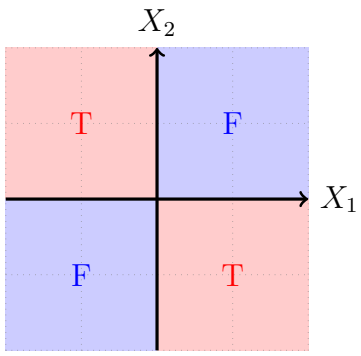
(b) Draw the region of classification on the (X_1, X_2) coordinate.

(c) For the training set given in the table above, what is the test accuracy if LOOCV method is applied? How does your answer change if you apply 2-NN or 3-NN?

Solution:

(a) The obs. 1 is the nearest neighbor to $(2, 1)$, with the distance being $|2-1|+|1-1| = 1$. So, $(2, 1)$ is predicted as **False**.

(b) The classification region is as below:



(c) 0%. For example, the nearest neighbor of $(1, 1)$ is either $(1, -1)$ or $(-1, 1)$ (excluding itself), both False. The 0% accuracy remains unchanged for 2-NN and 3-NN.