# Machine Learning for Finance (FIN 570)
# Midterm Exam

Instructor: Jaehyuk Choi  Student Name and ID: _____

2022-23 Module 3 (2023. 4. 4.)

The acronyms are defined same as in the class. For example, machine learning (**ML**), logistic regression (**LR**), principal component analysis (**PCA**), support vector machine (**SVM**), neural network (**NN**), etc. If you are not sure, please ask.

1. (9 points) Multiple choice questions about the bias and variance tradeoff. You do not have to explain your answer.
   (a) Different models trained using different training datasets derived from same population has high accuracy or make similar accurate predictions, then the models are said to have ____. **A. low bias** B. high bias
   (b) A model having high variance will most likely be suffering from ____.
   A. underfitting **B. overfitting**
   (c) Unpruned decision trees trained on different training datasets derived from same population usually have ____. A. low variance **B. high variance**
   (d) In case the models behave differently (different performance) when trained with different dataset derived from the same population, the models are said to have ____. A. low variance **B. high variance**
   (e) Linear regression models trained on different training datasets derived from same population usually have ____. A. low bias **B. high bias**
   (f) A model having high bias will most likely be suffering from ____.
   **A. underfitting** B. overfitting
   (g) Ideally, the model should have ____ bias, ____ variance.
   A. high, low **B. low, low** C. high, high D. low, high
   (h) Different models trained using different training datasets derived from same population has lower accuracy but make similar or consistent predictions, then the models are said to have ____.
      **A. high bias, low variance**
      B. low bias, low variance
      C. high bias, high variance
      D. low bias, high variance
   (i) If models trained on different training datasets derived from same population have different performance or make distinctly different predictions on unseen dataset, the models are said to have ____.
   A. low variance **B. high variance**

**Solution:** Source: vitalflux.com

2. (6 points) Consider a (one-dimensional) linear regression model without an intercept:

$$Y \sim X\beta.$$

   (a) (3 points) Find $\hat{\beta}$ that minimize the MSE from the given training dataset, $(x_i, y_i)$ for $1 \le i \le n$.

   (b) (3 points) What is the prediction value at a new data point $x_*$ (i.e., $\hat{y}_* = x_*\hat{\beta}$) ?. Express $\hat{y}_*$ in the form of the kernel regression:

$$\hat{y}_* = \sum_{i=1}^{n} K(x_i, x_*)\, y_i.$$

   What is the functional form of $K(x, y)$?

**Solution:** This question is modified from Exercise 3.5 (p. 121).

(a) The RSS and its derivative are

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 \quad \text{and} \quad \mathrm{RSS}'(\beta) = -2\sum_{i=1}^{n} x_i\, y_i + 2\beta\sum_{i=1}^{n} x_i^2.$$

The optimal $\hat{\beta}$ satisfies $\mathrm{RSS}'(\hat{\beta}) = 0$:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2}$$

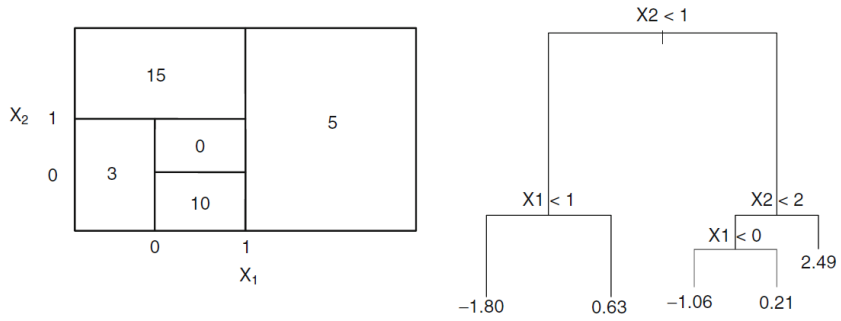
(b) The prediction value at $x_*$ is

$$\hat{y}_* = x_*\,\hat{\beta} = \frac{\sum_{i=1}^{n}(x_* x_i) y_i}{\sum_{i=1}^{n} x_i^2}.$$

Therefore, the kernel function is given by

$$K(x_i, x_*) = \frac{x_* x_i}{\sum_{i=1}^{n} x_i^2} \quad \text{or} \quad K(x, y) = \frac{x\, y}{\sum_{i=1}^{n} x_i^2}.$$

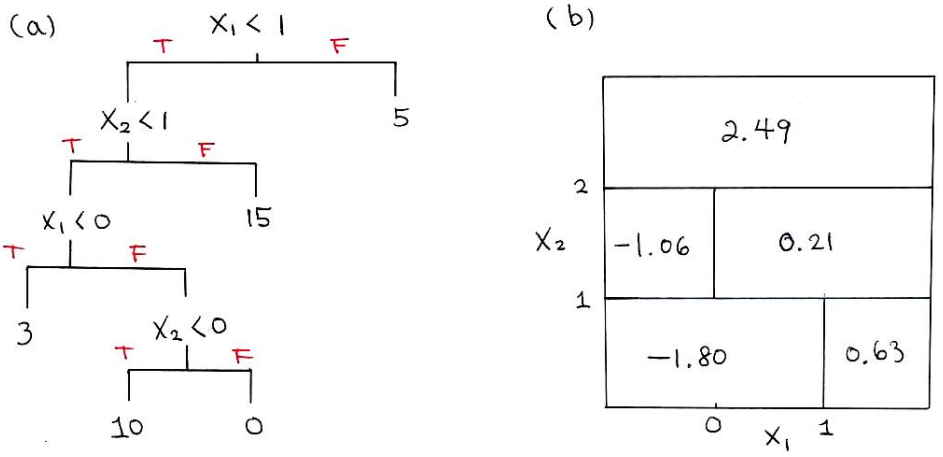
3. (4 points) This question is about the tree regression.



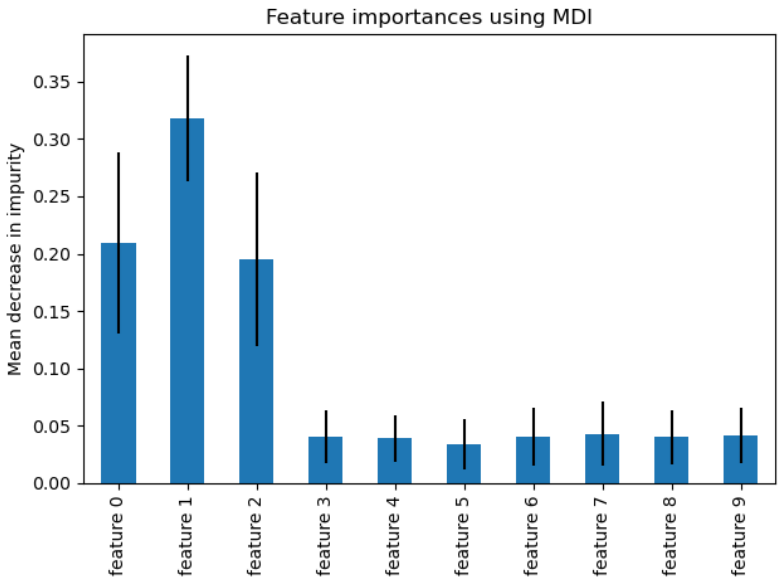

(a) (2 points) Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of the figure. The numbers inside the boxes indicate the mean of $Y$ within each region.

(b) (2 points) Create a diagram similar to the left-hand panel of the figure, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.

**Solution:**




4. (6 points) Although not mentioned in the class, it is possible to measure **feature importance** from random forest algorithm. Below is an example of the feature importance obtained from **RF**:




Source: scikit-learn.org.

(a) (3 points) Describe how the feature importance is computed. (You may rephrase the corresponding section from the **PML** textbook.)

(b) (3 points) Therefore, **RF** can be used as a feature selection mechanism. In the example figure, we can select `feature1` as it is the most important feature, if we must select only one feature. This feature can be used as an input to other ML algorithms. However, the feature selection based on the feature importance is not perfect when some features are highly correlated. Now assume that `feature0` and `feature2` are perfectly correlated features (i.e., 100% correlation). First, explain why the feature importance of the two is not identical (although very close). Second, what would be the approximate feature importance value of `feature0` if we remove the redundant `feature2` from **RF**?

**Solution:**

(a) In RF, the feature importance is measured as the averaged impurity decrease computed from all decision trees in the forest, without making any assumptions about whether our data is linearly separable or not.

(b) The feature importance of `feature0` and `feature2` are not exactly same because the RF algorithm is *random*. Because RF randomly selects features, the number of trees in which `feature0` and `feature2` are selected are not identical. Therefore, the feature importance values are not identical.

If `feature2` were removed from RF, the feature importance of `feature2` would moved to `feature0`. Therefore, the feature importance of `feature0` would be close to 40%.

5. (5 points) Here we explore the maximal margin classifier on a toy data set. We are given n = 7 observations in p = 2 dimensions. For each observation, there is an associated class label.

| Obs. | $X_1$ | $X_2$ | Y |
|---|---|---|---|
| 1 | 3 | 4 | Red |
| 2 | 2 | 2 | Red |
| 3 | 4 | 4 | Red |
| 4 | 1 | 4 | Red |
| 5 | 2 | 1 | Blue |
| 6 | 4 | 3 | Blue |
| 7 | 4 | 1 | Blue |

(a) (3 points) Sketch the optimal separating hyperplane (mark the points on $(x, y)$ axis), and provide the equation for this hyperplane in the form of:

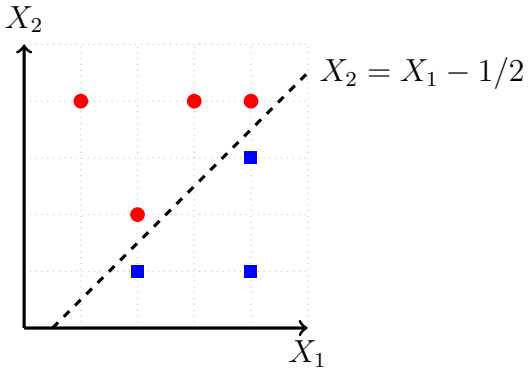
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0.$$

Provide the values for $\beta_0$, $\beta_1$, and $\beta_2$. Also, describe the classification rule for the maximal margin classifier (e.g., classify **red** if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$.)

(b) (2 points) Which points are the support vectors for the maximal margin classifier you found in (a). What is the margin of the classifier (i.e., Euclidean distance from the support vectors to the hyperplane)?

**Solution: (ISLR Exercise 9.3)**

(a) The red and blue points are shown below.




The hyperplane is given by

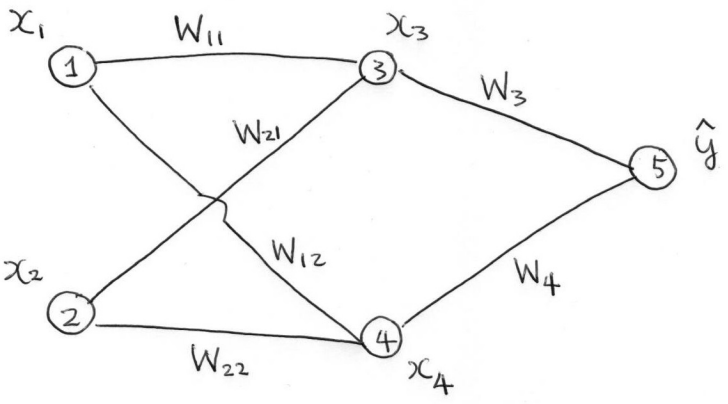
$$X_1 - X_2 - 1/2 = 0 \quad (\beta_1 = 1,\ \beta_2 = -1,\ \beta_0 = -1/2).$$

and the classification rule is

$$\textbf{Red} \text{ if } \quad X_1 - X_2 - 1/2 < 0$$
$$\textbf{Blue} \text{ if } \quad X_1 - X_2 - 1/2 > 0$$

(b) The two red points, $(2, 2)$ and $(4, 4)$, and the two blue points, $(2, 1)$ and $(4, 3)$, are the support vectors. The margin of the classifier is $\frac{1}{2\sqrt{2}}$.

6. (5 points) **(Backpropagation and ReLU)** Consider an NN model consist of $2 \times 2 \times 1$ nodes and sigmoid activation function.




**Node 1, 2 (input layer)** : $x_1$ and $x_2$

**Node 3 (hidden layer)** : $a_3 = x_1 w_{11} + x_2 w_{21}, \quad x_3 = \mathrm{ReLU}(a_3)$

**Node 4 (hidden layer)** : $a_4 = x_1 w_{12} + x_2 w_{22}, \quad x_4 = \mathrm{ReLU}(a_4)$

**Node 5 (output)** : $a_5 = x_3 w_3 + x_4 w_4, \quad \hat{y} = \phi(a_5),$

Here, $\phi(t)$ is the sigmoid function. (The NN is same as 2021 exam Q7, except that we use ReLU for the activation for the node 3 and 4 in the hidden layer.) The loss function for a sample $(x_1, x_2)$ and $y$ is given by

$$J(w_{11}, \cdots, w_3, w_4) = -y \log \phi(a_5) - (1-y) \log(1 - \phi(a_5)).$$

Assume that a data point and response are given by

$$x_1 = 1, \quad x_2 = -1, \quad \text{and} \quad y = 0,$$

and that the weights are initialized by

$$w_{11} = 0.5,\ w_{21} = -0.5,\ w_{12} = -0.5,\ w_{22} = 0.5,\ w_3 = 1,\ w_4 = 1.$$

(a) (3 points) Find the derivative $\dfrac{\partial J}{\partial w}$ for each $w$ (i.e., $w_3$, $w_4$, $w_{11}$, $w_{21}$, $w_{12}$, and $w_{22}$). (You may modify the answers from 2021 exam. You may use $\phi(0.5) = 0.62$, $\phi(1) = 0.73$, or $\phi(1.5) = 0.82$ for calculation.)

(b) (2 points) Which $w$ do you need to update in which direction (up or down)?

**Solution:**

(a) First, we obtain $a_i$ and $x_i$:

$$\begin{aligned} a_3 &= 1, \quad x_3 = \mathrm{ReLU}(a_3) = 1 \\ a_4 &= -1, \quad x_4 = \mathrm{ReLU}(a_4) = 0 \\ a_5 &= 1, \quad \hat{y} = \phi(a_5) = 0.73. \end{aligned}$$

Second, we obtain the "error" $\delta_i$ of each node:

$$\begin{aligned}
\delta_5 &= \hat{y} - y = 0.73, \\
\delta_3 &= \delta_5 w_3 \text{ReLU}'(a_3) = 0.73 \times 1 \times 1 = 0.73, \\
\delta_4 &= \delta_5 w_4 \text{ReLU}'(a_4) = \delta_5 w_4 \times 0 = 0.
\end{aligned}$$

Finally, the derivatives w.r.t. $w$ are given by

$$\begin{aligned}
\frac{\partial J}{\partial w_3} &= \delta_5 x_3 = 0.73 \times 1 = 0.73 \\
\frac{\partial J}{\partial w_4} &= \delta_5 x_4 = 0.73 \times 0 = 0 \\
\frac{\partial J}{\partial w_{11}} &= \delta_3 x_1 = 0.73 \times 1 = 0.73 \\
\frac{\partial J}{\partial w_{12}} &= \delta_4 x_1 = 0 \times 1 = 0. \\
\frac{\partial J}{\partial w_{21}} &= \delta_3 x_2 = 0.73 \times (-1) = -0.73 \\
\frac{\partial J}{\partial w_{22}} &= \delta_4 x_2 = 0 \times (-1) = 0.
\end{aligned}$$

(b) Based on the derivatives obtained in (a), $w_3$ and $w_{11}$ need to be updated down, and $w_{21}$ need to be updated up.