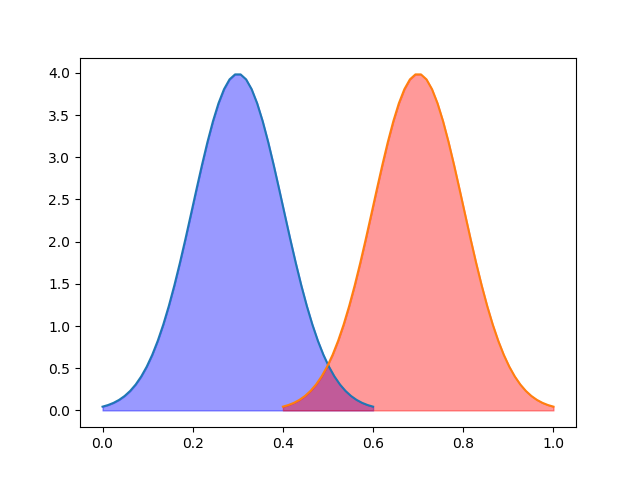
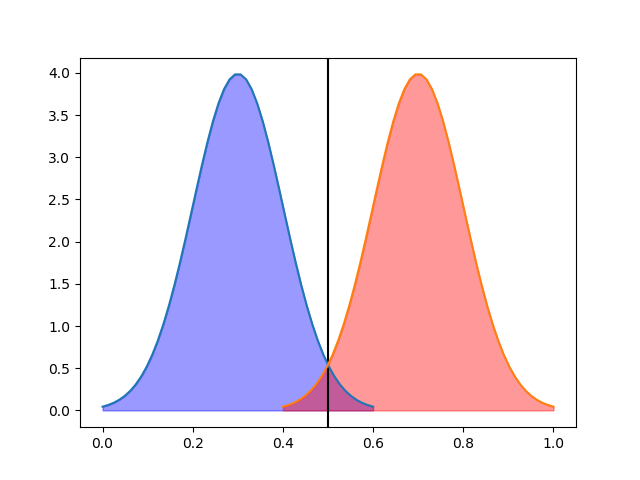
**COMP 642 Assignment --- MODULE 6**

1. Suppose we have a dataset of credit card transaction information and for each transaction record we have a label indicating whether it is fraudulent. We build a Logistic Regression model that can classify whether the transaction is fraudulent or not.

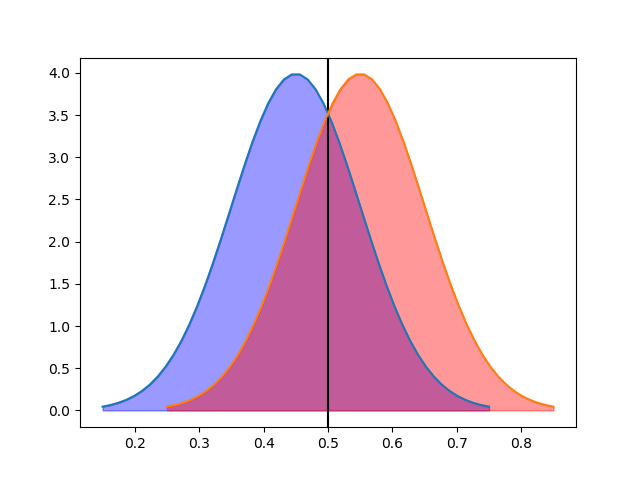
The distribution of the prediction is shown below, the x-axis represents the predicted probability, the y-axis represents the observations (like a histogram). Each blue and red pixel represents a transaction for which you want to predict the fraud status. There are 300 red pixels transactions that were actually fraud, and 300 blue pixels transactions that are not fraud.



1. The output of logistic regression will output a probability. In order to make the classification we need to choose a decision boundary. On the left hand side of the decision boundary the prediction will be all position (fraud), and on the right hand set the prediction will be all negative (not fraud). If the line is drawn at 0.5, which is a common choice in binary classification, 250 transactions are classified as fraud and 250 are classified as non-fraud. The chart is shown below. Calculate the True Positive Rate, False Positive Rate and the position of this decision boundary on ROC curve.



1. Draw the possible ROC curve for this model.
2. If we change the hyper parameters of the logistic regression model, the result prediction probability is shown below. The decision boundary is still 0.5 but 200 data are classified as fraud and 200 are classified as non-fraud. Calculate the same items as in part a.



1. Draw the possible ROC curve for this model.
2. Which model do you think is better and why?
3. Is making the decision boundary equal to 0.5 a good choice for this binary classification problem? Explain why or why not.
4. The line below shows the degree of complexity of a machine learning model. Add to it annotations of: (more/less) bias, variance. (try to not look at the lecture notes), complexity.

< --------------------------------------------------------------------------------------------------------------------------- >

less flexible (lower dimensionality) more flexible (higher dimensionality)

.

1. Execute the hw\_6.ipynb, fill the code cells and answer the questions in notebook. Include the answers in your homework document submission as well as in the Jupyter notebook.