Modern Machine Learning
Midterm Exam

1. *Logistic Regression:* Suppose that you are given the following classification task: predict the target $Y \in \{0, 1\}$ given two real valued features $X_1 \in \mathbb{R}$ and $X_2 \in \mathbb{R}$. Suppose that the decision rule obtained after training is the following:

   Predict $Y = 1$ **if** $w_1 X_1 + w_2 X_2 + w_0 \geq 0$,

   Predict $Y = 0$ **otherwise**,

   where $w_1 = 5$, $w_2 = 3$, and $w_0 = -15$.

   (a) Plot the decision boundary and label the region where the classifiers predicts $Y = 1$ and $Y = 0$.

   (b) Suppose that we learned the above weights using **logistic regression**. Find the mathematical expression for $P(Y = 0|X_1, X_2)$

2. *Bias and Variance:* Figure 1 depicts the training and validation curves of a learner as a function of the model complexity

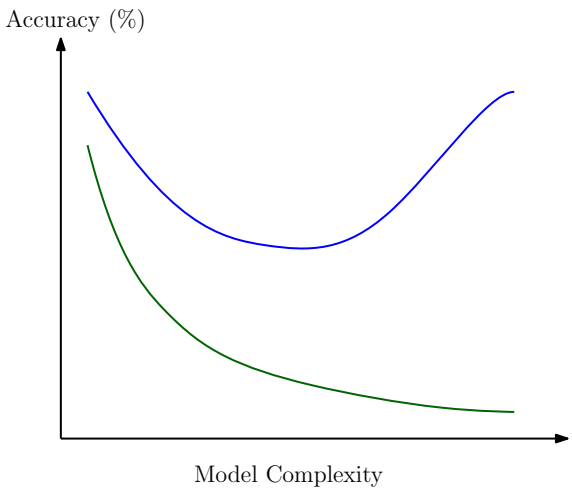

Figure 1: Classification Accuracy (%) as a function of Model Complexity

(a) Which of the curves is more likely to be the training error and which is more likely to be the validation error? Indicate on the graph

(b) In which regions of the graph are bias and variance low and high? Indicate on the graph with four labels: ?low variance?, ?high variance?, ?low bias?, ?high bias?

(c) In which regions does the model overfit or underfit? Indicate on the graph by labeling ?overfit? and ?underfit?

3. Consider the lasso regression problem given by

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

(a) Why does lasso promote sparsity in the solution?

(b) Suppose that you want to perform linear regression and your data consists of N-feature vectors. How would you use lasso regression to select the most relevant $m < N$ features for this model?

4. *Gradient Descent:* Consider a linear regression problem $y = w_1 x + w_0$, with a training set having m examples $(x_1, y_1), \ldots, (x_m, y_m)$. Suppose that we would like to minimize the following cost function:

$$Loss = \frac{1}{m}\sum_{i=1}^{m}(y_i - w_1 x_i - w_0)^3$$

(a) Calculate the gradient with respect to the parameters w_1 and w_0

(b) Write down the pseudo-code to minimize the loss function for this problem using gradient descent.