# CS6375.001 Machine Learning (Fall 2018)
# Midterm Practice Exam

September 28, 2018

**Name:** _____

This exam contains 10 pages (including this cover page) and 5 problems. Check to see if any pages are missing. Enter all requested information on the top of this page, and **put your initials on the top of every page**, in case the pages become separated.

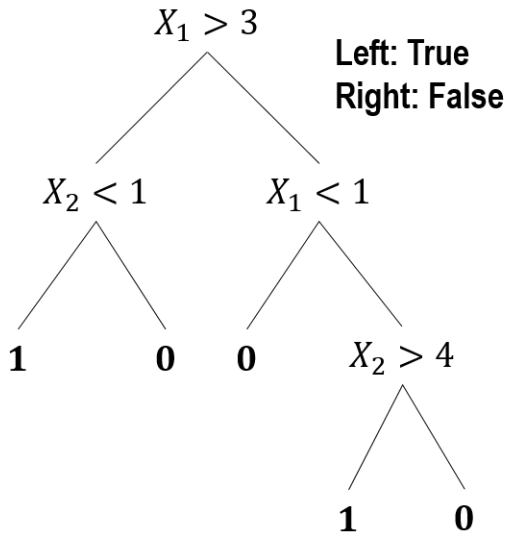
You may **NOT** use books, notes, or any electronic devices on this exam. Examinees found to be using any materials other than a pen or pencil **will receive a zero on the exam** and face possible disciplinary action.

The following rules apply:

- Organize your work, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit.
- To ensure maximum credit on short answer/algorithmic questions, be sure to EXPLAIN your solution.
- Problems/subproblems are not necessarily ordered by difficulty. Be sure to read each of the questions carefully. Do not write in the table to the right.

| Problem | Points | Score |
|---|---|---|
| 1 | 20 | |
| 2 | 15 | |
| 3 | 15 | |
| 4 | 30 | |
| 5 | 20 | |

1. (**Decision Trees**, 20 points) Consider the following decision tree:




a. Draw the decision boundaries defined by this tree. Each leaf is labeled with a number. Write this number in the corresponding region.

b. Give another tree that is syntactically different but defines the same decision boundaries.

c. When a decision tree is grown to full depth, it is more likely to fit the noise in the data. True or False? Explain.

2. (**Naïve Bayes**, 15 points) Consider the following data set, where the classification task is to predict if a car is going to be bought.

| Color | Type | Origin | Buy |
|---|---|---|---|
| Red | Sports | Domestic | Yes |
| Red | Sports | Domestic | No |
| Red | Sports | Domestic | Yes |
| Yellow | Sports | Domestic | No |
| Yellow | Sports | Imported | Yes |
| Yellow | SUV | Imported | No |
| Yellow | SUV | Imported | Yes |
| Yellow | SUV | Domestic | No |
| Red | SUV | Imported | No |
| Red | Sports | Imported | Yes |

a. Assume the Naïve Bayes condition and estimate the parameters using maximum likelihood estimation.

b. For the same domain, apply Laplace correction and estimate the parameters.

c. The logistic function is given by

$$\sigma(x) = \frac{1}{1+e^{-x}}.$$

Show that $\frac{d}{dx}\sigma(x) = \sigma(x) \cdot (1 - \sigma(x))$.

3. (**Evaluation**, 15 points)

a. Let the following predictions be the output of a probabilistic classifier. Draw the ROC curve for this prediction task.

| Predicted Probability | Correct Class |
|---|---|
| 0.96 | + |
| 0.80 | + |
| 0.76 | - |
| 0.66 | + |
| 0.52 | - |
| 0.4 | - |

b. What is the AUC-ROC for this classifier?

4. (**Short Questions**, 30 points) For the following questions, select your answer and **explain your reasoning**. **You will get no points without an explanation**.

a. Suppose you are given a data set of cellular images from patients with and without cancer. If you are required to train a classifier that predicts the probability that the patient has cancer, you would prefer to use decision trees over logistic regression. **True** or **False**?

b. Suppose the dataset in the previous question had 900 cancer-free images and 100 images from cancer patients. If I train a classifier which achieves 85% accuracy on this dataset, it is it a good classifier. **True** or **False**?

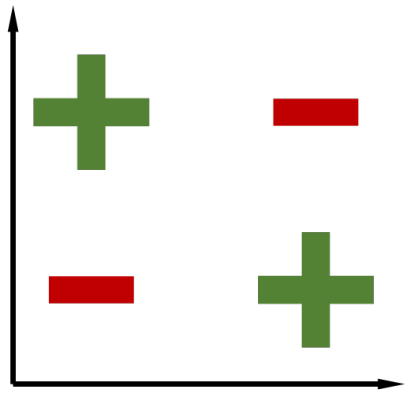
c. A classifier that attains 100% accuracy on the training set and 70% accuracy on test set is better than a classifier that attains 70% accuracy on the training set and 75% accuracy on test set. **True** or **False**?

d. In logistic regression, we model the odds ratio ($\frac{p}{1-p}$) as a linear function. **True** or **False**?

e. If you train a linear regression estimator with only half the data, its bias is smaller. **True** or **False**?

f. Because decision trees learn to classify discrete-valued outputs instead of real-valued functions, it is impossible for them to overfit. **True** or **False**?

g. Which of the following classifiers can perfectly classify the following data set? (i) support vector machines, (ii) logistic regression, (iii) perceptron, (iv) decision trees.



h. When the hypothesis space is richer, over fitting is more likely. **True** or **False**?

5. (**Support Vector Machines**, 20 points) Consider the data set below, with features $x_1$, $x_2$ and labels $y$:

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| 1 | 0 | +1 |
| -2 | 2 | −1 |
| 0 | 3 | −1 |
| 0 | -2 | +1 |
| 2 | 2 | −1 |

a. Find the linear SVM classifier for the data set given above in the form $w_1x_1 + w_2x_2 + b = 0$. (*Hint: Consider plotting the data set and finding a geometric solution*).

b. What is the margin of the linear SVM classifier from the previous question? What are the support vectors for this classifier?

c. One of the most commonly used kernels is the Gaussian kernel

$$\kappa(\mathbf{x}, \mathbf{z}) = \exp\left(-\gamma\|\mathbf{x} - \mathbf{z}\|^2\right).$$

Consider three points $\mathbf{x}$, $\mathbf{z}_1$ and $\mathbf{z}_2$. Geometrically, we know that $\mathbf{x}$ and $\mathbf{z}_1$ are **very far from each other**, and that $\mathbf{x}$ and $\mathbf{z}_2$ are **very close to each other**. Which one of the following is true, and why?

(a) $\kappa(\mathbf{z}_1, \mathbf{x})$ will be close to 1 and $\kappa(\mathbf{z}_2, \mathbf{x})$ will be close to 0.

(b) $\kappa(\mathbf{z}_1, \mathbf{x})$ will be close to 0 and $\kappa(\mathbf{z}_2, \mathbf{x})$ will be close to 1.

(c) $\kappa(\mathbf{z}_1, \mathbf{x})$ will be close to $c_1$, $c_1 \gg 1$ and $\kappa(\mathbf{z}_2, \mathbf{x})$ will be close to $c_2$, $c_2 \ll 0$, where $c_1$, $c_2 \in \mathbb{R}$.

(d) $\kappa(\mathbf{z}_1, \mathbf{x})$ will be close to $c_1$, $c_1 \ll 0$ and $\kappa(\mathbf{z}_2, \mathbf{x})$ will be close to $c_2$, $c_2 \gg 1$, where $c_1$, $c_2 \in \mathbb{R}$.