# Midterm: CS 6375
# Spring 2015

**The exam is closed book. You are allowed a one-page cheat sheet. Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, use an additional sheet (available from the instructor) and staple it to your exam.**

- **NAME** ______________________________
- **UTD-ID if known** ______________________________

| Question | Points | Score |
|---|---|---|
| Decision Trees | 10 | |
| Linear Classifiers | 10 | |
| Neural Networks | 20 | |
| Support Vector Machines | 10 | |
| Point Estimation | 10 | |
| Naive Bayes | 10 | |
| Instance Based Learning | 10 | |
| Extra Credit: Linear Regression | 20 | |
| Total: | 100 | |

## Question 1: Decision Trees (10 points)

Consider the training dataset given below. $A$, $B$, and $C$ are the attributes and $Y$ is the class variable.

| A | B | C | Y |
|---|---|---|---|
| 0 | 1 | 0 | Yes |
| 1 | 0 | 1 | Yes |
| 0 | 0 | 0 | No |
| 1 | 0 | 1 | No |
| 0 | 1 | 1 | No |
| 1 | 1 | 0 | Yes |

(a) (2 points) Can you draw a decision tree having 100% accuracy on this training set? If you answer is yes, draw the decision tree in the space provided below. If your answer is no, explain why?

(b) (3 points) Which attribute among $A$, $B$ and $C$ has the highest information gain? Explain your answer.

(c) (3 points) You are given a collection of datasets that are linearly separable. Is it always possible to construct a decision tree having 100% accuracy on such datasets? True or False. Explain your answer.

(d) (2 points) You are given two decision trees that represent the same target concept. In other words, given the same input the two decision trees will always yield the same output. Does it imply that the two decision trees have the same number of nodes? True or False. Explain your answer.

## Question 2: Linear Classifiers (10 points)

Recall that a linear threshold function or a linear classifier is given by: If $(w_0 + \sum_i w_i x_i) > 0$ then class is positive, otherwise it is negative. Assume that 1 is true and 0 is false.

(a) (5 points) Consider a function over $n$ Binary features, defined as follows. If at least $k$ variables are false, then the class is **positive**, otherwise the class is **negative**. Can you represent this function using a linear threshold function. If your answer is **YES**, then give a precise numerical setting of the weights. Otherwise, clearly explain, why this function cannot be represented using a linear threshold function.

(b) (5 points) Consider a linear threshold function over $n$ real-valued features having $w_0 = 0$ (namely the bias term is zero). Can you always represent such a function using a linear threshold function having only $n - 1$ features? Answer **YES** or **NO** and briefly explain your answer. Note that no credit will be given if your explanation is incorrect.

## Question 3: Neural Networks (20 points)

(a) (10 points) Draw a neural network that represents the function $f(x_1, x_2, x_3)$ defined below. You can only use two types of units: linear units and sign units. Recall that the linear unit takes as input weights and attribute values and outputs $w_0 + \sum_i w_i x_i$, while the sign unit outputs $+1$ if $(w_0 + \sum_i w_i x_i) > 0$ and $-1$ otherwise.

| $x_1$ | $x_2$ | $x_3$ | $f(x_1, x_2, x_3)$ |
|---|---|---|---|
| 0 | 0 | 0 | 10 |
| 0 | 0 | 1 | -5 |
| 0 | 1 | 0 | -5 |
| 0 | 1 | 1 | 10 |
| 1 | 0 | 0 | -5 |
| 1 | 0 | 1 | 10 |
| 1 | 1 | 0 | 10 |
| 1 | 1 | 1 | 10 |

Note that to get full credit, you have to write down the precise numeric weights (e.g., $-1$, $-0.5$, $+1$, etc.) as well as the precise units used at each hidden and output node.

(b) (10 points) Derive a gradient descent training algorithm for the following "special" unit. The "special" unit takes as input a vector $(x_1, \ldots, x_n)$ of feature values and outputs $o$, where $o$ is given by the following equation:

$$o = w_0 + \sum_{i=1}^{n} w_i(x_i + x_i^2 + x_i^3)$$

Here, $w_0, w_1, \ldots, w_n$ are the parameters which you have to learn from the training dataset $D$. Use the batch gradient descent approach. Use the following notation: $x_{i,d}$ denotes the value of the $i$-th attribute (feature) in the $d$-th example in the training set $D$.

## Question 4: Support Vector Machines (10 points)

Consider the training data given below ($X$ is the attribute, and $Y$ is the class variable)

| X | Y |
|---|---|
| -2 | -1 |
| -1 | +1 |
| 1 | +1 |
| 3 | -1 |

(a) (3 points) Assume that you are using a linear SVM. Let $\alpha_1, \alpha_2, \alpha_3$ and $\alpha_4$ be the lagrangian mutlipliers for the four data points. Write the precise expression for the lagrangian dual optimization problem that needs to be solved in order to compute the values of $\alpha_1, \ldots, \alpha_4$ for the dataset given above.

(b) (2 points) Do you think, you will get zero training error on this dataset if you use linear SVMs (Yes or No)? Explain your answer.

The dataset is replicated here for convenience ($X$ is the attribute, and $Y$ is the class variable).

| X | Y |
|---|---|
| -2 | -1 |
| -1 | +1 |
| 1 | +1 |
| 3 | -1 |

(c) (3 points) Now assume that you are using a quadratic kernel: $(1 + x_i^T x_j)^2$. Again, let $\alpha_1, \alpha_2, \alpha_3$ and $\alpha_4$ be the lagrangian mutlipliers for the four data points. Write the precise expression for the lagrangian dual optimization problem that needs to be solved in order to compute the values of $\alpha_1, \ldots, \alpha_4$ for the dataset and the quadratic kernel given above.

(d) (2 points) Do you think, you will get zero training error on this dataset if you use the quadratic kernel (Yes or No)? Explain your answer.

## Question 5: Point Estimation (10 points)

Given that it is virtually impossible to find a suitable "date" for boring, geeky computer scientists, you start a dating website called "www.csdating.com." Before you launch the website, you do some tests in which you are interested in estimating the failure probability of a "potential date" that your website recommends. In order to do that, you perform a series on experiments on your classmates (friends). You ask them to go on "dates" until they find a suitable match. The number of failed dates, $k$, is recorded.

(a) (5 points) Given that $p$ is the failure probability, what is the probability of $k$ failures before a suitable "match" is found by your friend.

(b) (5 points) You have performed $m$ independent experiments of this form (namely, asked $m$ of your friends to go out on dates until they find a suitable match), recording $k_1, \ldots, k_m$. Estimate the most likely value of $p$ as a function of $m$ and $k_1, \ldots, k_m$.

## Question 6: Naive Bayes (10 points)

Consider the following training dataset with two real-valued inputs $X_1$ and $X_2$ and a class variable $Y$ that takes two values, $+$ and $-$.

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 0 | 5 | $+$ |
| 2 | 9 | $+$ |
| 1 | 3 | $-$ |
| 2 | 4 | $-$ |
| 3 | 5 | $-$ |
| 4 | 6 | $-$ |
| 5 | 7 | $-$ |

We assume that the data is generated by a Gaussian naive Bayes model, and we will use the data to develop a naive Bayes classifier.

(a) (5 points) Assuming that the variance is independent of the class, estimate the parameters of the Gaussian Naive Bayes model from the given dataset.

(b) (5 points) Assuming that the variance is independent of the features $X_1$ and $X_2$, estimate the parameters of the Gaussian Naive Bayes model from the given dataset.

## Question 7: Instance Based Learning (10 points)

Consider the training data given below. $x$ is the attribute and $y$ is the class variable.

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | A | A | A | A | B | A | A | A | A | B | B | B | B | A | B | B | B | B |

(a) (2 points) What would be the classification of a test sample with $x = 4.2$ according to 1-NN ?

(b) (2 points) What would be the classification of a test sample with $x = 4.2$ according to 3-NN ?

(c) (3 points) What is the "leave-one-out" cross validation error of 1-NN. If you need to choose between two or more examples of identical distance, make your choice so that the number of errors is maximized.

(d) (3 points) What is the "leave-one-out" cross validation error of 17-NN. If you need to choose between two or more examples of identical distance, make your choice so that the number of errors is maximized.

## Question 8: Extra Credit: Linear Regression (20 points)

**Attempt this question, only after answering the rest of the questions. This question has three parts and each part will be graded on a binary scale; either you will get full points or no points at all.**

Consider fitting the model: $y = w_0 + w_1 x$ ($x$ is 1-dimensional) using the squared loss function that we discussed in class:

$$J(w_0, w_1) = \sum_{i=1}^{m} (y_i - (w_0 + w_1 x_i))^2$$

where $[(x_1, y_1), \ldots, (x_m, y_m)]$ are the data points.

Unfortunately we did not keep the original data, but we did store the following five quantities (statistics) that we computed from the data:

1. $\overline{x}^{(m)} = \frac{1}{m}\sum_{i=1}^{m} x_i$
2. $\overline{y}^{(m)} = \frac{1}{m}\sum_{i=1}^{m} y_i$
3. $C_{xx}^{(m)} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \overline{x}^{(m)})^2$
4. $C_{xy}^{(m)} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \overline{x}^{(m)})(y_i - \overline{y}^{(m)})$
5. $C_{yy}^{(m)} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \overline{y}^{(m)})^2$

(a) (5 points) What are the minimal set of statistics that we need to estimate $w_1$. Namely, which of the above five statistics do you need to estimate $w_1$. Explain your answer by giving a precise expression for $w_1$ in terms of your chosen statistics. No credit without correct explanation.

(b) (5 points) What are the minimal set of statistics that we need to estimate $w_0$. Namely, which of the above five statistics do you need to estimate $w_0$. Explain your answer by giving a precise expression for $w_0$ in terms of your chosen statistics. No credit without correct explanation.

(c) (10 points) Suppose a new data point $(x_{m+1}, y_{m+1})$ arrives, and we want to update our sufficient statistics without looking at the old data, which we have not stored. Give precise expression for the new statistics in terms of the old statistics and the new data point. (This is useful for online learning.)