# CS178 Midterm Exam

Machine Learning & Data Mining: Winter 2016

**Thursday February 11th, 2016**

**Your name:**

**Your UCINetID (e.g., myname@uci.edu):**

**Your seat (row and number):**

- Total time is 80 minutes. READ THE EXAM FIRST and organize your time; don't spend too long on any one problem.
- Please **write clearly** and **show all your work**.
- If you need clarification on a problem, please raise your hand and wait for the instructor to come over.
- You may use one sheet of your own, handwritten notes for reference.
- Turn in any scratch paper with your exam

(This page intentionally left blank)

## Problem 1: (10 points) Bayes Classifiers

In this problem you will use Bayes Rule: $p(y|x) = p(x|y)p(y)/p(x)$ to perform classification. Suppose we observe some training data with two binary features $x_1$, $x_2$ and a binary class $y$. After learning the model, you are also given some validation data.

Table 1: Training Data

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |

Table 2: Validation Data

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

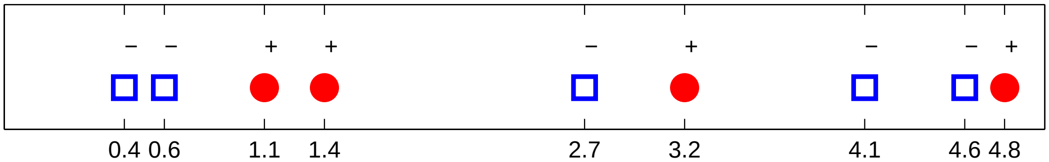
In the case of any ties, we will prefer to predict class 0.

(a) Give the predictions of a joint Bayes classifier on the validation data. What is the validation error rate?

(b) Give the predictions of a naïve Bayes classifier on the validation data. What is the validation error rate?

(c) **True** or **False** : In a naïve Bayes model, the features $x_i$ are independent, i.e., $p(x_1, x_2) = p(x_1)\, p(x_2)$.

## Problem 2: (9 points) Nearest Neighbor Classification




Given the above data with one scalar feature $x$ (whose values are given below each data point) and a class variable $y \in \{-1, +1\}$, with filled circles indicating $y = +1$ and squares $y = -1$ (the sign is also shown above each data point for redundancy), we use a k-nearest neighbor classifier to perform prediction; in the case of ties, we prefer class -1. Answer the following:

(a) Compute the training error rate of a 1-Nearest-Neighbor classifier trained on these data.

(b) Compute the leave-one-out cross-validation error rate of a 1-Nearest-Neighbor classifier on these data.

(c) Compute the training error for a 3-Nearest-Neighbor classifier on these data.

## Problem 3: (9 points) Gradient Descent

Suppose that we have training data $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$, where $x^{(i)}$ is a scalar feature and $y^{(i)} \in \{-1, +1\}$, and we wish to train a linear classifier, $\hat{y} = \text{sign}[\, a + bx \,]$, with two parameters $a$, $b$. In order to train the model, we use gradient descent on a smooth surrogate loss called the *exponential loss*:

$$J(X, Y) = \frac{1}{m} \sum_i \exp\left( y^{(i)} (a + bx^{(i)}) \right)$$

(a) Write down the gradient of our surrogate loss function.

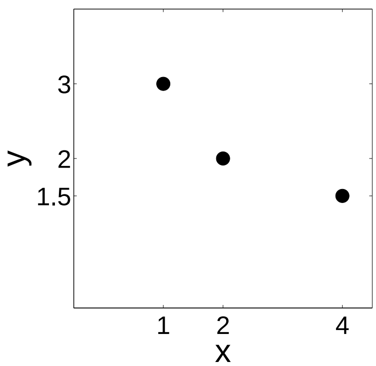
(b) Give pseudocode for a (batch) gradient descent function `theta = train(X,Y)`, including all necessary elements for it to work.

## Problem 4: (10 points) Linear Regression, Cross-validation

Consider the following data points, copied in each part. We wish to perform linear regression to minimize the mean squared error (MSE) of our predictions.

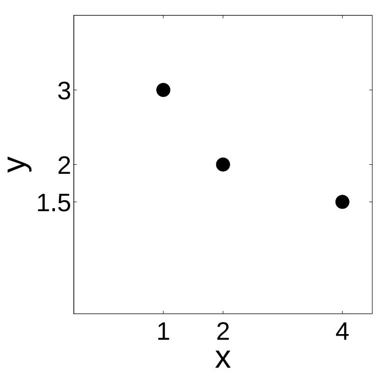
(a) Compute the **leave-one-out** cross-validation error of a zero-order (constant) predictor,

$$\hat{y}(x) = \theta_0$$




(b) Compute the **leave-one-out** cross-validation error of a first-order (linear) predictor,

$$\hat{y}(x) = \theta_0 + \theta_1 \, x$$



## Problem 5: (14 points) Multiple Choice

For the following questions, assume that we have $m$ data points $y^{(i)}$, $x^{(i)}$, $i = 1 \ldots m$, each with $n$ features, $x^{(i)} = [x_1^{(i)} \ \ldots \ x_n^{(i)}]$.

**Circle one answer for each:**

Suppose that we are training a linear classifier (perceptron). Before training, we decide to remove (throw away) 10% of our features (selected at random). This is most likely to make it **more** **equally** **less** likely to overfit the data.

Suppose that, when training a linear classifier, we double the amount of data available for training. This is most likely to decrease the **bias** **variance** **both** **neither** of our learned model.

When training a $k$-nearest neighbor model, we decide to increase the value of $k$. This will most likely make our model **more** **equally** **less** likely to overfit the data.

**True** or **false**: if the VC dimension of a model is $H$, then the model can shatter any set of $H$ training points.

**True** or **false**: Linear regression can be solved using either matrix algebra or gradient descent.

**True** or **false**: Increasing the regularization of a linear regression model will decrease the variance.

Before training a linear classifier, we transform one of our features by exponentiating it, i.e., `X[:,1] = np.exp(X[:,1]);`. This is likely to **increase** **not change** **decrease** the model's VC dimension.

## Problem 6: (9 points) Short answer

Consider the two possible decision boundaries (indicated by Line 1 and Line 2) for the binary classification problem shown in Figure 1. For each algorithm below, will it possibly produce boundary 1, boundary 2, or both? Please give a concise explanation of your choice.

**Perceptron Algorithm** :

**Logistic Regression** :

**Support Vector Machine (hard-margin)** :

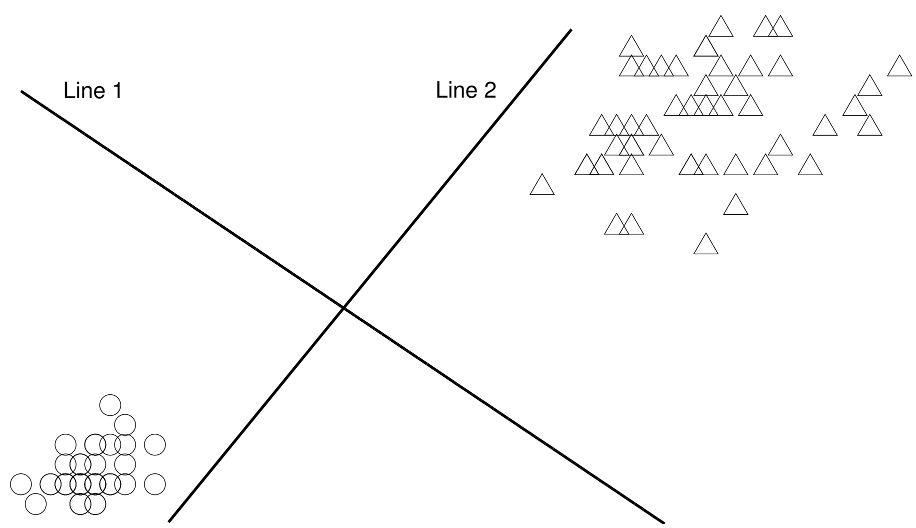
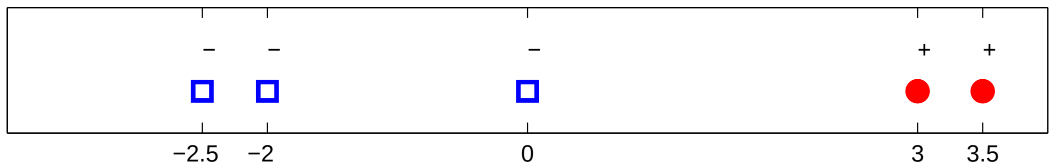



Figure 1: Possible linear decision boundaries.

**Problem 7: (10 points) Support Vector Machines**




Using the above data with one feature $x$ (whose values are given below each data point) and a class variable $y \in \{-1, +1\}$, with filled circles indicating $y = +1$ and squares $y = -1$ (the sign is also shown above each data point for redundancy), answer the following:

(a) Sketch the solution (decision boundary) of a linear SVM on the data, and identify the support vectors.

(b) Give the solution parameters $w$ and $b$, where the linear form is $wx + b$.

(c) Give one advantage that the dual (kernel) form of SVMs have over the primal form. (When would it be preferable to use the dual form?)

(d) In contrast, give one advantage that the primal SVM form has over the dual form. (When would it be preferable to use the primal form?)

## Problem 8: (10 points) VC Dimension

Consider the following classifier, parameterized by a single scalar parameter $a$ and operating on a scalar feature $x$:

$$f(x\ ;\ a) = \begin{cases} +1 & x \leq a \text{ or } a+1 < x \leq a+2 \\ -1 & \text{otherwise} \end{cases}$$

In this problem, we will show the VC dimension of $f(x\ ;\ a)$ is 3.

(a) Show by example that $f(x\ ;\ a)$ can shatter three points. Hint: place your points at $x^{(1)} = 0$, $x^{(2)} = 0.75$, $x^{(3)} = 1.5$.

(b) Argue that $f(x\ ;\ a)$ cannot shatter four points. (Which target pattern cannot be reproduced?)