Midterm: CS 7301
Spring 2017

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
Naive Bayes	10	
Decision Trees	10	
True/False and Short Questions	8	
Neural Networks and Perceptrons	12	
Support Vector Machines	10	
Total:	50	

Question 1: Naive Bayes (10 points)

Consider the training dataset given below. X_1, X_2 and X_3 are the features and Y is the class variable.

Y	X_1	X_2	X_3
0	0	0	1
1	1	0	0
1	0	1	1
0	0	1	0
1	1	1	0
0	1	1	1

(a) (2 points) Compute the parameters of the Naive Bayes model when trained (without Laplace correction) on the dataset given above.

Solution:
For brevity, for each variable $X \in \{Y, X_1, X_2, X_3\}$, let x denote the assignment $X = 1$ and $\neg x$ denote the assignment $X = 0$. The parameters are:

$$P(y) = \frac{\text{Count}(y)}{\text{\#examples}} = \frac{3}{6} = \frac{1}{2}$$

$$P(x_1|y) = \frac{\text{Count}(x_1, y)}{\text{Count}(y)} = \frac{2}{3}; \quad P(x_1|\neg y) = \frac{\text{Count}(x_1, \neg y)}{\text{Count}(\neg y)} = \frac{1}{3}$$

$$P(x_2|y) = \frac{\text{Count}(x_2, y)}{\text{Count}(y)} = \frac{2}{3}; \quad P(x_2|\neg y) = \frac{\text{Count}(x_2, \neg y)}{\text{Count}(\neg y)} = \frac{2}{3}$$

$$P(x_3|y) = \frac{\text{Count}(x_3, y)}{\text{Count}(y)} = \frac{1}{3}; \quad P(x_3|\neg y) = \frac{\text{Count}(x_3, \neg y)}{\text{Count}(\neg y)} = \frac{2}{3}$$

(b) (5 points) Now assume that the parameters associated with $X_1 = 1$ are constrained by the following prior distribution. The parameters can take one of the following three values: $(0.3, 0.5, 0.7)$ with probabilities $(1/3, 1/6, 1/2)$ respectively. Write down the parameters for the Naive Bayes model given the dataset and prior on $X_1 = 1$.

Solution:
We have to compute $P(X_1|Y = 1)$ and $P(X_1|Y = 0)$.
For $Y = 0$, we have one $1s$ and two $0s$ for X_1:

$$\begin{aligned} P(X_1 = 1|Y = 0) &= \arg\max_{\theta \in \{1/3,1/6,1/2\}} P(\theta)P(D|\theta) \\ &= \arg\max(1/3 \cdot 0.3 \cdot 0.7 \cdot 0.7, 1/6 \cdot 0.5 \cdot 0.5 \cdot 0.5, 1/2 \cdot 0.7 \cdot 0.3 \cdot 0.3) \end{aligned}$$

Since the first term is the largest, we get $P(X_1 = 1|Y = 0) = 0.3$.
For $Y = 1$, we have two $1s$ and one $0s$ for X_1.

$$\begin{aligned} P(X_1 = 1|Y = 1) &= \arg\max_{\theta \in \{1/3,1/6,1/2\}} P(\theta)P(D|\theta) \\ &= \arg\max(1/3 \cdot 0.3 \cdot 0.3 \cdot 0.7, 1/6 \cdot 0.5 \cdot 0.5 \cdot 0.5, 1/2 \cdot 0.7 \cdot 0.7 \cdot 0.3) \end{aligned}$$

Since the third term is the largest, we get $P(X_1 = 1|Y = 1) = 0.7$.

(c) (3 points) Compute $P(Y = 0|X_1 = 1, X_2 = 0, X_3 = 1)$ using the two Naive Bayes models (one without a prior and the second with a prior)?

Question 2: Decision Trees (10 points)

Consider the training dataset given below. X_1, X_2 are the features and Y is the class variable.

X_1	0	1	1.5	2	3	1	2	2.5	3	4
X_2	0	0	0	0	0	2	2	2	2	2
Y	+1	+1	-1	+1	+1	-1	-1	+1	-1	-1

Hint: *It might be helpful to plot the dataset. Also, recall that when a decision tree splits on a real-valued feature, it uses a threshold of the form* $x_1 < 1$.

(a) (2 points) What is the training error rate of a decision stump (decision tree with max-depth 1 and exactly two leaves) trained on the dataset given above?

(b) (2 points) What is the training error rate of a full decision tree (no maximum depth or pruning) trained on the dataset given above?

(c) (3 points) What is the leave one out cross-validation error rate of a decision stump (decision tree with max-depth 1 and exactly two leaves) trained on the dataset given above?

(d) (3 points) What is the leave one out cross-validation error rate of a full decision tree (no maximum depth or pruning) trained on the dataset given above?

Question 3: True/False and Short Questions (8 points)

(a) (2 points) L2 regularization is an ideal feature selection method and is much better than L1 regularization (for feature selection) because unlike L1 regularization, L2 will force several parameters to zero. Is the statement True? Explain your answer in one or two sentences. No credit if the explanation is incorrect.

(b) (2 points) As we increase k, the k-nearest neighbors algorithm overfits the training data. Is the statement True? Explain your answer in one or two sentences. No credit if the explanation is incorrect.

(c) (2 points) If the data is discrete and linearly separable, then the (discrete) Naive Bayes algorithm will yield the same solution as the Logistic Regression Classifier. Is the statement True? Explain your answer in one or two sentences. No credit if the explanation is incorrect.

(d) (2 points) The predictions of the k-nearest neighbors classifier will not be affected if we pre-process the training data by rescaling each feature (e.g., take a feature expressed in centimeters and convert it to inches). Is the statement True? Explain your answer in one or two sentences. No credit if the explanation is incorrect.

Question 4: Neural Networks and Perceptrons (12 points)

(a) (7 points) Draw a neural network that represents the following function. 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.

y_1 and y_2 are outputs and x_1 and x_2 are inputs. Therefore, your neural network will have two output nodes.

x_1	x_2	y_1	y_2
0	0	4	12
0	1	6	10
1	0	6	10
1	1	4	12

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) (5 points) Consider a dataset having one input variable x, one output variable y and m examples given by: $(x_1, y_1), \ldots, (x_m, y_m)$ where x_i and y_i are the values assigned to x and y respectively in the i-th example. Derive a gradient descent training algorithm for the following non-linear regression model having two parameters w_0 and w_1.

$$R(x) = w_0 \exp(x + w_1) \tag{1}$$

We decide to use the following evaluation function to train our model:

$$E = \frac{1}{m} \sum_{i=1}^{m} (y_i - R(x_i))^2$$

Use the **batch gradient descent approach.**

Question 5: Support Vector Machines (10 points)

Consider the dataset given below (x_1, x_2) are the attributes and y is the class variable):

x_1	x_2	y
1	0	-1
-3	1	$+1$
3	1	$+1$

(a) (6 points) Find the linear SVM classifier for the dataset given above. Do your optimization either using the primal problem or the dual problem. Provide a precise setting of the weights **w** and the bias term b. Also, identify the support vectors (Hint: The primal problem seems to be easier than the dual).

(b) (4 points) In class, we saw the primal form as well as the dual form for SVMs. Give one advantage of the dual form over the primal form. Similarly give one advantage of the primal form over the dual form.