CS 419: Mid Semester Exam

September 14, 2018.
13:30 To 15:30 PM

Roll: ______________________________

Name: ______________________________

The exam is open (your) notes and closed book. Answer to the point. Irrelevant or grossly wrong answers may fetch negative marks.

The following training dataset on two classes will be used to answer the questions below.

x_1	x_2	y
10	0	-1
0	1	+1
2	1	+1
6	0	-1
4	0	+1
6	1	+1
8	0	-1
12	1	-1

1. Suppose we try to train a two class naive Bayes classifier on this dataset. Assume attribute x_1 follows a Gaussian distribution and attribute x_2 is binary taking values 0 or 1 and follows the Bernoulli distribution.

   (a) Write down the unbiased maximum likelihood estimates of the following parameters in each of the class.

Parameter	Class “+1”	Class ”-1”
mean (μ) of x_1		
variance (σ^2) of x_1		
Bernoulli parameter (p) of x_2		
Class prior $\Pr(y)$		

   ..4

Parameter	Class ‘‘+1’’	Class ’’-1’’
mean (μ) of x_1	3	9
variance (σ^2) of x_1	20/3	20/3
Bernoulli parameter (p) of x_2	3/4	1/4
Class prior	1/2	1/2

   (b) Next we will use the above parameter values to write the equation of the decision surface. Provide values of coefficients w_0, w_1, w_2, w_3, w_4 if $f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2$ where $f(x) \geq 0$ implies that the predicted class label for x is “+1”. Is the decision surface linear or non-linear? ..4

$$\begin{aligned}
&\log \Pr(y=1|x) \geq \log \Pr(y=+1|x) \\
\rightarrow\ & \frac{-(x_1-u_1)^2}{2*\sigma_1^2} - \log\sigma_1 + x_2\log p_1 + (1-x_2)\log(1-p_1) + \log\Pr(y=+1) \\
\geq\ & \frac{-(x_1-u_2)^2}{2*\sigma_2^2} - \log\sigma_2 + x_2\log p_2 + (1-x_2)\log(1-p_2) + \log\Pr(y=-1) \\
&\rightarrow \frac{(2*3*x_1-3*3)}{2*20/3} + x_2\log 3/4 + (1-x_2)\log 1/4 \\
&\geq \frac{(2*9*x_1-9*9)}{2*20/3} + x_2\log 1/4 + (1-x_2)\log 3/4 \\
&\rightarrow -27/40 + 243/40 - log3 + (9/20-27/20)x_1 + 2*\log 3x_2 \geq 0
\end{aligned}$$

This implies that $w_3 = w_4 = 0, w_0 = -27/40 + 243/40 - log3, w_1 = (9/20-27/20), w_2 = 2*\log 3$

(c) What is the probability that an instance $(x_1, x_2, y) = (3, 1, +1)$ will be correctly predicted by the above classifier? [You do not need to simplify the final numerical expression.] ..2

$$\Pr(y=+1|(3,1)) = \frac{\frac{1}{\sqrt{2*pi\sigma}}e^{-(3-3)^2/2*\sigma}*3/4*1/2}{\frac{1}{\sqrt{2*pi\sigma}}(e^{-(3-3)^2/2*\sigma}*3/4*1/2+e^{-(3-9)^2/2*\sigma}*1/4*1/2)} = \frac{3}{3+e^{-36*3/20}} \approx 1$$

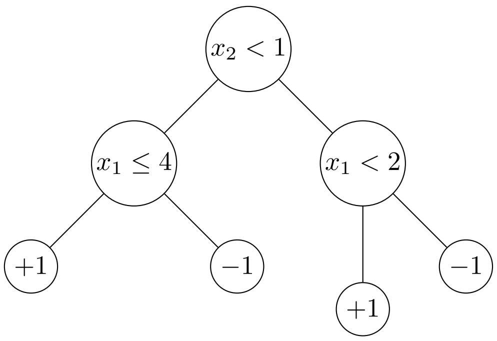
2. Next we will try to train a decision tree classifier on this dataset. Write down the Gini values for the best split along each of attribute x_1 and x_2 at the top-node of the tree. ..3

For Gini on x_1, x_2 we sort the data on these attribute.

x_1	x_2	y
0	1	+1
2	1	+1
4	0	+1
6	1	+1
6	0	-1
8	0	-1
10	0	-1
12	1	-1

x_1	x_2	y
0	1	+1
2	1	+1
6	1	+1
12	1	-1
6	0	-1
8	0	-1
10	0	-1
4	0	+1

3. For the tree below, suppose we apply bottom-up tree pruning with the dataset above as the validation set. Mark out the branches that will be pruned, and draw the new pruned tree alongside.



..2

4. Next we will try to apply the loss regularization framework to fit a linear classifier on the above dataset. Our linear classifier will be denoted as $f(\mathbf{x}) = w_1x_1 + w_2x_2 + w_0$, and predicted class label for an instance $\mathbf{x}$ will be sign$(f(\mathbf{x}))$. The w_1, w_2, w_0 are classifier parameters.

   (a) Suppose we use $w_0 = -1, w_1 = 0, w_2 = 2$ so $f(\mathbf{x}) = 2x_2 - 1$.. For this choice, specify the value of the total Hinge loss on this dataset by filling the table below. ..2

x_1	x_2	y	Hinge loss
10	0	-1	
0	1	+1	
2	1	+1	
6	0	-1	
4	0	+1	
6	1	+1	
8	0	-1	
12	1	-1	
Total loss			

x_1	x_2	y	Hinge loss
10	0	-1	0
0	1	+1	0
2	1	+1	0
6	0	-1	0
4	0	+1	3
6	1	+1	0
8	0	-1	0
12	1	-1	3
Total loss			

   (b) Next, while keeping $w_0 = -2, w_2 = 4$, we look for an optimal value for w_1 in the range $[a, b] = [-1, 1]$ using the Dichotomos binary search algorithm discussed in class. For that we will pick two values of w_1 in the midpoint — say one at 0 and the other a small ϵ distance away. Calculate and show if the optimum w_1 will lie to the left or right of 0. [No marks without proper calculation] ..4

   Since ϵ is small, only the loss of the two instances with non-zero loss will be affected by it. The loss is $\max(0, 3 - 4\epsilon) + \max(0, 3 + 12\epsilon)$. This term is greater than 6 when ϵ is greater than 0. Thus, w_1 will be to the left.

   (c) Now say we apply the logistic loss on the above dataset to train the parameters w_1, w_2, w_0. Write the value of the first order Taylor expansion of the total loss function at $[w_0, w_1, w_2] = [0, 0, 0]$. Note the first order expansion will be an expression of the form $C + C_0w_0 + C_1w_1 + C_2w_2$ where you need to replace C, C_0, C_1, C_2 by constant values. ..4

   The value of $e^{-y_i f(\mathbf{x}^i)}$ is 1 for instances at $[w_0, w_1, w_2] = [0, 0, 0]$.
   The logistic loss value is thus $8\log(2)$.
   The gradient along w_0 is $\sum_i \frac{e^{-y_i f(\mathbf{x}^i)}}{1+e^{-y_i f(\mathbf{x}^i)}}(-y_i) = 0$
   The gradient along w_1 is $\sum_i \frac{e^{-y_i f(\mathbf{x}^i)}}{1+e^{-y_i f(\mathbf{x}^i)}}(-x_1^i y_i) = -(-24/2) = 12$
   The gradient along w_2 is $\sum_i \frac{e^{-y_i f(\mathbf{x}^i)}}{1+e^{-y_i f(\mathbf{x}^i)}}(-x_2^i y_i) = -2$
   The first order Taylor expansion is then

$$8\log(2) + 12w_1 - 2w_2$$

5. Consider a training dataset for one dimensional data $D = \{(2, -1), (-2, 1)\}$. Assume the classifier is $C(x) = \text{sign}(wx)$. For each of the training objective below provide the w that is optimal. When multiple w values are optimal, state them all to get full marks.

   (a) Logistic loss with no regularization. ..2

   The objective is $2\log(1 + e^{2w})$ This is minimized when $w = -\infty$.

   (b) Logistic loss with L1 regularization , that is, we add $\lambda|w|$ to the training objective. [Hint: first argue why you can rewrite this specific training objective as a differentiable function]. ..3

   Since we know that $w < 0$ is better for the loss term, the training objective becomes $2\log(1 + e^{2w}) - \lambda w$.

   Now we can solve for the gradient being equal to zero and get the solution as $w = \frac{1}{2}\log(\frac{\lambda}{4-\lambda})$

   (c) Hinge loss with L2 regularization, that is, we add a λw^2 to the training objective. ..4

   The objective is $2\max(0, 1+2w)+\lambda w^2$ The loss is zero for any $w \leq -1/2$. With regularizer we need only consider $w \geq -1/2$.

   For $w \geq -1/2$ the loss can be written as

   $2(1 + 2w) + \lambda w^2$ When we write the gradient and solve for it we get:

   $4 + 2\lambda w = 0$. This implies that $w = -2/\lambda$. As long as $\lambda > 4$, we get this value as the optimum. Else, the optimum is $w = -1/2$.

6. Suppose you have a linear binary classifier on d attributes $C(\mathbf{x})$ as $\text{sign}(f(\mathbf{x}))$ where $f(\mathbf{x}) = w_1x_1 + \ldots + w_dx_d + w_0$. Assume all x_js take values only between 0 and 1. For example, when $w_1 = 1, w_2 = -1, b = -1/2$. The classifier will predict class "1" for all $\mathbf{x}$ for which $x_1 - x_2 - 1/2 < 0$, and class +1 otherwise. For example, point $\mathbf{x} = (1, 0)$ is assigned positive class and point $\mathbf{x} = (1/2, 1/2)$ is assigned negative class.

   Assume that infinite amount of training dataset is uniformly sampled from the d-dimensional unit cube, and then labeled using $C(\mathbf{x})$. Thereafter, a new classifier is trained on this dataset to mimic $C(\mathbf{x})$.

   (a) Is it possible to perfectly mimic $C(\mathbf{x})$ using a bounded depth decision tree? [Justify brielfy] ..2

   No, since a decision tree with split on single attributes at a time will create jagged decision boundary.

   (b) Is it possible for a LDA classifier to mimic $C(\mathbf{x})$ [Justify brielfy] ..2

   No, since there might be unequal volume of points in the two regions. Even though the decision surface is linear, there is no guarantee that it will be the same linear function.

   (c) Is it possible for a naive Bayes classifier to mimic $C(\mathbf{x})$ with any choice of distribution on each attribute? [Justify brielfy]. ..2

   No, since variance for an attribute in the two classes could be different leading to a non-linear decision boundary.

Total: 40