CS-419m: Practice question set 2

1. State true or false with a brief reason. You will only get marks if your reasons are correct.

   (a) During tree construction, if you reach a node where the maximum information gain over all splits is zero, then all training examples at that node have the same label. ..1

   `False. Consider example with two binary attributes and class label set as XOR of the two attributes. In this case, info. gain is zero for all possible splits at the top node but we can achieve zero training error.`

   (b) When a linear separator $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ is trained on data that is perfectly separable via square loss, the classifier is guaranteed to achieve zero error on the training data. ..1

   False. Let $d = 1$, there are many negative instances with $x_1 = -1, y = -1$, many positive instances with $x_1 = 1, y = 1$ and a single positive instance with $x_1 = -0.1, y = 1$. The least square minimizer is $w_1 = 1, w_0 = 0$ which does not correctly classify the last positive instance at $x_1 = -0.1$.

   (c) If $f(x)$ is convex in x, then $\log f(x)$ is also convex in x. ..1

   False. $f(x) = x$ is convex but $\log x$ is concave since its double differential is $-\frac{1}{x^2}$ which is always less than zero.

   (d) Whenever a set S of labeled instances is split into two sets S_1 and S_2, the average entropy will always decrease, irrespective of the split attribute or the split point. Give reasons if TRUE, or a contradictory example if FALSE. ..2 `The average entropy will either decrease or stay the same, it will never increase.`

2. The following dataset will be used to learn a decision tree for predicting whether a person is Happy (H) or Sad (S) based on the color of their shoes, whether they wear a wig and the number of ears they have.

Color	Wig	Num. Ears	Emotion
G	Y	2	S
G	N	2	S
G	Y	2	S
B	N	2	S
B	N	2	H
R	N	2	H
R	N	2	H
R	N	2	H
R	Y	3	H

   (a) What is Entropy(Emotion | Wig=Y)? ..1

   $-\frac{1}{3} * \log(\frac{1}{3}) + -\frac{2}{3} * \log(\frac{2}{3}) = \frac{1}{3} * \log(3) + \frac{2}{3} * \log(\frac{3}{2}) = 0.918$

   (b) Which attribute would the decision-tree building algorithm choose to use for the root of the tree (assume no pruning). ..1

   `Information gains on various attributes are as follows:`

$IG(Color) = 0.76885$
$IG(Wig) = 0.07278$
$IG(Ears) = 0.10219$

Hence, assuming entropy based information gain to be the goodness of an attribute, we would use color for a split at the root of the tree.

(c) Draw the full decision tree that would be learned for this data (assume no pruning). ..1

The first split is on color. For branch corresponding to color B, neither wig nor number of ears offers any further information gain. Hence, we can assign a label of 0 or 1 to the leaf at the end of the branch for color B, and stop building the tree.

(d) What would be the training set error for this dataset? Express your answer as the percentage of records that would be misclassified. ..1

One example in the training set would be misclassified (either 4^{th} or 5^{th} example in the dataset depending on the tree), leading to a train error of 11.11 percent.

3. Give an example of a dataset where the greedy decision tree construction method discussed in class creates a tree that is larger than the smallest required to get perfect accuracy. ..3

$d = 2, k = 3$, 10000 points of class 1 uniformly distributed in the box $x_1 = [0..0.5], x_2 = [0.1..1]$
10000 points of class 2 uniformly distributed in the box $x_1 = [0.5..1], x_2 = [0.1..1]$
10 points of class 3 uniformly distributed in the box $x_1 = [0..1], x_2 = [0..0.1]$

Smallest tree with two internal nodes will be created by first splitting on $x_2 = 0.1$ and then on $x_1 = 0.5$.

But the greedy method will first split on $x_2 = 0.5$ since the information gain is much higher than splitting first on $x_2 = 0.1$. Then, in each of the two branches it will have to split on $x_2 = 0.1$ leading to a tree with three internal nodes.

4. Consider the logistic classifier. If two features f_k and f_j are replicas of each other, state with a brief reason how the parameters w_k and w_j are related in the following two versions of the objectives used for training logistic classifiers: Assume the training data consists of instances of the form $(\mathbf{x}^i, y_i)$ where $y_i \in \{-1, 1\}$

(a) Unregularized: minimize $\sum_i \log(1 + e^{-\mathbf{w}^T\mathbf{x}^i y_i})$ ..1

There is no unique solution. All possible values of w_j and w_k whose sum is a constant γ will give rise to the same optimum value.

(b) Regularized with L2: minimize $C\sum_i \log(1 + e^{-\mathbf{w}^T\mathbf{x}^i y_i}) + \sum_j w_j^2$ where C is a constant. ..2

In this case w_j and w_k will be equal because for a fixed value of the sum, the L2 regularizer is minimized when the two parameter values are equal.

(c) Regularized with L1: minimize $C\sum_i \log(1 + e^{-\mathbf{w}^T\mathbf{x}^i y_i}) + \sum_j |w_j|$ where C is a constant. ..2

In this case also there is no unique solution. All possible values of w_j and w_k such w_j and w_k have the same sign but whose sum is a constant will give the same objective value.

5. Suppose we are trying to construct a decision tree on data with only numerical attributes. We decide to select the split condition based on the minimization of the perceptron loss

instead of minimizing average entropy. Assume binary class labels with one attribute x_1 and let us denote the data as: $(x_1^1, y^1), \ldots, (x_1^N, y^N)$

(a) Consider the example dataset $(0,-1),(1,1),(2,-1),(4,1),(5,1)$ What is the value of the loss for the split condition $x_1 \leq 0.5$ using $f(x) = x_1 - 0.5$ ..1

`The perceptron loss is` $\sum_i \max(-y_i f(x_1^i), 0)$. `The single misclassified example (2,1) has loss: 1.5`

(b) What is the value of v in the split condition $x_1 \leq v$ for which the perceptron loss is minimized for the above example? ..1

`Loss function in terms of` v `is` $\sum_i \max(-y_i(x_1^i - v), 0)$.

`Any value between 1 and 2 will be optimal and will yield a loss of 1.`

`E.g.` $x_1 \leq 2$. `Loss=1.`

(c) We need to design an efficient algorithm for finding the optimum split condition v assuming the data is sorted on attribute x_1. As in the case of entropy, we make one pass of the data to calculate the value of loss for increasing values of x_1 and choose the least loss value v^*. For this, what aggregate statistics will you maintain as you scan the data? For example, for the case of entropy, we maintained,

$$n_k(v) = \sum_{i, x_1^i \leq v, y^i = k} 1, \qquad \forall k = 1, 2. \tag{1}$$

where v is the value of x_1 of the current scanned instance. Using this and the aggregate statistics n_k over the entire data, we could calculate the average entropy of splitting at v as:

$$\text{Entropy}(v) = -\frac{N(v)}{N} \sum_k \frac{n_k(v)}{N(v)} \log \frac{n_k(v)}{N(v)} - \frac{N - N(v)}{N} \sum_k \frac{(n_k - n_k(v))}{N - N(v)} \log \frac{n_k - n_k(v)}{N - N(v)} \tag{2}$$

where $N(v) = \sum_k n_k(v)$. Write down the equivalent of equations 1 and 2 for choosing the split with minimum perceptron loss ..3

$$d_k(v) = \sum_{i, x_1^i \leq v, y^i = k} x_1^i. \tag{3}$$

$$\text{Loss}(v) = \min(d_1(v) + d_2 - d_2(v), d_2(v) + d_1 - d_1(v)) \tag{4}$$

`In the above, we measure the minimum of the loss of calling the part` $\leq v$ `as class 1 or class 2.`

(d) Show an example dataset where the best split based on entropy is different from one based on the loss function. ..2

`Choose one negative at 1000, two positives at 1,1, two negatives at -1, -1,. Entropy will choose` v `between -1 and 1 whereas the loss function will pick a` v `between 1000 and 1.`

(e) Show a training dataset where a linear discriminate that minimizes the Perceptron error leads to suboptimal 0/1 error on the training set.

$d = 1$. `Positive instances in` $D = \{-100, 1, 2, 3\}$ `Negative instances in` $D = \{(100, -1, -2, -3\}$
Optimal separator for minimizing 0/1 error is $f(x_1, x_2) = x_1$ where 0/1 loss is 2 whereas perceptron loss is 200. If instead we pick, $f(x_1, x_2) = -x_1$, the preceptron loss is 12 whereas the 0/1 loss is six. ..2

(f) Consider the following loss function $L(x, y, f) = \frac{1}{1+exp(yf(x))}$ called the Sigmoid loss.

i. In a x-y plot show the loss versus $yf(x)$ ..1

ii. Is the loss function concave, convex, or neither? Justify briefly.

`Neither convex nor concave. Find double differential of` $g(v) = \frac{1}{1+\exp(v)}$. `This turns out to be` $g''(v) = \frac{\exp(v)(1-\exp(v))}{(1+\exp(v))^3}$. `This quantity is` > 0 `for some values of` v `and` < 0 `for others. Using the results of Q3 we can show that the loss function is neither concave nor convex in w.` ..2

iii. State one major advantage of the sigmoid loss above the logistic loss.

`The maximum loss of any instance is 1 whereas it increases almost linearly for Logistic loss. This implies that for noisy training instances, of the kind in Question 2, Logistic loss will give rise to a bad separator.` ..2

(g) Consider a linear binary classification problem with one dimensional points ($d = 1$). Say the classifier is of the form sign(wx). Show a labeled dataset of 100 instances where there exists a separator w for which the 0/1 error is only 10% but for all these convex loss functions (Hinge, logistic, perceptron), the w that minimizes that loss function will have an error of 90% [Hint: The dataset just needs four types of instances.] ..4

`Consider the following four clusters`

i. `5 points with` $x = 1, y = -1$,

ii. `5 points with` $x = -1, y = 1$,

iii. `45 points with` $x = \epsilon, y = 1$,

iv. `45 points with` $x = -\epsilon, y = -1$,

`As` ϵ `goes to zero, 0/1 error is minimized with` $w = 1$ `where the first two groups only are misclassified. In contrast, logistic, perceptron, hinge are all minimized with some` $w < 0$ `which misclassifies the last two groups of points.`