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

   (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

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

   ..1

   (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

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

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

..1

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

..1

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

..1

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

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

(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

(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

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 $(0,1), (1,2), (2,1), (4,2), (5,2)$ What is the value of the loss for the split condition $x_1 \leq 0.5$

   ..1

   (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

   (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:

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

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

..2

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

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.

..2

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

..2

(g) Consider a linear binary classification problem with one dimensional points ($d = 1$). Say the classifier is of the form $\text{sign}(wx)$. 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