Instructions

- If you have any question about problems mail to mlt4717@yahoo.com with the subject in the format: QHW(homework number)-P(question number) (e.g., QHW(2)-P(2)).
- Send your assignments with subject HW(homework number)-(Student-ID) to mlt40717@gmail.com.
- Send your codes before 12:00 (noon) of the due date.

Problem 1: Logistic Regression and Naïve Bayes

1.1. [3 Pts] Obtain the cost function of the Logistic Regression classifier where this cost function is defined as 
   \[- \log P(D_y|D_x, w)\]
   and Bernoulli distribution for the conditional probabilities \(P(y|x)\).

1.2. Consider a classification problem with two classes. Assume that the class conditional distributions are 
   \(p(x|C_1) \sim N(\mu_1, \sigma_1^2)\) and \(p(x|C_2) \sim N(\mu_2, \sigma_2^2)\) and the prior \(p(C_1) = 0.3\):
   a) [3 Pts] Calculate \(p(C_1|x)\) and show it in the logistic formulation \(P(C_1|x) = \frac{1}{1+e^{-a(x)}}\).
   b) [4 Pts] What is the Bayesian decision rule in this case? Is it a Naïve Bayes classifier?

1.3. Consider a binary classification problem and the Logistic Regression classifier:
   a) [2 Pts] Describe equations used to find the parameters and also the decision rule for classifying a new sample.
   b) [4 Pts] How can we avoid overfitting in this method? How would the weight update rule be changed?

1.4. We intend to compare the results of the Logistic regression classifier with those of the Naïve Bayes classifier on a text classification task. The data set “doc_data.zip” (obtained based on 20 Newsgroup data set) is uploaded for this purpose. The features for classifying a document will be the count of each word (of the vocabulary) in the given document and the documents must be assigned to one of the two classes.
   a) [2 Pts] In this problem, our generative Naïve Bayes classifier assumes that the words in the documents of each class are drawn independently from a multinomial distribution. Let \(p_{i,y}\) denote the conditional probability of occurrence of the \(i\)-th word in the samples of class \(y \in \{0,1\}\).
   Assume that we estimate \(p_{i,y} \approx \frac{n_{i,y}}{n_y}\) where \(n_{i,y}\) shows the total number of occurrence of the \(i\)-th word in the samples of class \(y\) and \(n_y = \sum_{i=1}^{M} n_{i,y}\) (\(M\) denotes the number of distinct words in the whole collection). First use the training data to calculate the prior probabilities \(p_y\) and the conditional probabilities \(p_{i,y}\).
   b) [6 Pts] Show that the probabilities introduced in Part (a) can be obtained using the maximum likelihood estimation.
c) [5 Pts] Describe a Naïve Bayes classifier using the estimated parameters in Part (a) and implement it. What is the minimum number of parameters that you need to know in order to classify an example using this classifier?

d) [2 Pts] When using the above Naïve Bayes classifier, what will happen if some words do not occur in any documents of a class?

e) [3 Pts] Revise the estimation of $p_{i|y}$ as $p_{i|y} \approx \frac{n_{i|y} + 1}{n_y + M}$ in the Naïve Bayes classifier implementation. To which estimation method does this revision correspond?

f) [7 Pts] Implement a Logistic regression classifier ($\eta = 0.0001$).

Problem 2: Perceptron

2.1. [2 Pts] The cost function of the Perceptron algorithm is defined as $J(w) = \sum_{i \in M} -y^{(i)} w^T x^{(i)}$ where $M$ shows the set of misclassified samples. Show that the empirical loss is equivalent to

$$\frac{1}{n} \sum_{i=1}^{n} \max (0, -y^{(i)} w^T x^{(i)})$$

2.2. [2 Pts] What is the main drawback of the Perceptron algorithm as a linear classifier?

2.3. [3 Pts] Can the Perceptron update rule be modified to find a hyper-plane that discriminates well the classes that are not linearly separable (to give appropriate convergence for non-separable data)? Discuss.

2.4. [3 Pts] Can the Perceptron update rule be modified to convergence to a unique hyper-plane when the data is linearly separable (independent of the initial parameters)? Discuss.

2.5. [8 Pts] Implement a batch Perceptron algorithm and report its result on the document classification task mentioned in Problem 1 and also the number of updates.

2.6. [5 Pts] Implement an online Perceptron algorithm. Repeat cycling through the training data until the accuracy on the training data has stopped improving. Report its results on the document classification task and also the number of updates.

2.7. [2 Pts] Compare results with those of the Logistic regression and Naïve Bayes classifier.

Problem 3: Decision Tree

3.1 Information Gain

Information Gain (IG) is one of the heuristics used to find the best attribute in nodes of a decision tree. Assume we have two discrete random variables $X$ and $Y$, and we define IG using the conditional entropy as $IG(X,Y) = H(Y) - H(Y|X)$.

a) [3Pts] Show that $IG(X,Y) = IG(Y,X) = H(X) + H(Y) - H(X,Y)$.

b) [4Pts] Kullback-Leibler (KL) divergence from a distribution $P(X)$ to a distribution $Q(X)$ is defined as:

$$KL(P||Q) = -\sum_{i=1}^{n} P(X = i) \log \frac{P(X = i)}{Q(X = i)}$$

and it can be considered as a measure of dissimilarity from $P$ to $Q$ (Bishop, Section 1.6). Show that $IG(X,Y) = KL(P(X,Y)||P(X)P(Y))$.

c) [2 Pts] Based on Part (a), how can we interpret information gain in terms of dependencies between random variables?

d) [3 Pts] Describe why is there a bias in $IG(X,Y)$ measure that favors variables with many values over those with few values?

3.2 ID3 Algorithm

Consider the training set in Table 1 containing 8 training samples.
Table 1. Training samples

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Weak</td>
<td>1</td>
</tr>
<tr>
<td>Mid</td>
<td>Weak</td>
<td>2</td>
</tr>
<tr>
<td>High</td>
<td>Weak</td>
<td>2</td>
</tr>
<tr>
<td>Low</td>
<td>Weak</td>
<td>1</td>
</tr>
<tr>
<td>Mid</td>
<td>Weak</td>
<td>2</td>
</tr>
<tr>
<td>Mid</td>
<td>Strong</td>
<td>1</td>
</tr>
<tr>
<td>High</td>
<td>Strong</td>
<td>2</td>
</tr>
<tr>
<td>Mid</td>
<td>Strong</td>
<td>2</td>
</tr>
</tbody>
</table>

a) [2 Pts] Compute the sample entropy $H(Y)$ (with logarithms base 2)?
b) [2 Pts] Calculate the information gains $IG(X_1, Y)$ and $IG(X_2, Y)$?
c) [3 Pts] Create decision tree based on the ID3 algorithm.
d) [4 Pts] Repeat Part (b) using the Gini index $Gini(X, Y) = 1 - \sum_{i=1}^{K} P(Y = i | X)^2$ (where $K = 2$).
e) [2 Pts] For the Decision tree you have made in Part (c) calculate the misclassification error over the following validation set.
f) [3 Pts] Use “error reduced pruning” algorithm to prune the decision tree. Would any node be pruned from this decision tree?
g) [2 Pts] Is the error on the validation set after pruning can be used to evaluate the generalization capability of the resulted decision tree?

Table 2. Validation samples

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mid</td>
<td>Weak</td>
<td>2</td>
</tr>
<tr>
<td>Low</td>
<td>Weak</td>
<td>1</td>
</tr>
<tr>
<td>High</td>
<td>Strong</td>
<td>2</td>
</tr>
<tr>
<td>Low</td>
<td>Strong</td>
<td>1</td>
</tr>
</tbody>
</table>