Natural Language Processing (CO3086)
NLP 242 - Lab 6: Linear - Losgistic Regression

HO CHI MINH UNIVERSITY OF TECHNOLOGY
Vietnam National University Ho Chi Minh

Problem 1

We are dealing with samples x where x is a single value. We would like to test two alternative regression models:

1. $y = ax + e$
2. $y = ax + bx^2 + e$

We make the same assumptions we had in class about the distribution of e ($e \sim N(0, s^2)$).

a) Assume we have n samples: $x_1, \ldots, x_n$ with their corresponding y values: $y_1, \ldots, y_n$. Derive the value assigned to b in model 2. You can use a in the equation for b.

b) Which of the two models is more likely to fit the *training* data better and Explain?

(a) model 1
(b) model 2
(c) both will fit equally well
(d) impossible to tell

c) Which of the two models is more likely to fit the *test* data better and Explain?

(a) model 1
(b) model 2
(c) both will fit equally well
(d) impossible to tell

Problem 2

a) Now assume we only observe a single input for each output (that is, a set of $\{x, y\}$ pairs). We would like to compare the following two models on our input dataset (for each one we split into training and testing sets to evaluate the learned model). Assume we have an unlimited amount of data:

A: $y = w^2 x$

B: $y = wx$

Which of the following is correct and Explain:

(a) There are datasets for which A would perform *better* than B.

(b) There are datasets for which B would perform *better* than A.

(c) Both 1 and 2 are correct.

(d) They would perform equally well on all datasets.

b) For the data above we are now comparing the following two models:

$$\textbf{A:} \quad y = w_1^2 x + w_2 x$$
$$\textbf{B:} \quad y = wx$$

Note that model A now uses two parameters (though both multiply the same input value, x). Again, we assume unlimited data. Which of the following is correct (choose the answer that best describes the outcome) and Explain:

(a) There are datasets for which A would perform *better* than B.

(b) There are datasets for which B would perform *better* than A.

(c) Both 1 and 2 are correct.

(d) They would perform equally well on all datasets.

Problem 3

We are given a set of two-dimensional inputs and their corresponding output pair: $\{x_{i,1}, x_{i,2}, y_i\}$. We would like to use the following regression model to predict y:

$$y_i = w_1^2 x_{i,1} + w_2^2 x_{i,2}$$

Derive the optimal value for w_1 when using least squares as the target minimization function (w_2 may appear in your resulting equation). Note that there may be more than one possible value for w_1.

Problem 4

You are asked to use regularized linear regression to predict the target $Y \in \mathbb{R}$ from the eight-dimensional feature vector $X \in \mathbb{R}^8$. You define the model $Y = w^T X$ and then you recall from class the following three objective functions:

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 \tag{4.1}$$

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 + \lambda \sum_{j=1}^{8} w_j^2 \tag{4.2}$$

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 + \lambda \sum_{j=1}^{8} |w_j| \tag{4.3}$$

a) Show regularization terms in the objective functions above.

b) For large values of λ in objective 4.2 the bias would increase, decrease or remain unaffected?

c) For large values of λ in objective 4.3 the variance would increase, decrease or remain unaffected?

d) The following table contains the weights learned for all three objective functions (not in any particular order):

	Column A	**Column B**	**Column C**
w_1	0.60	0.38	0.50
w_2	0.30	0.23	0.20
w_3	-0.10	-0.02	0.00
w_4	0.20	0.15	0.09
w_5	0.30	0.21	0.00
w_6	0.20	0.03	0.00
w_7	0.02	0.04	0.00
w_8	0.26	0.12	0.05

Problem 5

Suppose you are given the following classification task: predict the target $Y \in \{0, 1\}$ given two real valued features $X_1 \in \mathbb{R}$ and $X_2 \in \mathbb{R}$. After some training, you learn the following decision rule:

$$\textbf{Predict } Y = 1 \textbf{ iff } w_1 X_1 + w_2 X_2 + w_0 \geq 0 \textbf{ and } Y = 0 \textbf{ otherwise}$$

where $w_1 = 3$, $w_2 = 5$, and $w_0 = -15$.

a) Plot the decision boundary and label the region where we would predict $Y = 1$ and $Y = 0$.

b) Suppose that we learned the above weights using logistic regression. Using this model, what would be our prediction for $P(Y = 1 \mid X_1, X_2)$? (You may want to use the sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$.)

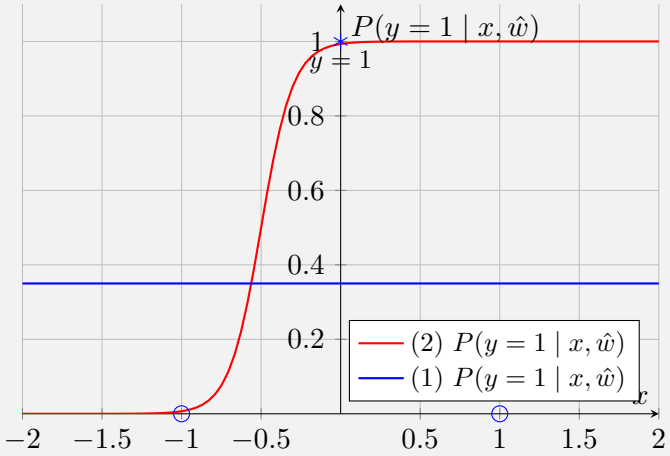
$$P(Y = 1 \mid X_1, X_2) =$$

Problem 6

Consider a simple one-dimensional logistic regression model

$$P(y = 1 \mid x, \mathbf{w}) = g(w_0 + w_1 x)$$

where $g(z) = \frac{1}{1+\exp(z)}$ is the logistic function. The following figure shows two possible conditional distributions $P(y = 1 \mid x; \mathbf{w})$, viewed as a function of x, that we can get by changing the parameters $\mathbf{w}$.



a) Please indicate the number of classification errors for each conditional given the labeled examples in the same figure.

b) One of the two classifiers corresponds to the maximum likelihood setting of the parameters w based on the labeled data in the figure, i.e., its parameters maximize the joint probability:

$$P(y=0 \mid x=-1; w) \quad P(y=1 \mid x=0; w) \quad P(y=0 \mid x=1; w)$$

Circle which one is the ML solution and briefly explain: **Classifier 1** or **Classifier 2**

c) Would adding a regularization penalty $|w_1|^2/2$ to the log-likelihood estimation criterion affect your choice of solution (Y/N)? (Note that the penalty above only regularizes w_1, not w_0.) Briefly explain why.

Problem 7

In many real-world scenarios, our data has millions of dimensions, but a given example has only hundreds of non-zero features. For example, in document analysis with word counts for features, our dictionary may have millions of words, but a given document has only hundreds of unique words. In this question, we will make l_2 regularized SGD efficient when our input data is sparse. Recall that in l_2 regularized logistic regression, we want to maximize the following objective **(in this problem we have excluded w_0 for simplicity)**:

$$F(\mathbf{w}) = \frac{1}{N}\sum_{j=1}^{N} l(x^{(j)}, y^{(j)}, \mathbf{w}) - \frac{\lambda}{2}\sum_{i=1}^{d} w_i^2$$

where $l(x^{(j)}, y^{(j)}, \mathbf{w})$ is the logistic objective function

$$l(x^{(j)}, y^{(j)}, \mathbf{w}) = y^{(j)}\left(\sum_{i=1}^{d} w_i x_i^{(j)}\right) - \ln\left(1 + \exp\left(\sum_{i=1}^{d} w_i x_i^{(j)}\right)\right)$$

and the remaining sum is our regularization penalty.mWhen we do stochastic gradient descent

on point $(x^{(j)}, y^{(j)})$, we are approximating the objective function as

$$F(\mathbf{w}) \approx l(x^{(j)}, y^{(j)}, \mathbf{w}) - \frac{\lambda}{2} \sum_{i=1}^{d} w_i^2$$

Definition of sparsity: Assume that our input data has d features, i.e., $\mathbf{x}^{(j)} \in \mathbb{R}^d$. In this problem, we will consider the scenario where $x^{(j)}$ is sparse. Formally, let s be the average number of nonzero elements in each example. We say that the data is sparse when $s \ll d$. In the following questions, **your answer should take the sparsity of $x^{(j)}$ into consideration when possible.**

Note: When we use a sparse data structure, we can iterate over the non-zero elements in $O(s)$ time, whereas a dense data structure requires $O(d)$ time.

a) Let us first consider the case when $\lambda = 0$. Write down the SGD update rule for $\mathbf{w}$, where $\lambda = 0$, using step size η, when the example $(x^{(j)}, y^{(j)})$ is given.

b) If we use a dense data structure, what is the average time complexity to update $\mathbf{w}_i$ when $\lambda = 0$? What if we use a sparse data structure? Justify your answer in one or two sentences.

c) Now let us consider the general case when $\lambda > 0$. Write down the SGD update rule for $\mathbf{w}_i$ when $\lambda > 0$, using step size η, given the example $(x^{(j)}, y^{(j)})$.

d) If we use a dense data structure, what is the average time complexity to update $\mathbf{w}_i$ when $\lambda > 0$?

e) Let $\mathbf{w}_i^{(t)}$ be the weight vector after t-th update. Now imagine that we perform k SGD updates on $\mathbf{w}$ using examples $(x^{(t+1)}, y^{(t+1)}), \ldots, (x^{(t+k)}, y^{(t+k)})$, where $x_i^{(j)} = 0$ for every example in the sequence. (i.e. the i-th feature is zero for all of the examples in the sequence). Express the new weight, $\mathbf{w}_i^{(t+k)}$, in terms of $\mathbf{w}_i^{(t)}$, k, η, and λ.

f) Using your answer in the previous part, come up with an efficient algorithm for regularized SGD when we use a sparse data structure. What is the average time complexity per example? (Hint: when do you need to update w_i?)

Algorithm 1: Sparse SGD Algorithm for Logistic Regression with Regularization

Initialize $c_i \leftarrow 0$ for $i \in \{1, 2, \ldots, d\}$
for $j \in \{1, 2, \ldots, n\}$ **do**
  $\hat{p} \leftarrow \frac{1}{1+\exp\left(-\sum_k w_k x_k^{(j)}\right)}$
  for i such that $x_i^{(j)} \neq 0$ **do**
    $k \leftarrow j - c_i$ {auxiliary variable c_i holds the index of last time we see $x_i^{(j)} \neq 0$}
    $w_i \leftarrow w_i(1 - \eta\lambda)^k$ {apply all the regularization updates}
    $w_i \leftarrow w_i + \eta x_i^{(j)}(y^{(j)} - \hat{p})$ {regularization is done in previous step}
    $c_i \leftarrow j$ {remember last time we see $x_i^{(j)} \neq 0$}
  end for
end for