CS584 Assignment 3: Report

Jingyu Zhu A20311779

1. Problem Statement

In this assignment, I try to classify the data via algorithms of discriminative learning. First, I implemented the logistic regression algorithm for two-class discrimination with linear features. Then I increased the dimension of the inputs and applied the algorithm again so that I could make a comparison for the performance. After that, I applied the softmax regression for the K-class discrimination and also measured the performance.

For the second part, I derived the backpropagation update equations by minimizing the error function instead of maximizing the likelihood function, from which we could see why in this case maximizing the likelihood function is better for the gradient descent algorithm.

2. Proposed Solution

The main idea to implement this algorithm is to use gradient descent. Before that, we should derive the parameter by differentiating the $h_\theta(x)$ function.

$$h_\theta(x) = \frac{1}{1 + \exp(-\theta^T x)}$$

Solve it and we can have following backpropagation update equation:

$$\theta_j \leftarrow \theta_{j-1} - \eta \sum_{i=1}^{m} (h_\theta(x^i) - y^i) \cdot x^i$$

After we converge by iterating the equation above, we get the final theta vector. Then we can calculate $h_\theta(x)$ for all testing examples. If $h_\theta(x) > 0.5$, it is classified to class 1. Otherwise, it will be classified to class 0.

In respect of the k-class discrimination, we can have a similar hypothesis function from softmax regression:

$$h_{\theta j}(x) = \frac{\exp\,(\theta_j^T x)}{\sum_{i=1}^{k} \mathrm{xp}\,(\theta_i^T x)}$$

And the backpropagation update equation:

$$\theta_j \leftarrow \theta_{j-1} - \eta \sum_{i=1}^{m} (h_{\theta_j}(x^i) - 1(y^i = j)) \cdot x^i$$

After we get the final theta, we can do the classification.

To implement a two-layer feedforward MLP for three-class classification, we can calculate the parameters by following equations:

$$v_j \leftarrow v_{j-1} - \eta \sum_{i=1}^{m} (\hat{y}^i - y^i) \cdot z^i$$

$$w_j^m \leftarrow w_{j-1}^m - \eta \sum_{i=1}^{m} (\hat{y}^i - y^i) \cdot v_m \cdot z_m^i \cdot (1 - z_m^i) \cdot x^i$$

After that, we can do the classification.

3. Implementation Details

To implement the logistic regression algorithm, we choose the sigmoid function as our hypothesis function.

$$h_\theta(x) = sigmoid(x) = \frac{1}{1 + \exp(-\theta^T x)}$$

Then we have following:

$$P(y = 1|x) = 1 - P(y = 0|x) = \frac{1}{1 + \exp(-\theta^T x)}$$

We define the likelihood function as follow:

$$l(\theta) = \prod_{x\ in\ class\ 1} P(y = 1|x) \cdot \prod_{x\ in\ class\ 0} P(y = 0|x)$$

The log likelihood functions is:

$$l(\theta) = \sum_{i=1}^{m} y^i \log\left(h_\theta\left(x^i\right)\right) + (1 - y^i)\log(1 - h_\theta(x^i))$$

Since we have following:

$$\frac{d}{dx} sigmoid(x) = sigmoid(x) \cdot (1 - sigmoid(x))$$

and

$$\frac{d}{dx} log(sigmoid(x)) = 1 - sigmoid(x)$$

We differentiate $l(\theta)$, make it equal to zero to compute the maximum and plug in the two equations above. Finally, we have:

$$\sum_{i=1}^{m}(y^i - h_\theta(x^i)) \cdot x^i = 0$$

Therefore,

$$\theta_j \leftarrow \theta_{j-1} - \eta \sum_{i=1}^{m}(h_\theta(x^i) - y^i) \cdot x^i$$

Similarly, from the equation:

$$h_{\theta j}(x) = \frac{\exp(\theta_j^T x)}{\sum_{i=1}^{k} \text{xp}(\theta_i^T x)}$$

,along with the property of the sigmoid function, we can derive the softmax backpropagation update equation:

$$\theta_j \leftarrow \theta_{j-1} - \eta \sum_{i=1}^{m}(h_{\theta_j}(x^i) - 1(y^i = j)) \cdot x^i$$

To implement a two-layer feedforward MLP, first we define the likelihood function as follow:

$$l(\theta) = \prod_{i=1}^{m}\prod_{j=1}^{k} P(y^i|x)^{1(y^i=j)}$$

Then we calculate the log likelihood function and differentiate it, we can get the backpropagation equations:

$$v_j \leftarrow v_{j-1} - \eta \sum_{i=1}^{m}(\hat{y}^i - y^i) \cdot z^i$$

$$w_j^m \leftarrow w_{j-1}^m - \eta \sum_{i=1}^{m}(\hat{y}^i - y^i) \cdot v_m \cdot z_m^i \cdot (1 - z_m^i) \cdot x^i$$

After that, we plug in these parameters into each sigmoid function and we can calculate the output of the testing data by evaluating the values of the sigmoid functions. With the output of MLP, we can classify the data.

4. Results and discussion

(1) Logistic Regression for two-class discrimination with linear features

```
> resultMatrix
     [,1] [,2]
[1,]   50    0
[2,]    0   50
> precisionOfClass0
[1] 1
> precisionOfClass1
[1] 1
> recallOfClass0
[1] 1
> recallOfClass1
[1] 1
> FMeasureOfClass0
[1] 1
> FMeasureOfClass1
[1] 1
> accuracy
[1] 1
```

(2) Logistic Regression for two-class discrimination with non-linear features

In this case, I add two features, which are feature1*feature2, feature3*feature4, and the result is as follow:

```
> resultMatrix
     [,1] [,2]
[1,]   50    0
[2,]    0   50
> precisionOfClass0
[1] 1
> precisionOfClass1
[1] 1
> recallOfClass0
[1] 1
> recallOfClass1
[1] 1
> FMeasureOfClass0
[1] 1
> FMeasureOfClass1
[1] 1
> accuracy
[1] 1
```

We can see from above that the performances for both cases are perfect. (No errors appeared)

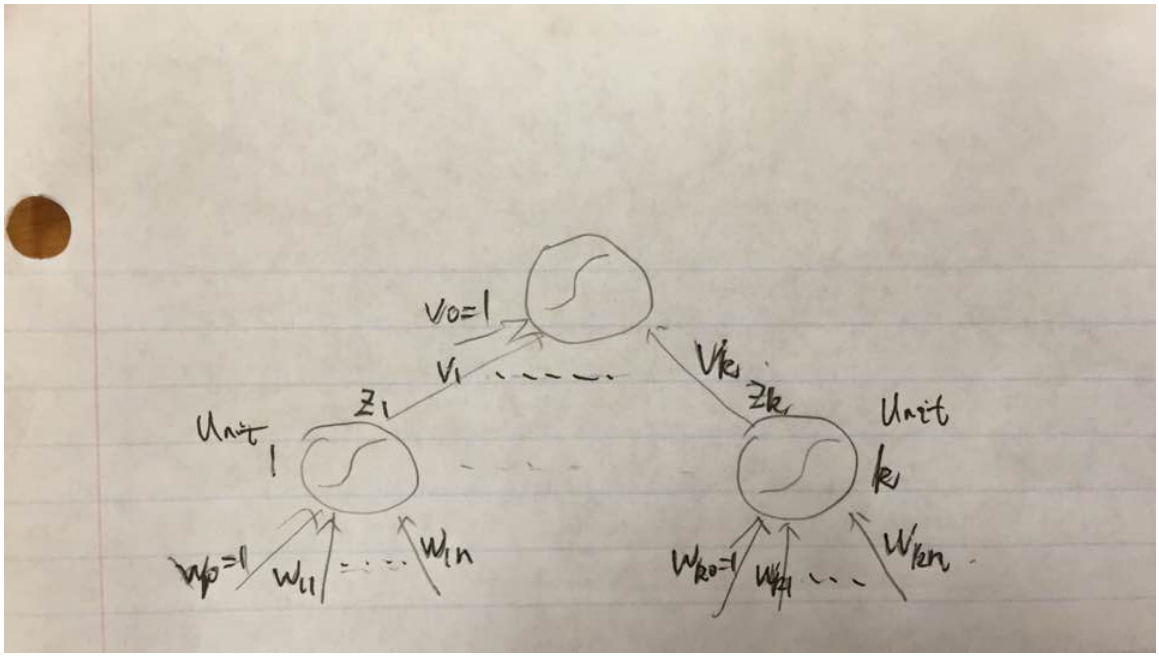
(3) Logistic Regression for k-class discrimination (k is three in this case)

```
> resultMatrix
     [,1] [,2] [,3]
[1,]   50    0    0
[2,]    0   48    0
[3,]    0    2   50
> precisionOfClass0
[1] 1
> precisionOfClass1
[1] 1
> precisionOfClass2
[1] 0.9615385
> recallOfClass0
[1] 1
> recallOfClass1
[1] 0.96
> recallOfClass2
[1] 1
> FMeasureOfClass0
[1] 1
> FMeasureOfClass1
[1] 0.9795918
> FMeasureOfClass2
[1] 0.9803922
> accuracy
[1] 0.9866667
```

We can see that I got two errors in this case, which is acceptable.

(4) Derive the backpropagation update equations of MLP

We can derive them by following the steps below:



The error function is:

$$E(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

By minimizing the error function, we have following:

$$\frac{\partial E}{\partial v} = \frac{\partial E}{\partial g}\cdot\frac{\partial g}{\partial v} = 2\cdot\frac{1}{2}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})\cdot\hat{y}^{(i)}(1-\hat{y}^{(i)})\cdot Z^{(i)}$$

$$= \sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})\cdot\hat{y}^{(i)}(1-\hat{y}^{(i)})\cdot Z^{(i)}$$

$$v \leftarrow v - \eta\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})\cdot\hat{y}^{(i)}(1-\hat{y}^{(i)})\cdot Z^{(i)}$$

$$\frac{\partial E}{\partial w_j} = \frac{\partial E}{\partial \hat{y}}\cdot\frac{\partial \hat{y}}{\partial z_j}\cdot\frac{\partial z_j}{\partial w_j} \qquad j = 1,\ldots,k$$

$$= 2\cdot\frac{1}{2}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})\cdot\hat{y}^{(i)}(1-\hat{y}^{(i)})\cdot v_j^{(i)}\cdot Z_j^{(i)}(1-Z_j^{(i)})X^{(i)}$$

$$\therefore w_j \leftarrow w_j - \eta \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) \cdot \hat{y}^{(i)}(1 - \hat{y}^{(i)}) \cdot v_j^{(i)} \cdot Z_j^{(i)}(1 - Z_j^{(i)}) X^{(i)}$$