Stat406 Spring 2010: homework 3

1 Gradient and Hessian of log-likelihood for logistic regression

1. Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Show that

$$\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a)) \tag{1}$$

2. Using the previous result and the chain rule of calculus, show that

$$\frac{d}{d\mathbf{w}} - \sum_{i=1}^{N} \log[\mu_i^{y_i} \times (1 - \mu_i)^{1-y_i}] = \sum_i (\mu_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\mu} - \mathbf{y}) \tag{2}$$

3. The Hessian can be written as $\mathbf{H} = \mathbf{X}^T\mathbf{S}\mathbf{X}$, where $\mathbf{S} \overset{\text{def}}{=} \text{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Show that $\mathbf{H}$ is positive definite. (You may assume that $0 < \mu_i < 1$, so the elements of $\mathbf{S}$ will be strictly positive, and that $\mathbf{X}$ is full rank.)

2 Regularizing separate terms in 2d logistic regression

(Source: Jaakkola)

1. Consider the data in Figure 1, where we fit the model $p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1x_1 + w_2x_2)$. Suppose we fit the model by maximum likelihood, i.e., we minimize

$$J(\mathbf{w}) = -\ell(\mathbf{w}, \mathcal{D}_{\text{train}}) \tag{3}$$

where $\ell(\mathbf{w}, \mathcal{D}_{\text{train}})$ is the log likelihood on the training set. Sketch a possible decision boundary corresponding to $\hat{\mathbf{w}}$. (Copy the figure first (a rough sketch is enough), and then superimpose your answer on your copy, since you will need multiple versions of this figure). Is your answer (decision boundary) unique? How many classification errors does your method make on the training set?

2. Now suppose we regularize only the w_0 parameter, i.e., we minimize

$$J_0(\mathbf{w}) = -\ell(\mathbf{w}, \mathcal{D}_{\text{train}}) + \lambda w_0^2 \tag{4}$$

Suppose λ is a very large number, so we regularize w_0 all the way to 0, but all other parameters are unregularized. Sketch a possible decision boundary. How many classification errors does your method make on the training set? Hint: consider the behavior of simple linear regression, $w_0 + w_1x_1 + w_2x_2$ when $x_1 = x_2 = 0$.

3. Now suppose we heavily regularize only the w_1 parameter, i.e., we minimize

$$J_1(\mathbf{w}) = -\ell(\mathbf{w}, \mathcal{D}_{\text{train}}) + \lambda w_1^2 \tag{5}$$

Sketch a possible decision boundary. How many classification errors does your method make on the training set?

4. Now suppose we heavily regularize only the w_2 parameter. Sketch a possible decision boundary. How many classification errors does your method make on the training set?

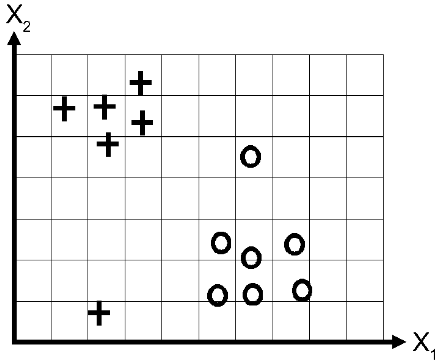

Figure 1: Data for logistic regression question.

3 Spam classification using logistic regression

Consider the email spam data set discussed on p300 of [**?**]. This consists of 4601 email messages, from which 57 features have been extracted. These are as follows:

- 48 features giving the percentage (0 to 100) of words in a given message which match a given word on the list. The list contains words such as “business”, “free”, “george”, etc. (The data was collected by George Forman, so his name occurs quite a lot.)
- 6 features giving the percentage (0 to 100) of characters in the email that match a given character on the list. The characters are ; ([! $ #
- Feature 55: The average length of an uninterrupted sequence of capital letters (max is 40.3, mean is 4.9)
- Feature 56: The length of the longest uninterrupted sequence of capital letters (max is 45.0, mean is 52.6)
- Feature 57: The sum of the lengts of uninterrupted sequence of capital letters (max is 25.6, mean is 282.2)

Load the data from `spamData.mat`, which contains a training set (of size 3065) and a test set (of size 1536). One can imagine performing several kinds of preprocessing to this data. Try each of the following separately:

1. Standardize the columns so they all have mean 0 and unit variance.
2. Transform the features using $\log(x_{ij} + 0.1)$.
3. Binarize the features using $\mathbb{I}(x_{ij} > 0)$.

For each version of the data, fit a logistic regression model. Use cross validation to choose the strength of the ℓ_2 regularizer. Report the mean error rate on the training and test sets. You should get numbers similar to this:

```
method   train   test
stnd     0.082   0.079
log      0.052   0.059
binary   0.065   0.072
```

(The precise values will depend on what regularization value you choose.) Turn in your code and numerical results.

References

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009. 2nd edition.