Due Date: 1399.09.11

Homework 5

Theoretical

1. **(15 points)** Answer the following questions and explain your answers.
   (a) Why isn't the SSE loss function suitable for classification?
   (b) What are the advantages of using the perceptron criterion instead of an error function that counts the number of misclassified datapoints?
   (c) What are the advantages of using the logistic regression error function instead of the perceptron criterion?
   (d) We can optimize a logistic regression classifier with gradient descent. Why do we sometimes use iterative reweighted least squares? What are the advantages?
   (e) Which one of logistic regression or probit regression are more sensitive to outliers? Why?

2. **(8 points)** Prove that decision regions of a K-class discriminant (with K linear functions) are convex.

3. **(8 points)** Suppose that we have a linearly separable dataset and there is a weight vector w^* that

$$\|w^*\| = 1,$$
$$\forall i \in \{1, \ldots, N\} \quad {w^*}^T x^{(i)} t^{(i)} \geq \gamma > 0.$$

   Besides, suppose that

$$\forall i \in \{1, \ldots, N\} \quad \|x^{(i)}\| \leq R.$$

   Show that non-batch perceptron algorithm will have at most $\frac{R^2}{\gamma^2}$ steps before convergence. (Let the initial value of w be $\mathbf{0}$)

4. **(8 points)** Consider a two-class, d-dimensional probabilistic Gaussian (shared covariance matrix) Bayes classifier. Form the likelihood function and using the maximum likelihood technique, find the estimates of prior class probabilities $(p(C_1), p(C_2))$, μ_1, μ_2, and Σ. Write all needed equations.

5. **(8 points)** Show that for a linearly separable dataset, the maximum likelihood solution for the logistic regression model is obtained by finding a vector w so that the decision boundary separates the classes and then taking the magnitude of w to infinity.

6. **(8 points)** What happens to a logistic regression classifier if the dataset has two identical features? For example, consider the following datasets

$$D_1 = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\},$$
$$D_2 = \{(x^{(1)}, x^{(1)}, y^{(1)}), (x^{(2)}, x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, x^{(N)}, y^{(N)})\}.$$

   How would the performance of a logistic regression classifier differ when trained on D_2 instead of D_1?

7. **(5 points)** Show that the probit function

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta$$

and the erf function

$$\text{erf}(a) = \frac{2}{\sqrt{\pi}} \int_{0}^{a} e^{-\theta^2} d\theta$$

are related by

$$\Phi(a) = \frac{1}{2}\left(1 + \text{erf}(\frac{a}{\sqrt{2}})\right).$$

8. **(Additional 8 points)** Write down expressions for the gradient of the log likelihood and the corresponding Hessian matrix, for the probit regression model.

Practical

In this part, we are going to classify our data samples by different classification methods you have learnt.

Dataset: You are going to work with Iris dataset[1] in this assignment. The dataset contains 3 classes (0, 1, 2) of 50 instances each, where each class refers to a type of iris plant. Each data sample has 4 real attributes.

Allowed packages: Pandas, matplotlib, and numpy. Sklearn is allowed only for getting the Iris dataset. You are not going to use any other predefined functions in this part.

Assignment: Hand in your report in pdf and your codes in Python. (You may also use Jupyter Notebooks instead.) For each classification method, you should split your dataset into two sets of training data (80% of the data) and test data (20% of the data) at first. You may shuffle the dataset before splitting.

1. **(Additional 20 points) Perceptron.** Now, you are going to classify data samples with Perceptron method. To this end, you should consider data samples from classes 0 and 1. After splitting the dataset, train a Perceptron classifier on your training data. In your report, plot the number of misclassified samples per iteration. Finally, you should report the accuracy of your classifier as well as a confusion matrix on test data.

2. **(40 points) Logistic Regression.**

   (a) In this part, you are going to train a logistic regression classifier on the whole 3-class training set. Report a plot showing the loss value of training data per iteration. Besides, report the accuracy of your classifier and its confusion matrix on test data.

   (b) You may know that logistic regression can exhibit severe overfitting for datasets that are linearly separable. Hence, add the L_2 regularizer ($||w||^2$) to the loss function and report accuracy and confusion matrix on test data. You should employ the regularizer term with these weights: $\{10^{-1}, 10^0, 10^1\}$. Compare the results of these three classifiers with the previous part.

Good Luck ;)

[1] Iris dataset