Homework Assignment #3

Assigned: Tuesday 03/31/2015; Due: Monday 04/13/2015 11:59pm (via Onourse).

(total: 105 points)

Problem 1. (5 points) Consider a logistic regression problem where $\mathcal{X} = \mathbb{R}^k$ and $\mathcal{Y} = \{-1, +1\}$. Derive the weight update rule that maximizes the likelihood.

Problem 2. (20 points) Consider a logistic regression problem with its initial solution obtained through the OLS regression, i.e. $\mathbf{w}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, in the context of the code provided in class (week 6). Recall that $\mathbf{x}$ had a Gaussian distribution and that $\dim\{\mathbf{x}\} = 2$ (before adding a column of ones) and that $y \in \{0, 1\}$. You probably noticed that the initial separation line is consistently closer to the data points of class 0.

a) (10 points) Why is this the case? Draw a picture (if possible) to support your argument.
b) (5 points) Devise a better initial solution by modifying the standard formula $\mathbf{w}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
c) (5 points) Now again consider the case where $y \in \{-1, +1\}$. What is the form of the modified solution from part (b) in this case?

Problem 3. (15 points) Consider the same situation as in previous question. We used logistic regression to solve this classification problem, but here we want to compare that solution to the optimal decision surface (line or curve).

a) (5 points) Find the optimal decision surface assuming that each class-conditional distribution is defined as a two-dimensional Gaussian distribution:

$$p(\mathbf{x}|y = i) = \frac{1}{(2\pi)^{k/2}} \cdot \frac{1}{|\mathbf{\Sigma}_i|^{1/2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \cdot \mathbf{\Sigma}_i^{-1} \cdot (\mathbf{x}-\mathbf{m}_i)}$$

where $i \in \{0,1\}$, $\mathbf{m}_0 = (1, 2)$, $\mathbf{m}_1 = (6, 3)$, $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $P(y = 0) = P(y = 1) = \frac{1}{2}$, and $|\mathbf{\Sigma}_i|$ is a determinant of $\mathbf{\Sigma}_i$. Use Chapter 1 (Section 1.5) from your textbook for hints on making optimal classification decisions.

b) (5 points) Generalize the solution from part (a) for the case when $\mathbf{m}_0 = (m_{01}, m_{02})$, $\mathbf{m}_1 = (m_{11}, m_{12})$, $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$, and $P(y = 0) \neq P(y = 1)$.

c) (5 points) Generalize the solution from part (b) to arbitrary covariance matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$. Discuss the shape of the optimal decision surface.

Problem 4. (20 points) We showed in class that the perceptron training rule guarantees convergence of the training algorithm when data sets are linearly separable. Your task in this exercise is to verify this claim experimentally and comment on what you observe. Start by implementing perceptron training, e.g. by using the code from class but modify the data generation code (say, from the EM algorithm code, logistic regression, etc.), to incorporate data of several different dimensionalities $2 \leq k \leq 100$ and data set sizes $100 \leq n \leq 10000$ (for simplicity use that the number of positive and negative examples are roughly identical). Using Gaussian class-conditional distributions will suffice but if you prefer to use other data generators, do not hesitate to do so. Use the logistic regression algorithm to keep only those data sets that are linearly separable (logistic regression may not perform this task perfectly, but it is good enough) and use its solution to obtain the set of "true" coefficients $\mathbf{w}_0$ and parameter ε from the class lecture notes #9.

a) (10 points) Record the number of weight updates (upon misclassification) and compare it with the theoretical limit $\ell_{\max}$ for each data set. Find an appropriate way to summarize (e.g. visualize) your recordings over many runs to support your claims. What do you observe? How tight is the theoretical limit? How does the total number of weight updates change with parameter k? Comment on all your findings.
b) (5 points) Repeat the process from step (a) for one chosen parameter $2 \leq k \leq 10$. Now explore the dependency of the number of weight updates with the maximum norm M of any data point in your data set. For each data set, construct a "normalized" replica with a controlled norm (smaller or larger) and run the perceptron training algorithm. Does the algorithm behave as expected. Comment on all your findings.
c) (5 points) Introduce the learning rate parameter $\eta \in (0,1]$ into the perceptron update rule. What changes are you noticing when η is used. Comment on all your findings.

Problem 5. (20 points) Consider a classification problem where $\mathcal{X} = \mathbb{R}^k$ and $\mathcal{Y} = \{0, 1\}$. Based on your understanding of the maximum likelihood estimation of weights in logistic regression, develop a linear classifier that models the posterior probability of the positive class as

$$P(y = 1|\mathbf{x}, \mathbf{w}) = \frac{1}{2} \cdot \left(1 + \frac{\mathbf{w}^T\mathbf{x}}{\sqrt{1 + (\mathbf{w}^T\mathbf{x})^2}}\right)$$

Implement (preferably in Matlab) the iterative weight update rule and compare the final decision line between logistic regression and this new linear classifier (use at least 10 different data sets to draw your conclusions).

Problem 6. (25 points) Ordinary Least Squares (OLS) linear regression vs. Generalized Linear Models (GLMs).

a) (5 points) Use data set crab.xlsx to develop a linear regression model for predicting the number of male satellites on a female's back. What are the sum-of-squares and mean-square-error of the OLS fit.

(20 points) Because the target variable (number of male satellites) is a non-negative integer, it may be beneficial to model it by a Poisson distribution, instead of Gaussian that is a base distribution in OLS linear regression. Read the Generalized Linear Models section from class notes (Instructor's Notes #6). Verify the correctness of the weight update rule for a loglinear link function with Poisson distribution and apply it to the crab data set. Compare both sum-of-squares and mean-square-error for the fits you achieved using OLS regression and GLM. Comment on the most important aspects of this fitting process.

Homework policies (read carefully):

Your assignment must be typed; for example, in Latex, Microsoft Word, Lyx, etc. Images may be scanned and inserted into the document if it is too complicated to draw them properly. Submit a single pdf document or if you are attaching your code submit your code together with the typed (single) document as one .zip file.

All code (if applicable) should be turned in when you submit your assignment. Use Matlab, Python or R.

Policy for late submission assignments: Unless there are legitimate circumstances, late assignments will be accepted up to 5 days after the due date and graded using the following rule:

on time:	your score $\times$ 1
1 day late:	your score $\times$ 0.9
2 days late:	your score $\times$ 0.7
3 days late:	your score $\times$ 0.5
4 days late:	your score $\times$ 0.3
5 days late:	your score $\times$ 0.1

For example, this means that if you submit 3 days late and get 80 points for your answers, your total number of points will be $80 \times 0.5 = 40$ points.

All assignments are individual, except when collaboration is explicitly allowed. All the sources used for problem solution must be acknowledged, e.g. web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously; for detailed information see Indiana University Code of Student Rights, Responsibilities, and Conduct.

Good luck!