CS 391D Data Mining: A Mathematical Perspective Fall 2009

Homework 3

Instructor: Inderjit Dhillon Date Due: October 27, 2009

Keywords: *Classification, Logistic Regression, Perceptron, Support Vector Machines*

Use Matlab for problem 3. Turn in your code along with your results in hard copy only. Note that the assignment is due IN CLASS.

1. (4 points) Given training instances (x_n, y_n) with $y_n \in \{0, 1\}$, consider the following error function for logistic regression:

$$E(w) = -\sum_{n=1}^{N}(y_n \log z_n + (1 - y_n)\log(1 - z_n)),$$

where $z_n = \sigma(w^T x_n)$, w specifies a hyperplane, and σ is the logistic sigmoid function defined by

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$

Prove that the error function $E(w)$ is a convex function and provide a condition on the input data so that $E(w)$ has a unique minimum.

2. (6 points) In this exercise, we will prove correctness and convergence of the Perceptron algorithm for linearly separable data.

Let w_t represent the hyperplane at step t and (x_t, y_t) represent an input instance with $y_t \in \{1, -1\}$. Note that the input data is padded with one, i.e. $x_t = \begin{bmatrix} x_t \\ 1 \end{bmatrix}$. Recall the update:

$$w_{t+1} = w_t + y_t x_t, \text{ if } y_t(w_t^T x_t) < 0, \text{ i.e., a mistake.}$$

Assume that all the input data points have bounded Euclidean norm, i.e., $\|x_t\| \leq R$ and are linearly separable with finite margin $\gamma > 0$, i.e., there exists a hyperplane specified by w^* such that:

$$y_t(w^{*T} x_t) \geq \gamma, \ \forall\ t.$$

   (a) (2 points) Prove that the following holds after t updates: $w^{*T} w_t \geq t\gamma$.

   (b) (2 points) Prove that: $\|w_t\|_2^2 \leq tR^2$.

   (c) (2 points) Using parts (a) and (b), prove that the Perceptron algorithm converges to a separating hyperplane after at most $\frac{R^2\|w^*\|_2^2}{\gamma^2}$ steps.

3. (6 points) In this exercise, we will compare the performance of least squares regression and logistic regression for a 3-class classification problem. The data set for this problem can be downloaded at `http://www.cs.utexas.edu/~wtang/cs391d/3gaussian.tar.gz`.

   (a) (2 points) Solve the classification problem by using least squares regression in Matlab. Plot the data points according to the predicted labels. What do you observe in your plot?

(b) (4 points) Solve the classification problem by using logistic regression by the Newton-Raphson method in Matlab. Plot the data points according to the predicted labels. What do you observe in your plot?

Hint: The Newton-Raphson update, for minimizing a function $E(\boldsymbol{w})$, takes the form

$$\boldsymbol{w}^{(new)} = \boldsymbol{w}^{(old)} - H^{-1}\nabla_{\boldsymbol{w}} E(\boldsymbol{w}),$$

where H is the Hessian, i.e., $H = \nabla^2_{\boldsymbol{w}} E(\boldsymbol{w})$ and both derivatives are evaluated at $\boldsymbol{w}^{(old)}$.

When we solve a K-class classification problem by logistic regression, the k-th element of prediction vector for point $\boldsymbol{x}_n$, $p(C_k|\boldsymbol{x}_n) = \boldsymbol{z}_k(\boldsymbol{x}_n) = \exp(a_k)/\sum_j \exp(a_j)$, where $a_k = \boldsymbol{w}_k^T \boldsymbol{x}_n$. The objective is to minimize

$$E(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} y_k(\boldsymbol{x}_n) \log z_k(\boldsymbol{x}_n),$$

where $y_k(\boldsymbol{x}_n)$ is the k-th element of target vector $\boldsymbol{y}(\boldsymbol{x}_n)$ for point $\boldsymbol{x}_n$.

The gradient of the error function w.r.t $\boldsymbol{w}_j$ is given by

$$\nabla_{\boldsymbol{w}_j} E(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) = \sum_{n=1}^{N} (z_j(\boldsymbol{x}_n) - y_j(\boldsymbol{x}_n))\boldsymbol{x}_n,$$

and the Hessian matrix H comprises blocks of size $(D+1) \times (D+1)$, where D is the dimensionality of the data points. The block i, j is given by

$$\nabla_{\boldsymbol{w}_i}\nabla_{\boldsymbol{w}_j} E(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) = \sum_{n=1}^{N} z_i(\boldsymbol{x}_n)(I_{ij} - z_j(\boldsymbol{x}_n))\boldsymbol{x}_n \boldsymbol{x}_n^T,$$

where I_{ij} is the (i, j)-th element of the $K \times K$ identity matrix I.

You can initialize $\boldsymbol{w}_0$ to be zero in your implementation.

4. (9 points) In this exercise, we will derive an algorithm for solving the SVM problem. Recall the dual formulation for the linearly-separable SVM:

$$\max_{\alpha} \quad W(\alpha), \qquad \text{where } W(\alpha) = \textstyle\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K_{ij}$$

$$\text{subject to} \quad \sum_{i=1}^{N} y_i \alpha_i = 0, \tag{1}$$

$$\alpha_i \geq 0, \quad i = 1, ..., N. \tag{2}$$

In the above problem, K_{ij} could be $\mathbf{x}_i^T \mathbf{x}_j$ or $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) = h(\mathbf{x}_i)^T h(\mathbf{x}_j)$. Note that the matrix K is positive semi-definite. The dual variables $\alpha_1, ..., \alpha_N$ are said to be feasible if (1) and (2) are satisfied. We will consider the following strategy for optimizing this problem: at each iteration, we start with a feasible α and then update exactly 2 α's at a time. The update must maintain feasibility. Assume without loss of generality that the variables to be updated are α_1 and α_2. In the following, you will derive an update to α_1 and α_2 that maximizes the dual problem given above when only α_1 and α_2 are allowed to change.

(a) (1 points) α_1 and α_2 are to be updated to $\bar{\alpha}_1$ and $\bar{\alpha}_2$. Using the constraints on α from the dual problem, show that if $y_1 = y_2$, then $\bar{\alpha}_2 \leq \alpha_1 + \alpha_2$, and if $y_1 \neq y_2$, then $\bar{\alpha}_2 \geq \alpha_2 - \alpha_1$.

(b) (2 points) Given that $y_1\alpha_1 + y_2\alpha_2 = \text{constant} = y_1\bar{\alpha}_1 + y_2\bar{\alpha}_2$, express this equivalently as $\alpha_1 + s\alpha_2 = \gamma$, where $s = y_1y_2$. Furthermore, let

$$v_i = \sum_{j=3}^{N} y_j\alpha_j K_{ij}, \quad i = 1, 2.$$

Write the dual objective as a function of α_1 and α_2 (fixing the other α variables as constants), then use the equation $\alpha_1 + s\alpha_2 = \gamma$ to express the dual as a function of only α_2, yielding

$$\begin{aligned} W(\alpha_2) &= \gamma - s\alpha_2 + \alpha_2 - \frac{1}{2}K_{11}(\gamma - s\alpha_2)^2 - \frac{1}{2}K_{22}\alpha_2^2 \\ &\quad -sK_{12}(\gamma - s\alpha_2)\alpha_2 - y_1(\gamma - s\alpha_2)v_1 - y_2\alpha_2 v_2 + \text{constant} \end{aligned}$$

(c) (3 points) Differentiate $W(\alpha_2)$ with respect to α_2 to calculate the maximizing $\bar{\alpha}_2$. Let $d_{12} = K_{11} - 2K_{12} + K_{22}$ for notational convenience. Justify why this solution is a maximum (not a minimum).

(d) (3 points) Let $E_i = f(\mathbf{x}_i) - y_i = (\sum_{j=1}^{N} \alpha_j y_j K_{ij} + w_0) - y_i$, i.e., the difference between the predicted value and the true class label. Simplify your result in part (c) to obtain the following:

$$\bar{\alpha}_2 = \alpha_2 + \frac{y_2(E_1 - E_2)}{d_{12}},$$

and then, using part (a), obtain the final solution for $\bar{\alpha}_2$ as:

$$\bar{\alpha}_2 := \begin{cases} \max(0, \min(\bar{\alpha}_2, \alpha_1 + \alpha_2)) & \text{if } y_1 = y_2, \\ \max(\bar{\alpha}_2, \alpha_2 - \alpha_1, 0) & \text{if } y_1 \neq y_2. \end{cases}$$

Furthermore, show that $\bar{\alpha}_1 = \alpha_1 + y_1y_2(\alpha_2 - \bar{\alpha}_2)$. This update results in a non-decreasing dual, and repeating over pairs of α eventually leads to global convergence of the SVM problem.