UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
Department of Electrical and Computer Engineering

ECE 544NA Pattern Recognition
Fall 2016

EXAM 1 SOLUTIONS

Tuesday, October 4, 2016

- This is a CLOSED BOOK exam. You may use one page, both sides, of handwritten notes.
- There are a total of 100 points in the exam. Plan your work accordingly.
- You must SHOW YOUR WORK to get full credit.

Problem	Score
1	
2	
3	
4	
5	
Total	

Name: ______________________________

Problem 1 (20 points)

Linear regression is defined by p-dimensional observation vectors, $\vec{x}_t$, and scalar targets, y_t, which can be arranged into matrices as

$$X = \begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_T^T \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}$$

The goal of linear regression is to find a weight vector $\vec{w} = [w_1, \ldots, w_p]^T$ to minimize $E = \|Y - X\vec{w}\|^2$. This can be done in closed form, as $\vec{w} = X^\dagger Y$, or using an iterative gradient descent algorithm, with iterations $\vec{w} \leftarrow \vec{w} - \eta \nabla_{\vec{w}} E$. Suppose that gradient descent requires m iterations, T is the number of training tokens, and p is the dimension of $\vec{x}_t$; in terms of m, T, and p, specify the computational complexity of the closed-form and gradient descent algorithms. Assume $T > p$.

(a) Closed-form:

SOLUTION: $\mathcal{O}\left\{Tp^2\right\}$ because $X^T X$ requires Tp^2, then inverting it requires p^3 but $p < T$.

(b) Gradient Descent:

SOLUTION: $\mathcal{O}\left\{mTp\right\}$ because there are m iterations each requiring the matrix-vector product of X^T with $(Y - X\vec{w})$.

Problem 2 (15 points)

A particular set of N swimmers is characterized by personality vectors $\vec{x}_n$, for $1 \le n \le N$. Each of the swimmers has tried T times to swim faster than a particular threshold time. Suppose that the variable $y_{nt} = 1$ if the n^{th} swimmer beat the target time on the t^{th} trial, otherwise $y_{nt} = 0$. A logistic regression model $\hat{y}_n = \vec{w}^T \vec{x}_n$ is trained in order to minimize

$$E = \frac{1}{2NT} \sum_{n=1}^{N} \sum_{t=1}^{T} (y_{nt} - \hat{y}_n)^2$$

Notice that $\hat{y}_n$ is a function of n, but not of t. Define $p_n = \frac{1}{T}\sum_{t=1}^{T} y_{nt}$ to be the fraction of victories achieved by the n^{th} swimmer. Find a formula for $\nabla_{\vec{w}} E$ that depends only on p_n, $\vec{w}$, and $\vec{x}_n$, and does not depend on y_{nt}.

SOLUUTION:

$$\nabla_{\vec{w}} E = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} (\hat{y}_n - y_{nt})\, \nabla_{\vec{w}} \hat{y}_n$$

$$= \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} (\hat{y}_n - y_{nt})\, g'(\vec{w}^T \vec{x}_n) \vec{x}_n,$$

where the $g'()$ was not graded because $g()$ was omitted from the problem statement. Simplifying,

$$\nabla_{\vec{w}} E = \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - p_n)\, g'(\vec{w}^T \vec{x}_n) \vec{x}_n$$

Problem 3 (15 points)

A support vector machine finds $\vec{w}$ in order to minimize

$$E = \frac{1}{2}\|\vec{w}\|^2 + CR_{data}$$

where

$$R_{data} = \sum_{t=1}^{T} \max\left(0, 1 - y_t\vec{w}^T\vec{x}_t\right)$$

where $1 \le t \le T$ is the token index, C is an arbitrary constant, $\vec{x}_t$ is the observation vector, and $y_t \in \{-1, 1\}$ is the target. Demonstrate that the optimum value of $\vec{w}$ (the value that sets $\nabla_{\vec{w}}E = 0$) can be expressed as a linear combination of some of the training vectors $y_t\vec{x}_t$.

SOLUTION:

$$\nabla_{\vec{w}}E = \vec{w} - \sum_{t\in I} y_t\vec{x}_t$$

which equals zero at

$$\vec{w} = \sum_{t\in I} y_t\vec{x}_t$$

where $I = \left\{t : 1 - y_t\vec{w}^T\vec{x}_T > 0\right\}$

Problem 4 (26 points)

The outputs $z_{jt}^{(L)}$ of a softmax function are defined in terms of its inputs $a_{jt}^{(L)}$ as

$$z_{jt}^{(L)} = \frac{e^{a_{jt}^{(L)}}}{\sum_{k=1}^{n} e^{a_{kt}^{(L)}}}$$

where $1 \le t \le T$ is the training token index, $1 \le j \le n$ is the output node number, and L is the number of layers in the neural network (thus layer number L is the last layer). The training corpus error may be defined as

$$E = -\sum_{t=1}^{T}\sum_{j=1}^{n} y_{jt}\log z_{jt}^{(L)}$$

where $y_{jt} \in \{0, 1\}$ is the training target.

(a) Define $\delta_{jt}^{(L)} = \partial E/\partial a_{jt}^{(L)}$. Give a formula for $\delta_{jt}^{(L)}$ in terms of $\vec{z}_{jt}^{(L)}$ and y_{jt}.

SOLUTION:

$$\frac{\partial E}{\partial a_{jt}^{(L)}} = -\sum_{i=1}^{n} \frac{y_{it}}{z_{it}^{(L)}}\frac{\partial z_{it}^{(L)}}{\partial a_{jt}^{(L)}}$$

where

$$\frac{\partial z_{it}^{(L)}}{\partial a_{jt}^{(L)}} = \begin{cases} z_{it}^{(L)}(1 - z_{it}^{(L)}) & i = j \\ -z_{it}^{(L)}z_{jt}^{(L)} & i \ne j \end{cases}$$

Therefore

$$\frac{\partial E}{\partial a_{jt}^{(L)}} = -y_{jt} + \sum_{i=1}^{n} y_{it} z_{jt}^{(L)}$$

(b) On Saturday October 1, 2016 in room 1005 of the Beckman Institute, Shuicheng Yang proposed that the fully-connected output layer of a CNN can be replaced by an average-pooling layer, defined similarly to the average-pooling final layer of a TDNN, thus:

$$a_{jt}^{(L)} = \sum_p \sum_q z_{jt}^{(L-1)}(p,q)$$

$$z_{jt}^{(L-1)}(p,q) = f(a_{jt}^{(L-1)}(p,q))$$

where p and q are the pixel indices in the $(L-1)^{\text{th}}$ layer, j is the channel index in both the $(L-1)^{\text{st}}$ and L^{th} layer, and $f()$ is a nonlinearity whose derivative is $f'()$. Define the back-prop errors to be

$$\delta_{jt}^{(L)} = \frac{\partial E}{\partial a_{jt}^{(L)}}, \quad \delta_{jt}^{(L-1)}(p,q) = \frac{\partial E}{\partial a_{jt}^{(L-1)}(p,q)}$$

Express $\delta_{jt}^{(L-1)}(p,q)$ in terms of of $\delta_{jt}^{(L)}$ and $f'(a_{jt}^{(L-1)}(p,q))$.

SOLUTION: $\delta_{jt}^{(L-1)}(p,q) = \delta_{jt}^{(L)} f'(a_{jt}^{(L-1)}(p,q))$.

Problem 5 (24 points)

Suppose we have a database of feature vectors $\vec{x}_t$ and associated labels $y_t \in \{-1, 1\}$, where $1 \le t \le T$.

- Define $\vec{z}_t$, for this problem only, to be the signed feature vector, $\vec{z}_t = y_t \vec{x}_t$.
- Define $\mathcal{W}_\infty$ to be the set of vectors $\vec{w}$ such that $\vec{w}^T \vec{z}_t > 0$ for all t.
- Assume linearly separable classes, which means that $\mathcal{W}_\infty$ is not an empty set.
- Define $\vec{w}_0 = \sum_{t=1}^{T} \vec{z}_t$

For each of the following statements, circle T if the statement is always true, circle F if the statement is sometimes false. If true, prove it. If false, disprove it (e.g., provide a training set $\{\vec{z}_1, \vec{z}_2\}$ that is linearly separable but disproves the claim; or you may use any other proof method).

(a) $\vec{w}_0^T \vec{w}_\infty > 0$, for all $\vec{w}_\infty \in \mathcal{W}_\infty$: T
Proof:

$$\vec{w}_0^T \vec{w}_\infty = \sum_{t=1}^{T} \vec{z}_t^T \vec{w}_\infty > 0$$

(b) $\vec{w}_0^T \vec{w}_\infty \ge 0$, for all $\vec{w}_\infty \in \mathcal{W}_\infty$: T
Proof: Part (a) implies part (b).

(c) The vector $\vec{w}_0$ is in the set $\mathcal{W}_\infty$: F?
Proof: A counter-example is the training dataset $\vec{z}_1 = [15, 1]^T$, $\vec{z}_2 = [-1, 1]^T$.

(d) $\vec{w} \in \mathcal{W}_\infty$ is unique (there is only one $\vec{w}$ such that $\vec{w}^T \vec{z}_t > 0$ for all t): F?
Proof: A counter-example is the training dataset $\vec{z}_1 = [15, 1]^T$, $\vec{z}_2 = [-1, 1]^T$.

In the following two subsections, define

$$\vec{w}_n = \vec{w}_{n-1} - \nabla_{\vec{w}_{n-1}} E_{n-1}$$

where

$$E_{n-1} = \sum_{t=1}^{T} \max\left(0, -\vec{w}_{n-1}^T \vec{z}_t\right)$$

(e) $\vec{w}_0^T \nabla_{\vec{w}_0} E_0 \le 0$: F
Proof: $\nabla_{\vec{w}_0} E_0 = -\sum_{t \in I} \vec{z}_t$, where $I = \left\{t : \vec{w}_0^T \vec{z}_t < 0\right\}$, therefore

$$\vec{w}_0^T \nabla_{\vec{w}_0} E_0 = -\sum_{t \in I} \vec{w}_0^T \vec{z}_t > 0$$

(f) $\vec{w}_1^T \vec{w}_\infty \ge 0$ for all $\vec{w}_\infty \in \mathcal{W}_\infty$: T
Proof:

$$\vec{w}_\infty^T \vec{w}_1 = \vec{w}_\infty^T \left(\vec{w}_0 + \sum_{t \in I} \vec{z}_t\right) = \sum_{t \notin I} \vec{w}_\infty^T \vec{z}_t + \sum_{t \in I} 2\vec{w}_\infty^T \vec{z}_t > 0$$