CS6501: Advanced Machine Learning **Spring 2017**

Quiz 1

Feb. 22^{th}, 2017

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam.
- This exam booklet contains **two** problems.
- You have 30 minutes to earn a total of 25 points.
- **Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.**

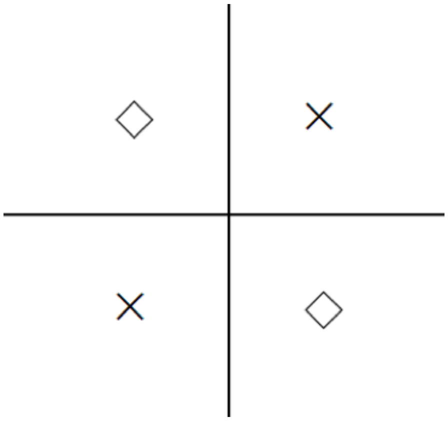
Good Luck!

Name (Computing ID): (1 Point)

Short Questions		/11
Multi-class classification		/13
Total		/25

Short Questions [11 points]

(a) [2 points] Circle all of the classifiers that will achieve zero training error on this data set.



- Logistic regression
- Linear-SVM
- 2-layer neural network (one hidden layer and one output layer)
- Perceptron

Ans. (c) (Note: the data is not linearly separable. The vanilla versions of (a), (b),(d) are all linear models.

(b) [4 points] In logistic regression, we assume the mapping between input $x \in R^n$ and the binary output $y \in \{1, -1\}$ is

$$P(y|x, w) = \frac{1}{1 + e^{-yw^T x}}. \tag{1}$$

Show that this is a valid probability assumption (i.e., show that $0 \leq P(y|x, w) \leq 1$ and $\sum_{y \in \{1,-1\}} P(y|x, w) = 1$).

Since $e^x \geq 0 \quad \forall x$, we have $1 \geq P(y|x, w) = \frac{1}{1+e^{-yw^T x}} \geq 0$
Using the given formula,

$$\begin{aligned} P(1|x, w) + P(-1|x, w) &= \frac{1}{1 + e^{-w^T x}} + \frac{1}{1 + e^{-(-1)w^T x}} \\ &= \frac{e^{w^T x}}{1 + e^{w^T x}} + \frac{1}{1 + e^{w^T x}} \\ &= \frac{1 + e^{w^T x}}{1 + e^{w^T x}} = 1 \end{aligned}$$

(c) [3 points] Give one application that can be modeled as a structured prediction problem. Please give a concrete description, including what are the input variables, what are the output variables, and what are the modeling assumptions?

Example task : Parts of speech tagging
Example input variables: sentences consisting of words $\{x_1, x_2, \ldots, x_n\}$ (e.g., I play football).
Example output variables: part of speech tag sequence for each sentence $\{y_1, y_2, \ldots, y_n\}$ (e.g., PN VBZ N).
Example assumptions: Markov assumption. $P(y_i|y_{i-1}, y_{i-2}, \ldots x_i, x_{i-1}, \ldots x_1) = P(y_i|y_{i-1}, x_i), \forall i \geq 2$

Multiclass [13 points]

Given a set of training data $D = \{x_i, y_i\}_{i=1}^N$, where $x_i \in R^n, y_i \in \{1, 2, \dots K\}$, we introduce the following multi-class SVM formulation in the class.

$$\begin{aligned} \min_{w_1, w_2, \cdots w_K, \xi_i} \quad & \frac{1}{2} \sum_k w_k^T w_k + C \sum_i \xi_i \\ \text{s.t.} \quad & w_{y_i}^T x - w_k^T x \geq \Delta(y_i, k) - \xi_i, \forall i, k \end{aligned} \tag{2}$$

where

$$\Delta(y, y') = \begin{cases} 1 & y \neq y' \\ 0 & y = y' \end{cases}$$

(a) [8 points] Please explain the underlying meanings of the following terms:
Example: (x_i, y_i). Ans: this represents a data point. x_i is a feture vector representing input, y_i is the output label.

i. w_k: The weight vector associated with class k.

ii. $w_{y_i}^T x$: This term denotes how close an instance x to the classifier hyperplane of class k, representing the score of assigning class k to the instance x.

iii. ξ_i: ξ_i is a slack variable, which represents the penalty added to the loss term when the constraint in Eq. (2) associated with x_i is not satisfied.

iv. Meaning of the constraint: $w_{y_i}^T x - w_k^T x \geq 1 - \xi_i$: The constraint basically states the score assigned to the correct class should be higher than the score assigned to other classes by at least 1. If the constraint is not satisfied, a penalty ξ_i is added to the loss term.

Given a set of training data $D = \{x_i, y_i\}_{i=1}^N$, where $x_i \in R^n, y_i \in \{1, 2, \ldots K\}$, we introduce the following multi-class SVM formulation in the class.

$$\begin{aligned} \min_{w_1, w_2, \cdots w_K, \xi_i} \quad & \frac{1}{2} \sum_k w_k^T w_k + C \sum_i \xi_i \\ \text{s.t.} \quad & w_{y_i}^T x - w_k^T x \geq \Delta(y_i, k) - \xi_i, \forall i, k \end{aligned} \tag{2}$$

where

$$\Delta(y, y') = \begin{cases} 1 & y \neq y' \\ 0 & y = y' \end{cases}$$

(b) [4 points] Using Keslers construction, we can represent $w_y^T x$ as $w^T \phi(x, y)$. Show that how we define w and $\phi(x, y)$ such that $w_y^T x = w^T \phi(x, y)$. Then, rewrite Eq. (2) using the Keslers construction.

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_K \end{bmatrix} \tag{3}$$

$$\phi(x, y) = \begin{bmatrix} 0 \\ \vdots \\ x \\ \vdots \\ 0 \end{bmatrix} \text{at} \quad y^{th} \quad \text{position} \tag{4}$$

Equation (2) can be rewritten as:

$$\min_w \quad \frac{1}{2} w^T w + C \sum_i \xi_i \tag{5}$$

s.t

$$w^T \phi(x, y_i) - w^T \phi(x, k) \geq \Delta(y_i, k) - \xi_i \ \forall i, k$$

$$\Delta(y_i, k) = \begin{cases} 1 & \text{if } y_i \neq k, \\ 0 & \text{otherwise.} \end{cases}$$

(c) [3 points] Recall that in binary SVM:

$$\begin{aligned} \min_{w, \xi_i} \quad & \frac{1}{2} w^T w + C \sum_i \xi_i \\ \text{s.t.} \quad & y_i w^T x_i \geq 1 - \xi_i, \forall i, \\ & \xi_i \geq 0, \forall i. \end{aligned} \tag{6}$$

We have constraints $\xi_i \geq 0$. But, we don't explicitly list these constraints in Eq. (2). Show that $\xi_i \geq 0, \forall i$ are implicitly listed in the constraints of (2). Hint: consider the case when $k = y_i$.

When $y_i = k$, $w_{y_i}^T x = w_k^T x$, and $\delta(y_i, k) = 0$. Therefore, the constraint

$$w_{y_i}^T x - w_k^T x \geq \Delta(y_i, k) - \xi_i$$

implies $\xi_i \geq 0$.