University of Illinois Spring 2018

CS 446: Machine Learning
Homework 5

Due on Tuesday, February 20, 2018, 11:59 a.m. Central Time

1. **[6 points]** Multiclass Classification Basics

   (a) Which of the following is the most suitable application for multiclass classification? Which is the most suitable application for binary classification?

   i. Predicting tomorrow's stock price;
   ii. Recognizing flower species from photos;
   iii. Deciding credit card approval for a bank;
   iv. Assigning captions to pictures.

   Your answer: 2 is the most suitable application for multi-class classification and 3 is the most suitable application for binary classification.

   (b) Suppose in an n-dimensional Euclidean space where $n \geq 3$, we have n samples $x^{(i)} = e_i$ for $i = 1...n$ (which means $x^{(1)} = (1, 0, ..., 0)_n, x^{(2)} = (0, 1, ..., 0)_n, ..., x^{(n)} = (0, 0, ..., 1)_n$), with $x^{(i)}$ having class i. What are the numbers of binary SVM classifiers we need to train, to get 1-vs-all and 1-vs-1 multiclass classifiers?

   Your answer: We need $(n-1)$ binary SVM classifiers to get 1-vs-all classifiers and $\frac{n(n-1)}{2}$ to get 1-vs-1 multi-class classifiers.

   (c) Suppose we have trained a 1-vs-1 multiclass classifier from binary SVM classifiers on the samples of the previous question. What are the regions in the Euclidean space that will receive the same number of majority votes from more than one classes? You can ignore samples on the decision boundary of any binary SVM.

   Your answer: Given the data set in b, there is no region with ties. This is because all decision hyper planes pass through the origin. This results is no enclosed hyper-volume, if there is no hyper-volume we will not find a region with multiple majority votes.

2. **[8 points]** Multiclass SVM

Consider the objective function of multiclass SVM as

$$\min_{w,\xi^{(i)}\geq 0} \frac{C}{2}\|w\|^2 + \sum_{i=1}^{n} \xi^{(i)}$$

$$\text{s.t.} \quad w_{y^{(i)}}\phi(x^{(i)}) - w_{\hat{y}}\phi(x^{(i)}) \geq 1 - \xi^{(i)} \quad \forall i = 1...n, \hat{y} = 0...K-1, \hat{y} \neq y_i$$

Let $n = K = 3$, $d = 2$, $x^{(1)} = (0, -1)$, $x^{(2)} = (1, 0), x^{(3)} = (0, 1), y^{(1)} = 0, y^{(2)} = 1, y^{(3)} = 2$, and $\phi(x) = x$.

(a) Rewrite the objective function with w being a Kd-dimensional vector $(w_1, w_2, w_3, w_4, w_5, w_6)^\top$ and with the specific choices of x, y and ϕ.

Your answer:

$$\min_{w,\xi^{(i)}\geq 0} \frac{C}{2}(w_1^2 + w_2^2 + w_3^2 + w_4^2 + w_5^2 + w_6^2) + \sum_{i=1}^{n} \xi^{i}$$

$$\begin{bmatrix} 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}^T \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \end{bmatrix} \geq \mathbf{1} - \begin{bmatrix} \xi_1 \\ \xi_1 \\ \xi_2 \\ \xi_2 \\ \xi_3 \\ \xi_3 \end{bmatrix}$$

(b) Rewrite the objective function you get in (a) such that there are no slack variables $\xi^{(i)}$.

Your answer:

$$\xi_1 = \max\{0, 1 + w_2 - w_4, 1 + w_2 - w_6\}$$

$$\xi_2 = \max\{0, 1 + w_1 - w_3, 1 + w_5 - w_3\}$$

$$\xi_3 = \max\{0, 1 + w_2 - w_6, 1 + w_4 - w_6\}$$

$$\min_{w,\xi^{(i)}\geq 0} \frac{C}{2}(w_1^2 + w_2^2 + w_3^2 + w_4^2 + w_5^2 + w_6^2) + \max\{0, 1 + w_2 - w_4, 1 + w_2 - w_6\} + \max\{0, 1 + w_1 - w_3, 1 + w_5 - w_3\} + \max\{0, 1 + w_2 - w_6, 1 + w_4 - w_6\}$$

$$\begin{bmatrix} 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}^T \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \end{bmatrix} \geq \begin{bmatrix} 1 - \max\{0, 1 + w_2 - w_4, 1 + w_2 - w_6\} \\ 1 - \max\{0, 1 + w_2 - w_4, 1 + w_2 - w_6\} \\ 1 - \max\{0, 1 + w_1 - w_3, 1 + w_5 - w_3\} \\ 1 - \max\{0, 1 + w_1 - w_3, 1 + w_5 - w_3\} \\ 1 - \max\{0, 1 + w_2 - w_6, 1 + w_4 - w_6\} \\ 1 - \max\{0, 1 + w_2 - w_6, 1 + w_4 - w_6\} \end{bmatrix}$$

(c) Let $w_t = (1, 1, 1, 2, 1, -1)^\top$. Compute the derivative of the objective function you get in (b) w.r.t. w_2, at w_t, where w_2 is the weight of second dimension on Class 0 (in case you used non-conventional definition of w in (a)).

Your answer:

$$\frac{\partial}{\partial w_2}\Bigg(\frac{C}{2}(w_1^2+w_2^2+w_3^2+w_4^2+w_5^2+w_6^2)+\max\{0,1+w_2-w_4,1+w_2-w_6\}$$
$$+\max\{0,1+w_1-w_3,1+w_5-w_3\}$$
$$+\max\{0,1+w_2-w_6,1+w_4-w_6\}\Bigg)(w_t)$$
$$=Cw_2+1$$

$$\begin{bmatrix}0&0&-1&0&0&0\\-1&-1&0&0&-1&0\\0&0&1&1&0&0\\1&0&0&0&0&-1\\0&0&0&-1&0&0\\0&1&0&0&1&1\end{bmatrix}^T\begin{bmatrix}w_1\\w_2\\w_3\\w_4\\w_5\\w_6\end{bmatrix}\geq\begin{bmatrix}1-\max\{0,1+w_2-w_4,1+w_2-w_6\}\\1-\max\{0,1+w_2-w_4,1+w_2-w_6\}\\1-\max\{0,1+w_1-w_3,1+w_5-w_3\}\\1-\max\{0,1+w_1-w_3,1+w_5-w_3\}\\1-\max\{0,1+w_2-w_6,1+w_4-w_6\}\\1-\max\{0,1+w_2-w_6,1+w_4-w_6\}\end{bmatrix}$$

(d) Prove that

$$\max_{\hat{y}}\left(1+w_{\hat{y}}^\top\phi(x)\right)=\lim_{\epsilon\to 0}\epsilon\ln\sum_{\hat{y}}\exp\left(\frac{1+w_{\hat{y}}^\top\phi(x)}{\epsilon}\right).$$

Your answer:

$$\max_{\hat{y}}(1+w_{\hat{y}}^T\phi(x))=\lim_{\epsilon\to 0}\epsilon\ln\sum_{\hat{y}}\exp\left(\frac{(1+w_{\hat{y}}^T\phi(x))}{\epsilon}\right)\qquad \textit{let } \epsilon=\frac{1}{p}$$
$$=\lim_{p\to\infty}\frac{1}{p}\ln\sum_{\hat{y}}\exp(1+w_{\hat{y}}^T\phi(x))^p$$
$$=\lim_{p\to\infty}\ln\left(\left(\sum_{\hat{y}}e^{(1+w_{\hat{y}}^T\phi(x))^p}\right)^{\frac{1}{p}}\right)$$
since $\ln(x)$ *is continuous on the range it is defined*
$$=\ln\left(\lim_{p\to\infty}\left(\sum_{\hat{y}}e^{(1+w_{\hat{y}}^T\phi(x))^p}\right)^{\frac{1}{p}}\right)$$
using the definition of the infinity norm
$$=\ln\left(\max_{\hat{y}}\left(e^{(1+w_{\hat{y}}^T\phi(x))}\right)\right)$$
$$=\max_{\hat{y}}(1+w_{\hat{y}}^T\phi(x))$$