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.

   Solution:
   (2 points) ii. is the most suitable for multiclass classification; iii. is the most suitable 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?

   Solution:
   (2 points) n or $n-1$ for 1-vs-all (depends on understanding); $n(n-1)/2$ for 1-vs-1.

   (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.

   Solution:
   (2 points) For the binary SVM trained on two samples $x^{(i)}, x^{(j)}$, it will classify any point x with $x_i < x_j$ as clss i, and any point x with $x_i > x_j$ as class j, where x_i and x_j is the i-th and j-th dimension of x.

   Given any x. Let $x_i = \min_j x_j$. Then $x_i < x_j$ for any $j \neq i$, if i is not on the decision boundary of any binary SVM. Then class i will receive $n-1$ votes, while any class $j \neq i$ will receive at most $n-2$ votes.

   Therefore, no region will receive receive the same number of 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.

Solution: (2 points)

$$\min_{w_j, \xi^{(i)} \geq 0} \frac{C}{2} \sum_{j=1}^{Kd} w_j^2 + \sum_{i=1}^{n} \xi^{(i)}, \quad \text{s.t.}$$

$$-w_2 + w_4 \geq 1 - \xi^{(1)}, -w_2 + w_6 \geq 1 - \xi^{(1)}$$
$$w_3 - w_1 \geq 1 - \xi^{(2)}, w_3 - w_5 \geq 1 - \xi^{(2)}$$
$$w_6 - w_2 \geq 1 - \xi^{(3)}, w_6 - w_4 \geq 1 - \xi^{(3)}$$

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

Solution: (2 points)
$\min_{w_j} \frac{C}{2} \sum_{j=1}^{Kd} w_j^2 + \max\{1 + w_2 - w_4, 1 + w_2 - w_6\} + \max\{1 - w_3 + w_1, 1 - w_3 + w_5\} + \max\{1 - w_6 + w_2, 1 - w_6 + w_4\}$

(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)).

Solution: (2 points) At $w = w_t$, we have $\max\{1+w_2-w_4, 1+w_2-w_6\} = 1+w_2-w_6$, $\max\{1-w_6+w_2, 1-w_6+w_4\} = 1-w_6+w_4$, so the derivative w.r.t. w_2 is $Cw_2 + w_2 = C + 1$.

(d) Prove that

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

Solution: (2 points) On one hand,

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

On the other hand, let K be the number of classes, then

$$\epsilon \ln \sum_{\hat{y}} \exp\left(\frac{1 + w_{\hat{y}}^\top \phi(x)}{\epsilon}\right) \leq \epsilon \ln \sum_{\hat{y}} \exp\left(\frac{\max_{\hat{y}} \left(1 + w_{\hat{y}}^\top \phi(x)\right)}{\epsilon}\right)$$

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

Let $\epsilon \to 0$, we get the equality we want to prove.