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:
   For multiclass classification: ii. Labels: each of the different species that can be recognized.
   For binary classification: iii. Labels: approved or not-approved.

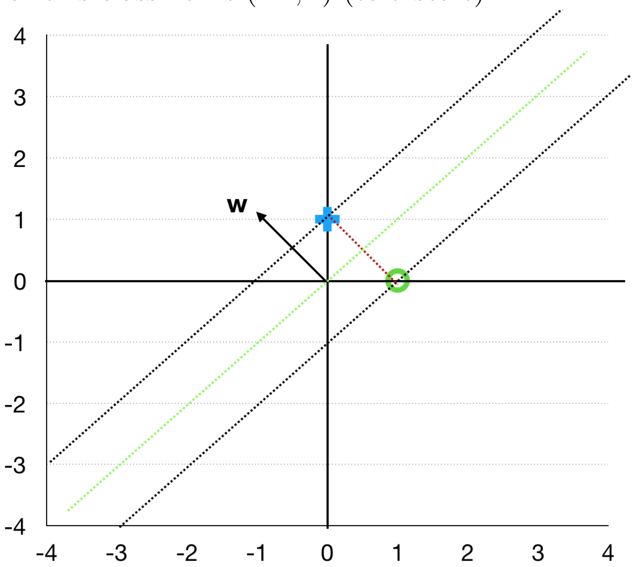
   (b) Suppose in an n-dimensional Euclidean space where $n \geq 3$, we have n samples $x^{(i)} = e_i$ for $i = 1 \ldots n$ (which means $x^{(1)} = (1, 0, \ldots, 0)_n, x^{(2)} = (0, 1, \ldots, 0)_n, \ldots, x^{(n)} = (0, 0, \ldots, 1)_n$), with $x^{(i)}$ having class i. What are the numbers of binary SVM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

   Your answer:
   For 1-vs-all: $n - 1$ classifiers.
   For 1-vs-1:

$$\frac{n\,(n-1)}{2}$$

   (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: Let $c_{i,j}$ be the classifier for classes i and j, $i, j = 0, 1, ..., K-1, i < j$. Note that we only have one training example for each class so every example will be in the margin boundary for every $c_{i,j}$. Below there is a plot of one of these classifiers ignoring all the dimensions that have zero for both points. The marked points are the support vectors and the green line is the decision boundary. It is easy to see that w for this classifier is $(-1, 1)$ (to a scale).



Now consider a test example $x = (x_1, x_2, ..., x_n)$; when tested against $c_{i,j}$ we only care about the components i, j of x. In particular, if $x_i > x_j$ then x will be classified as class i.

Since in particular case is symmetric, e.g. we can take any permutation of the classes and the problem won't change, without loss of generality lets assume that

$$x_1 \geq x_2 \geq ... \geq x_n$$

Note that we can actually assume strict inequality since we are ignoring points in decision boundaries. So,

$$x_1 > x_2 > ... > x_n$$

This means that $c_{1,2}, c_{1,3}, ..., c_{1,n}$ will all vote for class 1. So *at least* $n-1$ classifiers will vote for class 1. Now note that $c_{2,3}, c_{2,4}, ..., c_{2,n}$ will all vote for class 2 so there are at least $n-2$ votes for class 2. By doing this for all i in $c_{i,j}$ we can see that there are exactly $n-i$ classifiers that will vote for class i. Therefore, there are **no** regions in the Euclidean space that will receive the same number of majority votes for more than one class other than the decision boundaries.

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:

Let w_1, w_2 be the weights of the classifier for class 0, w_3, w_4 for class 1, and, w_5, w_6 for class 2.

Let's now rewrite the constraints using the given setup.

For $i = 1, \hat{y} = 1$:

$$\begin{aligned} w_{y^{(1)}}\phi\left(x^{(1)}\right) - w_{\hat{y}}\phi\left(x^{(1)}\right) &= (0w_1 + (-1)w_2) - (0w_3 + (-1)w_4) \\ &= -w_2 + w_4 \\ &\geq 1 - \xi^{(1)} \end{aligned}$$

Similarly, for

$$\begin{aligned} i = 1, \hat{y} = 2 &\rightarrow -w_2 + w_6 \geq 1 - \xi^{(1)} \\ i = 2, \hat{y} = 0 &\rightarrow w_3 - w_1 \geq 1 - \xi^{(2)} \\ i = 2, \hat{y} = 2 &\rightarrow w_3 - w_5 \geq 1 - \xi^{(2)} \\ i = 3, \hat{y} = 0 &\rightarrow w_6 - w_2 \geq 1 - \xi^{(3)} \\ i = 3, \hat{y} = 1 &\rightarrow w_6 - w_4 \geq 1 - \xi^{(3)} \end{aligned}$$

We know that $||w||^2 = w_1^2 + w_2^2 + w_3^2 + w_4^2 + w_5^2 + w_6^2$ so the objective function can be rewritten as

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

s.t. the six constraints above.

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

Your answer: By solving for $\xi^{(i)}$ on each of the constraints and taking the max for each of the ξ we have that the function can be written as

$$\begin{aligned}\min_{w_1,\ldots,w_6} \frac{C}{2}\left(w_1^2+w_2^2+w_3^2+w_4^2+w_5^2+w_6^2\right) \\ &+\max\left(1+w_2-w_4, 1+w_2-w_6\right) \\ &+\max\left(1+w_1-w_3, 1+w_5-w_3\right) \\ &+\max\left(1+w_2-w_6, 1+w_4-w_6\right)\end{aligned}$$

(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: Note that in 2(b) $\max\left(1+w_1-w_3, 1+w_5-w_3\right)$ does not depend on w_2 so its derivative will be zero.
Also note that the partial w.r.t w_2 of $\max\left(1+w_2-w_4, 1+w_2-w_6\right)$ will be 1.
We now need to plug w_t in the

$$\max\left(1+w_2-w_6, 1+w_4-w_6\right) = \max\left(1+1-(-1), 1+2-(-1)\right)$$

Since the second term is higher the partial w.r.t. w_2 of this term is zero. Finally we have that:

$$\frac{\partial F}{\partial w_2} = Cw_2 + 1 + 0 + 0 = C+1$$

(d) Prove that

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

Your answer: Lets start by remembering that the Chebyshev norm is given by

$$||\mathbf{z}||_\infty = \lim_{p\to\infty}\left(\sum_i z_i^p\right)^{\frac{1}{p}} = \max_i\left(z_i\right)$$

Since $f(z) = \exp(z)$ is a monotone function we have that

$$\max_i(z_i) = \ln\max_i\left(\exp\left(z_i\right)\right)$$

Combining these two results it follows that

$$\begin{aligned}\max_{\hat{y}}\left(1 + w_{\hat{y}}^\intercal\phi\left(x\right)\right) &= \ln\max_{\hat{y}}\left(\exp\left(1 + w_{\hat{y}}^\intercal\phi\left(x\right)\right)\right)\\ &= \ln\lim_{p\to\infty}\left(\sum_{\hat{y}}\exp\left(1 + w_{\hat{y}}^\intercal\phi\left(x\right)\right)^p\right)^{\frac{1}{p}}\\ &= \lim_{p\to\infty}\ln\left(\sum_{\hat{y}}\exp\left(1 + w_{\hat{y}}^\intercal\phi\left(x\right)\right)^p\right)^{\frac{1}{p}}\\ &= \lim_{\epsilon\to 0}\ln\left(\sum_{\hat{y}}\exp\left(1 + w_{\hat{y}}^\intercal\phi\left(x\right)\right)^{\frac{1}{\epsilon}}\right)^{\epsilon}\\ &= \lim_{\epsilon\to 0}\epsilon\ln\sum_{\hat{y}}\exp\left(1 + w_{\hat{y}}^\intercal\phi\left(x\right)\right)^{\frac{1}{\epsilon}}\\ &= \lim_{\epsilon\to 0}\epsilon\ln\sum_{\hat{y}}\exp\frac{1}{\epsilon}\left(1 + w_{\hat{y}}^\intercal\phi\left(x\right)\right)\\ &= \lim_{\epsilon\to 0}\epsilon\ln\sum_{\hat{y}}\exp\left(\frac{1 + w_{\hat{y}}^\intercal\phi\left(x\right)}{\epsilon}\right) \blacksquare\end{aligned}$$