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:
   ii. Recognizing flower species from phtos is suitable application for multiclass classification.
   iii. Deciding credit card approval for a bank is 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:
   1-vs-all: we need $n - 1$ binary classifiers.
   1-vs-1: we need $\frac{n(n-1)}{2}$ binary 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:
   There is no such region existing.
   First I have try this with Python as follows:

```
X = np.eye(100)
y = np.arange(100)
from sklearn.multiclass import OneVsOneClassifier
from sklearn.svm import LinearSVC
OneVsOneClassifier(LinearSVC(random_state=0)).fit(X, y).predict(X)
```

   The output of predict is from 0 to 99 without repeated class. This is a way to explicitly prove the statement.
   Furthermore, we can treat all the observations as an Identity Matrix in the previous question. So the 'similarity' of those observations is 0, in another way, they are independent of each other. So there will be no such region in N-dimensional space. Also, when we use 1-vs-1 to classify samples, it may not be transitive.

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:

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

Constraints are as follows:

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

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

Your answer:

$$\min_{w,\xi^{(i)}\geq 0} \frac{C}{2}(w_1, w_2, w_3, w_4, w_5, w_6) \cdot (w_1, w_2, w_3, w_4, w_5, w_6)^\intercal + \sum_{i\in(1,2,3)} \max_{\hat{y}}(1 - w_{y^{(i)}}\phi(x^{(i)}) + w_{\hat{y}}\phi(x^{(i)}))$$

(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:
What I get in part (b) is:

$$\min_{w,\xi^{(i)}\geq 0} \frac{C}{2}(w_1,w_2,w_3,w_4,w_5,w_6)\cdot(w_1,w_2,w_3,w_4,w_5,w_6)^\intercal$$

$$+\sum_{i\in(1,2,3)} \max_{\hat{y}}(1-w_{y^{(i)}}\phi(x^{(i)})+w_{\hat{y}}\phi(x^{(i)}))$$

So if we intend to take the derivative of w_2, we need unfold each part of the objective function and take the derivative of w_2. Furthermore, we need plug in the value of w_t with $w_t=(1,1,1,2,1,-1)^\intercal$

$$\frac{\partial \sum_{i\in(1,2,3)} \max_{\hat{y}}(1-w_{y^{(i)}}\phi(x^{(i)})+w_{\hat{y}}\phi(x^{(i)}))}{\partial w_2} =$$

$$\frac{\partial \max(1+w_2-w_4,1+w_2-w_6)}{\partial w_2}+$$

$$\frac{\partial \max(1-w_3+w_1,1-w_3+w_5)}{\partial w_2}+$$

$$\frac{\partial \max(1-w_6+w_2,1-w_6+w_4)}{\partial w_2}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_2} &= \frac{\partial \frac{C}{2}||\mathbf{w}||^2}{\partial w_2}+\frac{\partial \sum_{i\in(1,2,3)} \max_{\hat{y}}(1-w_{y^{(i)}}\phi(x^{(i)})+w_{\hat{y}}\phi(x^{(i)}))}{\partial w_2}\\ &= C+\frac{\partial \max(1+w_2-w_4,1+w_2-w_6)}{\partial w_2}\\ &= C+1\end{aligned}$$

(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:

Proof:

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

In conclusion,

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