CS 446 / ECE 449 Homework 5

Naman Shukla

TOTAL POINTS

14 / 14

QUESTION 1

Q1 6 pts

1.1 (a) 2 / 2

- ✓ **- 0 pts** Correct
- **- 1 pts** Partial Credit (1/2)
- **- 2 pts** Incorrect

1.2 (b) 2 / 2

- ✓ **- 0 pts** Correct
- **- 1 pts** Partial credit (1-vs-all correct)
- **- 1 pts** Partial credit(1-vs-1 correct)
- **- 1 pts** Parital credit (both parts)
- **- 2 pts** Incorrect
- **- 0.5 pts** Typo

1.3 (c) 2 / 2

- ✓ **- 0 pts** Correct
- **- 0.5 pts** Undefined notation
- **- 0.5 pts** Minor mistake
- **- 1 pts** Partial credit
- **- 2 pts** Incorrect / unclear

Correct, but for future reference, not that the null set is denoted \emptyset, not \Phi.

QUESTION 2

Q2 8 pts

2.1 (a) 2 / 2

- ✓ **- 0 pts** Correct
- **- 0.5 pts** Minor mistake
- **- 0.5 pts** Missing 1 condition
- **- 1 pts** Missing 2 or more condtions
- **- 1 pts** Partial credit
- **- 2 pts** Incorrect / unclear

2.2 (b) 2 / 2

- ✓ **- 0 pts** Correct
- **- 0.5 pts** The objective function is incorrect
- **- 0.5 pts** Did not plug in the given values
- **- 0.5 pts** Did not simplify
- **- 2 pts** Incorrect
- **- 0 pts** please select pages for answer

2.3 (c) 2 / 2

- ✓ **- 0 pts** Correct
- **- 1 pts** The result is not correct
- **- 1 pts** Did not use the given values
- **- 2 pts** Incorrect

2.4 (d) 2 / 2

- ✓ **- 0 pts** Correct
- **- 0.5 pts** Incorrect, steps missing
- **- 0.5 pts** Incorrect
- **- 1 pts** Incorrect, steps missing
- **- 2 pts** Incorrect

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:

   i. Recognizing flower species from photos $\rightarrow$ Multiclass Classification.

   ii. Deciding credit card approval for a bank $\rightarrow$ 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:

   i. 1-vs-all multiclass classifiers $\rightarrow$ N classifiers.

   ii. 1-vs-1 multiclass classifiers $\rightarrow$ $\frac{N(N-1)}{2}$ 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.

1.1 (a) **2 / 2**

✓ **- 0 pts** Correct

- 1 pts Partial Credit (1/2)

- 2 pts Incorrect

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:

i. Recognizing flower species from photos $\rightarrow$ Multiclass Classification.

ii. Deciding credit card approval for a bank $\rightarrow$ 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:

i. 1-vs-all multiclass classifiers $\rightarrow$ N classifiers.

ii. 1-vs-1 multiclass classifiers $\rightarrow$ $\frac{N(N-1)}{2}$ 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.

1.2 (b) 2 / 2

✓ **- 0 pts** Correct

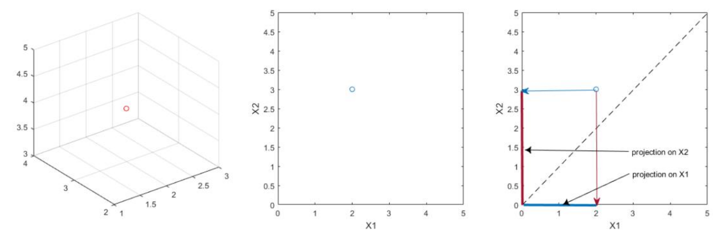
- 1 pts Partial credit (1-vs-all correct)

- 1 pts Partial credit(1-vs-1 correct)

- 1 pts Parital credit (both parts)

- 2 pts Incorrect

- 0.5 pts Typo



Your answer:
Claim : The euclidean space is empty (Φ) for the same number of votes.
Proof :
For 1-vs-1 multiclass classification, let $Classify(\mathbf{P}^{x_1,x_2}_{(x_t)})$ denotes the classification of test data x_t on the projected space (x_1, x_2). Then the classifier would give the result based on the closest distance. In the figure, the dotted line is the decision boundary which separates two decision region. Therefore, $\mathbf{P}^{x_1}_{(x_t)}$ is the projection (score) on x_1 and $\mathbf{P}^{x_2}_{(x_t)}$ is the projection (score) on x_2. Therefore, the better score among the two will decide the classifiers result.
Lets assume there is a clash in majority. So,

$$Classify(\mathbf{P}^{x_1,x_2}_{(x_t)}) = x_2 \text{ as the score } x_2 > x_1$$

$$Classify(\mathbf{P}^{x_1,x_3}_{(x_t)}) = x_1 \text{ as the score } x_1 > x_3$$

Now in order for clash to happen, $Classify(\mathbf{P}^{x_2,x_3}_{(x_t)})$ must classify as x_3. But,

$$x_2 > x_1 \text{ and } x_1 > x_3 \;\; \rightarrow \;\; x_2 > x_3$$

So the classifier will classify the test data as x_2 (due to better score).
Therefore, we contradicted our clash and hence this situation will never occur. The same argument is valid in n^{th} dimension as well because of the linear independency of the classes in the Euclid's Space. Hence, The space will always be empty (Φ).

1.3 (c) 2 / 2

✓ - 0 pts Correct

- 0.5 pts Undefined notation

- 0.5 pts Minor mistake

- 1 pts Partial credit

- 2 pts Incorrect / unclear

Correct, but for future reference, not that the null set is denoted \emptyset, not \Phi.

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:
For given

$$\phi(x) = x$$

we have,

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \end{bmatrix} \quad \text{and} \quad \psi(x^{(i)}, y^{(i)}) = \begin{bmatrix} \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \end{bmatrix} \delta(y^{(i)} = 0) \\ \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \end{bmatrix} \delta(y^{(i)} = 1) \\ \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \end{bmatrix} \delta(y^{(i)} = 2) \end{bmatrix}$$

For $(i = 1)$, first two sets of the constraint equations are :

$$\begin{bmatrix} w_1, w_2, w_3, w_4, w_5, w_6 \end{bmatrix} \begin{bmatrix} \left(\begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \end{bmatrix} \right) \left(\begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{bmatrix} \right) \end{bmatrix} \geq 1 - \xi^1$$

This gives us following two equations,

$$-w_2 + w_4 \geq 1 - \xi^1 \text{ and } - w_2 + w_6 \geq 1 - \xi^1$$

Similarly for $(i = 2)$ and $(i = 3)$ we have following equations,

$$w_3 - w_1 \geq 1 - \xi^2 \text{ and } w_3 - w_5 \geq 1 - \xi^2$$

$$w_6 - w_2 \geq 1 - \xi^3 \text{ and } w_6 - w_4 \geq 1 - \xi^3$$

Finally we have the objective function as :

$$\min_{w, \xi^{(i)} \geq 0} \frac{C}{2} \|w\|^2 + \sum_{i=1}^{3} \xi^{(i)}$$

Such that,

$$-w_2 + w_4 \geq 1 - \xi^1 \text{ and } - w_2 + w_6 \geq 1 - \xi^1$$
$$w_3 - w_1 \geq 1 - \xi^2 \text{ and } w_3 - w_5 \geq 1 - \xi^2$$
$$w_6 - w_2 \geq 1 - \xi^3 \text{ and } w_6 - w_4 \geq 1 - \xi^3$$

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

2.1 (a) **2 / 2**

- ✓ **- 0 pts** Correct
- **- 0.5 pts** Minor mistake
- **- 0.5 pts** Missing 1 condition
- **- 1 pts** Missing 2 or more condtions
- **- 1 pts** Partial credit
- **- 2 pts** Incorrect / unclear

Your answer:

$$\min_{\mathbf{w}} \frac{C}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{3} max_{\hat{y}}(1 + \mathbf{w}^\intercal \psi(x^{(i)}, \hat{y})) - \mathbf{w}^\intercal \psi(x^{(i)}, y^{(i)})$$

$$\min_{\mathbf{w}} \frac{C}{2}\|\mathbf{w}\|_2^2 + max(1-w_4, 1-w_6) + w_2 + max(1+w_1, 1+w_5) - w_3 + max(1+w_2, 1+w_4) - w_6$$

(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:

$$\frac{\partial F}{\partial w_2} = \frac{\partial(\frac{C}{2}\|\mathbf{w}\|_2^2)}{\partial w_2} + \frac{\partial(max(1-w_4, 1-w_6) + w_2)}{\partial w_2} + \frac{\partial(max(1+w_1, 1+w_5) - w_3)}{\partial w_2} + \frac{\partial(max(1+w_2, 1+w_4) - w_6)}{\partial w_2}$$

$$\frac{\partial F}{\partial w_2} = C \cdot w_2 + 1 + \lambda$$

Where λ is given by $\lambda = 0$ if $w_2 > w_4$ and $= 1$ if $w_2 \leq w_4$

For $w_t = (1, 1, 1, 2, 1, -1)^\intercal$ we have $\lambda = 0$

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

(d) Prove that

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

2.2 (b) **2 / 2**

- ✓ **- 0 pts** Correct
- **- 0.5 pts** The objective function is incorrect
- **- 0.5 pts** Did not plug in the given values
- **- 0.5 pts** Did not simplify
- **- 2 pts** Incorrect
- **- 0 pts** please select pages for answer

Your answer:

$$\min_{\mathbf{w}} \frac{C}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{3} max_{\hat{y}}(1 + \mathbf{w}^\intercal \psi(x^{(i)}, \hat{y})) - \mathbf{w}^\intercal \psi(x^{(i)}, y^{(i)})$$

$$\min_{\mathbf{w}} \frac{C}{2}\|\mathbf{w}\|_2^2 + max(1-w_4, 1-w_6) + w_2 + max(1+w_1, 1+w_5) - w_3 + max(1+w_2, 1+w_4) - w_6$$

(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:

$$\frac{\partial F}{\partial w_2} = \frac{\partial(\frac{C}{2}\|\mathbf{w}\|_2^2)}{\partial w_2} + \frac{\partial(max(1-w_4,1-w_6)+w_2)}{\partial w_2} + \frac{\partial(max(1+w_1,1+w_5)-w_3)}{\partial w_2} + \frac{\partial(max(1+w_2,1+w_4)-w_6)}{\partial w_2}$$

$$\frac{\partial F}{\partial w_2} = C \cdot w_2 + 1 + \lambda$$

Where λ is given by $\lambda = 0$ if $w_2 > w_4$ and $= 1$ if $w_2 \leq w_4$

For $w_t = (1, 1, 1, 2, 1, -1)^\intercal$ we have $\lambda = 0$

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

(d) Prove that

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

2.3 (c) 2 / 2

✓ - 0 pts Correct

- 1 pts The result is not correct

- 1 pts Did not use the given values

- 2 pts Incorrect

Your answer: For any vector v:

$$0 \leq \frac{1}{r} \ln \sum_{i=1}^{n} \exp(r\mathbf{v}_i) - \|v\|_\infty \leq \frac{\ln(n)}{r}$$

Replace variable $\frac{1}{r} \rightarrow \epsilon$ and take limit,

$$\lim_{\epsilon \to 0} 0 \leq \lim_{\epsilon \to 0} (\epsilon \ln \sum_{i=1}^{n} \exp(\frac{\mathbf{v}_i}{\epsilon}) - \|v\|_\infty) \leq \lim_{\epsilon \to 0} (\epsilon \ln(n))$$

$$\lim_{\epsilon \to 0} (\epsilon \ln \sum_{i=1}^{n} \exp(\frac{\mathbf{v}_i}{\epsilon}) - \|v\|_\infty) = 0$$

From the definition of infinity norm,

$$\lim_{\epsilon \to 0} (\epsilon \ln \sum_{i=1}^{n} \exp(\frac{\mathbf{v}_i}{\epsilon}) = \max_i (\mathbf{v})$$

we have $\mathbf{v} = \left(1 + w_{\hat{y}}^\top \phi(x)\right)$,

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

alternate:

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

$$= \lim_{p \to \infty} \ln \max \exp\left(1 + w_{\hat{y}}^\top \phi(x)\right)$$

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

$$= \max_{\hat{y}} \left(1 + w_{\hat{y}}^\top \phi(x)\right)$$

2.4 (d) 2 / 2

- ✓ - 0 pts Correct
- - 0.5 pts Incorrect, steps missing
- - 0.5 pts Incorrect
- - 1 pts Incorrect, steps missing
- - 2 pts Incorrect