



THE CHINESE UNIVERSITY OF HONG KONG, SHENZHEN

DDA 2020

Assignment2 Report

Author: Qiyu Zhang

Date Last Edited: April 3, 2022

Contents

1	Written Questions	2
2	Programming Question	5
2.1	Feature Engineering	5
2.2	Model algorithm	5
2.3	Linear SVM	5
2.4	Linear SVM with slack variable	6
2.5	Linear SVM with slack variables and kernel	7

1 Written Questions

Exercise 8.7 Regularizing separate terms in 2d logistic regression

(Source: Jaaakkola.)

- a. Consider the data in Figure 8.13, where we fit the model $p(y = 1 | \mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$. Suppose we fit the model by maximum likelihood, i.e., we minimize

$$J(\mathbf{w}) = -\ell(\mathbf{w}, D_{\text{train}}) \quad (8.133)$$

where $\ell(\mathbf{w}, D_{\text{train}})$ is the log likelihood on the training set. Sketch a possible decision boundary corresponding to \mathbf{w} . (Copy the figure first (a rough sketch is enough), and then superimpose your answer on your copy, since you will need multiple versions of this figure). Is your answer (decision boundary) unique? How many classification errors does your method make on the training set?

- b. Now suppose we regularize only the w_0 parameter, i.e., we minimize

$$J_0(\mathbf{w}) = -\ell(\mathbf{w}, D_{\text{train}}) + \lambda w_0^2 \quad (8.134)$$

Suppose λ is a very large number, so we regularize w_0 all the way to 0, but all other parameters are unregularized. Sketch a possible decision boundary. How many classification errors does your method make on the training set? Hint: consider the behavior of simple linear regression, $w_0 + w_1 x_1 + w_2 x_2$ when $x_1 = x_2 = 0$.

- c. Now suppose we heavily regularize only the w_1 parameter, i.e., we minimize

$$J_1(\mathbf{w}) = -\ell(\mathbf{w}, D_{\text{train}}) + \lambda w_1^2 \quad (8.135)$$

Sketch a possible decision boundary. How many classification errors does your method make on the training set?

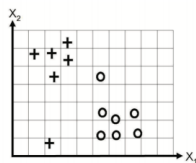
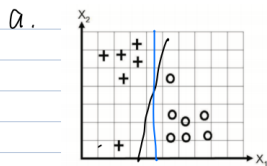


Figure 8.13 Data for logistic regression question.

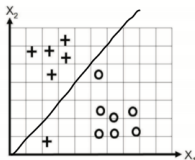
- d. Now suppose we heavily regularize only the w_2 parameter. Sketch a possible decision boundary. How many classification errors does your method make on the training set?



No, it's not unique.
Zero error that my methods has made
boundary $w_0 + w_1 x_1 + w_2 x_2 = 0$

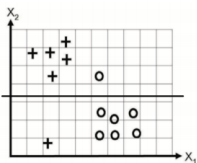
- b. If λ is very large, then the minimization force $w_0 = 0$

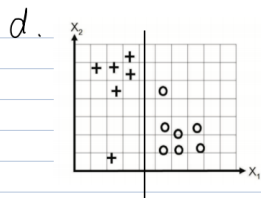
So if $x_1 = x_2 = 0$, we get $w_0 + w_1 x_1 + w_2 x_2 = 0$ origin on the boundary.



Then the possible decision boundary is in the picture.
One error happens in any cases. 1 error

- c. Then we force $w_1 = 0$, the boundary $w_0 + w_2 x_2 = 0 \Rightarrow x_2 = -\frac{w_0}{w_2}$.
So it's a horizontal line on the graph. 2 classification error still.





Then we force $w_2 = 0, x_1 = -\frac{w_0}{w_1}$
So the graph is a vertical line on the graph

No classification error at this time

Exercise 14.1 Fitting an SVM classifier by hand

(Source: Jaakkola.) Consider a dataset with 2 points in 1d: $(x_1 = 0, y_1 = -1)$ and $(x_2 = \sqrt{2}, y_2 = 1)$. Consider mapping each point to 3d using the feature vector $\phi(x) = [1, \sqrt{2}x, x^2]^T$. (This is equivalent to using a second order polynomial kernel.) The max margin classifier has the form

$$\min ||w||^2 \quad \text{s.t.} \quad (14.97)$$

$$y_1(w^T \phi(x_1) + w_0) \geq 1 \quad (14.98)$$

$$y_2(w^T \phi(x_2) + w_0) \geq 1 \quad (14.99)$$

- Write down a vector that is parallel to the optimal vector w . Hint: recall from Figure 7.8 (12Apr10 version) that w is perpendicular to the decision boundary between the two points in the 3d feature space.
- What is the value of the margin that is achieved by this w ? Hint: recall that the margin is the distance from each support vector to the decision boundary. Hint 2: think about the geometry of 2 points in space, with a line separating one from the other.
- Solve for w , using the fact the margin is equal to $1/||w||$.
- Solve for w_0 using your value for w and Equations 14.97 to 14.99. Hint: the points will be on the decision boundary, so the inequalities will be tight.
- Write down the form of the discriminant function $f(x) = w_0 + w^T \phi(x)$ as an explicit function of x .

$$a. \phi(x_1) = [1, 0, 0]^T, \phi(x_2) = [1, 2, 2]^T$$

$$\phi(x_2) - \phi(x_1) = [0, 2, 2]^T$$

Notice that the plain $w^T \phi(x) + w_0 = 0$

is vertical with $\phi(x_2) - \phi(x_1)$ to maximize the margin. w is also perpendicular with the boundary. So the vector is $[0, 2, 2]^T$

$$b. \text{ The value is } \frac{1}{2} ||\phi(x_2) - \phi(x_1)|| = \sqrt{2}$$

$$c. \frac{1}{\sqrt{2}} = \frac{1}{||w||}, ||w|| = \frac{1}{\sqrt{2}}, w = (0, k, k)^T / [0, 2, 2]^T$$

$$\text{So } 2k^2 = \frac{1}{2}, k = \frac{1}{2}, w = (0, \frac{1}{2}, \frac{1}{2})^T$$

$$d. (0, \frac{1}{2}, \frac{1}{2}) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + w_0 = -1$$

$$(0, \frac{1}{2}, \frac{1}{2}) \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} + w_0 = 1, w_0 = -1$$

$$e. f(x) = \frac{1}{2}x + \frac{1}{2}x^2 - 1$$

Exercise 14.2 Linear separability

(Source: Koller.) Consider fitting an SVM with $C > 0$ to a dataset that is linearly separable. Is the resulting decision boundary guaranteed to separate the classes?

No.

$$\min \frac{1}{2} ||w||^2 + C \sum \xi_i \quad \text{s.t. } 1 - \xi_i - y_i(w^T x_i + b) \leq 0, \xi_i \geq 0, \forall i$$

$$\text{in the Dual problem, it's } \max \sum \alpha_i \quad \text{s.t. } \sum \alpha_i y_i x_i^T x_j \leq C \quad \text{s.t. } \sum \alpha_i y_i \leq 0, 0 \leq \alpha_i \leq C \quad \forall i$$

If we set C to be very small, very close to zero, then there will be no penalty for ξ_i and the constraints will be meaningless

So we just $\min \frac{1}{2} ||w||^2 \Rightarrow w$ is close to 0, which obviously can't guarantee to separate data.

- Given a binary data set: (1 point)

$$\text{Class -1: } \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{Class +1: } \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

- Find the svm classifier of this given dataset;
- Specify which of given data points are supporting vectors;
- Predict the label of $[1; 2]$.

$$b) \max \sum_{i=1}^4 \alpha_i - \sum_{i,j=1}^4 \alpha_i \alpha_j x_i^T x_j y_i y_j$$

$$\text{s.t. } \sum_{i=1}^4 \alpha_i y_i = 0, \alpha_i \geq 0$$

$$\Rightarrow \max \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \alpha_1^2 - \alpha_2^2 - \alpha_3^2 - \alpha_4^2$$

$$\text{s.t. } -y_1 - y_2 + y_3 + y_4 = 0$$

$$\Rightarrow \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{1}{2}$$

$$\text{So } w = \sum_{i=1}^4 \alpha_i x_i y_i = \frac{1}{2} ((-1, 0) + (0, -1) + (-1, 0) + (0, -1)) = (-1, -1)$$

$$b = \frac{1}{4} \sum_{i=1}^4 (y_i - \sum_{j=1}^4 \alpha_j y_j x_i^T x_j) = 0$$

$$\text{So we get } f(x) = w^T x + b = [-1, -1] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -x_1 - x_2. f < 0 \Rightarrow \text{Class -1}, f > 0 \Rightarrow \text{Class +1}$$

2. Since all the $\alpha_i > 0$, so all points are supporting vectors. $[0, 1]^T, [0, 1]^T, [-1, 0]^T, [0, -1]^T$

3. $-1 - 2 = -3 < 0$, so it's in Class -1

5. Given a binary data set: (2 points)

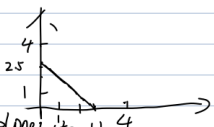
$$\text{Class -1: } \begin{bmatrix} (1 & 0) \\ (0 & 1) \\ (-1 & 0) \\ (0 & -1) \end{bmatrix} \quad \text{Class +1: } \begin{bmatrix} (2 & 0) \\ (0 & 2) \\ (-2 & 0) \\ (0 & -2) \end{bmatrix}$$

(1) Can you find a svm classifier (without slack variable) for this data set? explain why; (1 point)

(2) Use SVM by expanding the original feature vector $\mathbf{x} = [x_1; x_2]$ to $\phi(\mathbf{x}) = [x_1^2; x_2^2]$, find the svm of this given data set and predict the label of $[1; 2]$. (1 point)

(1) No. Assume $\exists w^T x + b$ s.t. $w^T x + b > 0$, then it's in class 1.
Then Notice that $w^T \begin{pmatrix} 1 \\ 0 \end{pmatrix} + b > 0$, $w^T \begin{pmatrix} -1 \\ 0 \end{pmatrix} + b > 0 \Rightarrow b > 0$.
 $w^T \begin{pmatrix} 0 \\ 1 \end{pmatrix} + b < 0$, $w^T \begin{pmatrix} 0 \\ -1 \end{pmatrix} + b < 0 \Rightarrow b < 0$, Contradiction. So we can't find it.

(2)
Class -1: $\begin{pmatrix} (1, 0) \\ (0, 1) \end{pmatrix}$ Class +1: $\begin{pmatrix} (4, 0) \\ (0, 4) \end{pmatrix}$
Obviously the SVM is $x_1 + x_2 - 2.5 = 0$,
So, it's $x_1, x_2^2 - \frac{5}{2} = 0$. $1^2 + 2^2 - \frac{5}{2} = \frac{5}{2}$, so it belongs to +1.



6. Show that the value γ of the margin for the maximum-margin hyperplane is given by

$$\frac{1}{\gamma^2} = \sum_{n=1}^N a_n$$

where $\{a_n\}$ are given by the following optimization problem

$$\begin{aligned} \max \quad & \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m) \\ \text{s.t.} \quad & \sum_{n=1}^N a_n t_n = 0 \\ & a_n \geq 0 \quad \forall n = 1, 2, \dots, N \end{aligned}$$

Note: You can treat the kernel function as $k(\mathbf{x}_n, \mathbf{x}_m) = \mathbf{x}_n^T \mathbf{x}_m$ if you are unfamiliar with kernel SVM (2 points)

$$W = \sum_{i=1}^M a_i y_i x_i \quad \gamma = W^T W = \sum_{i=1}^M \sum_{j=1}^M a_i a_j y_i y_j \ker(x_i, x_j)$$

Notice that $\sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$
of $\min \frac{1}{2} \|W\|^2$
s.t. $\sum_{n=1}^N a_n t_n = 0, a_n \geq 0$ is the dual problem

which shown in the below report
They share the same optimal value and

$$\text{So } \sum_{n=1}^N a_n - \frac{1}{2} \|W\|^2 = \frac{1}{2} \|W\|^2$$

$$\text{So } \sum_{n=1}^N a_n = \|W\|^2 = \frac{1}{\gamma^2}$$

2 Programming Question

2.1 Feature Engineering

The training and testing dataset, including 120 training data and 30 testing data, respectively, have been given. It covers 3 classes, corresponding to setosa, versicolor, virginica, denoted as Class 1, 2, 3 respectively. They are derived from the Iris dataset (<https://archive.ics.uci.edu/ml/datasets/iris>), and contains 3 classes of 50 instances each, where each class refers to a type of iris plant.

The error is defined as:

$$Error = \frac{True_Prediction}{n_Sample}$$

2.2 Model algorithm

we will use one-vs-rest strategy with svc, which can be implemented by manually aggregating two classes into one and run svc function. For example, suppose there are 3 classes and we want to calculate the class 0 vs rest, then we aggregate the data from class 1 and 2 into a single class, then use SVC to solve this problem. So we will get 3 errors for both the training set and the test set. They will be shown in the form of tables. The final model is defined as Figure 1:

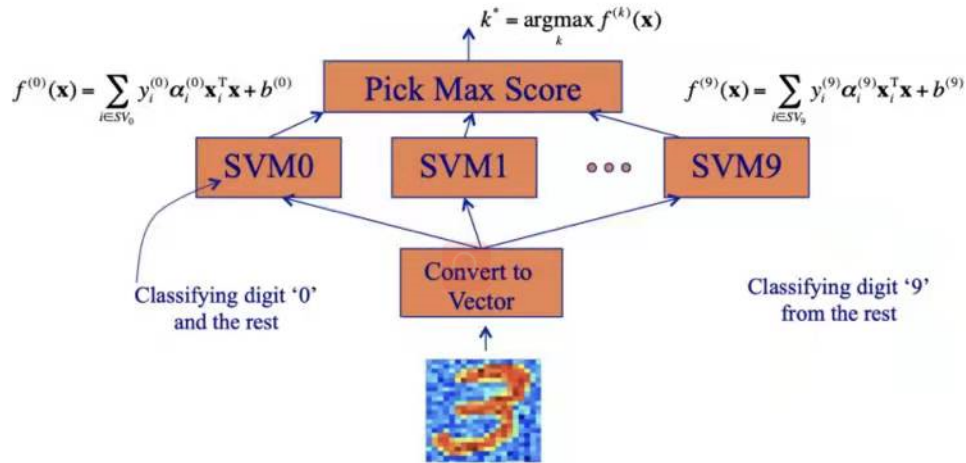


Figure 1: Model Algorithm

Here we build 3 SVM and compare their outputs. We choose the index corresponding to the largest outputs as the predicted class. Also the total error is given, according to the final prediction.

2.3 Linear SVM

The basic formula of the linear SVM is:

$$\begin{aligned} \min_{w,b} & \frac{1}{2} \|w\|^2 \\ \text{s.t.} & 1 - y_i(w^T x + b) \leq 0, \forall i \end{aligned}$$

For the convenience of the computer programming and to escape error in code running, we change the formula into: ($c = 10^5$)

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_i^m \xi_i$$

$$s.t. \quad 1 - y_i(w^T x + b) \leq \xi_i, \quad \forall i$$

Notice that in Q3, we have proven that when C is infinity, then the problem is equal to the standard SVM. Here $c = 10^5$ is applied so that the difference is tiny and omitted. Actually, the package of SVC from sklearn can help users solve the optimization problem. The errors are:

Table 1: Error of SVM

Error type	Class1 vs rest	Class2 vs rest	Class3 vs rest	Final model
Train error	0	0.217	0.017	0.0417
Test error	0	0.367	0	0

The class1 vs rest is linearly separable, since the zero error in the training set means there is no classification error. The other two are not separable since there is a classification error. **So only the setosa is separable.**

2.4 Linear SVM with slack variable

The basic formula of the linear SVM with slack variable is:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_i^m \xi_i$$

$$s.t. \quad 1 - y_i(w^T x + b) \leq \xi_i, \quad \forall i$$

The meaning of the formula is that, if the dataset is not separable, then we can have those $1 - y_i(w^T x + b) > 0$. The ξ_i is set for a tolerance that how positive $1 - y_i(w^T x + b)$ can be, while the large ξ_i will make the objective function large, as a penalty. To solve the problem, we retrieve the KKV conditions:

$$L = \frac{1}{2} \|w\|^2 + \left(\sum_{i=1}^m \xi_i + \sum_{i=1}^m [\alpha_i (1 - \xi_i - y_i(w^T x_i + b) - u_i \xi_i)] \right)$$

$$\alpha_i, u_i \geq 0$$

To get the minimum of L , we need: $\frac{\partial L}{\partial w} = 0$, $\frac{\partial L}{\partial b} = 0$, $\frac{\partial L}{\partial \xi_i} = 0$, The result shows that:

$$w - \sum_{i=0}^m \alpha_i y_i x_i = 0$$

$$\sum_{i=0}^m \alpha_i y_i = 0$$

$$\alpha_i = c - u_i, \quad \forall i$$

$$\xi_i = 0, \quad \forall i$$

Then we substitute w in the formula and $\xi_i = 0$, $\forall i$:

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=0}^m \alpha_i y_i b$$

By use $\sum_{i=0}^m \alpha_i y_i = 0$ and $\alpha_i = C - u_i$, $u_i \geq 0$:

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\sum_{i=1}^m \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, \forall i$$

The above problem is the dual problem of the standard problem of SVM with slack variables. After we solve the dual and get the optimal solution, we can find w in the formula. And $S: \{i | \alpha_i > 0\}$. For those data points with the index in S , we call them support vectors. For $\forall j \in S$, we have

$$y_j(w^T x_j + b) = 1, \quad y_i\left(\sum_{i=1}^m \alpha_i y_i x_i^T x_j + b\right) = 1$$

By multiply y_i in both sides, we can get $b = \frac{1}{|S|} \sum_{j \in S} (y_i - \sum_{i=1}^m \alpha_i y_i x_i^T x_j)$. After that, we can get $f = w^T x + b$, which is the decision function of SVM with input x .

Actually, the package of SVC from sklearn can help use solve the optimization problem. The example of $C=1$ is used and the errors are:

Table 2: Error of SVM with slack variables($C=1$)

Error type	Class1 vs rest	Class2 vs rest	Class3 vs rest	Final model
Train error	0	0.25	0.025	0.05
Test error	0	0.367	0.033	0.067

The following train error and test error with $C=0.1-0.9$ will be given:

Table 3: Error of SVM with slack variables

C	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Train error	0.125	0.058	0.050	0.050	0.050	0.050	0.050	0.050	0.050
Test error	0.233	0.167	0.133	0.100	0.100	0.100	0.100	0.100	0.067

2.5 Linear SVM with slack variables and kernel

The kernel function $k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, means we substitute x_i with $\varphi(x_i)$, to make the nonlinear boundary. After the substitution, the rest part would be the same. The dual would become:

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j \varphi(x_i)^T \varphi(x_j)$$

$$\sum_{i=1}^m \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, \forall i$$

In this way, we have the below result similarly:

$$b = \frac{1}{|S|} \sum_{j \in S} [y_i - \sum_{i=1}^m \alpha_i y_i \varphi(x_i)^T \varphi(x_j)]$$

$$w = \sum_{i=0}^m \alpha_i y_i \varphi(x_i)$$

And the following boundary would be $w^T \varphi(x) + b = 0$, and the decision function $f = \sum_{i=0}^m \alpha_i y_i \varphi(x_i)^T \varphi(x) + b = \sum_{i=0}^m \alpha_i y_i K(x_i, x) + b$, as the output of the SVM. Here we apply three kinds of kernel:

$$K_1(x_i, x_j) = (x_i^T x_j)^2, \quad K_2(x_i, x_j) = (x_i^T x_j)^3, \quad K_3(x_i, x_j) = \tanh\left(\frac{1}{N} x_i^T x_j\right), \quad K_4(x_i, x_j) = \exp(-2\|x_i - x_j\|^2)$$

K_1, K_2 : polynomial with degree 2, 3; K_3 : sigmoid; K_4 : RBF The chosen C is 1, and the errors tables are below:

Table 4: Error of SVM with kernel K_1

Error type	Class1 vs rest	Class2 vs rest	Class3 vs rest	Final model
Train error	0	0.033	0.025	0.025
Test error	0	0	0	0

Table 5: Error of SVM with kernel K_2

Error type	Class1 vs rest	Class2 vs rest	Class3 vs rest	Final model
Train error	0	0.008	0.008	0.008
Test error	0	0	0.033	0

Table 6: Error of SVM with kernel K_3

Error type	Class1 vs rest	Class2 vs rest	Class3 vs rest	Final model
Train error	0.333	0.333	0.333	0.825
Test error	0.333	0.333	0.333	0.767

Table 7: Error of SVM with kernel K_4

Error type	Class1 vs rest	Class2 vs rest	Class3 vs rest	Final model
Train error	0	0.033	0.025	0.033
Test error	0	0.033	0.033	0.033