1.

Question 1

Suppose you have trained an SVM classifier with a Gaussian kernel. You suspect the SVM is overfitting your dataset. Then which **two** of the following would be the reasonable approaches to take?

* C is the cost parameter, e.g., it's the same "C" in eq (4) on p3 of "SVMinR\_JSS2006.pdf" (in the Code folder at the Resources page on Piazza).
* sigma is the parameter in the Gaussian Radial Basis Function (RBF) kernel; see eq (16) on p6 of "SVMinR\_JSS2006.pdf"

1 point



Decrease C.



Decrease sigma.



Increase C.



Increase sigma.

2.

Question 2

Suppose you have a training data of *p*=10 features and *n*=5000 samples for a binary classification task. You've tried logistic regression but find it underfits your data. Then which **two** of the following might be promising steps to take? Circle all that apply.

1 point



Use an SVM with a Gaussian kernel.



Run logistic regression with AIC to select a subset of the 10 features.



Use an SVM with a linear kernel.



Drop half of the training samples.



Use a logistic regression model including additional quadratic features.

3.

Question 3

We are given n = 7 observations in p = 2 dimensions. For each observation, there is an associated class label.

* *X*=(3,4),*Y*=1;
* *X*=(2,2),*Y*=1;
* *X*=(4,4),*Y*=1;
* *X*=(1,4),*Y*=1;
* *X*=(3,1),*Y*=−1;
* *X*=(4,3),*Y*=−1;
* *X*=(4,1),*Y*=−1;

We fit a linear SVM on this data set and have found that the 2nd, 3rd, and 6th obs are the support vectors. The corresponding lambda values for the the 2nd and 3rd observations are

* *λ*2​=1,*λ*3​=3.

Write the corresponding decision rule as: if *b*1​*X*1​+*b*2​*X*2​+1>0, then predict Y=1*Y*=1. Then **b\_{1}=[b\_{1}], b\_{2}=[b\_{2}]*b*1​=[*b*1​],*b*2​=[*b*2​]. (b\_{1}*b*1​ and b\_{2}*b*2​ take integer values.)**

Hint: Check p2 of lec\_svm\_intro.pdf. You need the following two equations:

* \sum\_i \lambda\_i y\_i = 0.∑*i*​*λi*​*yi*​=0.
* the coefficient vector is proportional to \sum\_i y\_i \lambda\_i X\_i∑*i*​*yi*​*λi*​*Xi*​.

[b\_{1}][*b*1​]=\_\_\_\_\_-2\_\_\_\_\_\_\_\_\_\_

1 point



4.

Question 4

[b\_{2}][*b*2​]=\_\_\_\_\_2\_\_\_\_\_\_\_\_\_\_

1 point



5.

Question 5

Which **two** of the following statements are correct?

1 point



If we decrease the cost parameter (i.e., the value of C), the 0/1 training error will decrease.



Suppose we use a (binary) SVM to classify the digits data (that have 10 classes) using the one-vs-one approach. Then we need to train 45 SVMs.



If the data are linearly separable, an SVM with a linear kernel will return the same decision boundary regardless of the value of C.



The maximum value of a Gaussian kernel is 1.

6.

Question 6

**For the remaining questions, use the `e1071` package in R.**

**Question 6 to Question 14 are related.** Train a **linear SVM** to classify the spam data. Use 1:100 and 1901:1960 samples as the test sample.

Report the number of support vectors, training error, and test error, for SVMs with different cost values. Here error refers to the number of misclassified emails. So all **your answers should be integers.**

* If cost=1 (the default choice), the resulting SVM uses **[s1]** support vectors, with training error **[tr1]** and test error **[te1]**
* If cost=10, the resulting SVM uses **[s2]** support vectors, with training error **[tr2]** and test error **[te2]**
* If cost=50, the resulting SVM uses **[s3]** support vectors, with training error **[tr3]** and test error **[te3]**

**[s1]**=\_\_\_904\_\_\_\_\_\_\_ (integer)

1 point



7.

Question 7

**[tr1]**=\_\_291\_\_\_\_\_\_\_\_ (integer)

1 point



8.

Question 8

**[te1]**=\_\_\_\_14\_\_\_\_\_\_\_ (integer)

1 point



9.

Question 9

**[s2]**=\_\_\_\_\_879\_\_\_\_\_\_ (integer)

1 point



10.

Question 10

**[tr2]**=\_\_\_288\_\_\_\_\_\_\_\_ (integer)

1 point



11.

Question 11

**[te2]**=\_\_\_\_14\_\_\_\_\_\_\_\_\_\_\_ (integer)

1 point



12.

Question 12

**[s3]**=\_\_\_\_\_ 861 \_\_\_\_\_\_ (integer)

1 point



13.

Question 13

**[tr3]**=\_\_\_\_ 291 \_\_\_\_\_\_\_\_ (integer)

1 point



14.

Question 14

**[te3]**=\_\_\_\_\_ 14 \_\_\_\_\_\_\_ (integer)

1 point



15.

Question 15

**Question 15 to Question 23 are related.** Train a **Gaussian SVM** to classify the spam data, use the default gamma value. Use 1:100 and 1901:1960 samples as the test sample.

Report the number of support vectors, training error, and test error, for SVMs with different cost values. Here error refers to the number of misclassified emails. So all your **answers should be integers**.

* If cost=1 (the default choice), the resulting SVM uses **[s1]** support vectors, with training error **[tr1]** and test error **[te1]**.
* If cost=10, the resulting SVM uses **[s2]** support vectors, with training error **[tr2]** and test error **[te2]**.
* If cost=50, the resulting SVM uses **[s3]** support vectors, with training error **[tr3]** and test error **[te3]**.

**[s1]**=\_\_\_1243\_\_\_\_\_\_\_\_\_

1 point



16.

Question 16

**[tr1]**=\_\_\_232\_\_\_\_\_\_\_\_

1 point



17.

Question 17

**[te1]**=\_\_\_\_\_14\_\_\_\_\_\_\_\_\_

1 point



18.

Question 18

**[s2]**=\_\_\_\_1008\_\_\_\_\_\_

1 point



19.

Question 19

**[tr2]**=\_\_\_\_146\_\_\_\_\_\_\_\_\_

1 point



20.

Question 20

**[te2]**=\_\_\_\_\_15\_\_\_\_\_\_\_\_\_\_

1 point



21.

Question 21

**[s3]**=\_\_\_\_\_956\_\_\_\_\_\_\_\_\_

1 point



22.

Question 22

**[tr3]**=\_\_\_90\_\_\_\_\_\_\_\_\_\_\_

1 point



23.

Question 23

**[te3]**=\_\_\_\_17\_\_\_\_\_\_\_\_\_\_

1 point