Benjamin Ye

CS/CNS/EE 156a: Learning Systems (Fall 2023)

December 1, 2023

**Final Exam**

|  |  |  |  |
| --- | --- | --- | --- |
| **Problem** | **Answer** | **Problem** | **Answer** |
| 1 | [e] | 11 |  |
| 2 |  | 12 |  |
| 3 |  | 13 |  |
| 4 |  | 14 |  |
| 5 |  | 15 |  |
| 6 |  | 16 |  |
| 7 |  | 17 |  |
| 8 |  | 18 |  |
| 9 |  | 19 |  |
| 10 |  | 20 |  |

**Nonlinear Transforms**

1. The polynomial transform of order applied to of dimension resuilts in a space of what dimensionality (not counting the constant coordinate or )?

**Answer: [e] None of the above**

For a second-order polynomial transformation, we have :

For a third-order polynomial transformation, we have :

And from Homework 5 Problem 3, we have for a fourth-order polynomial transformation.

By interpolating the dimensionalities of these lower-order examples, it is clear to see that

That is, each additional order adds terms. Therefore,

**Bias and Variance**

1. Recall that the average hypothesis was based on training the same model on different data sets to get , and taking the expected value of with respect to to get . Which of the following models could result in ?

**Answer:**

**Overfitting**

1. Which of the following statements is false?

**Answer:**

1. Which of the following statements is true?

**Answer:**

**Regularization**

1. The regularized weight is a solution to:

where is a Tikhonov matrix. If , where is the linear regression solution, then what is ?

**Answer:**

1. Soft-order constraints that regularize polynomial models can be

**Answer:**

**Regularized Linear Regression**

We are going to experiment with linear regression for classification on the processed U.S. Postal Service Zip Code data set from Homework 8. Download the data (extracted features of intensity and symmetry) for training and testing:

<http://www.amlbook.com/data/zip/features.train>

<http://www.amlbook.com/data/zip/features.test>

(The format of each row is: **digit, intensity, symmetry**.) We will train two types of binary classifiers; one-versus-one (one digit is class +1 and another digit is class , with the rest of the digits disregarded), and one-versus-all (one digit is class +1 and the rest of the digits are class ). When evaluating and of the resulting classifier, use binary classification error. Implement the regularized least-squares linear regression for classification that minimizes

where includes .

1. Set and dop not apply a feature transform (i.e., use ). Which among the following classifiers has the lowest ?

**Answer:**

1. Now, apply a feature transform , and set . Which among the following classifiers has the lowest ?

**Answer:**

1. If we compare using the transform versus not using it, and applying that to “0 versus all” through “9 versus all”, which of the following statements is correct for ?

**Answer:**

1. Train the “1 versus 5” classifier with with and . Which of the following statements is correct?

**Answer:**

**Support Vector Machines**

1. Consider the following training set generated from a target function , where :

Transform this training set into another two-dimensional space :

Using geometry (not quadratic programming), what values of (without ) and specify the separating plane that maximizes the margin in the space? The values of , , and are:

**Answer:**

1. Consider the same training set of the previous problem, but instead of explicitly transforming the input space , apply the hard-margin SVM algorithm with the kernel

(which corresponds to a second-order polynomial transformation). Set up the expression for and solve for the optimal (numerically, using a quadratic programming package). The number of support vectors you get is in what range?

**Answer:**

**Radial Basis Functions**

We experiment with the RBF model, both in regular form (Lloyd + pseudo-inverse) with centers:

(notice that there is a bias term), and in kernel form (using the RBF kernel in hard-margin SVM):

The input space is with uniform probability distribution, and the target is

which is slightly nonlinear in the space. In each run, generate 100 training points at random using this target, and apply both forms of RBF to these training points. Here are some guidelines:

* Repeat the experiment for as many runs as needed to get the answer to be stable (statistically away from flipping to the closest competing answer).
* In case a data set is not separable in the “ space” by the RBF kernel using hard-margin SVM, discard the run but keep track of how often this happens, if ever.
* When you use Lloyd’s algorithm, initialize the centers to random points in and iterate until there is no change from iteration to iteration. If a cluster becomes empty, discard the run and repeat.

1. For , how often do you get a data set that is not separable by the RBF kernel (using hard-margin SVM)? Hint: Run the hard-margin SVM, then check if the solution has .

**Answer:**

1. If we use for regular RBF and take , how often does the kernel form beat the regular form (excluding runs mentioned in Problem 13 and runs with empty clusters, if any) in terms of ?

**Answer:**

1. If we use for regular RBF and take , how often does the kernel form beat the regular form (excluding runs mentioned in Problem 13 and runs with empty clusters, if any) in terms of ?

**Answer:**

1. Now we focus on regular RBF only, with . If we go from clusters to clusters (only 9 and 12), which of the following 5 cases happens most often in your runs (excluding runs with empty clusters, if any)? Up or down means strictly so.

**Answer:**

1. For regular RBF with , if we go from to (only 1.5 and 2), which of the following 5 cases happens most often in your runs (excluding runs with empty clusters, if any)? Up or down means strictly so.

**Answer:**

1. What is the percentage of time that regular RBF achieves with and (excluding runs with empty clusters, if any)?

**Answer:**

**Bayesian Priors**

1. Let be the unknown probability of getting a heart attack for people in a certain population. Notice that is just a constant, not a function, for simplicity. We want to model using a hypothesis . Before we see any data, we assume that is uniform over (the prior). We pick one person from the population, and it turns out that he or she had a heart attack. Which of the following is true about the posterior probability that given this sample point?

**Answer:**

**Aggregation**

1. Given two learned hypotheses and , we construct the aggregate hypothesis given by for all . If we use the mean-squared error, which of the following statements is true?