ANLY512: Homework 4

Due Apr 7 11:59pm

Problem 1: - Explore the characteristics of the ROC curve

Suppose a logistic model has been fitted to a data set and the ROC curve has been plotted. If the point with sensitivity = α and specificity = β is on the curve, what are the smallest and largest values for the area under the curve?

Problem 2: - Distinguish between digits in MNIST using logistic regression

We'll use the MNIST image classification data, available as mnist_all.RData. For this problem we want to distinguish between `1` and `3`. Filter the train data to only include digits 1 and 3. Remove all variables (pixels) that have zero variance, i.e. pixels that have the same value for both digits. Repeat this for the test data. It's recommended to write a function that you can run on both datasets.

In this problem, you will do two forward selection steps for finding good logistic models.

part a: Find the pixel that gives the best logistic model for the training data, using the area under the ROC curve as a criterion. Do this with a complete search. Do not show the output of all logistic models!

part b: Now find one more pixel such that the resulting logistic model using the pixel from part a together with the new one has the best area under the ROC curve. Do this with a complete search. Minimize the output.

part c: Use the test data to decide whether the second model is really better than the first one.

part d: How many logistic models altogether have you examined? How many will you have to examine if you want to continue this process and make the best logistic model with 10 pixels?

Problem 3: - Exploring MNIST data using neural networks

In this problem we want to distinguish between `4` and `7`. Extract the relevant train data and test data.

part a: Pick two features (variables) that have large variances and low correlation. Fit a logistic regression model with these two features. Evaluate the model with the AUC score.

part b: Create a neural net with one unit in the hidden layer. Train the neural net with the same two features as the previous part and evaluate the model with AUC. Compare to the results from (a) and explain.

part c: With the same two features, train three different neural nets, each time using more units in the hidden layer. How do the results improve, using the AUC?

part d: Is there evidence for overfitting in your results in (c)? Use the test data, also available in mnist_all.RData, to find out.

Problem 4 - Explaining the structure of shallow neural networks

Below is the output from nnet after we fit a model. Let's assume we used a `tanh()` activation function throughout. Let `xi`, `i = 1, 2, . . .` be the input variables and let `h1, h2, . . .` be the output from the hidden layer.

```
a 2-2-1 network with 9 weights
options were - linear output units
b->h1 i1->h1 i2->h1
1.2 4.2 -0.5
b->h2 i1->h2 i2->h2
-30 20 -40
b->o h1->o h2->o
5 -8 1.5
```

part a: Draw a diagram of this neural network architecture. Label all the edges with the corresponding weights.

part b: Provide an expression for the output value of the first hidden unit as a function of the values of the input features. This should have the form h1 = f(x1, x2, . . .) for a suitable explicit function f.

part c: Provide an expression for the value at the output node as a function of the values at the hidden units. This should have the form z = g(h1, h2, . . .) for a suitable explicit function g.

part d: Provide an expression for the value at the output node as a function of the input values. This should have the form z = F(x1, x2, . . .) for a suitable explicit function F.

Problem 5 - Analyze the crime rates in the `Boston` dataset using logistic, LDA, and QDA models

The goal of this problem is to predict the crime rate of neighborhoods using classification methods. In order to convert your quantitative variable `crim` to a binary outcome, make `1` or `high_crime` when `crim` is greater than the median `crim`. Use ten-fold cross-validation for each model type - logistic, LDA, and QDA.

part a: Logistic regression method

part b: LDA method

part c: QDA method

part d: Compare the results of all three models by plotting all three ROC curves in a single plot. Also report the area under the curve (AUC) for each method in the plot. Comment on which model should be chosen.

Problem 6 - Characteristics of LDA and QDA

part a: Suppose that the form of Bayes decision boundary is linear, which is the better performing model on the training set, LDA or QDA? What about the testing set?

part b: If instead the Bayes decision boundary is non-linear, which model, LDA or QDA, will perform better on the training set? What about the testing set?

part c: In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

part d: True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.