Homework 2: Linear Models

CS412

Released: February 4th

Due: February 13th, 11:30pm on Gradescope

1 Feature Extraction

Use the same training set that you used for HW1. Use your ML package to extract two features using Primary Component Analysis (PCA) and graph the training data in 2D space with the axes as the first and second components. You do not need to label your axes. **Label this Figure 2.1 in your report**.

Similarly, use Linear Discriminate Analysis (LDA) to extract two features. Plot the data in 2D space using these features as your axes. You do not need to label your axes. **Label this Figure 2.2 in your report**

a) Compare the above figures with the features you selected from HW1. Do either of these extracted feature 2D problems seem to better separate the data?

b) Give the explained variance ratio for each of the two feature extractions given above. Is there one of the methods which explains more variance than the other? Is this what you expect? Explain your answer.

THIS resource should be very helpful.

2 Logistic Regression

Use the two features that you created for your 2D graph in HW1. Use a logistic regression classifier to classify your data. In the **sklearn documentation for the Logistic Regression classifier** , notice the attribute C, which is the inverse of the regularization strength. Unless specified, use L2 regularization.

Plot the decision region for your 2D space with a logistic regression solver where c is 0.01. You should reuse the region plotting code from HW1. **Label this figure 2.3**

Plot the decision region for your 2D space as above but with C = 2.0. **Label this figure 2.4**

Perform 10-fold cross-validation and report the error for 100 values of c from 0.01 to 1.00. Graph these numbers with the x-axis as the value for c and the y axis should be the cross validation error. **Label this Figure 2.5** This may take some time to run.

a) For what value of C is the cross validation error lowest? Is this what you expect?

b) **Graduate student question:** Repeat the experiment for Figure 2.5 using L1 regularization. Does this regularization method make it more or less likely for the model to overfit the data. If you don't think there is any overfitting, defend your answer.

Extra Credit

Run a logistic regression classifier on the 256D data. Find the optimum level of regularization and give the reported cross validation error. Use the system clock to time the number of milliseconds it takes to run the 10-fold cross validation for the 256D fit. Compare this to the time it takes to fit the 2D data from above. Give the cross validation error for the two models and discuss the tradeoff between runtime and accuracy. Use data to support your conclusions as needed.

Making your report

When you submit, there will be two submissions on gradescope. One for your pdf report and another for your zip file containing all your code. If you use a language other than R or Python, include a list of any packages you downloaded in your report.