Homework 2: Linear + Kernel Models

CS412

Released: February 18th

Due: February 27th, 11:30pm on Gradescope

1 Feature Extraction

Use the same training set that you used for HW1. Use packages to extract two features using Kernel Primary Component Analysis (kPCA) and graph the training data in 2D space with the axes as the first and second components.. Do this for a polynomial kernel with degree 1. **Label this Figure 2.1 in your report**.

a) Compare kernel PCA features with the extracted features from HW1. Do these features seem to better separate the data?

b) Repeat the process above with kernel of degree 3. **Label this Figure 2.2 in your report**.

THIS resource should be very helpful.

2 Logistic Regression

Use the two features 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**

Graduate student question: Repeat the experiment for Figure 2.3 and 2.4 using L1 regularization. Does this regularization method make it more or less likely for the model to overfit the data? Does L1 vs L2 regularization have an impact here? If you don't think there is any overfitting, defend your answer.

3 Support Vector Machines

In this section, you will compare two different types of SVM. **HERE** is the python library for the SVC. The first is just a linear, soft margin SVM. On your 2D data from HW1, find the cross-validation errors for all values of c from 0.01 to 100. Your x axis should have logarithmic scale and you should have at least 20 data points. **Label this figure 2.5**

Repeat the above experiment for the 256D degree data. **Label this figure 2.6**
Find the value of c from Figure 2.4 which has the lowest cross validation error and plot the decision region in your 2D space. **Label this 2.7** Your x and y variables should be the same ones from HW1.

Now, repeat the experiment for the polynomial kernel model for your 2D data. Let the degree of the kernel be each of 2,5,10,20. Give the value of c for each of these kernel degrees which gives the lowest cross validation error. Explain the tradeoff between the 'degree' and 'c' as far as fit is concerned.

Using your 2D data from HW1, find the values for 'degree' and 'c' that you believe finds the best fit for the model. Plot the decision region for this model (From figure 2.8) and explain why you think this model is best. Support your conclusion with data.

Graduate student question: Provide two graphs of decision regions for SVC models in the 2D space. One should have evidence of overfitting, and the other should have evidence of underfitting. Explain the paramaters that lead to each of the graphs and their cross-validation errors.

Extra Credit

For the radial kernel support vector classifiers, examine the parameter gamma. Conduct experiments for different values of gamma, c and degree and try to determine the relationship between gamma and over/under fitting for the model. Find the optimal radial SVC model in the 2D space. Support your explanation with data and figures.

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 jupyter notebook, you need only submit the pdf of the notebook itself.