COM S 474/574 Spring 2022
Introduction to Machine Learning

Homework 7

1 Directions:

- **Due: Thursday April 7, 2022 at 9pm.** Late submissions will be accepted for 24 hours with a 15% penalty. (the enforcement is strict, beginning at 9:01pm, except for extreme situations; having a poor wifi connection or minor computer problems is not sufficient for the penalty to be waived.)
- Upload the homework to Canvas as a single pdf file.
- If the graders cannot easily read your submission (writing is illegible, image is too dark, or if the contrast is too low) then you might receive a zero or only partial credit.
- Any non-administrative questions must be asked in office hours or (if a brief response is sufficient) Piazza.

2 Problems

Problem 1. [30 points] For this problem you will fit classifiers to three simple data-sets and visually examine the classification boundaries.

(a). Download the example script `https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html` and run it (make sure there are no errors).

(b). Change the classifiers used to the following:

- logistic regression
- LDA
- QDA
- Max Margin (use `SVC` with a linear kernel and hyper-parameter `C=1000`)
- Support vector classifier
- Support vector machine with polynomial kernel of degree 2
- Support vector machine with polynomial kernel of degree 5
- Support vector machine with a radial basis function kernel

You may need to do a little reading of the sklearn documentation to use these. Some hyper-parameters are named differently than in class. For instance, for SVC `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html` the hyper-parameter "C" parameter is not the "budget" of mistakes as in lecture or the book. Except as noted above, use the default values of hyper-parameters.

(c). Make the following changes:

- For test samples, add

  ```
  marker="*", s=100
  ```

  to the plot arguments
- set the image resolution to 150dpi, such as using one of

  ```
  plt.rcParams['figure.dpi'] = 150
  plt.rcParams['savefig.dpi'] = 150
  ```
- For the data set named `linearly_separable`, comment out code adding uniform random noise

With those changes, run the script, examine the plots, and answer the following questions. Bear in mind that we did not optimize hyper-parameters (such as for regularization penalties).

A. Include a copy of your plot.

B. Based on discussions in lecture, which methods do you expect to fit models that would result in linear decision boundaries? Was that the case? Briefly describe whether the methods that have linear boundaries appear to find the same/similar/different boundaries.

C. Repeat the same for methods known to have boundaries defined by quadratic equations.

D. For methods whose boundaries can be non-linear, do they appear linear for the "linearly separable" data set?

E. How do LDA and QDA compare for these data sets?

F. How do the different support vector methods compare?