Thai Nguyen, Gahyun (Susie) Kim, Colin Stern
CS390MB
A2, Part 1: Supervised Classification

1. In the activity recognition task, what are the latent variables and what are the known variables?
    - **Latent** variables:
        - Action (biking, running, walking)
    - **Known** variables:
        - Max speed (mph)
        - Average speed (mph)

2. Describe what happens when you vary the number of data points in the generated dataset. Could a linear classifier achieve 100% accuracy on held out generated test data, if given arbitrarily sufficient (e.g. approaching infinite) training data? Explain.
    - In general, increasing the number of data points increases the accuracy of the classifier. However, a linear classifier cannot achieve 100% accuracy on some data sets, even with nearly infinite training data, because some data cannot be separated neatly by a linear model. For example, a linear classifier would never get better at separating data in the shape of a ring.

3. In the script, compute the precision and recall from the confusion matrix. Why is accuracy not a sufficient metric for reporting classification performance? What crucial information do precision and recall capture that accuracy does not? Briefly describe a scenario where high accuracy may be misleading.
    - Accuracy is not a sufficient metric for reporting classification performance because it ignores the amount of data collected in each class and we lose data on the prediction accuracy of each class. Precision captures the proportion of the predicted positive cases that were correct and recall captures the proportion of positive cases that were correctly identified. Additionally, precision and recall calculates the accuracy of prediction for each class. There is an imbalance in the average calculation and we have an inflated accuracy. A scenario where high accuracy may be misleading is when our test data has 90 cases of running and 10 of walking, and we correctly identify all of the running cases but fail to label a single walking case. However, the accuracy would be 90%. Clearly, accuracy fails to capture information that precision and recall do.
4. What happens if you shuffle the data points in the cross validation step? Compare your results both quantitatively and qualitatively to the case where shuffle=False.
    - Shuffle increases the accuracy from 0.87 to 0.9. When shuffling, we reduce bias in the data because it randomly selects data points without regard to their original positions. For example, if we have 100 sample and the first 90 data points are running and 10 is biking. If we choose the first 90, there would be no data about biking. Shuffling balances both the training and test sets.