DSC 255: Machine Learning
Homework 7

Mathematical and conceptual exercises

1. A linear predictor is used to solve a classification problem with three classes. The data are two–dimensional and the linear functions for each class are:

   - Class 1: $w_1 = (1, 1),\ b_1 = 0$
   - Class 2: $w_2 = (1, 0),\ b_2 = 1$
   - Class 3: $w_3 = (0, 1),\ b_3 = -1$

   Draw the resulting decision boundary and clearly mark the region corresponding to each class.

2. **Plant recognizer.** Suppose you are building a plant recognition system that takes as input a photograph of a plant and outputs the name of that plant. Your intention is that this will be used throughout California. When picking a training set for this task, which of the following options would best satisfy the statistical learning framework?

   (a) Do a web search on `American plants` and download a subset of the images you find.

   (b) Obtain a collection of photos from a region of the US whose flora is similar to that of California.

   (c) Go to your favorite Californian city and take photos of the plants you encounter.

   Give a brief explanation of your answer.

3. **A shortage of data.** We have a binary classification problem with very high-dimensional data: there are a million features. Unfortunately, we only have 1000 training points. Nonetheless, we train a support vector machine classifier and find that it works well in practice. What is a possible explanation for why we are able to find a good model despite the shortage of labeled data?

4. **Distribution shift.** Suppose we are building a document classification system that categorizes news articles according to topic: sports, politics, business, and so on. To train this system, we use a corpus of *New York Times* articles from the past decade. However, by test time the distribution has changed somewhat. Each of the following scenarios is an example of either *covariate shift* or *label shift*. Say which is which, with a brief explanation.

   (a) There are fewer articles on sports and more on politics.

   (b) The important public figures, and thus the proper nouns in the articles, have changed.

Programming exercises

For this week you will need the data file `data0.txt`, which you can download from the course web site.

1. **Multiclass Perceptron.** Implement the multiclass Perceptron algorithm from class.

   (a) Load the data set *data0.txt*. This file contains 2-d data in four classes (coded as 0,1,2,3). Each row consists of the two coordinates of a point followed by its label.

(b) Run the multiclass Perceptron algorithm to learn a classifier. Create a plot that shows all data points (using different colors and shapes for different labels) as well as the decision regions.

2. Multiclass SVM. In this problem we will use support vector machines on the same data set, `data0.txt`.

- Learn a linear SVM classifier using `sklearn.svm.LinearSVC`. Set `loss='hinge'` and `multi_class='crammer_singe`
  Try $C \in \{0.01,\ 0.1,\ 1.0,\ 10.0\}$.

(a) For each value of C, plot the decision boundary (no need to show the margins).

(b) What do you notice as C increases? Briefly comment.