Exploring classification tools in sklearn

M. Hassell

July 31, 2017

1 Introduction

In the context of machine learning, *classification problems* ask for a model to assign to some input a discrete class label. This label could be anything from a 0 or 1 (perhaps corresponding to a yes or no answer to some question), a color, or a digit representing a numeral, as is the case with the MNIST data set we will explore here. Classification differs from another major ML strategy, *regression* whereby a (generally) continuous predictor is assigned to some data. This could be something like to price of a house given the number of bedrooms, bathrooms, and zip code. One could also predict the amount of electricity that needs to be generated at a given time of the day for a certain area, given the season, population, predicted weather, and historical usage. We will focus for now on classification problems, where our desired outcomes, or *labels* take values in a discrete set.

2 Jumping in

We'll start by using the MNIST handwriting data that is included with `sklearn` in the `datasets` module. This is convenient since the data has already been acquired and put into a standardized format, which is no small task on its own. The data set includes 1797 images of the digits 0-9, stored as 8×8 arrays with pixel intensities 0-255. When we perform any model training or prediction on elements of this data set, we will need to reshape these arrays into 64×1 vectors, as elements of $\mathbb{R}^{64}$. Our first attempt at training a model for this dataset will make use of a support vector classifier (we'll see this as SVC within the sklearn environment).

3 What's it all mean?

A support vector classifier for *two* classes seeks to find a separating hyperplane (if possible) that separates the two classes with the greatest possible margin (check the Wikipedia for some nice pictures of this). Recall that the definition of a hyperplane is given by

$$\{x \in \mathbb{R}^n \mid w \cdot x - b = 0,\ w \in \mathbb{R}^n,\ b \in \mathbb{R}\}$$

Note that this is an affine set when $b \neq 0$ (it does not contain the origin). If w is not a unit vector we can divide through by its norm to get the more canonical form

$$\frac{1}{\|w\|} w \cdot x - \frac{b}{\|w\|} = 0$$

where $\frac{b}{\|w\|}$ is now more clearly our offset from the origin.

Supposing that we are in the linearly separable case, we consider data of the form (x_i, y_i) for $i = 1, \ldots, n$ where $y_i = 1$ or $y_i = -1$ for data taking values on one side of the hyperplane or the other. There are then two parallel hyperplanes that separate the data, with equations

$$w \cdot b - 1 = 0$$

and

$$w \cdot b + 1 = 0.$$

The region between is called the margin, and the hyperplane that lies halfway between them is called the *maximum margin hyperplane.*

Equations here

discussion of the equations

4 A natural generalization

Now that we've understood the case of two classes in the linearly separable case, we can look into the two natural generalizations. The first obvious generalization is to consider more than two classes. What if we have n distinct classes that are linearly separable. We can imagine this like one of the variety pack cheesecakes from Barnes and Noble. There are eight slices of cheesecake and four potential flavors, each occupying one quarter of the cheesecake. We seek to divide the cheesecake into quarters, where each quarter contains two slices