Building Models with Distance Metrics

This chapter will cover the following topics:

- Using KMeans to cluster data
- Optimizing the number of centroids
- Assessing cluster correctness
- Using MiniBatch KMeans to handle more data
- Quantizing an image with KMeans clustering
- Finding the closest objects in the feature space
- Probabilistic clustering with Gaussian Mixture Models
- Using KMeans for outlier detection
- Using k-NN for regression

Introduction

Clustering is often grouped together with unsupervised techniques. These techniques assume that we do not know the outcome variable. This leads to ambiguity in outcomes and objectives in practice, but nevertheless, clustering can be useful. As we'll see, we can use clustering to "localize" our estimates in a supervised setting. This is perhaps why clustering is so effective; it can handle a wide range of situations, and often, the results are for the lack of a better term, "sane". We'll walk through a wide variety of applications in this chapter; from image processing to regression and outlier detection. Through these applications, we'll see that clustering can often be viewed through a probabilistic or optimization lens. Different interpretations lead to various trade-offs. We'll walk through how to fit the models here so that you have the tools to try out many models when faced with a clustering problem.