12.1 The Challenge of Unsupervised Learning

Supervised Learning

- In the *supervised* learning setting, we have
  - some set of p features $X = X_1, X_2, \ldots, X_p$ measured on n observations
  - a response Y measured on those same n observations
- We wish to use X to predict Y.

Unsupervised Learning

- In the *unsupervised* learning setting, we don't have a response variable Y.
- Instead, we wish to discover interesting things about the measurements on X.
  - Is there an informative way to visualize the data?
  - Can we find meaningful subgroups among the variables/observations?
- We will discuss two methods:
  - Principal components analysis
  - Clustering

The Challenge

- Unsupervised learning is much more subjective, as there is no simple goal like prediction.
- Techniques for unsupervised learning are of growing importance:
  - subgroups of breast cancer patients grouped by gene expression measurements
  - groups of shoppers characterized by their browsing and purchase histories
  - movies grouped by the ratings assigned by movie viewers

An Advantage

- Can be easier to obtain *unlabeled data* from an instrument or computer, as it often requires no human intervention
- For example, difficult to automatically assess the overall sentiment of a movie review.
  - Some human needs to go through and manually tag all those reviews!