Lecture 23: Unsupervised learning

CSI 5v93: Introduction to machine learning

Baylor University
Computer Science Department

Dr. Greg Hamerly
`http://cs.baylor.edu/~hamerly/`

Questions?

Introduction

Until now, we have discussed the problem of *supervised* learning.

Supervised learning = having a “training signal” or correct answer for the training examples. This comes in two forms:

- class labels
- regression value

What if we didn’t have the training signal for the examples? Can we learn anything?

Answer: yes – we now turn to unsupervised learning.

Unsupervised learning

Unsupervised learning is learning without a training signal.

It can be very different than supervised learning

- no “correct answer”
- goal is different – the concept to learn is not given with the training data
- overfitting and underfitting are still possible, but it’s not as clear when they occur

Several views on unsupervised clustering

Summarization: Finding a concise representation of given data.

Density estimation: finding the probability distribution of the data.

Classification: Dividing a dataset into distinct subsets.

Pattern discovery: identifying relationships in the data that were unknown.

All this without a training signal – given only the input features of the data.

Types of unsupervised learning

Finding association rules. For example: “If a customer buys lunch meat and cheese, he will probably also buy bread.” (14.2)

Clustering (aka grouping or segmenting) data according to its similarity or dissimilarity. (14.3)

Finding low-dimensional, simple surfaces in high-dimensional data. (14.5)

Data clustering (4.3)

Several views/motivations:

- identify and group items which are similar into several groups
- identify and divide items which are different into several groups
- summarize the data with several prototypes

We can do all these with the k-means algorithm.

Data clustering (4.3)

Several views/motivations:

- identify and group items which are similar into several groups
- identify and divide items which are different into several groups
- summarize the data with several prototypes

We can do all these with clustering algorithms.

There are several types of clustering algorithms:

- hierarchical bottom-up
- hierarchical top-down
- iterative “flat” clustering (e.g. k-means and Gaussian EM)
- spectral clustering
- density-seeking (e.g. mean-shift)

Most clustering algorithms are iterative in some fashion.

The k-means algorithm

A "flat", iterative improvement clustering algorithm.

Very popular: easy to implement, fairly fast algorithm, provides good results.

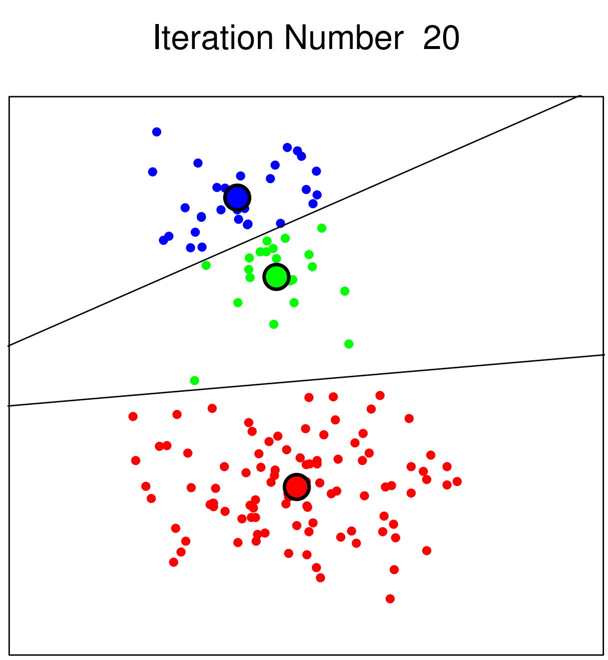
Input: a set of n data points in d dimensions.

Goal: find k prototypes, or centers, that represent the data "well".

The centers are *not* constrained to be part of the input.

Each center represents the data that is nearest to it.

The k-means algorithm



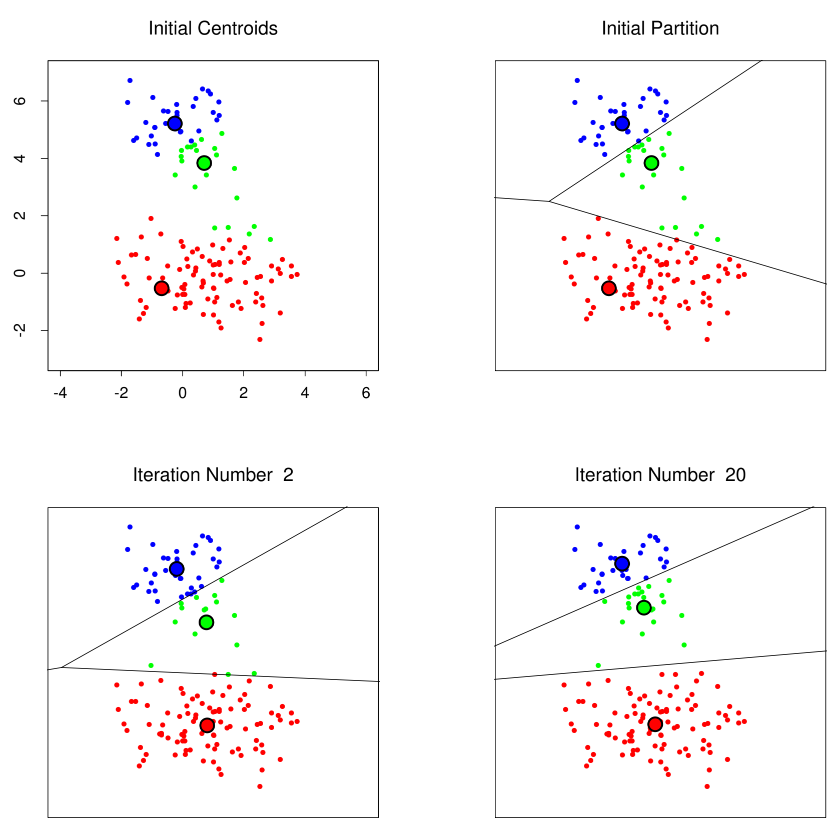
This is an example of using k-means with $k = 3$ clusters on 2-dimensional data.

The k-means algorithm

Basic algorithm:

- choose the number of cluster centers k
- position the k centers
- iterate until no change:
  - assign each data point to the cluster center nearest to the point
  - move each cluster center to the average of the points assigned to it

Example: k-means with $k = 3$ on 2d data



Demonstration of k-means

What k-means is actually doing

k-means is actually minimizing the within-cluster sum-of-squared-distances.

Let $\delta(i, j)$ indicate that datapoint x_i belongs to cluster center c_j. Then k-means minimizes

$$\begin{aligned} F(X, C) &= \sum_{i=1}^{n} \sum_{j=1}^{k} ||x_i - c_j||^2 \delta(i, j) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} \delta(i, j) \sum_{m=1}^{d} (x_{im} - c_{jm})^2 \end{aligned}$$

Some points:

- only finds a *local* minimum of $F(X, C)$
- may get stuck in poor solutions
- starting with some solution, always finds a solution that is better (or equal)
- because of the last point, it will always terminate

2-minute journal

Please write a response to the following on a piece of paper and hand it in immediately. Please make it anonymous (no names). Write about:

- major points you learned today
- areas not understood or requiring clarification