Announcements

Next on the agenda --- study some commonly used models

(Naive Bayes, clustering (e.g. GMM), temporal clustering (e.g. HMM))

Learning these models from data will motivate the EM algorithm for inference.

(Last part of the course will develop sampling based inference methods).

Naive Bayes

Suppose categories indexed by c, and features represented in a vector **x**.

Assume features in **x** are independent given the category.

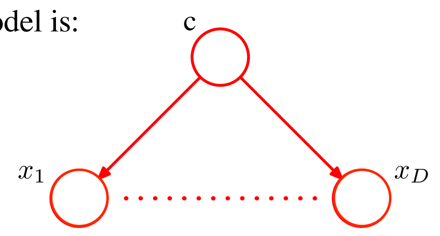
(Feature independence is the "naive" part).

$$p(\mathbf{x}|c) = \prod_i p(x_i|c)$$

Naive Bayes

$$p(\mathbf{x},c) = \mathrm{p}(c)\prod_i p(x_i|c)$$

Graphical model is:



Naive Bayes

$$p(\mathbf{x},c) = \mathrm{p}(c)\prod_i p(x_i|c)$$

Note that:

The forms of $p(x_i|c)$ need not all be the same (but often are)

If $p(\mathbf{x}|c)$ is a Gaussian, then it has diagonal covariance matrix. (This simplifying assumption is nearly always needed with Gaussians if the dimension, D, is large).

Naive Bayes

Typically, $p(x_i|c)$ come from training data linked to known (labeled) classes (supervised learning).

Example (1) fit a univariate Gaussian to each variable, x_i, for each class, c.

Example (2), record a histrogram for each variable, x_i, for each class, c.

Inference using Naive Bayes

$$p(\mathbf{x}|c) = \prod_i p(x_i|c) \qquad \text{(forward model)}$$

$$p(c|\mathbf{x}) \propto p(\mathbf{x}|c)\,p(c) \qquad \text{(the Bayes part)}$$

This leaves us with simple, and often very effective model and associated inference. We combine the likelihood $p(\mathbf{x}|c)$ with the prior $p(c)$ over categories.

Naive Bayes for face identification

- Example features
  - Location, color, texture, of left eye
  - Location, color, texture, of right eye
  - Location, color, texture, of mouth
  - Location, color, texture, of nose
- We can imagine training these with different facial expressions, lighting conditions, etc.
- Notice that these are not independent.
- This sort of thing often works pretty well anyway.
- Possible explanation is that, while the model allows for the eyes to be different, this rarely occurs in training or testing data.

Clustering

Clustering is the canonical case of “unsupervised” learning.

Given the data, what are the categories (clusters), c?

(Given a cluster, the features might be independent like Naive Bayes, or they might not be).

We will focus on clustering based on statistical models, but first review clustering in general.

Why is clustering hard?

Main reason

- The number of possible clusterings is exponential in the number of data points

Other important issues

- The number of clusters (and a good way to check) is usually **not** known
- A good distance function between points may not be known
- A good model explaining the existence of clusters may not be known.
- High dimensionality

Clustering based on distance measure

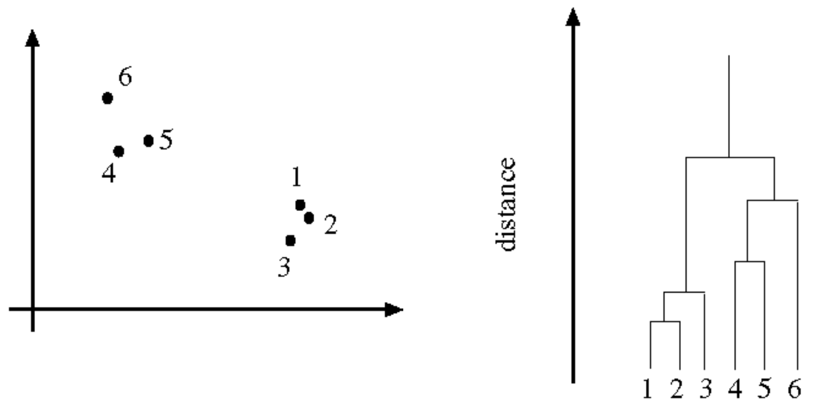
- Most common data representation is an N dimensional “feature” vector.
- Most common distance is Euclidian distance.
- Be careful with scaling and units!
- Probabilistic models can finesse scaling and multiple modalities
- Problems with correlated variables can be mitigated using transformations and data reduction methods such as PCA, ICA.

Clustering approaches

- Agglomerative clustering
  - initialize: every item is a cluster
  - attach item that is “closest” to a cluster to that cluster
  - repeat
- Divisive clustering (e.g., ncuts)
  - split cluster along best boundary
  - repeat recursively
- Clustering by optimizing a heuristic criterion (e.g., K-means)
  - e.g., small within variance
- Probabilistic clustering
  - define a probabilistic grouping model
  - use statistical inference to fit the model

Simple agglomerative approaches

- Point-Cluster or Cluster-Cluster distance
  - single-link clustering (minimum distance from point to points in clusters or among pairs of points, one from each cluster)
  - complete-link clustering (maximum)
  - group-average clustering (average)
  - (terms are not important, but concepts are worth thinking about)
- Dendrograms
  - classic picture of output as clustering process continues



K-Means

- Choose a fixed number of clusters ("K")
- Choose cluster centers (**means**) and point-cluster allocations (membership) to minimize the error

$$\sum_{i \in \text{clusters}} \left\{ \sum_{j \in \text{elements of i'th cluster}} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2 \right\}$$

- **x**'s could be any set of features for which we can compute a distance (careful with scaling)

K-Means

- Want to minimize

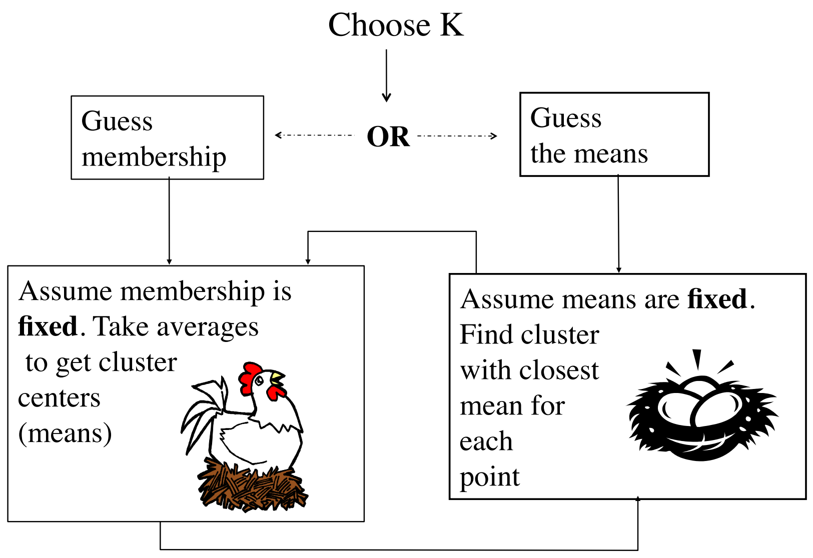
$$\sum_{i \in \text{clusters}} \left\{ \sum_{j \in \text{elements of i'th cluster}} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2 \right\}$$

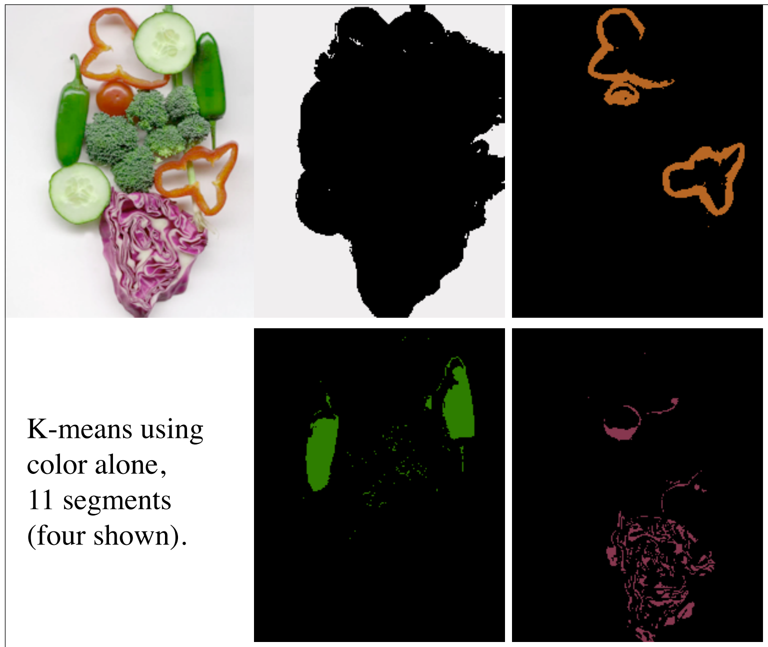
- **Cannot** do this optimization by search, because there are too many possible allocations.
- Standard difficulty which we handle with an iterative process (chicken and egg)

K-Means algorithm (intuition)

- If we know the cluster centers, the best cluster for each point is easy to compute
  - Just compute the distance to each to find the closest
- If we know the best cluster for each point, the cluster centers are also easy to compute
  - Just average the points in each cluster
- Algorithm
  - 1) Guess one of the two.
  - 2) Alternatively re-compute the values for each

K-means flow chart


Choose K
Guess membership
OR
Guess the means
Assume membership is **fixed**. Take averages to get cluster centers (means)
Assume means are **fixed**. Find cluster with closest mean for each point



K-means using color alone, 11 segments (four shown).

Notes on K-Means

- K-means is “hard” clustering. This means that each point is completely in exactly one cluster
- What you get is a function of starting “guess”
- The error goes down with every iteration
  - This means you get a local minimum, but not necessarily a global one.
- Unfortunately, the dimension of the space is usually large, and high-dimensional space has lots of room for local maximum (standard problem!)
  - Dimensionality here is K*dim(**x**)
- Finding the global minimum for a real problem is very optimistic!

you should be able to argue why this is true