Clustering and Dimensionality Reduction

Preview

- Clustering
  - K-means clustering
  - Mixture models
  - Hierarchical clustering
- Dimensionality reduction
  - Principal component analysis
  - Multidimensional scaling
  - Isomap

Unsupervised Learning

- Problem: Too much data!
- Solution: Reduce it
- Clustering: Reduce number of examples
- Dimensionality reduction:
  Reduce number of dimensions

Clustering

- Given set of examples
- Divide them into subsets of "similar" examples
- How to measure similarity?
- How to evaluate quality of results?

K-Means Clustering

- Pick random examples as initial means
- Repeat until convergence:
  - Assign each example to nearest mean
  - New mean = Average of examples assigned to it

K-Means Works If . . .

- Clusters are spherical
- Clusters are well separated
- Clusters are of similar volumes
- Clusters have similar numbers of points

Mixture Models

$$P(x) = \sum_{i=1}^{n_c} P(c_i) P(x|c_i)$$

Objective function: Log likelihood of data

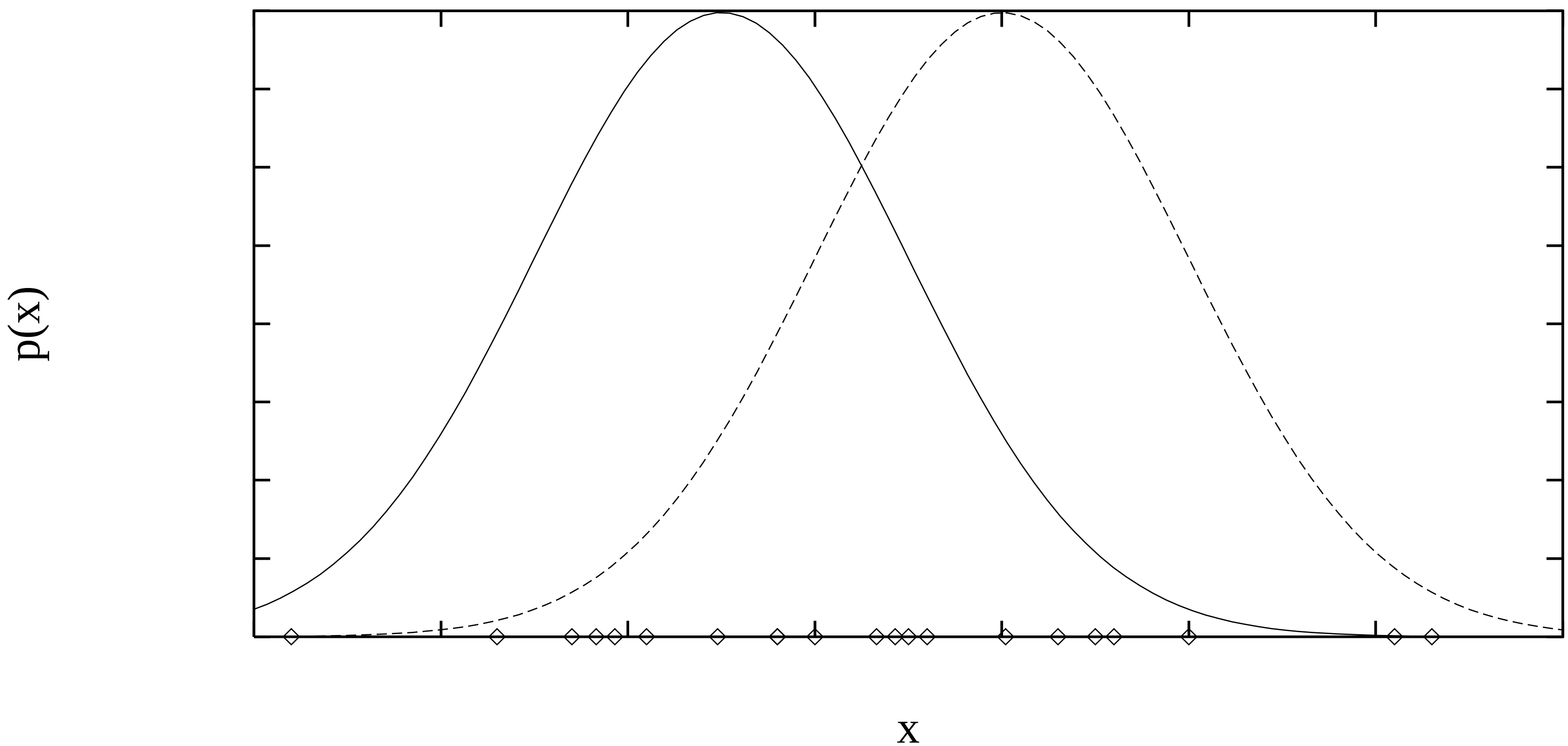
Naive Bayes: $P(x|c_i) = \prod_{j=1}^{n_d} P(x_j|c_i)$

AutoClass: Naive Bayes with various x_j models

Mixture of Gaussians: $P(x|c_i) =$ Multivariate Gaussian

In general: $P(x|c_i)$ can be any distribution

Mixtures of Gaussians



$$P(x|\mu_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma}\right)^2\right]$$

The EM Algorithm

Initialize parameters ignoring missing information

Repeat until convergence:

E step: Compute expected values of unobserved variables, assuming current parameter values

M step: Compute new parameter values to maximize probability of data (observed & estimated)

(Also: Initialize expected values ignoring missing info)

EM for Mixtures of Gaussians

Initialization: Choose means at random, etc.

E step: For all examples x_k:

$$P(\mu_i|x_k) = \frac{P(\mu_i)P(x_k|\mu_i)}{P(x_k)} = \frac{P(\mu_i)P(x_k|\mu_i)}{\sum_{i'} P(\mu_{i'})P(x_k|\mu_{i'})}$$

M step: For all components c_i:

$$\begin{aligned}
P(c_i) &= \frac{1}{n_e} \sum_{k=1}^{n_e} P(\mu_i|x_k) \\
\mu_i &= \frac{\sum_{k=1}^{n_e} x_k \, P(\mu_i|x_k)}{\sum_{k=1}^{n_e} P(\mu_i|x_k)} \\
\sigma_i^2 &= \frac{\sum_{k=1}^{n_e} (x_k - \mu_i)^2 \, P(\mu_i|x_k)}{\sum_{k=1}^{n_e} P(\mu_i|x_k)}
\end{aligned}$$

Mixtures of Gaussians (cont.)

- K-means clustering $\prec$ EM for mixtures of Gaussians
- Mixtures of Gaussians $\prec$ Bayes nets
- Also good for estimating joint distribution of continuous variables

Hierarchical Clustering

- Agglomerative clustering
  - Start with one cluster per example
  - Merge two nearest clusters
    (Criteria: min, max, avg, mean distance)
  - Repeat until all one cluster
  - Output dendrogram
- Divisive clustering
  - Start with all in one cluster
  - Split into two (e.g., by min-cut)
  - Etc.

Dimensionality Reduction

- Given data points in d dimensions
- Convert them to data points in $r < d$ dimensions
- With minimal loss of information

Principal Component Analysis

Goal: Find r-dim projection that best preserves variance

1. Compute mean vector μ and covariance matrix Σ of original points
2. Compute eigenvectors and eigenvalues of Σ
3. Select top r eigenvectors
4. Project points onto subspace spanned by them:

$$y = A(x - \mu)$$

where y is the new point, x is the old one, and the rows of A are the eigenvectors

Multidimensional Scaling

Goal: Find projection that best preserves inter-point distances

x_i Point in d dimensions

y_i Corresponding point in $r < d$ dimensions

δ_{ij} Distance between x_i and x_j

d_{ij} Distance between y_i and y_j

- Define (e.g.) $E(\mathbf{y}) = \sum_{i,j} \left(\frac{d_{ij} - \delta_{ij}}{\delta_{ij}} \right)^2$
- Find y_i's that minimize E by gradient descent
- Invariant to translations, rotations and scalings

Isomap

Goal: Find projection onto *nonlinear* manifold

1. Construct neighborhood graph G:
   For all x_i, x_j
   If distance$(x_i, x_j) < \epsilon$
   Then add edge (x_i, x_j) to G
2. Compute shortest distances along graph $\delta_G(x_i, x_j)$
   (e.g., by Floyd's algorithm)
3. Apply multidimensional scaling to $\delta_G(x_i, x_j)$

Summary

- Clustering
  - K-means clustering
  - Mixture models
  - Hierarchical clustering
- Dimensionality reduction
  - Principal component analysis
  - Multidimensional scaling
  - Isomap