CS 730/830: Intro AI

Unsupervised Learning

Overview

modeling = predicting = understanding = compressing

eg, clustering

finding 'structure' in data

Bottom-Up Unsupervised Learning

explain the data all-at-once vs piece-by-piece?

repeat
  repeat
    make a model to explain a minimal amount of data
    check how much of the total data the model explains
  until model fits a decent amount of the data
  if model, remove explained data from the set
until hard to find a decent model or not enough data left

Random Sample Consensus (RANSAC)

given data, find a set of explanatory models:

repeat
  repeat many times
    randomly pick minimum data to fit model
    find inliers
    repeat until no change
      fit model to inliers
      find new inliers
  if best model has enough inliers
    record model
    remove inliers from data
until best model not good enough or not enough data left

Break

- asst 10
- asst 11
- posters: two weeks from Friday (noon–2pm Kingsbury)
- papers: four weeks from yesterday! (2pm my office)
  - 2 hardcopies, 1 copy of code (2 pages per page)
  - PDF of paper and tarball/zip of code via email

k-Means Clustering

Naive Bayes model: choose class, generate attributes independently

mixture model: choose class, generate data

$$P(x|\theta) = \sum_k P(C = k|\theta_k)P(x|C = k, \theta_k)$$

eg, for mixture of Gaussians,

$$P(x|C = k, \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\sigma_k^2\pi}} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right)$$

An Algorithm

Means represent the center of a cluster/class
Values for the means are the model
Model changes based on the classes assigned to the data

init the k means somehow
repeat until cluster assignments do not change:
- Assign each data point to the mean nearest to it
- Calculate new means for the data assigned to each cluster

An Algorithm

Means represent the center of a cluster/class
Values for the means are the model
Model changes based on the classes assigned to the data

init the k means somehow
repeat until cluster assignments do not change:
  Assign each data point to the mean nearest to it
  Calculate new means for the data assigned to each cluster

Example

An Algorithm

Means represent the center of a cluster/class
Values for the means are the model
Model changes based on the classes assigned to the data

init the k means somehow
repeat until cluster assignments do not change:
  Assign each data point to the mean nearest to it
  Calculate new means for the data assigned to each cluster

Example

Is the classification optimal?
What is it optimizing?

Expectation-Maximization

model parameters θ (eg, $\mu, \sigma^2, P(C = k)$)
observed variables x_j
hidden variables C_j

init the θ_k somehow
repeat until done:
  E: compute expected values of hidden vars: $P(C_j = k|x_j, \theta_k)$
    eg by $\alpha P(C = k)P(x_j|C = k, \theta_k)$
  M: maximize data likelihood using current estimates:
    θ_k, with each x_j weighted by $P(C_j = k|x_j)$, eg by

$$\theta \leftarrow \operatorname*{argmax}_{\theta} \sum_{z} P(Z = z|x, \theta)P(x, Z = z|\theta)$$

Expectation-Maximization

model parameters θ (eg, $\mu, \sigma^2, P(C = k)$)
observed variables x_j
hidden variables C_j

init the θ_k somehow
repeat until done:
  E: compute expected values of hidden vars: $P(C_j = k|x_j, \theta_k)$
    eg by $\alpha P(C = k)P(x_j|C = k, \theta_k)$
  M: maximize data likelihood using current estimates:
    θ_k, with each x_j weighted by $P(C_j = k|x_j)$, eg by

$$\theta \leftarrow \operatorname*{argmax}_{\theta} \sum_{z} P(Z = z|x, \theta)P(x, Z = z|\theta)$$

greedy increase of data likelihood

Expectation-Maximization

Features

- Probabilistic clustering
- Explicit model
- Locally optimal

Issues

- Number of classes (means, Gaussians, etc.)
- Local maxima

Agglomerative Clustering

dendrogram
$O(n^2)$ vs $O(kn)$
AutoClass

Summary

supervised learning: learning a function or a density
unsupervised learning: explaining data
reinforcement learning: learning how to act

EOLQs

- What question didn't you get to ask today?
- What's still confusing?
- What would you like to hear more about?

Please write down your most pressing question about AI and put it in the box on your way out.
Thanks!