About this class

Unsupervised learning

k-means Clustering

Expectation Maximization

Unsupervised Learning

Build a model for your data. Which datapoints are similar?

Nowadays lots of work on using unlabeled data to improve the performance of supervised learning

k-means Clustering

Problem: given m data points, break them up into k clusters, where k is pre-specified

Objective: minimize $\sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||^2$

where μ_j is the cluster mean

Algorithm:

Initialize $\mu_1, \ldots, \mu_k$ randomly

Repeat until convergence:

1. Assign each x_i to the cluster with the closest mean

2. Calculate the new mean for each cluster

$$\mu_j \leftarrow \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Always terminates at a local minimum

Bad clustering examples for $k = 2$ (circles) and $k = 3$ (bad initialization leads to bad results)

Issues with k-means: how to choose k and how to initialize

Possible ideas: Use multiple runs with different random start configurations? Pick starting points far apart from each other?

Expectation Maximization

(EM developed by Dempster, Rubin & Laird, 1977. These notes mostly from Tom Mitchell's book, with some other references thrown in for good measure)

Let's do away with the "hard" assignments and maximize data likelihood!

Suppose points on the real line are drawn from one of two Gaussian distributions using the following algorithm:

1. One of the two Gaussians is selected

2. A point is sampled from the selected Gaussian and placed on the real line

Assume the two Gaussians have the same variance σ and unknown means μ_1 and μ_2. What are the maximum likelihood estimates of μ_1 and μ_2?

How do we think about this problem? Start by thinking about each data point as a tuple (x_i, z_{i1}, z_{i2}) where the zs indicate which of the distributions the points were drawn from (but they are unobserved).

Now apply the EM algorithm. Start with arbitrary values for μ_1 and μ_2. Now repeat until we have converged to stationary values for μ_1 and μ_2:

1. Compute each expected value $E[z_{ij}]$ assuming the means of the Gaussians are actually the current estimates of μ_1 and μ_2

$$E[z_{i1}] = \frac{f(x = x_i|\mu = \mu_1)}{f(x = x_i|\mu = \mu_1) + f(x = x_i|\mu = \mu_2)}$$

$$= \frac{\exp(-\frac{1}{2\sigma^2}(x_i - \mu_1)^2)}{\exp(-\frac{1}{2\sigma^2}(x_i - \mu_1)^2) + \exp(-\frac{1}{2\sigma^2}(x_i - \mu_2)^2)}$$

2. Compute updated (maximum likelihood) estimates of μ_1 and μ_2 using the expected values $E[z_{ij}]$ from step 1.

$$\mu_i = \frac{\sum_{i=1}^{m} E[z_{ij}]x_i}{\sum_{i=1}^{m} E[z_{ij}]}$$

EM in General

Define:

1. θ: parameters governing the data (what we're trying to find ML estimates of
2. X: observed data
3. Z: unobserved data
4. $Y = X \cup Z$

We want to find $\widehat{\theta}$ that maximizes $E[\ln \Pr(Y|\theta)]$

The expectation is taken because Y itself is a random variable (the Z part is unknown!)

But we don't know the distribution governing Y, so how do we take the expecation?

EM uses the current estimate of θ, call it h, to estimate the distribution governing Y

Define $Q(h'|h)$ that gives the expected log probability above, *assuming that the data were generated by* h

$$Q(h'|h) = E[\ln \Pr(Y|h')|h, X]$$

Now EM consists of repeating the next two steps until convergence

1. Estimation (E) step: Calculate $Q(h'|h)$ using the current estimate h and the observed data X to estimate the probability distribution over Y
2. Maximization (M) step: Replace h by the h' that maximizes Q

Again, only guaranteed to converge to a local minimum

Deriving Mixtures of Gaussians

Let's do this for k Gaussians

First, let's derive an expression for $Q(h'|h)$

$$f(y_i|h') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2} \sum_{j=1}^{k} z_{ij}(x_i - \mu'_j)^2$$

$$\sum_{i=1}^{m} \ln f(y_i|h') = \sum_{i=1}^{m} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^{k} z_{ij}(x_i - \mu'_j)^2 \right)$$

Taking the expectation and using $E[f(z)] = f(E[z])$ when f is linear:

$$Q(h'|h) = \sum_{i=1}^{m} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^{k} E[z_{ij}](x_i - \mu'_j)^2 \right)$$

And the expectation of z_{ij} is computed as before, *based on the current hypothesis*:

$$E[z_{ij}] = \frac{\exp(-\frac{1}{2\sigma^2}(x_i - \mu_j)^2)}{\sum_{n=1}^{k} \exp(-\frac{1}{2\sigma^2}(x_i - \mu_n)^2)}$$

E-step defines the Q-function in terms of the expectations generated by the previous estimate.

Then the M-step chooses a new estimate to maximize the Q-function, which is equivalent to finding the μ'_j that minimize:

$$\sum_{i=1}^{m} \sum_{j=1}^{k} E[z_{ij}](x_i - \mu'_j)^2$$

This is just a maximum likelihood problem with the solution described earlier, namely:

$$\frac{\sum_{i=1}^{m} E[z_{ij}] x_i}{\sum_{i=1}^{m} E[z_{ij}]}$$