Lecture 15 - EM Algorithm

The Expectation-Maximization (EM) algorithm

The expectation maximization algorithm is a natural generalization of maximum likelihood estimation to the incomplete data case. – Chuong B Do & Serafim Batzoglou. What is the expectation maximization algorithm? Nature Biotechnology. 2008.

Overall

- A general framework, introduced by Laird and Rubin (1977) to encompass a variety of problems:
  - Filling in missing data
  - Infering Latent Variables
  - Estimating HMM parameters
  - Estimating Parameters for mixture models
  - Unsupervised cluster learning

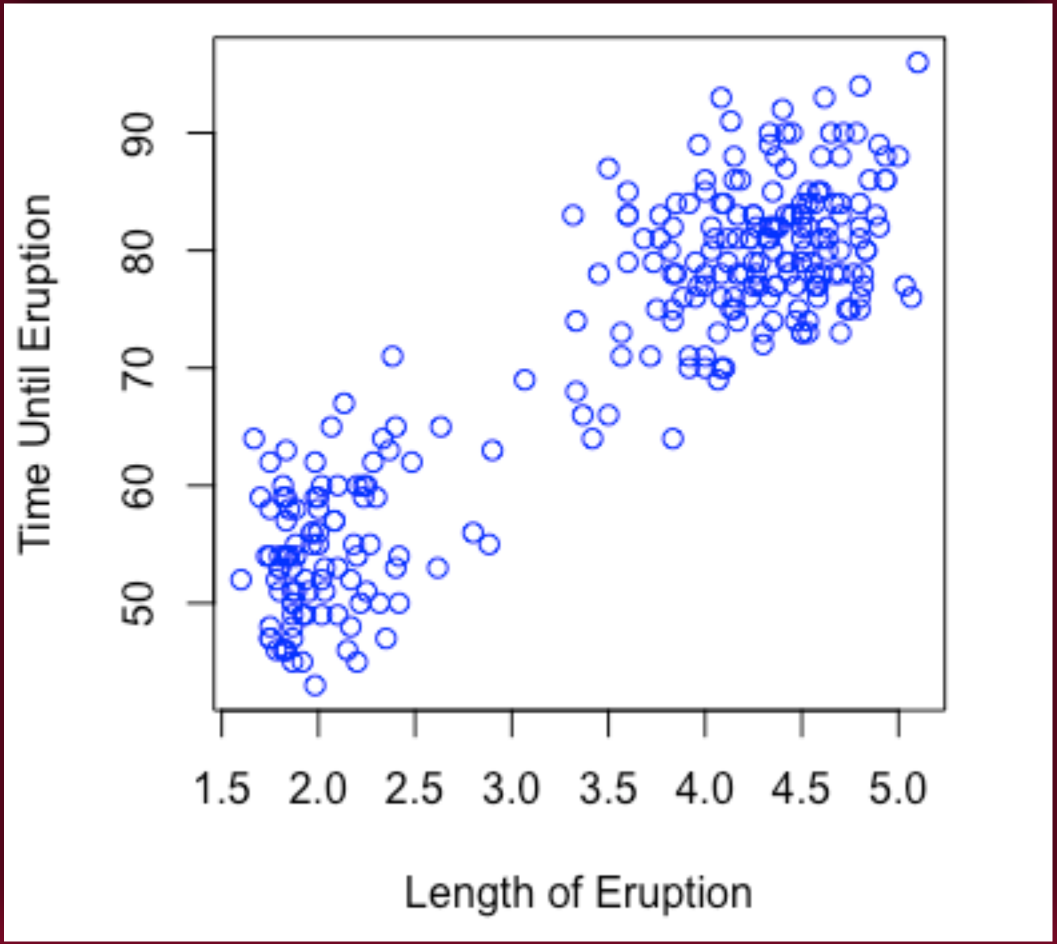
EM had been around for a while before it was formally ‘christened’…

Newcomb (1887)
McKendrick (1926)
Hartley (1958)
Baum et al. (1970)

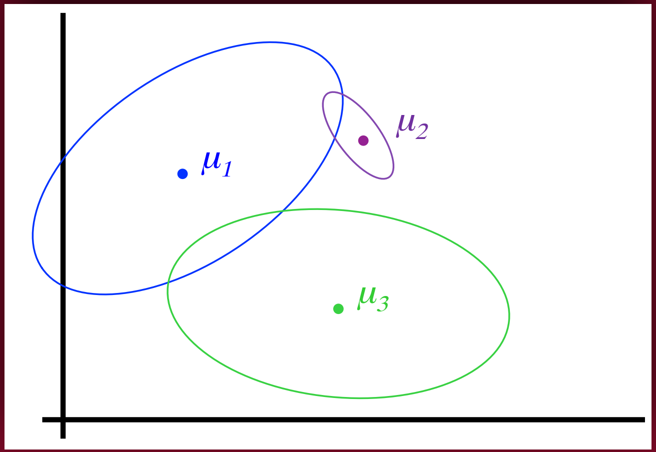
Canonical example

- Have a mixture distribution.
- Don’t know which points belong to which components of the mixture.
- Wish to estimate parameters of underlying mixture (or assign points probabilistically to the mixture components).

Old Faithful



Gaussian Mixture Model example



Underlying assumptions:

1. There are k components
2. Each point is generated by:
   - Sampling a component Y with prob. P(Y)
   - Sampling from $N(\mu_i, \sigma_i^2)$

Probability model

- So, in general, for each datapoint, x, we will have:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \underset{\substack{\uparrow \\ \text{Mixing coefficient}}}{\pi_k} \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

Maximum likelihood

- The EM algorithm, seeks to maximize the marginal likelihood:

$$\text{argmax}_\theta \prod_j P(x_j) = \text{argmax}_\theta \prod_j \sum_{k=1}^{K} P(Y_j=k, x_j)$$

where Y_j is the cluster to which x_j belongs.

E-M algorithm

- The algorithm alternates between two steps:
  - Expectation - compute expectations to fill in missing (unobserved) values conditional on the current parameters.
  - Maximization - re-estimate the parameters conditional on the current **probabilistic** assignments (using maximum likelihood).

EM algorithm - mixture models

- E step: Calculate $P(Y_j{=}k \mid x_j,\theta)$.
- M step:
  Set $\theta = \operatorname{argmax}_\theta \sum_j \sum_k P(Y_j{=}k \mid x_j, \theta) \log P(Y_j{=}k, x_j \mid \theta)$

Xj = the datapoints
Yj = cluster to which xj is assigned (probabilistic!)
θ = parameters (Normal means and variances here)
K = cluster index

EM algorithm - mixture models

- E-step:
  - Compute 'expected' clusters of all datapoints

$$P\left(Y_j = k \middle| x_j, \mu_1 \ldots \mu_K\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left\| x_j - \mu_k \right\|^2\right) \mathrm{P}\left(Y_j = k\right)$$

EM algorithm - mixture models

- M-step:
  - Compute most likely cluster means given current assignments:

$$\mu_k = \frac{\sum_{j=1}^{m} P\left(Y_j = k \middle| x_j\right) x_j}{\sum_{j=1}^{m} P\left(Y_j = k \middle| x_j\right)}$$

EM Algorithm

- See examples in EM_Algorithm repo from week 15.
- Note: If we did 'hard' rather than 'soft' assignments, the algorithm becomes equivalent to k-means clustering.
-

Final exam

- When: Friday May 13th at 10am.
- What: 5-7 minute presentation each. (Can join with 1 or 2 others if you prefer.)
- Where: Zoom (room details will be on Blackboard under ZoomPro tab).
- I will be strict with timing so that we are not there all day (with apologies).
- Each person will share their screen when presenting
- Please have fun with it!

END