LECTURE 27 - INTRODUCTION TO EXPECTATION MAXIMIZATION AND GAUSSIAN MIXTURE MODEL CLUSTERING

1. EM

- EM is a general algorithm that can be applied to a variety of problems (not just mixture model clustering).
- With MLE, we define a likelihood and maximize it to find parameters of interest.
- With MAP, we maximize the posterior to find parameters of interest.
- The goal of EM is to also find the parameters that maximize your likelihood function.
- **The 1st step** is to define your likelihood function (defines your objective)
- Originally introduced by Dempster, Laird, and Rubin in 1977 - "Maximum Likelihood from Incomplete Data via the EM Algorithm"
- EM is a method to simplify difficult maximum likelihood problems.
- Suppose we observe $\mathbf{x}_1, \ldots, \mathbf{x}_N$ i.i.d. from $g(\mathbf{x}_i|\Theta)$
- We want: $\hat{\Theta} = argmax L(\Theta|X) = \prod_{i=1}^{N} g(\mathbf{x}_i|\Theta)$
- But suppose this maximization if very difficult. EM simplifies it by expanding the problem to a bigger easier problem - "demarginalization"

$$g(x|Theta) = \int_z f(x, z|\Theta)dz \tag{1}$$

Main Idea: Do all of your analysis on f and then integrate over the unknown z's.

1.1. Censored Data Example.

- Suppose we observe $\mathbf{y}_1, \ldots, \mathbf{y}_N$ i.i.d. from $f(\mathbf{y}|\Theta)$
- Lets say that we know that values are censored at $\geq a$
- So, we see: $\mathbf{y}_1, \ldots, \mathbf{y}_m$ (less than a) and we do not see $\mathbf{y}_{m+1}, \ldots, \mathbf{y}_N$ which are censored and set to a.
- Given this censored data, suppose we want to estimate the mean if the data was uncensored.
- Our observed data likelihood in this case would be:

$$L = \prod_{i=1}^{m} [1 - F(a|\theta)]^{n-m} f(\mathbf{y}_i|\theta) \tag{2}$$

$$= \prod_{i=1}^{m} f(\mathbf{y}_i|\theta) \prod_{j=m+1}^{n} \int_a^{\infty} f(\mathbf{y}_i|\theta)dy_j \tag{3}$$

where $F(\cdot)$ is the cumulative distribution function and $f(y|\theta) = N(y|\theta)$, for example.

- So, the observed data likelihood would be very difficult to maximize to solve for θ
- In EM, we introduce *latent variables* (i.e., "hidden variables") to simplify the problem
- **The second step**: Define the *complete likelihood* by introducing variables that simplify the problem.
- Going back to the censored data example, if we had observed the missing data, the problem would be easy to solve! It would simplify to a standard MLE. For this example, the complete data likelihood is:

$$L^c = \prod_{i=1}^{m} f(y_i|\theta) \prod_{i=m+1}^{N} f(z_i|\theta) \tag{4}$$

where z_i are the latent, hidden variables.

- Note: you cannot just use a for the censored data, it would skew the results!
- The complete data likelihood would be much much simplier to optimize for θ if we had the zs...
- Consider the Gaussian Mixture Model example
  - $p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$
  - *How would you draw from this?*
  - $p(z_k = 1) = \pi_k$, $0 \leq \pi_k \leq 1$, $\sum_k \pi_k = 1$
  - $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$
  - $p(x|z_k = 1) = N(x|\mu_k, \Sigma_k)$
  - So, suppose we are given $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where each $\mathbf{x}_i$ is a sample from one of the K Gaussians in our mixture model. We want to estimate π_k, μ_k, Σ_k given X.
  - So, we want to maximize the following data likelihood:

$$\hat{\Theta} = argmax_{\Theta} \prod_{i=1}^{N} g(\mathbf{x}_i|\Theta) \tag{5}$$

  - *What is $g(\mathbf{x}_i|\Theta)$ for this problem?*
- It is difficult to maximize! We should try to simpler version in which we add latent variables to simplify the problem (and apply EM).
- *What latent variables can we add to simplify this problem?*
- In this example, a hidden/latent/missing varabile can be the label of the Gaussian from which $\mathbf{x}_i$ was drawn

$$\mathbf{x}, z \sim f(\mathbf{x}, z|\theta) \tag{6}$$

- *What would be the complete data likelihood once we add the latent variables?*

- In this example, the complete data likelihood is:

$$L^c = \prod_{i=1}^{N} p(\mathbf{x}_i|z_i, \theta)p(z_i) \tag{7}$$

$$= \prod_{i=1}^{N} N(\mathbf{x}_i|\mu_{z_i}, \theta_{z_i}) \tag{8}$$

- Again, the above problem is much simpler if we know the zs.
- So, we do not know the zs. We can learn them too and put them in the parameter set.
- *how could you find z_i?*
- You could take the average! (integrate over all possible values of z_i)

$$p(z_i = 1|\mathbf{x}_i) \ln p(\mathbf{x}_i|z_i = 1, \theta) + p(z_i = 2|\mathbf{x}_i) \ln p(\mathbf{x}_i|z_i = 2, \theta) + ... \tag{9}$$

- This is the *Expected Value* of z_i