10-701 Recitation 6

Expectation-Maximization

K-Gaussians Mixture Model

- Say you believe your data come from K Gaussians in a D-dimensional space

- In other words, each data point x_1, …, x_N is "generated" via the following process
  - Randomly pick one of the K Gaussians
  - Randomly draw a sample from the chosen Gaussian

K-Gaussians Mixture Model

- Let's introduce notation to describe this:
  - Let $z_1, \ldots, z_N$ be K-dimensional indicator vectors denoting the respective Gaussians chosen by $x_1, .., x_N$
    - An indicator vector has exactly one '1', and is '0' everywhere else
  - Let $\mu_1, \ldots, \mu_K$ be the D-dimensional means of each Gaussian
  - Let $\Sigma_1, \ldots, \Sigma_K$ be the D-by-D covariance matrices of each Gaussian
  - Let π be the K-dimensional multinomial prior probability of choosing each Gaussian

- The "generative process" described on the previous slide boils down to:
  - For i in 1…N:
    - Draw $z_i \mid \pi \sim \text{Multinomial}(\pi)$
    - Draw $x_i \mid z_i, \mu, \Sigma \sim \text{Normal}(\mu_{z_i}, \Sigma_{z_i})$

K-Gaussians Mixture Model

- The "generative process" described on the previous slide boils down to:
  - For i in 1…N:
    - Draw $z_i \mid \pi \sim \text{Multinomial}(\pi)$
    - Draw $x_i \mid z_i, \mu, \Sigma \sim \text{Normal}(\mu_{z_i}, \Sigma_{z_i})$

- This translates to the following probabilities:

$$p(z_i) = \prod_{k=1}^{K} (\pi_k)^{z_i^k}$$

$$p(x_i \mid z_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_{z_i}|^{1/2}} \exp\left\{ -\frac{1}{2} (x_i - \mu_{z_i})^\top (\Sigma_{z_i})^{-1} (x_i - \mu_{z_i}) \right\}$$

Unsupervised Learning

- We know $x_1, \ldots, x_N$ but not z, μ, Σ
- This is an unsupervised learning problem: we have data points x, but no labels z
- We want to learn which x_i should share the same label

So how do we learn z, π, μ, Σ?

- First attempt: Let's try Maximum Likelihood Estimation (MLE)
  - We used this in Linear Regression and Naïve Bayes

- So, we write down the complete log-likelihood function…

$$
\begin{aligned}
\ell_c(\pi, \mu, \Sigma; x, z) &= \log p(x, z \mid \pi, \mu, \Sigma) \\
&= \log \prod_{i=1}^{N} p(x_i, z_i \mid \pi, \mu, \Sigma) \\
&= \sum_{i=1}^{N} \log p(x_i \mid z_i, \mu, \Sigma)\, p(z_i \mid \pi) \quad \text{(by defn of conditional probability)}
\end{aligned}
$$

So how do we learn z, π, μ, Σ?

$$
\begin{aligned}
\ell_c(\pi, \mu, \Sigma; x, z) &= \log p(x, z \mid \pi, \mu, \Sigma) \\
&= \log \prod_{i=1}^{N} p(x_i, z_i \mid \pi, \mu, \Sigma) \\
&= \sum_{i=1}^{N} \log p(x_i \mid z_i, \mu, \Sigma)\, p(z_i \mid \pi) \quad \text{(by defn of conditional probability)}
\end{aligned}
$$

- And now, maximize it wrt π, μ, Σ!

- … Except that we're forgetting something here

We don't know z!

- We have 2 types of random variables, x and z
  - And we don't know z

- What should we do?

- Fortunately, we have marginal probability:

$$p(A) = \begin{cases} \int_b p(A, B = b)\, db & \text{if } B \text{ continuous} \\ \sum_b p(A, B = b) & \text{if } B \text{ discrete} \end{cases}$$

Defining an incomplete log-likelihood

- Second attempt: Use marginal probability to define an "incomplete" log-likelihood function
  - This function won't require knowledge of z:

$$
\begin{aligned}
\ell(\pi, \mu, \Sigma; x) &= \log p(x \mid \pi, \mu, \Sigma) \\
&= \log \prod_{i=1}^{N} p(x_i \mid \pi, \mu, \Sigma) \\
&= \log \prod_{i=1}^{N} \sum_{k=1}^{K} p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) \quad \text{(marginal probability)} \\
&= \sum_{i=1}^{N} \log \sum_{k=1}^{K} p\left(x_i \mid z_i^k = 1, \mu, \Sigma\right) p\left(z_i^k = 1 \mid \pi\right) \quad \text{(conditional probability)}
\end{aligned}
$$

Can we now learn z, π, μ, Σ?

$$
\begin{aligned}
\ell(\pi, \mu, \Sigma; x) &= \log p(x \mid \pi, \mu, \Sigma) \\
&= \log \prod_{i=1}^{N} p(x_i \mid \pi, \mu, \Sigma) \\
&= \log \prod_{i=1}^{N} \sum_{k=1}^{K} p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) \quad \text{(marginal probability)} \\
&= \sum_{i=1}^{N} \log \sum_{k=1}^{K} p\left(x_i \mid z_i^k = 1, \mu, \Sigma\right) p\left(z_i^k = 1 \mid \pi\right) \quad \text{(conditional probability)}
\end{aligned}
$$

- So here's the new plan: learn π, μ, Σ by maximizing the incomplete log-likelihood wrt them
  - Then, find Maximum a Posteriori (MAP) values for z_i
  - Use Bayes Rule to compute $p(z_i|x_i)$ for all possible choices of z_i, then pick the choice that gives the largest $p(z_i|x_i)$

- Alas, there's another problem…

Can you find the maximum?

$$
\begin{aligned}
\ell(\pi, \mu, \Sigma; x) &= \log p(x \mid \pi, \mu, \Sigma) \\
&= \log \prod_{i=1}^{N} p(x_i \mid \pi, \mu, \Sigma) \\
&= \log \prod_{i=1}^{N} \sum_{k=1}^{K} p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) \quad \text{(marginal probability)} \\
&= \sum_{i=1}^{N} \log \sum_{k=1}^{K} p\left(x_i \mid z_i^k = 1, \mu, \Sigma\right) p\left(z_i^k = 1 \mid \pi\right) \quad \text{(conditional probability)}
\end{aligned}
$$

- This function has no closed-form solution for its maximum wrt π, μ, Σ
  - You can't find the maximum by differentiating
- Side note: the function is not even concave in π, μ, Σ
  - Gradient ascent works, but you'll only get a local maximum

Enter the EM algorithm

- Let's try a somewhat different approach
  - We'll maximize a lower bound on the incomplete log-likelihood
- Why a lower bound?
  - At the lower bound's maximum, we know the "true" incomplete log-likelihood will be higher

Enter the EM algorithm

- Let's derive that lower bound to the incomplete log-likelihood:

$$
\begin{aligned}
\ell(\pi, \mu, \Sigma; x) &= \sum_{i=1}^{N} \log \sum_{k=1}^{K} p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) \\
&= \sum_{i=1}^{N} \log \sum_{k=1}^{K} q\left(z_i^k = 1\right) \frac{p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right)}{q\left(z_i^k = 1\right)} \quad \text{(this is true for any probability distribution } q()\text{)} \\
&\geq \sum_{i=1}^{N} \sum_{k=1}^{K} q\left(z_i^k = 1\right) \log \frac{p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right)}{q\left(z_i^k = 1\right)} \quad \text{(Jensen's inequality)} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} q\left(z_i^k = 1\right) \log p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) - \sum_{i=1}^{N} \sum_{k=1}^{K} q\left(z_i^k = 1\right) \log q\left(z_i^k = 1\right) \\
&= F(q, \pi, \mu, \Sigma) + H_q
\end{aligned}
$$

Enter the EM algorithm

$$
\begin{aligned}
\ell(\pi, \mu, \Sigma; x) &= \sum_{i=1}^{N} \log \sum_{k=1}^{K} p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) \\
&= \sum_{i=1}^{N} \log \sum_{k=1}^{K} q\left(z_i^k = 1\right) \frac{p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right)}{q\left(z_i^k = 1\right)} \quad \text{(this is true for any probability distribution } q()\text{)} \\
&\geq \sum_{i=1}^{N} \sum_{k=1}^{K} q\left(z_i^k = 1\right) \log \frac{p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right)}{q\left(z_i^k = 1\right)} \quad \text{(Jensen's inequality)} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} q\left(z_i^k = 1\right) \log p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) - \sum_{i=1}^{N} \sum_{k=1}^{K} q\left(z_i^k = 1\right) \log q\left(z_i^k = 1\right) \\
&= F(q, \pi, \mu, \Sigma) + H_q
\end{aligned}
$$

- What can we do with this lower bound F() + H_q?
- Notice that it's a function of 2 things: the distributions q(), and the model parameters π, μ, Σ
- We can't maximize it wrt q(), π, μ, Σ all at once
  - But we can maximize it alternatingly: first wrt to the q()'s, then wrt π, μ, Σ together. We repeat both steps until convergence.
  - This is different from gradient ascent; here we go directly to the maximum wrt some variable

Maximizing the lower bound wrt q()

- First, what are the $q(z_i)$ really?
  - Notice there's one $q(z_i)$ function for each indicator $z_1, \ldots, z_N$
- z_i's are indicator vectors with K possibilities
  - So $q(z_i)$ is really a discrete probability mass function, i.e. a multinomial
  - Let's call $q(z_i)$'s parameter ϕ_i, which is a K-dimensional vector
- Thus, we've defined $q(z_i)$ to be

$$q(z_i) = \prod_{k=1}^{K} \left(\phi_i^k\right)^{z_i^k}$$

Maximizing the lower bound wrt q()

$$q(z_i) = \prod_{k=1}^{K} \left(\phi_i^k\right)^{z_i^k}$$

- What did we get out of this definition?

- The maximization problem over functions $q(z_i)$ is really one over multinomial parameters ϕ_i!

Maximizing the lower bound wrt q()

- So we maximize the lower bound

$$
\begin{aligned}
F(q, \pi, \mu, \Sigma) - H_q &= \sum_{i=1}^{N} \sum_{k=1}^{K} q\left(z_i^k = 1\right) \log p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) - \sum_{i=1}^{N} \sum_{k=1}^{K} q\left(z_i^k = 1\right) \log q\left(z_i^k = 1\right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \phi_i^k \log p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) - \sum_{i=1}^{N} \sum_{k=1}^{K} \phi_i^k \log \phi_i^k
\end{aligned}
$$

wrt $\phi_1, \ldots, \phi_N$, while holding π, μ, Σ fixed

- Maximize by setting the differential to zero
  - It's similar to finding the MLE for a multinomial
  - Do it for one parameter $\phi_1, \ldots, \phi_N$ at a time

Maximizing the lower bound wrt q()

- I'm not going to derive the maximizer here
  - You should try it yourself

- I'll just tell you what the maximizer is:

$$
\begin{aligned}
q\left(z_i^k = 1\right) \equiv \left\langle z_i^k \right\rangle_q &\equiv \phi_i^k \qquad \text{(the 3 expressions on this line are equivalent)} \\
&= p\left(z_i^k = 1 \mid x_i, \pi, \mu, \Sigma\right) \\
&= \frac{p\left(x_i \mid z_i^k = 1, \mu, \Sigma\right) p\left(z_i^k = 1 \mid \pi\right)}{\sum_{j=1}^{K} p\left(x_i \mid z_i^j = 1, \mu, \Sigma\right) p\left(z_i^j = 1 \mid \pi\right)} \qquad \text{(Bayes Rule)} \\
&= \frac{\mathrm{N}\left(x_i \mid \mu_k, \Sigma_k\right) \pi_k}{\sum_{j=1}^{K} \mathrm{N}\left(x_i \mid \mu_j, \Sigma_j\right) \pi_j}
\end{aligned}
$$

where N() is defined as

$$
\mathrm{N}\left(x_i \mid \mu_k, \Sigma_k\right) = \frac{1}{(2\pi)^{D/2} \left|\Sigma_k\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(x_i - \mu_k\right)^{\top} \left(\Sigma_k\right)^{-1} \left(x_i - \mu_k\right)\right\}
$$

What about π, μ, Σ?

$$
\begin{aligned}
F(q, \pi, \mu, \Sigma) - H_q &= \sum_{i=1}^{N} \sum_{k=1}^{K} q\left(z_i^k = 1\right) \log p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) - \sum_{i=1}^{N} \sum_{k=1}^{K} q\left(z_i^k = 1\right) \log q\left(z_i^k = 1\right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \phi_i^k \log p\left(x_i, z_i^k = 1 \mid \pi, \mu, \Sigma\right) - \sum_{i=1}^{N} \sum_{k=1}^{K} \phi_i^k \log \phi_i^k
\end{aligned}
$$

- We previously maximized the lower bound wrt q()
- Recall we're doing alternating maximization
  - So we now hold q() fixed, and maximize wrt π, μ, Σ
  - This is similar to MLE parameter estimation when both x and z are observed
  - In particular, you can ignore H_q since it's not a function of π, μ, Σ

What about π, μ, Σ?

- The maximizers wrt π, μ, Σ are as follows:

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \left\langle z_i^k \right\rangle_q$$

$$\mu_k = \frac{\sum_{i=1}^{N} \left\langle z_i^k \right\rangle_q x_i}{\sum_{i=1}^{N} \left\langle z_i^k \right\rangle_q}$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} \left\langle z_i^k \right\rangle_q (x_i - \mu_k)(x_i - \mu_k)^\top}{\sum_{i=1}^{N} \left\langle z_i^k \right\rangle_q}$$

- You should try deriving these yourself!

3-point Summary of EM

- Apply Jensen's inequality to get a lower bound on the incomplete log-likelihood

- This lower bound is a function of two things:
  - Distributions $q(z_1)$, …, $q(z_N)$ over each hidden z_i
  - The model parameters π, μ, Σ

- EM alternates between:
  - Maximizing the lower bound wrt $q(z_1)$, .., $q(z_N)$
  - Maximizing the lower bound wrt π, μ, Σ