- Math stuff
- Gaussian Mixture Model (GMM)
- Expectation-Maximization Algo (EM)
- EM applied to GMM
- demo

Section notes on EM Algo + Gaussian mixtures

6/13/17

Douglas Rubin

Math stuff

① Indicator functions: $\mathbb{1}\{\text{condition}\} = \begin{cases} 1 & \text{if condition is t} \\ 0 & \text{if condition is fa} \end{cases}$

examples: $\mathbb{1}\{2 \neq 5\} = 1$

$\mathbb{1}\{-370\} = 0$

- convenient in summations and integrals bc sometimes they "kill" the sum/integral
- Also convenient for counting things, for example the # of misclassifications:

$$\text{misclassification rate} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{\hat{y}(x^{(i)}) \neq y^{(i)}\}$$

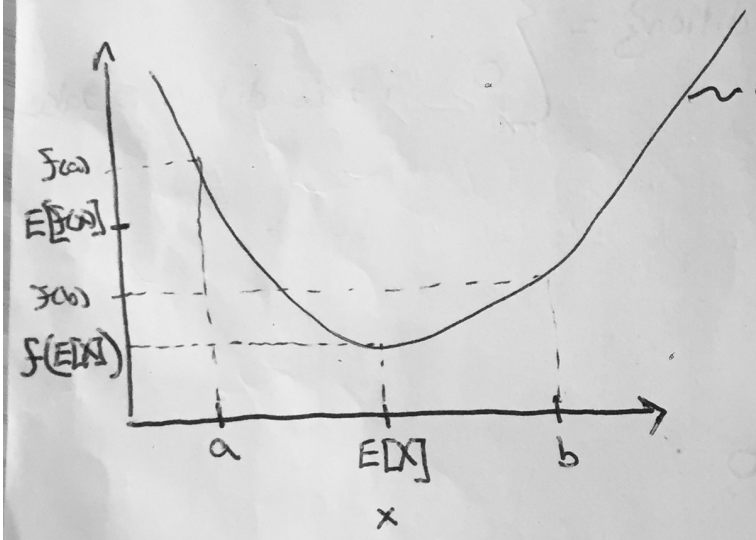
② Jensen's inequality

Let $f: \mathbb{R}^n \to \mathbb{R}$ and $X \in \mathbb{R}^n$ be a random variable. Then

$$E[f(X)] \geq f(E[X]) \quad (\text{if } f \text{ is convex})$$

$$E[f(X)] \leq f(E[X]) \quad (\text{if } f \text{ is concave})$$

Moreover, if f is strictly convex/concave $E[f(X)] = f(E[X])$ holds only if X is a constant.



convex function

let $P(a) = P(b) = .5$

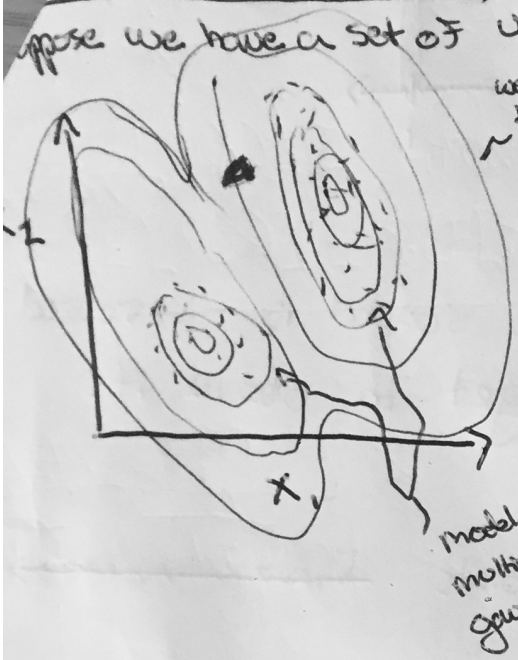
clearly $E[f(x)] \geq f(E[X])$

eg let P be a prob. dist., then

$$\sum_{j=1}^{N} P_j \log(X_j) \leq \log \sum_{j=1}^{N} P_j X_j \quad (\log \text{ is strictly concave})$$

ssian Mixtures

ppose we have a set of unlabeled data $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$; $x^{(i)} \in \mathbb{R}^n$



we are trying to learn this underlying density field

- GMM achieves 2 things
  1. estimates underlying density field (possibly good for outlier detection)
  2. clusters the data
     - does so in a "soft" way; i.e., assigns probs. of being associated to each cluster. may be useful in some scenarios.

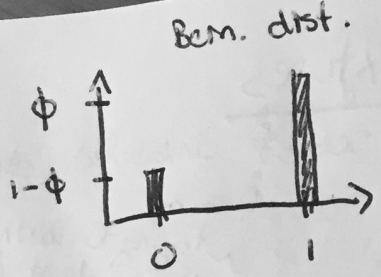
* note: we will work through the case of $K=2$ clusters, but the generalization to arbitrary K is very straight forward.

let's model this

1. Flip an unfair coin
2. if heads (class 1) draw a data point from Gaussian 1
   " tails (class 0) " " " " 0

→

in math:



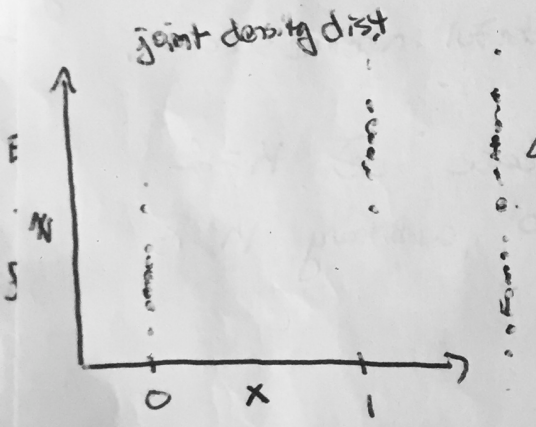
$$z^{(i)} \sim \text{Bern}(\phi)$$

$$x^{(i)} \mid z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

$$\mu_j \in \mathbb{R}^n, \quad \Sigma_j \in \mathbb{R} S_+^n$$

called a "latent" r.v. that is a r.v. we suspect is involved in the process, but we don't actually get to observe it.

Consider a 1D observed set of data points



Let $\Theta = \{\phi, \mu_0, \mu_1, \Sigma_0, \Sigma_1\}$ (For convenience)

Lets write down all the prob. dists:

$$P(z^{(i)}; \Theta) = \phi^{z^{(i)}} (1-\phi)^{1-z^{(i)}}$$

$$P(x^{(i)} \mid z^{(i)} = j) = \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma^{-1} (x^{(i)} - \mu_j)\right\}$$

$$P(x^{(i)}, z^{(i)}; \Theta) = P(x^{(i)} \mid z^{(i)}; \Theta) P(z^{(i)}; \Theta) \quad \text{(By defn. of cond. prob)}$$

$$P(z^{(i)} \mid x^{(i)}; \Theta) = \frac{P(x^{(i)} \mid z^{(i)}; \Theta) P(z^{(i)}; \Theta)}{\sum_{j=1}^{k} P(x^{(i)} \mid z^{(i)} = j) P(z^{(i)} = j; \Theta)} \quad \text{(Bayes rule)}$$

solve for all the parameters Θ in the usual way:

MML

$$\ell(\Theta) = \log \prod_{i=1}^{m} p(x^{(i)}; \Theta)$$

$$= \sum_{i=1}^{m} \log p(x^{(i)}; \Theta)$$ ← we need to introduce our model here, so we realize that this is the joint dist. marginalized over the zs.

$$= \sum_{i=1}^{m} \log \sum_{z^{(i)}=0}^{k} p(x^{(i)}, z^{(i)}; \Theta)$$

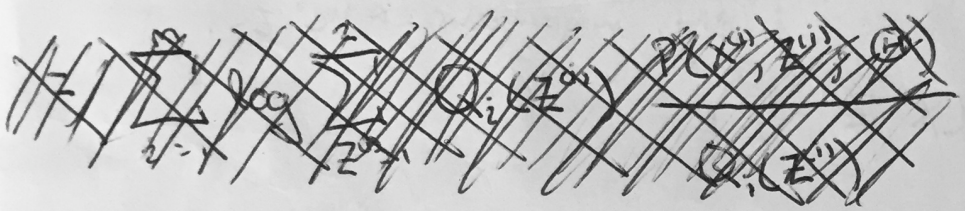
Problem - plugging in the dists. and trying to maximize $\ell(\Theta)$ by taking derivatives wrt Θ, we find that this is analytically intractable, mostly due to summation in the log. We need to try to push the summation out of the log (let's try Jensen's ineq.)

→

EM algo

- in a somewhat unmotivated fashion I will introduce m different prob. dists. $Q_1(Z^{(i)}), Q_2(Z^{(i)}), \ldots, Q_m(Z^{(i)})$

$\left(\sum_{Z^{(i)}=0}^{Q_i} Q_i(Z^{(i)}) = 1, \quad Q_i(Z^{(i)}) \geq 0\right)$, where I have yet to specify the functional form of the Q_i's! → note it is important that each term have its own Q_i, as you will see later



$$= \log\left\{ \sum_{Z^{(i)}=0}^{Q_1} Q_1(Z^{(i)}) \frac{P(X^{(i)}, Z^{(i)}, \Theta)}{Q_1(Z^{(i)})} \right\} + \ldots + \log\left\{ \sum_{Z^{(i)}=0}^{Q_1} Q_m(Z^{(i)}) \frac{P(X^{(m)}, Z^{(i)}, \Theta)}{Q_m(Z^{(i)})} \right\}$$

$$\geq \sum_{Z^{(i)}=1}^{2} Q_1(Z^{(i)}) \log \frac{P(X^{(i)}, Z^{(i)})}{Q_1(Z^{(i)})} \qquad \geq \sum_{Z^{(i)}=1}^{2} Q_m(Z^{(i)}) \log \frac{P(X^{(m)}, Z^{(i)}, \Theta)}{Q_m(Z^{(i)})}$$

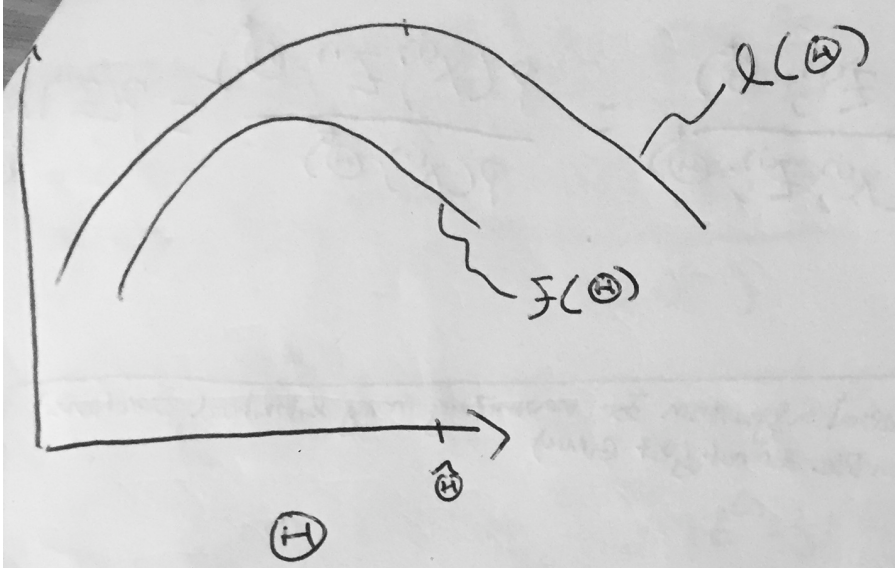
From Jensen's ineq.

$$\Rightarrow \geq \sum_{Z^{(i)}=0}^{1} Q_1(Z^{(i)}) \log \frac{P(X^{(i)}, Z^{(i)})}{Q_1(Z^{(i)})} + \ldots + \sum_{Z^{(i)}=0}^{1} Q_m(Z^{(i)}) \log \frac{P(X^{(i)}, Z^{(i)})}{Q_m(Z^{(i)})}$$

$$= \sum_{i=1}^{m} \sum_{Z^{(i)}=0}^{1} Q_i(Z^{(i)}) \log \frac{P(X^{(i)}, Z^{(i)}; \Theta)}{Q_i(Z^{(i)})} \Bigg\} \equiv f(\Theta)$$

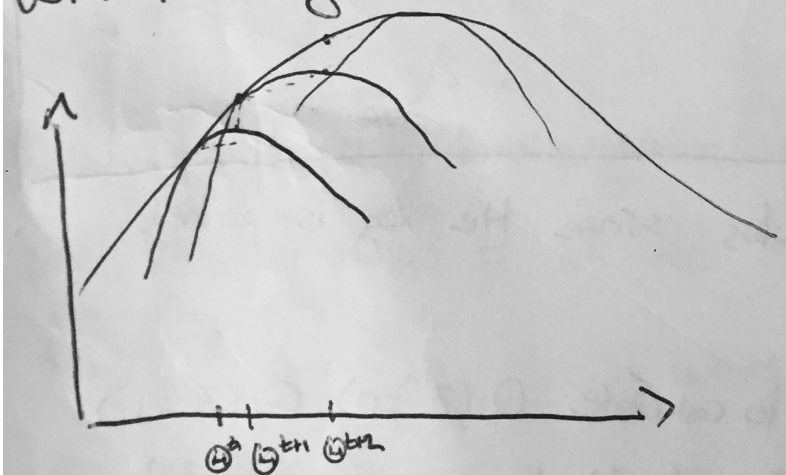
(notice that we have been able to push the log inside the sum, so we can maximize Θ, but at the expense of an inequality sign)

...rally (if Θ is 1D)



- goal is to find $\hat{\Theta}$

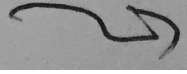
we still need to specify all the Q_i's. To get tightest lower bound on $\ell(\Theta)$ lets construct the Q_i's to hold w/ equality at a current "best" guess of Θ, Θ^t. This (we are allowed to do this w/ Jensen's inequality) will naturally lead to a method to find $\hat{\Theta}$:



- construct $Q_i(Z^{(i)}; \Theta^t)$
- maximize F_Θ over Θ
- construct $Q_i(Z^{(i)}; \Theta^{t+1})$

Constructing the Q_i's:

- Remember Jensen's inequality only holds with equality if the RV is a constant



$$\frac{P(X^{(i)}, Z^{(i)}; \theta^*)}{Q_i(Z^{(i)})} = C$$

$$\Rightarrow Q_i(Z^{(i)}) = \frac{P(X^{(i)}, Z^{(i)}; \theta^t)}{\sum_{Z^{(i)}} P(X^{(i)}, Z^{(i)}; \theta^t)} = \frac{P(X^{(i)}, Z^{(i)}; \theta^t)}{P(X^{(i)}; \theta^t)} = P(Z^{(i)} | X^{(i)}; \theta^t)$$

EM Algo

(general algorithm for maximizing many likelihood functions) (can use for not just GMM)

① initialize θ^*

② repeat until convergence {

(E-step) ②a $Q_i(Z^{(i)}) := P(Z^{(i)} | X^{(i)}; \theta^*)$

(M-step) ②b $\theta^* := \arg\max_{\theta} \sum_i \sum_{Z^{(i)}} Q_i(Z^{(i)}) \log \frac{P(X^{(i)}, Z^{(i)}; \theta)}{Q_i(Z^{(i)})}$

}

- Note 2b is now tractable w/ calculus since the log is pushed inside the sum

- Note that for GMM, we need to calculate $Q_i(Z^{(i)}=0)$, $Q_i(Z^{(i)}=1)$ for each data point, then plug this into 2b to sum over the $Z^{(i)}$'s.

EM applied to GMM

- need to find $\phi^*, \mu_1^*, \mu_2^*, \Sigma_1^*, \Sigma_2^*$ by maximizing

$$\sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$ w.r.t. these variables.

$$= \sum_{i=1}^{m} \sum_{j=1}^{0} Q_i(z^{(i)}=j) \log p(x^{(i)} | z^{(i)}=j; \theta)\, p(z^{(i)}=j) - Q_i(z^{(i)}=j) \log Q_i(z^{(i)}=j)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{0} Q_i(z^{(i)}=j) \log \left\{ \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp\left\{ \frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma^{-1} (x^{(i)}-\mu_j) \right\} \times \phi^{1\{z^{(i)}=1\}} (1-\phi)^{1\{z^{(i)}=0\}} \right\} - Q_i \log Q_i(z^{(i)}=j)$$

take the derivatives of this expression w/ $\frac{\partial}{\partial \phi}$, ∇_{μ_ℓ}, ∇_{Σ_ℓ}, and set the expressions to 0, and solve to find:

$$\phi_* = \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}$$

where $w_j^{(i)} \equiv Q_i(z^{(i)}=j)$ ← evaluated at old values of θ

$$\mu_j^* = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$$

$$\Sigma_j^* = \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

- notice that if we replace $w_j^{(i)}$ w/ $\mathbb{1}\{z^{(i)}=j\}$ we get the sample ϕ, sample μ_js and sample Σ_js.

- thus the MML parameters closely resemble the sample estimates but w/ the hard (0,1) $\mathbb{1}$ indicator function replaced w/ a soft $w_j^{(i)} \in [0,1]$ probability w/ $w_j^{(i)}$ being close to 1 if it is more likely to be in cluster 1, and closer to 0 if it is less likely to be in cluster 1.