COL774

Machine Learning

Oct 26, 2021

Last class:-

K-Means

$\{x^{(i)}\}_{i=1}^{m}$

↳ $\{c^{(i)}\}_{i=1}^{m}$ given no. of clusters

↳ $\{\mu_1 \dots \mu_K\}$ cluster centroids

given $\{c^{(i)}\}_{i=1}^{m}$

GMM

$P(x^{(i)};\theta)$

$P(x^{(i)}, z^{(i)};\theta)$ ↳ Sudoku for labels

↳ E-step ↳ $Q_i(z^{(i)}) = P(z^{(i)} | x^{(i)};\theta)$

↳ $\theta = \arg\max_{\theta} P(x^{(i)}, z^{(i)};\theta)$

M-step given $z^{(i)}$'s — filled in missing values

Do {

E-step: fill in missing values (give params)

M-step: estimate the parameters / given filled in values

} while (not converged)

↳ A more general framework for dealing with the problem of missing data

$\{x^{(i)}\}_{i=1}^{m}$

$P(x^{(i)}, z^{(i)};\theta)$:- underlying distribution

↳ Sudoku

⇒ our optimal set of θ parameters which describe the observed data

in some parametric form

$$\theta^* = \arg\max_{\theta} LL(\theta)$$

$$\hookrightarrow \log\left[\prod_{i=1}^{m} P(x^{(i)};\theta)\right]$$

$$= \sum_{i=1}^{m} \log P(x^{(i)};\theta)$$

$$LL(\theta) = \sum_{i=1}^{m} \log \sum_{z^{(i)}} \left[P(x^{(i)}, z^{(i)};\theta)\right]$$

e how do we optimize over θ:-

⇒ Directly gradient ascent

$$\arg\max_{\theta} LL(\theta)$$

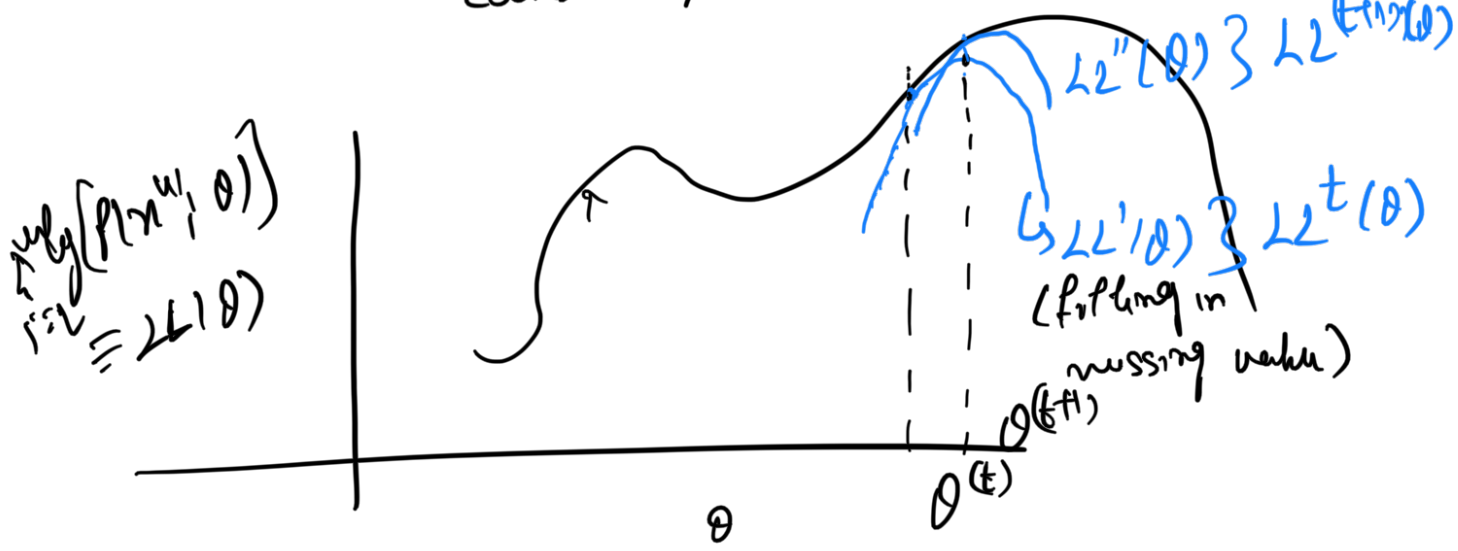
To $LL(\theta)$:- perform gradient ascent

Computing this very intractable

Assumption- it is easy to optimize

$$\log\left[\prod_{i=1}^{m} \left[P(x^{(i)}, z^{(i)};\theta)\right]\right]$$

⇒ if we knew $z^{(i)}$'s then we could optimize the Likelihood efficiently



① $LL(\theta) \geq LL'(\theta) \quad \forall \theta$

② $LL(\theta)|_{\theta^{(t)}} = LL'(\theta^{(t)})$

⇒ (A) E-step:- estimate $LL(\theta)$ (filling in of missing values)

(B) M-step:- $\arg\max_{\theta} LL'(\theta)$

⇒ convergence $\|\theta^{(t+1)} - \theta^{(t)}\| \leq \epsilon_0$ ✓

At convergence:- local optimal of $LL(\theta)$

$t = 0$

$\theta^{(t)} \leftarrow$ init();

do{

E-step $LL^t(\theta) \leftarrow$ construct lower bound ($LL(\theta)$);

M-step $\theta^{(t+1)} \leftarrow \arg\max_{\theta} LL^t(\theta)$; }

$t \leftarrow t+1$;

} (while ! converged)

E-M Expectation Maximization to deal with missing data

⇒ In contrast, with directly optimizing $LL(\theta)$ using gradient ascent (may not be tractable)

Next:-

① How to construct the "lower" bound

② "Prove" convergence

③ EM algorithm can be seen as:- an instance of Block coordinate ascent over a suitably defined

objective function

$$LL(\theta) = \sum_{i=1}^{m} \log p(x^{(i)}; \theta)$$

$\Rightarrow$ $\hat{LL}(\theta, \theta_1)$ performing Block coordinate descent

$$\arg\max_{\theta} \arg\max_{\theta_1} [\hat{LL}(\theta, \theta_1)] \quad \text{over} \quad \hat{LL}(\theta, \theta_1)$$

max max $\hat{LL}(\theta, \theta_1)$
θ θ_1 do{

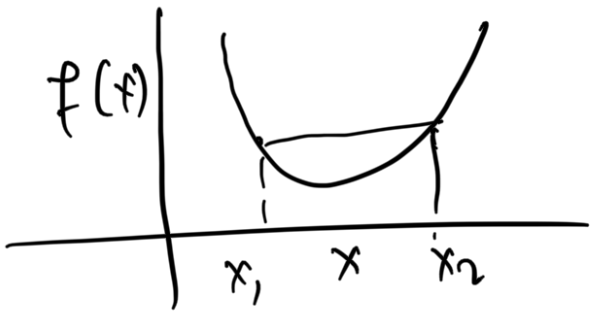
↳ optimize over θ_1 given θ (E-step)

↳ optimize over θ given θ_1 (M-step)

} (while ! converged)

Jensen's Inequality:-

$f: R^n \rightarrow R$ be convex



$f(\alpha x_1 + (1-\alpha)x_2)$ ✓

$\leq \alpha f(x_1) + (1-\alpha) f(x_2)$ — (1)

$0 \leq \alpha \leq 1$

$\Rightarrow$ $f(E[X]) \leq E[f(X)]$

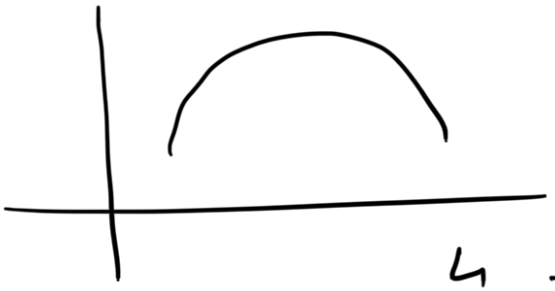
$P(X)$:- over X

computed using

Think about: how to prove Jensen's inequality

Hint:- Generalize Eq (1) to a convex combination of K point (K>2)

For concave functions :- $f: \mathbb{R}^n \to \mathbb{R}$ concave



$$f(\alpha x_1 + (1-\alpha) x_2) \geq \alpha f(x_1) + (1-\alpha) f(x_2)$$

$$\hookrightarrow f(E[X]) \geq E[f(X)]$$

$\Rightarrow$ If $f: \mathbb{R}^n \to \mathbb{R}$ is strictly convex (concave)

(strictly) Convex!:-

$$f(E[X]) < E[f(X)]$$

unless X is a constant (given $p(x)$)

$$f(E[C]) = E[f(C)] \quad (C)$$

$$\downarrow f(C)$$

(strictly) Concave!:-

$$f(E[X]) > E[f(X)]$$

unless X is a constant (given $p(x)$)

$$f(E[C]) = E f(C))$$

$$\downarrow f(C)$$

Coming back:-

$$LL(\theta) = \log \prod_{i=1}^{N} [p(x^{(i)}; \theta)] = \sum_{i=1}^{N} \log p(x^{(i)}; \theta)$$

$\log[P(x^{(i)};\theta)] = \log\left[\sum_{z^{(i)}} P(x^{(i)}, z^{(i)};\theta)\right]$ some explan...

$\Rightarrow$

$$LL(\theta) = \sum_{i=1}^{m} \log\left[\sum_{z^{(i)}} [P(x^{(i)}, z^{(i)};\theta)]\right]$$

↳ strictly concave

Let $Q_i(z^{(i)})$: be some distribution over $z^{(i)}$. (non-zero)

$$= \sum_{i=1}^{m} \log \sum_{z^{(i)}} \left[\frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \cdot Q_i(z^{(i)})\right]$$

$$= \sum_{i=1}^{m} \log \sum_{z^{(i)}} \left[\frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right] Q_i(z^{(i)})$$

$$= \sum_{i=1}^{m} \log E_{Q_i(z^{(i)})}\left[\frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]$$

log concave

$$\geq \sum_{i=1}^{m} E_{Q_i(z^{(i)})} \log \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

↳ Jensen's inequality

$$= \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log\left[\frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]$$

$= LL'(\theta)$ tractable to maximize

$$LL(\theta) \geq LL'(\theta)$$