PGM - Week 2

Latent Variable Models and EM

the clustering problem

Consider a dataset $\{x_1, \dots, x_N\} \subseteq \mathbb{R}^d$. Our task is to partition this set into K subsets (K known for now). Intuitively, each subset (from now on referred to a cluster) of points should share some common patterns.

A natural solution for this is to define K prototypes $\{\mu_1, \mu_2, \dots, \mu_K\} \subseteq \mathbb{R}^d$ and determine the assignment of each datapoint x_n to each prototype μ_k according to a given criterion.

To solve this optimisation prob. we can define a set of binary variables r_{nk}, where $r_{nk} = 1$ iff x_n assigned to μ_k, and $r_{nk} = 0$ if not

Then, the objective can be

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

Goal: find r_{nk}, μ_k optimizing J. This is

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0 & \text{if } n \neq l \end{cases}$$

$$\mu_k = \frac{\sum_i r_{nk} x_n}{\sum r_{nk}}$$

$\Longrightarrow$ We found K-means: the rules above are applied iteratively - do they ensure convergence?

- Discuss: discrete / local minima / # of steps.

Drawbacks

- slow (to compute r_{nk} costs $O(nk)$)
- other dissimilarity.
- hard assignments

Uses: - decision making
- compression
-

Gaussian mixture model

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k)$$

We also introduce $z \in \{0,1\}^K$, $\sum_k [z]_k = 1$ as well as

define

$$p(x, z) = p(x|z)\, p(z)$$

and also consider

$$p(z_k = 1) = \pi_k \qquad (0 \le \pi_k \le 1, \ \Sigma \pi_k = 1)$$

the marginal pmf is

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k},$$

likewise

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(\mu_k, \Sigma_k)^{z_k}$$

$$p(x) = \sum_z p(z)\, p(x|z) = \sum \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

★) We have found an equivalent formulation of GMM with explicit latent variable

Another important quantity is $p(z|x)$

$$\gamma(z_k) \equiv p(z_k=1|x) = \frac{p(x|z_k=1)\, p(z_k=1)}{\sum_j p(x|z_j=1)\, p(z_j=1)}$$

$$= \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_j \pi_j N(x|\mu_j, \Sigma_j)}$$

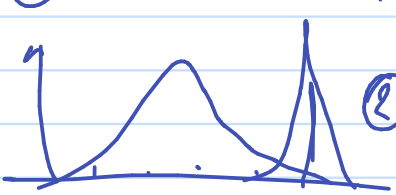
we will refer to this as the 'responsibility' of component to explain the obs x.

- Ancestral sampling to generate $x \sim GMM$

The loglike of GMM is $\theta = \{\pi, \mu, \Sigma\}$

$$\ell(\theta) = \log p(X|\mu, \pi, \Sigma) \qquad (*)$$

$$= \sum_n \log \left(\sum_k \pi_k N(x_n|\mu_k, \Sigma_k) \right)$$

① 

② Symmetries → ill posed

log applies to Σ, not to log → stand.

EM for GMMs. (we'll see this for the particular case of GMM, then in gral)

Set the first order opt cond for ⊛

$$\mu_k = \frac{1}{N_k} \sum_n \underbrace{\gamma(z_{nk})}_{\text{"}} x_n \quad , \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\Sigma_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T$$

$\pi_k = N_k / N$ – for this Lagrange mult

Exercise: derive these

Remark: this is not a closed-form soln. since since we don't have $\gamma(z_{nk})$. Workaround: 2 steps

E: Compute $\gamma(z_{nk}) = p(z_{nk}=1 \mid x)$

M: Use $\gamma(z_{nk})$ to apply μ, Σ, π (ML)

To implement: initialise with K-means.

Alternative View of EM

Let us now leave the GMM aside. In more general, perhaps abstract, terms, the goal of EM is to find ML solns for models involving latent variables.

Recall our notation X, Z

$$\ln p(X|\theta) = \ln \int p(X|\theta, Z)\, p(Z|\theta)\, d\theta$$

When can we optimize this? or even before that. in which cases can this be calculated? In very limited cases. Mix tures mix well with log, even in the exp. family case this is not. Mention Linear Gaussian

Let us consider the hypothetical scenario where we know the values of Z_i (along with X)

Def: $\{X, Z\}$ will be the complete dataset, while $\{X\}$ will be the observed dataset. (incomplete)

The related complete-data log likelihood would be

$$\log(p(X, Z|\theta)) = \log \prod p(x_i, z_i|\theta)$$
$$= \log \prod p(x_i|z_i, \theta)\, p(z_i)$$
$$= \log \prod \pi_{L(z)}.$$

: easy to work

This is impractical as Z unknown. However one interesting interpretation of this is to see the likelihood as a random var, as it depends on Z which is random.

A good approx of this would be to take it's expectation and optimise that. (E)

Then with an explicit expression, we maximise (M). So $\theta^{old} + E + M = \theta^{new}$ for θ.

E step: Use θ^{old} to compute $p(Z|X,\theta^{old})$,
compute expectation of the complete-data Likelihood

$$Q(\theta,\theta^{old}) = \sum_{Z} \ln p(X,Z|\theta) \cdot p(Z|X,\theta^{old})$$

M step: $$\theta^{new} = \arg\max_{\theta} Q(\theta,\theta^{old})$$

This can also be use for MAP (instead of ML)
in which case $Q(\theta,\theta^{old}) + \log p(\theta)$ ← prior.

Back to GMM

Complete-data likelihood

$$p(X,Z|\mu,\Sigma,\pi) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} p(x_n|\mu_k,\Sigma_k)^{z_{nk}}$$

$$\rightarrow \ln(\theta) = \sum_{n}\sum_{k} z_{nk}\left(\log \pi_k + \log N(x_n|\mu_k,\Sigma_k)\right)$$

$\longrightarrow$ Tractable

This is much easier: only one term in the K-sum is non-zero → μ, Σ one-simple estimates.

further more $\pi_k = \frac{1}{N} \sum_n z_{nk}$ (portion of data assigned to that cluster)

Now let's see 2:

$$p(z|X,\theta) \propto p(X,z|\theta) = \prod_k \prod_n \underbrace{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)^{z_{nk}}}_{\propto p(z_{nk}|x_n,\theta)^{z_{nk}}}$$

→ it factorizes over n → z_n all indep

so it looks like

$$p(z|X,\theta) = \prod_k \prod_n p(z_{nk}|x_n,\theta)$$

Now the expectation (binary var)

$$\mathbb{E}[z_{nk}] = 1 \cdot p(z_{nk}=1|X,\theta) + 0 \cdot p(z_{nk}=0|X,\theta)$$

$$= p(z_{nk}=1|X,\theta)$$

$$= \gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)}$$

(recall)

Finally

$$\to \mathbb{E}_z \log p(X,z|\theta) = \sum_k \sum_n \gamma(z_{nk}) \left[\log \pi_k + \log \mathcal{N}(x_n|\mu_k,\Sigma_k)\right]$$

Intuition says that EM recovers K-means

Exercise: Show that K-means is recovered

1) choosing $p(x|\mu_k, \Sigma_k) = N(x|\mu_k, I\epsilon)$ → $\epsilon > 0$ fixed.

to show a soft-assignment version of K-means

Take $\epsilon \to 0$ and see what happens.

Recommendation: See the example for EM for Bayesian linear reg and mixture of Bernoulli for completion..

EM in general

$$p(X|\theta) = \sum_z p(X, Z|\theta)$$

difficult ↑ → easy

Introduce a distribution $q(Z)$ over Z. For any choice of $q(\cdot)$ the following holds

$$\log p(X|\theta) = \sum q(Z) \cdot \log p(X|\theta)$$

→

$$\log p(X|\theta) = \sum_k q(z_k) \log p(X|\theta)$$

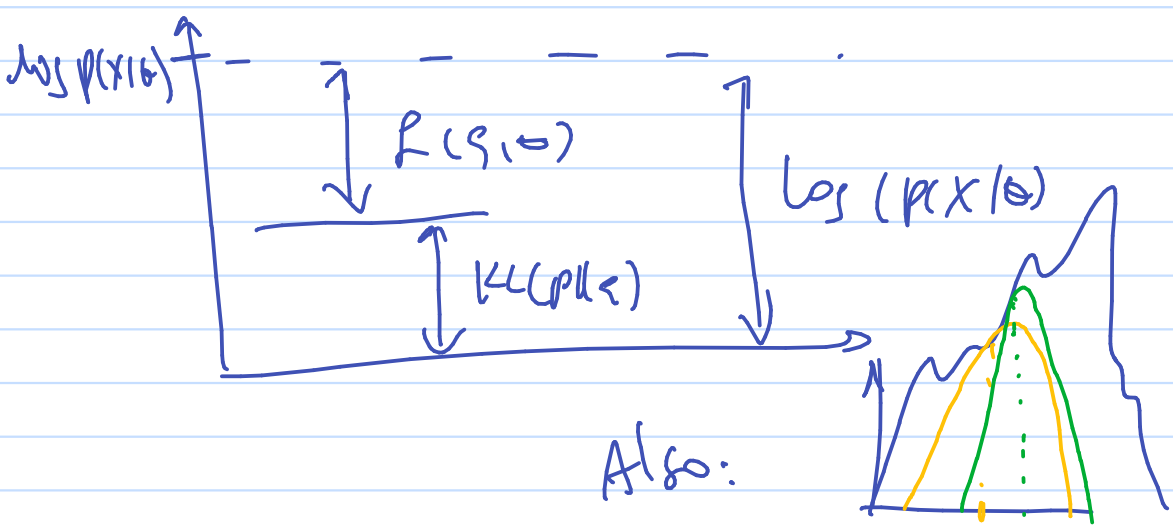
$$= \sum_k q(z_k) \log \left(\frac{\overbrace{p(X|\theta)\, p(z|X,\theta)}^{p(X,z|\theta)} \cdot q(z)}{p(z|X,\theta) \cdot q(z)} \right)$$

$$= \sum_k q(z_k) \log \left(\frac{p(X,z|\theta)}{q(z)} \cdot \left(\frac{p(z|X,\theta)}{q(z)} \right) \right)$$

$$= \sum_k q(z_k) \log \left(\frac{p(X,z|\theta)}{q(z)} \right)$$

$$- \sum_k q(z_k) \log \left(\frac{q(z)}{p(z|X,\theta)} \right)$$

$$= \mathcal{L}(q,\theta) + KL(p\|q)$$



Also:

EM Breaks the prob of ML into two simpler problems

- compute an approx to $\log p(X|\theta)$
- maximize this approximation.

in this problem, this actually achieves ML (locally)

Next: $q(z) = p(z|X,\theta)$ is unattainable, so we'll try to get closer.