Generative models

I For classification

II For Clustering

III Approximate learning with ELBO

Intro : $\mathcal{Y} = \{1, \dots, K\}$

$\mathcal{X} = \mathbb{R}^d$ observed data

Discriminative : $\mathbb{P}(Y = k \mid X = x)$

→ Logistic regression $\Theta = \{W, b\}$, $W \in \mathbb{R}^{d \times k}$, $b \in \mathbb{R}^k$

$$\mathbb{P}(k \mid x) = \frac{e^{w_{:k}^T x + b}}{\sum_{\ell=1}^{k} e^{w_{:\ell}^T x + b}}$$

Direct

Generative : Model x $\neq$ for each k

$$\mathbb{P}_k(x) \rightarrow \mathbb{P}(X = x, Y = k)$$

$= \mathbb{P}(X = x \mid Y = k)$ assumed to be normal $\underline{x \sim \mathcal{N}(\mu_k, \Sigma)}$

$$= \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k)}$$

(LDA) Linear Discriminant Analysis

Posterior

Classification: $\mathbb{P}(Y = k \mid X = x) = \frac{(\mathbb{P}(X = x \mid Y = k)) \times (\mathbb{P}(Y = k))}{\mathbb{P}(X = x)}$ ① ② ③

① $= \mathbb{P}_k(x)$ density function

② Prior $= \pi_k = \mathbb{P}(Y = k)$

③ Evidence

$$\mathbb{P}(X=x) = \sum_{\ell=1}^{K} \mathbb{P}(y=\ell) * \mathbb{P}(X=x \mid y=\ell)$$

LDA assumes $x \sim \mathcal{N}(\mu_k, \Sigma)$

$$\Theta = \{ \underset{\downarrow \in \mathbb{R}^d}{\mu_1}, \dots, \mu_k, \underset{\downarrow \in \mathbb{R}^{d\times d}}{\Sigma}, \pi_1, \dots, \underset{\downarrow \in \mathbb{R}}{\pi_k} \}$$

Largest $\mathbb{P}_\Theta(Y=k \mid X=x)$

$$= \underset{k}{\operatorname{argmax}} \left(\log \pi_k * \mathbb{P}_k(x) \right)$$

$$= \underline{\quad\quad} \underbrace{\left[\log(\pi_k) + x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k \right]}_{\text{Discriminant score } \delta_k(x)}$$

$k_1, k_2 \quad \delta_{k_1}(x) = \delta_{k_2}(x) \quad (\_ - \_)x + (\_) = 0$

Learning Θ ? $= \frac{1}{n} \sum_{i=1}^{n} \log \mathbb{P}_\Theta(x^{(i)}, y^{(i)}) = \mathcal{L}(\Theta)$

MLE $\quad \underset{\Theta}{\operatorname{Argmax}} \ \mathcal{L}(\Theta)$

$$\mathbb{P}_\Theta(x,y) = \mathbb{P}_\pi(y) * \mathbb{P}_{\mu,\Sigma}(x \mid y)$$

Systematic derivation

$$\frac{\partial}{\partial \pi_\ell}\left(\mathbb{P}_\Theta(x,y)\right) = 0$$

$$\frac{\partial}{\partial \mu_\ell}\left(\mathbb{P}_\Theta(x,y)\right) = 0$$

$\dots \mathbb{P}_\Theta(x^{(i)}, y^{(i)})$

$y^{(i)} = k \quad \mathbb{P}_\pi(y^{(i)} = k) = \pi_k$

Frequency : $\pi_k = \left(\sum_{\substack{i=1 \\ y^{(i)}=k}}^{n} 1 \right) / n \quad \left(\sum 1 \right) = n_k$

$$\mu_k = \frac{1}{n_k} \sum_{\substack{i=1 \\ y^{(i)}=k}}^{n} x^{(i)} \qquad \Sigma = \frac{1}{n} \sum_{k=1}^{K} \sum_{\substack{i=1 \\ y^{(i)}=k}}^{n} (x^{(i)}-\mu_k)(x^{(i)}-\mu_k)^T$$

QDA : Quadratic DA

for k, $x \sim \mathcal{N}(\mu_k, \Sigma_k)$

$$\Sigma_k = \frac{1}{n_k} \sum_{\substack{i=1 \\ y^{(i)}=k}}^{n} (x^{(i)}-\mu_k)(x^{(i)}-\mu_k)^T$$

$$\delta_k(x) = \log(\pi_k) + x^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k$$

new term $\left(-\frac{1}{2} x^T \Sigma_k^{-1} x \right.$

$$\delta_{k_1}(x) = \delta_{k_2}(x) \longrightarrow \_x^2 + \_x + \_ = 0$$

LDA versus Logistic regression ?

QDA = LDA + interaction between coordinates of x

LDA : Assumption of normality

- ✓ more efficient
- ✗ LR does MLE $\pi(Y=k \mid x)$
  ↳ it's more efficient

Generative versus Discriminative

- Include prior knowledge
- Density function → Sample
  → Detect anomalies → missing values

II Clustering K clusters

GMM : Gaussian Mixture Model

$$P_\theta(x,z) = P_\theta(x|z)\, P_\theta(z)$$

Assume generative process (i)

$$z^{(i)} \sim \text{Multi}(1, \dots K)$$

$$x^{(i)} \mid z^{(i)} = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

- You know $z^{(i)}$ → QPA !
- You don't know $z^{(i)}$ → Clustering

Observations $x^{(i)}$

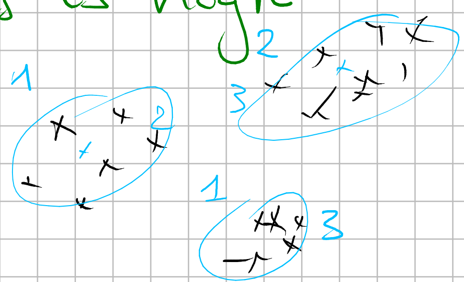
Latent variables $z^{(i)}$

$$\theta = \{\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K\}$$

MLE $\quad \arg\max_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \log P_\theta(x^{(i)}) \right)$ → Unsupervised learning

K! $\quad \log P_\theta(x^{(i)}) = \log\left(\sum_{k=1}^{K} \underbrace{P_\theta(x^{(i)}, z^{(i)} = k)} \right)$

at least one k for which this is high



Difficulty:

- no closed formed solution
- not identifiable

Reasoning for smarter learning:

(~ classif with QDA)

1 - $z^{(i)}$ are latent;

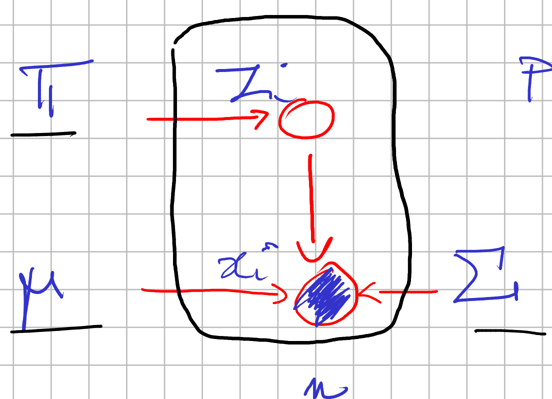
But if we know them, easy to find Θ

2 - Estimating the $z^{(i)}$?

Expectation - Maximization (EM)

1. Find the expected clusters (E-step) from Θ
2. Maximize the expected likelihood (M-step) for Θ

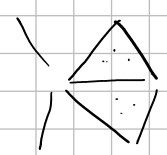
EM: general estimation for incomplete data.



PGM: probabilistic Graphical Model

- K-means → finding clusters → Hard assignments
  finding new centers → Isotropic gaussian

$\Sigma = \sigma^2 Id$



- EM, for $t = 0 \ldots$ until convergence

1. E-step; for each $x^{(i)}$

$$\mathbb{E}[z^{(i)} \mid x^{(i)}] \rightarrow \mathbb{P}_{\Theta^t}(z \mid x^{(i)})$$

$$\mathbb{P}_{\Theta^t}(z = k \mid x^{(i)}) = \frac{\pi_k * \mathbb{P}_{\Theta_k}(x^{(i)})}{\sum_{l=1}^{K} \pi_l \, \mathbb{P}_{\Theta_l}(x^{(i)})}$$

2 - M-step: Compute θ^{t+1}

$$\theta^{t+1} = \arg\max_{\theta} \sum_{i=1}^{n} \mathbb{E}_{z^{(i)} \sim \mathbb{P}_{\theta^t}(z \mid x^{(i)})} \left[\log \mathbb{P}_\theta(x^{(i)}, z^{(i)}) \right]$$

→ Solve this, for example with GD

- Why EM? → Very simple / easy to implement
  → Guaranteed to converge

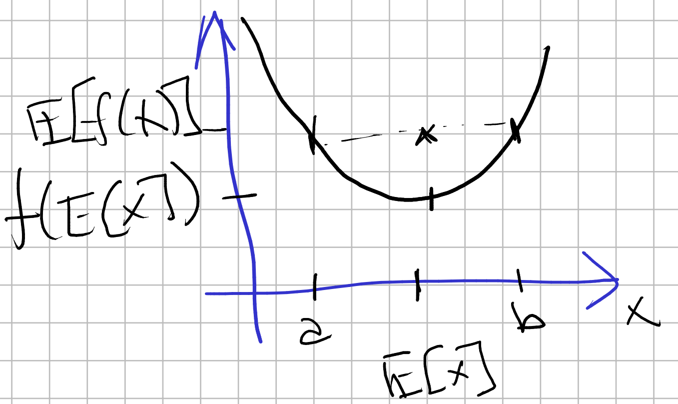
Issue: local optima

III Why does EM work?

→ More general formulation: ELBO (Evidence Lower Bound).

Jensen's inequality: f convex

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$



Equality $\iff$ x constant

$\nearrow \frac{1}{2}$ a

$\searrow \frac{1}{2}$ b

$$x, z \to \log \mathbb{P}_\theta(x) = \log \sum_{k=1}^{K} \mathbb{P}_\theta(x, z=k)$$

In GMM: $\mathbb{P}_\theta(x, z=k) = \mathbb{P}_\theta(z=k) \times \mathbb{P}_\theta(x \mid z=k)$

Easy!

→ We don't know how to write $\mathbb{P}_\theta(x, z=k)$

→ We parametrize the value of z

$$\Theta = \{\pi, \mu, \Sigma\}$$

$$\phi = \{z\}$$

Distribution (proposal) $Q_\phi(z=k) = \phi_k$

$$\sum_{k=1}^{K} \phi_k = 1$$

$$\log \sum_{k=1}^{K} \mathbb{P}_\Theta(x, z=k) = \log \sum_{k=1}^{K} Q_\phi(z=k) * \frac{\mathbb{P}_\theta(x, z=k)}{Q_\phi(z=k)}$$

$$= \log\left(\mathbb{E}_{z \sim Q_\phi}\left[\frac{\mathbb{P}_\theta(x, z)}{Q_\phi(z)}\right]\right)$$

Apply Jensen's :

$$\log \mathbb{P}_\Theta(x) \geq \mathbb{E}_{z \sim Q_\phi}\left[\log\left(\frac{\mathbb{P}_\Theta(x, z)}{Q_\phi(z)}\right)\right]$$

$=?$

$\forall \phi, \Theta, x$, → Lower bound on my evidence

→ Choose Q : sample z from Q

→ What Q_ϕ should we choose ?

We want the bound to be tight

$\rightarrow$ Equality $\Leftrightarrow$ $\dfrac{\mathbb{P}_\theta(x,z)}{Q_\phi(z)} = C \Leftrightarrow Q_\phi(z) \propto \mathbb{P}_\theta(x,z)$

But $\sum_{k=1}^{K} Q_\phi(z=k) = 1 \rightarrow Q_\phi(z=k) = \dfrac{\mathbb{P}_\theta(x, z=k)}{\underbrace{\sum_{\ell=1}^{K} \mathbb{P}_\theta(x, z=\ell)}_{\mathbb{P}_\theta(x)}}$

$Q_\phi(z=k) = \mathbb{P}_\theta(z=k \mid x)$

posterior !

$$\mathrm{ELBO}(x, \phi, \theta) = \mathbb{E}_{z \sim Q_\phi}\left[\log\left(\frac{\mathbb{P}_\theta(x,z)}{Q_\phi(z)}\right)\right]$$

- EM ? E-step : setting $Q_\phi(z) = \mathbb{P}_\theta(z \mid x)$

  $\rightarrow \log \mathbb{P}_\theta(x) = \mathrm{ELBO}(x, \phi, \theta)$

  M-step : $\operatorname{Argmax}_\theta \mathrm{ELBO}(x, \phi, \theta)$

- Many examples $x^{(1)}, \dots, x^{(n)}$

  $Q_{\phi^{(i)}}(z=k) = \phi_k^{(i)} \rightarrow$ We have $n \times K$ parameters.

$$\sum_{i=1}^{n} \log \mathbb{P}_\theta(x^{(i)}) \geq \sum_{i=1}^{n} \mathrm{ELBO}(x^{(i)}, \phi^{(i)}, \theta)$$

- Generalized EM:

Block Coordinate Ascent

- $\forall i, \quad \phi^{(i)\,t+1} = \operatorname{argmax}_{\phi \in \Delta(K)} \mathrm{ELBO}(x^{(i)}, \phi, \theta^t)$

- $\theta^{t+1} = \operatorname{argmax}_\theta \sum_{i=1}^{n} \mathrm{ELBO}(x^{(i)}, \phi^{(i)\,t+1}, \theta)$

- Show convergence : $\ell(\theta^t) \leq \ell(\theta^{t+1})$?

$$\ell(\theta^t) = \sum_{i=1}^{n} \log \mathbb{P}_{\theta^t}(x^{(i)})$$

(Estep + Jensen) $$= \sum_{i=1}^{n} ELBO(x^{(i)}, \phi^{(i)t}, \theta^t)$$

$$\Big\downarrow \text{max w.r.t } \theta$$

$$\ell(\theta^{t+1})$$

- For GMM :

M-step : $\forall p \in \Theta$

Optim : $$\frac{\partial}{\partial p} \sum_{i=1}^{n} ELBO(x^{(i)}, \phi^{(i)}, \theta^t) = 0$$

$$\to p^{t+1}$$

- Interpret

$$\log \mathbb{P}_\theta(x) - ELBO(x, \phi, \theta) = D_{KL}(Q_\phi \,\|\, \mathbb{P}_{\theta\, z|x})$$

- What's next ?

- E-step $\to$ $$Q_\phi(z) = \frac{\mathbb{P}_\theta(z)\, \mathbb{P}_\theta(x|z)}{\left(\sum_{k=1}^{K} \mathbb{P}_\theta(z=k)\, \mathbb{P}_\theta(x|z=k)\right)}$$

1 What when it's intractable

Mean field assumption :

$$Q_\phi(z) = \prod_{k=1}^{K} Q_\phi^k(z_k)$$ (Independent latent coordinates)

$\to$ Sigmoid Belief Network $\to$ 2^K

2 Assume $z \in \mathbb{R}^k$ is not discrete

z discrete $\to$ Q multinomial (π)

z continuous $\to$

$$\phi = \{\varepsilon, \psi\} \qquad Q_\phi = \mathcal{N}\left(q_\varepsilon(x), \operatorname{diag}\left(v_\psi(x)\right)^2\right)$$

$z \sim Q$

$\begin{pmatrix} q_\varepsilon \\ v_\psi \end{pmatrix} \to$ Neural Networks

$\to$ Variational Auto-encoders.

$\hookrightarrow$ Encoder

$$x \mid z \sim \mathcal{N}\left(g_\theta(z), \sigma^2 I_d\right)$$

$\downarrow$ decoder

Learning:

E step: Evaluate $Q^{(i)}(z^{(i)})$ $\to$ Compute the density (encoder)

M step: $\mathbb{E}_{z^{(i)} \sim Q^{(i)}}$ $\to$ sample from $Q^{(i)}$

$\downarrow$

ELBO

$\theta = \theta + \eta \nabla_\theta \text{ELBO}$ $\to$ easy

$\left.\begin{array}{l} \varepsilon = \ldots \\ \psi = \ldots \end{array}\right\} \to$ Difficult

$\to$ reparametrization trick.