→ Sign-up for a presentation slot (random assignment after tonight 11/16 midnight)

→ HW4 Dec 1st, Project

—

Last time: Generative assumptions on the data

→ ML estimation, ELBO objective

→ EM algorithm (today)

→ Variational Inference

Summary: Learn data distribution

Variational

Parametric assumption on the distribution and learn the unknown parameters that best fit the data

E.g. ML *, (today)

* VI (Bayesian) (today)

Sampling

Input: data
Incomplete description of distribution

↳ Generate more samples from the distribution

E.g.

Ⓐ → * MCMC
Markov Chain Monte Carlo variants

Ⓑ → * GAN, * VAE
(NN representations of distributions)

Ⓒ → Stein Variational Gradient Descent (SVGD) and particle-based deterministic/stochastic algorithms *

Ⓓ → Variational perspective "Transport" methods
Optimal transport, Normalizing flows etc

Ⓔ → Score-generative models

LDA Linear Discriminant Analysis

Introduces idea of making generative assumptions on the data

Setting: binary classification
$h(x) = 1$ or -1

Bayes optimal classifier

Assumption: • $P(Y) = \begin{cases} 1/2 & Y = 1 \\ 1/2 & Y = -1 \end{cases}$ (Uniform)

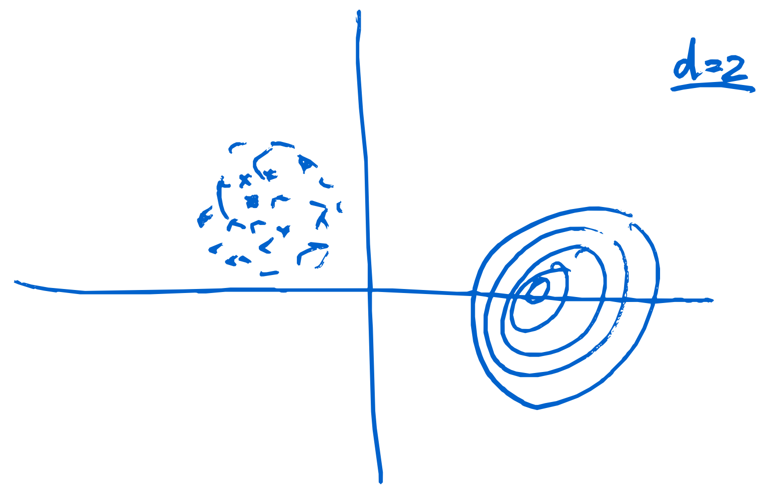
Conditional Gaussian

• $$P(X = x \mid Y = y) = \frac{e^{-(x-\mu_y)^T \Sigma^{-1} (x-\mu_y)}}{(2\pi)^{d/2} |\Sigma|^{1/2}}$$

$x \in \mathbb{R}^d$ $y = \pm 1$

Box B

$$\boxed{\begin{aligned} h_{Bayes}(x) &= sgn(w \cdot x + b) \\ w &= (\mu_1 - \mu_{-1})^T \Sigma^{-1} \\ b &= \tfrac{1}{2}(\mu_{-1}^T \Sigma^{-1} \mu_{-1} - \mu_1^T \Sigma^{-1} \mu_1) \end{aligned}}$$



$$h_{Bayes}(x) = \arg\max_y \{ P(Y=y \mid X=x) \}$$

$$P(Y=y \mid X=x) = \frac{P(Y=y)\, P(X=x \mid Y=y)}{P(X=x)}$$

$$h_{Bayes}(x) = \arg\max_y \{ P(Y=y) P(X=x \mid Y=y) \}$$

$(\because P(Y=y) = 1/2 \quad y = \pm 1)$

$$= \arg\max_y \{ P(X=x \mid Y=y) \}$$

$$h_{Bayes}(x) = sgn\left(\log \frac{P(X=x \mid Y=1)}{P(X=x \mid Y=-1)} \right)$$

$$= sgn\left(\log \frac{e^{-(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)}}{e^{-(x-\mu_{-1})^T \Sigma^{-1} (x-\mu_{-1})}} \right)$$

$$= sgn\left(-(x-\mu_1)^T \Sigma^{-1} (x-\mu_1) + (x-\mu_{-1})^T \Sigma^{-1} (x-\mu_{-1}) \right)$$

(to get box B)

Takeaway: Can make parametric generative assumptions on data distribution and solve for Y "easily"

Gaussian Mixture model

Generative assumption / Probalistic model

$$P_\theta(X, Z) = \sum_{j=1}^{k} \frac{\pi_j \, e^{-(x-\mu_j)^T \Sigma_j^{-1} (x-\mu_j)}}{\mathcal{Z}}$$

set of parameters   observed   Latent variable   normalization constant

$x \in \mathbb{R}^d$

$Z \in [k]$ k : no of components

$$\theta = (\underset{\mathbb{R}^d}{\mu_1}, \ldots, \mu_k, \underset{\mathbb{R}^{d \times d}}{\Sigma_1}, \ldots, \Sigma_k, \pi_1 \ldots \pi_k)$$

Z: Latent variables

$$P_\theta(X \mid Z = j) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-(x-\mu_j)^T \Sigma^{-1} (x-\mu_j)}$$

Z: clustering

Maximum Likelihood estimation of parameters

$$\ell(\theta) = \max_{\theta} \log P_\theta(X_1, \ldots, X_m)$$

iid data $(X_1, \ldots, X_m) = (x_1, x_2, \ldots, x_m)$

$$\ell(\theta) = \max_{\theta} \log \prod_{i=1}^{m} P_\theta(X_i)$$

$$= \max_{\theta} \sum_{i=1}^{m} \log P_\theta(X_i = x_i)$$

Sources of error

→ non-convex, only solved approximately

→ P_θ model could be wrong

Variational perspective

Latent variable Z

Data X

Parameters θ that describe $P_\theta(X,Z)$

—

Want Variational Inference

$$\underset{q \in Q}{\arg\max} \; -D_{KL}(q \,\|\, P_\theta(\cdot|x))$$

Q → set of probability distributions over Z

$$D_{KL}(q \,\|\, P_\theta(\cdot|x)) = \underset{q}{E} \log q - \underset{q}{E} \log P_\theta(z|x)$$

$$= \underset{q}{E} \log q - \underset{q}{E} \log \frac{P_\theta(z,x)}{P_\theta(x)}$$

$$= \underset{q}{E} \log q - \underset{q}{E} \log P_\theta(z,x) + \underset{q}{E} \log P_\theta(x)$$

$$\left(\because \underset{q}{E} \log P_\theta(x) = \sum_z q(z) \log P_\theta(x) = \log P_\theta(x)\right)$$

$$= \underset{q}{E} \log q - \underset{q}{E} \log P_\theta(z,x) + \log P_\theta(x)$$

$$ELBO(q,\theta,x) = \underset{q}{E} \log P_\theta(z,x) - \underset{q}{E} \log q$$

Variational objective function

$$D_{KL}(q \,\|\, P_\theta(\cdot|x)) = -ELBO(q,\theta,x) + \log P_\theta(x)$$

$$\log P_\theta(x) = ELBO(q,\theta,x) + \underbrace{D_{KL}(\quad)}_{\geq 0}$$

For any q,

$$\log P_\theta(x) \geq ELBO(q,\theta,x).$$

EM algorithm

An algorithm for ML estimation in the presence of Latent variables and a probabilistic model

$$\ell(\theta) = \sum_{i=1}^{m} \log P_\theta(x_i)$$

Fix x

$$\ell(\theta, x) = \log P_\theta(x) \hookrightarrow \text{Marginal of } x$$

$$= \log \sum_{j=1}^{k} P_\theta(x, z = z_j)$$

$$= \log \sum_{j=1}^{k} P_\theta(z = z_j) P_\theta(x \mid z_j)$$

Iterative algorithm:

E-step: $q_{t+1,i}(z) = P_{\theta_t}(z \mid x_i)$

M-step: $\theta_{t+1} = \underset{\theta}{\text{argmax}} \sum_{i=1}^{m} \text{ELBO}(q_{t+1,i}, \theta, x_i)$

$$\text{ELBO}(q, \theta, x) = -\underset{q}{E} \log q + \underset{q}{E} \log P_\theta(x, z)$$

Why does EM algorithm amount to $\max_\theta \sum_{i=1}^{m} \ell(\theta, x_i)$

- For $q(z) = P_\theta(z \mid x)$:

$$\max_\theta \ell(\theta, x) = \max_\theta \text{ELBO}(q, \theta, x)$$

$$\text{ELBO}(q, \theta, x) = -\underset{q}{E} \log q + \underset{q}{E} \log P_\theta(x) P_\theta(z \mid x)$$

$$= \underbrace{-\underset{q}{E} \log q + \underset{q}{E} \log P_\theta(z \mid x)} + \log P_\theta(x)$$

$$= \log P_\theta(x)$$

- For a fixed θ,

$q(z) = P_\theta(z \mid x)$ is

$$\underset{q \in Q}{\text{argmax}}\ \text{ELBO}(q, \theta, x)$$

EM algorithm always increasing the $\log P_\theta(x)$

$$\ell(\theta_{t+1}) \geqslant \ell(\theta_t)$$

$\downarrow$

$$\sum_{i=1}^{m} \ell(\theta_{t+1}, x_i)$$

$$\ell(\theta_{t+1}) \overset{\text{(E-step)}}{=} \text{ELBO}(q_{t+2}, \theta_{t+1})$$

$$\overset{\text{(E-step)}}{\geqslant} \text{ELBO}(q_{t+1}, \theta_{t+1})$$

$$\overset{\text{(M-step)}}{\geqslant} \text{ELBO}(q_{t+1}, \theta_t)$$

$$= \ell(\theta_t)$$