Generative Models

Discriminative vs. **Generative**

Discriminative:
- care about classification
- don't care about underlying structure of data

Generative:
- try to learn underlying structure
- learn $P(X, y)$
- can be used to generate data according to model

Generative Model Framework

1. ESTIMATE the parameters of model that "best" fit your data
2. PREDICT classification or generate/sample data from your model

Examples of Generative Models

- Multinomial / Categorical Distr. (language/Shakespeare example in lecture)

likelihood of data $P(D|\theta) = \prod_{w \in W} \theta_w^{n(w)}$

where $n(w) :=$ # occurrences of w in D

- Spherical Gaussian Distr. $N(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|x-\mu\|^2\right)$

square euclidean distance

likelihood of data $S_n = \{ x^{(t)} | t = 1 \ldots n \}$ $\quad x \in \mathbb{R}^d$

$$L(S_n; \mu, \sigma^2) = \prod_{i=1}^{n} p(x^{(i)}|\theta) = \prod_{i=1}^{n} N(x^{(i)}; \mu, \sigma^2)$$

Maximum Likelihood Estimation

* See lecture notes for language model multinomial distr. ML estimation example

- ML Estimation for Spherical Gaussian

$$\max_{\mu, \sigma^2} L(S_n; \mu, \sigma^2) \Rightarrow \max_{\mu, \sigma^2} l(S_n; \mu, \sigma^2) = \max_{\mu, \sigma^2} \sum_{i=1}^{n} \log N(x^{(i)}; \mu, \sigma^2 I)$$

$$l(S_n; \mu, \sigma^2) = \sum_{i=1}^{n} \log \left[(2\pi\sigma^2)^{-d/2} \exp\left(-\frac{1}{2\sigma^2}\|x^{(i)}-\mu\|^2\right)\right]$$

$$= -\frac{nd}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \|x^{(i)}-\mu\|^2$$

To find $\max_{\mu,\sigma^2} \ell(S_n; \mu, \sigma^2)$ just solve $\frac{d\ell}{d\mu} = 0$

and $\frac{d\ell}{d\sigma^2} = 0$

solve for $\hat{\mu}$: $\frac{d\ell}{d\mu} = 0 - \frac{1}{\cancel{2\sigma^2}} \cdot \cancel{2} \sum_{i=1}^{n} (x^{(i)} - \hat{\mu}) = 0$

$$\left(\sum_{i=1}^{n} x^{(i)}\right) - n\hat{\mu} = 0$$

$$n\hat{\mu} = \sum_{i=1}^{n} x^{(i)}$$

sample mean $\longrightarrow$ $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}$

solve for $\hat{\sigma}^2$: $\frac{d\ell}{d\hat{\sigma}^2} = -\frac{nd}{\cancel{2}} \cdot \frac{1}{\sigma^2} + \frac{1}{\cancel{2}\sigma^4} \sum_{i=1}^{n} \|x^{(i)} - \mu\|^2 = 0$

$$-nd\sigma^2 + \sum_{i=1}^{n} \|x^{(i)} - \mu\|^2 = 0$$

sample variance $\rightarrow$ $\hat{\sigma}^2 = \frac{1}{nd} \sum_{i=1}^{n} \|x^{(i)} - \hat{\mu}\|^2$

ML estimation for multinomial ✓

Gaussian ✓

Gaussian mix model ← up next!

GMM — clustering

k Gaussians/components, each Gaussian has own $\mu^{(j)}, \sigma_j^2$

mixing proportions $p_1, p_2 \ldots p_k$

log likelihood of data $\ell(S_n \mid \theta) = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} p_j N(x^{(i)}; \mu^{(j)}, \sigma_j^2)\right]$

Still use ML estimation, except hard to directly optimize!

use EM

1. Initialize parameters (initialize random, or similar to kmeans)
2. E step: calculate soft assignments (posterior probabilities)

$$p(j \mid t) = \frac{p_j N(x^{(t)}; \mu^{(j)}, \sigma_j^2)}{\sum_{\ell=1}^{k} p_\ell N(x^{(t)}; \mu^{(\ell)}, \sigma_\ell^2)}$$

(partial assignments, unlike hard assignments in kmeans)

3. M step: estimate parameters

soft counts $\hat{n}_j = \sum_{t=1}^{n} p(j|t)$

$$\hat{p}_j = \frac{\hat{n}_j}{n}$$

$$\hat{\mu}_j = \frac{1}{\hat{n}_j} \sum_{t=1}^{n} p(j|t)\, x^{(t)}$$

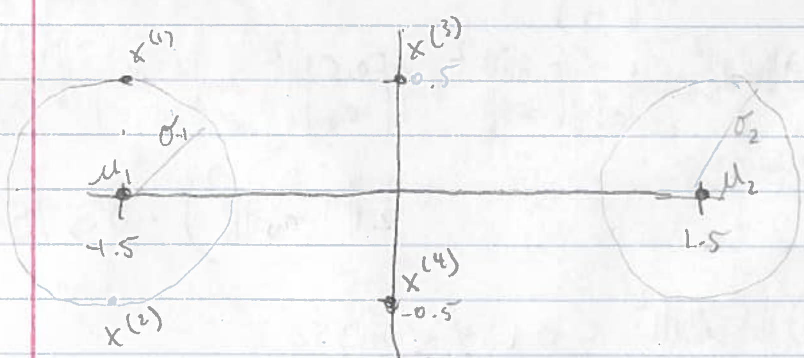
$$\hat{\sigma}_j^2 = \frac{1}{d\hat{n}_j} \sum_{t=1}^{n} \| x^{(t)} - \hat{\mu}_j \|^2$$

4. Repeat E and M steps until convergence

Convergence criterion

- # of iterations
- loglikelihood$_{new}$ $\leq$ loglikelihood$_{OLD}$ $+ \epsilon$ for small $\epsilon > 0$

EM example



$\mu_1 = (-1.5, 0)$

$\mu_2 = (1.5, 0)$

$\sigma_1^2 = \sigma_2^2 = 0.5^2$

$p_1 = p_2 = 0.5$

1. E step: calc $p(\text{Gauss} \mid \text{point})$ for every (Gauss, point) pair

$$p(G_1|x^{(1)}) = \frac{p(x^{(1)}|G_1)\,p(G_1)}{\sum_{g=G_1 \text{ or } G_2} p(x^{(1)}|g)\,p(g)} = \frac{N(x^{(1)}; \mu_1, \sigma_1^2)\,p_1}{N(x^{(1)}; \mu_1, \sigma_1^2)\,p_1 + N(x^{(1)}; \mu_2, \sigma_2^2)\,p_2}$$

$$P(x^{(1)}|G_1) = \frac{1}{2\pi 0.25} \exp\left(-\frac{1}{2 \cdot 0.25} \left\| \begin{bmatrix}-1.5\\0.5\end{bmatrix} - \begin{bmatrix}-1.5\\0\end{bmatrix} \right\|^2\right) = 0.386$$

$$P(x^{(1)}|G_2) = \frac{1}{2\pi \cdot 1/4} \exp\left(-\frac{1}{2 \cdot 1/4} \left\| \begin{bmatrix}-1.5\\0.5\end{bmatrix} - \begin{bmatrix}1.5\\0\end{bmatrix} \right\|\right) = 5.88 \times 10^{-9} \approx 0$$

$\Rightarrow P(G_1|x^{(1)}) \approx 1, \quad P(G_2|x^{(1)}) \approx 0$

$P(G_1|x^{(2)}) \approx 1$ $\quad P(G_1|x^{(3)}) = 0.5$ $\quad P(G_1|x^{(4)}) = 0.5$

$P(G_2|x^{(2)}) \approx 0$ $\quad P(G_2|x^{(3)}) = 0.5$ $\quad P(G_2|x^{(4)}) = 0.5$

by symmetry

M step

$\hat{n}_1 = \sum_x P(G_1|x) = 1+1+0.5+0.5 \approx 3$ $\quad \hat{n}_1 + \hat{n}_2 = n = 4$

$\hat{n}_2 = \sum_x P(G_2|x) = 1$

$\hat{p}_1 = 3/4 \quad \hat{p}_2 = 1/4$ $\quad$ since $\hat{p}_j = \frac{\hat{n}_j}{n}$

$$\hat{\mu}_1 = \frac{1}{\hat{n}_1} \sum_x P(G_1|x)\, x = \frac{1}{3}\left(1 \cdot \begin{bmatrix} -1.5 \\ 0.5 \end{bmatrix} + 1 \begin{bmatrix} -1.5 \\ -0.5 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 0 \\ -0.5 \end{bmatrix} \right)$$

$$\approx \frac{1}{3} \begin{bmatrix} -3 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

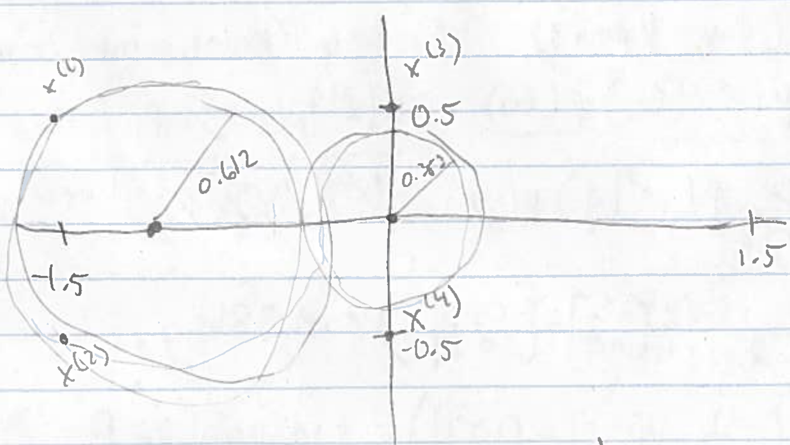
$$\hat{\mu}_2 = \frac{1}{\hat{n}_2} \sum_x P(G_2|x) = 1 \left[0 + 0 + \frac{1}{2} \begin{bmatrix} 0 \\ 1/2 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 0 \\ -1/2 \end{bmatrix} \right]$$

$$\approx \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\hat{\sigma}_1^2 = \frac{1}{2\hat{n}_1} \sum_x P(G_1|x) \|x - \hat{\mu}_1\|^2 = \frac{1}{6} \left(\left\| \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} \right\|^2 + \left\| \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix} \right\|^2 + \frac{1}{2} \left\| \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} \right\|^2 \right.$$

$$\left. + \frac{1}{2} \left\| \begin{bmatrix} 1 \\ -0.5 \end{bmatrix} \right\|^2 \right) = 0.375 = 0.612^2$$

$$\hat{\sigma}_2^2 = \frac{1}{d\hat{n}_2} \sum_x P(G_2|x) \|x - \hat{\mu}_2\|^2 = 0.125 = 0.354^2$$



After one iteration!

Somethings to note!

1. Initialization → sometimes w/ bad initialization can become stuck in local optim.
   → what happens if initialize the same?

2. How to select # of components (k)?
   - Can use validation on validation set