SPR Day 13

Mixtures of Gaussians

Reading: Bishop Sec 9.2 "Mixtures of Gaussians"

Bishop Sec 2.3.9 (for basic motivation)

Topics:

Mixture models: Why?

Capture heterogeneous patterns in data

Easy way to make flexible distributions from simple ones

Gaussian Mixture Models (GMM)

Two "views" of same model for data x

- No assignment variables
- With assignment variables "z"

Two tasks

- Estimate parameters π, μ, σ
- Estimate latent variable posteriors $p(z_n | x_n)$

Two ways to estimate parameters

- Gradient descent
- Expectation Maximization

Why mixture models?

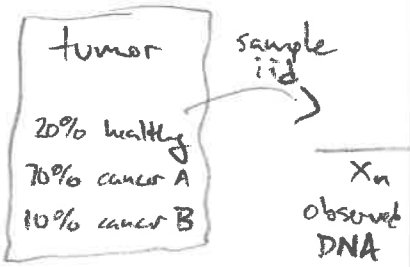
Goal of mixture model:

define distribution (PMF/PDF) over data where we mix together several simpler parametric distributions each of which conceptually represents a "type" or "cluster" of data examples

Motivating examples:

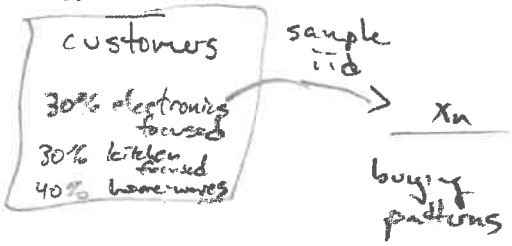
computational biology

When we sequence DNA from tumor cells, we don't know if we have a
- healthy cell
- mutated (cancer) cell



advertising

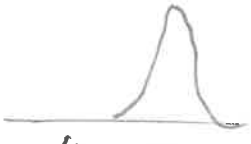
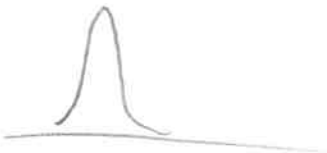
When customer walks into store, they could be of many subtypes based on their (unobservable) needs



Big Idea

Entire distribution

can be written as

	cluster 1	cluster 2	...	cluster K
appearance proba.	$\pi_1 = 20\%$	$\pi_2 = 40\%$	...	$\pi_K = 5\%$
cluster specific data-generating distribution	 μ_1, τ_1	+  μ_2, τ_2	+	 μ_K, τ_K

Gaussian mixture distribution

Random variable: $x_n \in \mathbb{R}^D$

Sample space: $\mathbb{R}^D$

Parameters: K = # clusters

μ_k = mean of kth cluster $\mu_k \in \mathbb{R}^D$

Σ_k = covariance of kth cluster Σ_k is sym. and positive definite

π_k = proba. of assigning cluster k to data $\pi_k \in [0,1]$.

PDF:

$$\text{GMMPDF}(x_n) = \sum_{k=1}^{K} \pi_k \, \text{NormPDF}(x_n \mid \mu_k, \Sigma_k)$$

Easy to prove this satisfies $\text{pdf}(x) \geq 0 \;\; \forall x$

$\int \text{pdf}(x)\,dx = 1$ as required.

so long as $\sum_k \pi_k = 1$ and $\pi_k \geq 0$.

Recall that π is a "probability vector". Can write $\pi \in \Delta^K$

For entire dataset of $\underline{N}$ examples, we write

$X = \{x_n\}_{n=1}^{N}$. $\pi = \{\pi_k\}_{k=1}^{K}$. $\mu = \{\mu_k\}_{k=1}^{K}$. $\Sigma = \{\Sigma_k\}_{k=1}^{K}$

$$p(X \mid \pi, \mu, \Sigma) = \prod_{n=1}^{N} \text{GMMPDF}(x_n \mid \pi, \mu, \Sigma)$$

Assumes <u>iid</u> from mixture distribution

Two ways to write GMM: With & without assignment random variables

Key insight: Have <u>same</u> marginal distribution over $p(x_n)$

(1) Without.

$$p(x_n) = \sum_{k=1}^{K} \pi_k \, \text{NormPDF}(x_n \mid \mu_k, \Sigma_k)$$

(2) Consider adding another random variable z_n to model.

z_n is a <u>one-hot</u> vector of size K

Interpretation: $z_{nk} = \begin{cases} 1 & \text{if example } n \text{ is assigned to cluster } k \\ 0 & \text{otherwise} \end{cases}$

Random variable is z_n with sample space $\text{onehot}(K)$

PMF is: $p(z_n) = \text{Categorical}(\pi_1, \pi_2, \ldots \pi_K)$

$= \prod_k \pi_k^{z_{nk}}$

Which means $p(z_{nk} = 1) = \pi_k$

Now, our joint model for z_n and data features x_n is

$$p(z_n, x_n) = p(z_n)\, p(x_n \mid z_n)$$

$$= \text{CatPMF}(z_n \mid \pi) \cdot \prod_{k=1}^{K} \text{NormPDF}(x_n \mid \mu_k, \Sigma_k)^{z_{nk}}$$

$$= \prod_{k=1}^{K} \pi_k^{z_{nk}} \, \text{NormPDF}(x_n \mid \mu_k, \Sigma_k)^{z_{nk}}$$

Can compute <u>marginal</u> of x_n under this joint model

$$p(x_n) = \sum_{z_n \in \text{onehot}(K)} p(x_n, z_n) = \sum_{z_n \in \text{onehot}(K)} \pi_k^{z_{nk}} \, \text{NormPDF}(x_n \mid \mu_k, \Sigma_k)^{z_{nk}}$$

$$= \sum_{k=1}^{K} \pi_k \, \text{NormPDF}(x_n \mid \mu_k, \Sigma_k)$$

Same as (1) above!

Two possible tasks for data analysis

Posterior Estimation of Latent assignments

Given: Known GMM parameters π, μ, Σ

data feature vector x_n

Goal: Estimate posterior probability that mixture cluster k generated data x_n

$$p(z_{nk}=1 \mid x_n) = \frac{p(z_{nk}=1)\, p(x_n \mid z_{nk}=1)}{p(x_n)} \quad \text{by Bayes}$$

$$= \frac{\pi_k \, \text{NormPDF}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \, \text{NormPDF}(x_n \mid \mu_j, \Sigma_j)}$$

Key Idea: Can write

$$p(z_n \mid x_n) = \text{Categorical}(r_{n1}, \cdots r_{nK}) \qquad r_n \in \Delta^K$$

non negative sums to one

So set $\tilde{r}_{nk} = \pi_k \cdot \text{NormPDF}(x_n \mid \mu_k, \Sigma_k)$

$$r_{nk} = \tilde{r}_{nk} \Big/ \sum_j \tilde{r}_{nj}$$

Take advantage of fact that $r_n \propto \tilde{r}_n$ (proportional)

Parameter Estimation

Given: N examples $\{x_n\}_{n=1}^{N}$

Goal: Estimate GMM parameters π, μ, Σ

Principle: Maximum likelihood

$$\max_{\substack{\pi \in \Delta^K \\ \mu \in \mathbb{R}^{K \times D} \\ \Sigma_k \in \text{valid cov. matrix}}} \sum_n \log \text{GMMPDF}(x_n \mid \pi, \mu, \Sigma) = \sum_n \log\left[\sum_k \pi_k \, \text{NormPDF}(x_n \mid \mu_k, \Sigma_k)\right]$$

Problems with Maximum Likelihood

As usual, with <u>small</u> <u>data</u>, ML can have problems.

Consider GMM for $N=1$ single data point, in 1D

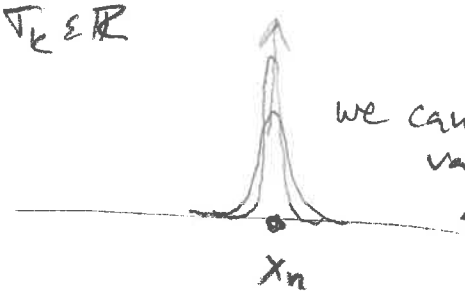
$x_n \in \mathbb{R}$, $\mu_k \in \mathbb{R}$, $\Sigma_k = \sigma_k^2 > 0$. Let $K=1$ (one cluster) with $\pi_k = 1$

ML problem:

$$\max_{\substack{\mu_k \in \mathbb{R} \\ \sigma_k \in \mathbb{R}}} \log \text{NormPDF}(x_n \mid \mu_k, \sigma_k^2)$$

$$= \log \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{1}{\sigma_k} \quad \text{if } \mu_k = x_n$$

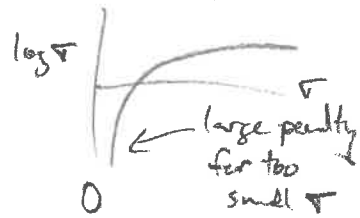
goes to inf as $\sigma \to 0$



we can shrink variance toward zero and make data "infinitely" <u>likely</u>

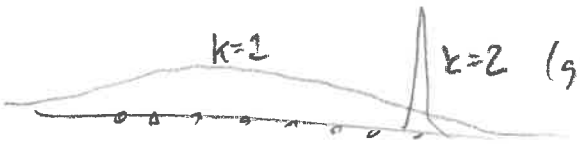
Can avoid this pathology by "penalized" ML

$$\max_{\pi, \mu, \sigma} \sum_n \log \text{GMMPDF}(x_n \mid \pi, \mu, \sigma) + \lambda \log \sigma$$



Note, can have this problem at any data size N when $K > 1$.

Just fit model with one cluster assigned to one data point



See Bishop Fig 9.7.

ML Estimation for GMMs

Method 1: Gradient Descent

(not discussed in Bishop!)

Objective:

$$\max_{\substack{\pi \in \Delta^K \\ \mu \in \mathbb{R}^{K \times D} \\ \Sigma_k \in \text{valid covariance matrix}}} \overbrace{\sum_n \log \text{GMMPDF}(x_n \mid \pi, \mu, \Sigma)}^{\mathcal{L}(\pi, \mu, \Sigma)}$$

Cannot do gradient descent when parameters are not admissable under any real value.

Why? Imagine $\pi_t \leftarrow \pi_{t-1} - \varepsilon \nabla_\pi \mathcal{L}(\pi)$

Our new π vector is no longer guaranteed to sum to one!

Solution: Reparameterize:

Let $s \in \mathbb{R}^K$, then define $\pi = \text{softmax}(s)$

Let $t \in \mathbb{R}^{K \times D}$, then define $\Sigma_k = \begin{bmatrix} e^{t_{k1}} & & \\ & \ddots & \\ & & e^{t_{kD}} \end{bmatrix}$

diagonal covariance

$e^{t_{kd}}$ ensures positive

Now can solve w/ 1st or 2nd order GD:

$$\max_{\substack{s \in \mathbb{R}^K \\ \mu \in \mathbb{R}^{K \times D} \\ t \in \mathbb{R}^{K \times D}}} \mathcal{L}(\text{softmax}(s), \mu, \exp(t))$$

element wise exponential

How to compute gradients? Can do by hand, or via automatic differentiation libraries

Concerns: step size selection (as usual)

need to include penalty term to avoid pathologies as $\sigma \to 0$

ML Estimation for GMMs:

Method 2: Expectation Maximization

Wish to solve:

$$\max_{\pi,\mu,\Sigma} \underbrace{\sum_n \log \sum_k \pi_k \,\text{NormPDF}(x_n \mid \mu_k, \Sigma_k)}_{\mathcal{L}_n(\pi,\mu,\Sigma)}$$

Let's consider what an optimal (π,μ,Σ) parameter would satisfy.

$$\vec{0} = \nabla_{\mu_j} \mathcal{L} = \sum_n \nabla_{\mu_j} \mathcal{L}_n(\pi,\mu,\Sigma)$$

$$= \sum_n \frac{1}{\sum_\ell \pi_\ell \,\text{NormPDF}(x_n \mid \mu_\ell, \Sigma_\ell)} \cdot \nabla_{\mu_j}\left[\sum_k \pi_k \,\text{NormPDF}(x_n \mid \mu_k, \Sigma_k)\right]$$

$$= \sum_n \frac{1}{\sum_\ell \pi_\ell \,\text{NormPDF}(x_n \mid \mu_\ell, \Sigma_\ell)} \cdot \pi_j \left[\nabla_{\mu_j} \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_j|^{1/2}} e^{\overbrace{-\frac{1}{2}(x_n-\mu_j)^T \Sigma_j^{-1}(x_n-\mu_j)}^{f(\mu)}}\right]$$

$$= \sum_n \frac{1}{p(x_n)} \cdot \pi_j \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_j|^{1/2}} e^{f(\mu)} \nabla_{\mu_j}\left[f(\mu)\right]$$

$$= \sum_n \frac{p(x_n, z_n = j)}{p(x_n)} \nabla_{\mu_j}\left[-\frac{1}{2}(x_n-\mu_j)^T \Sigma_j^{-1}(x_n-\mu_j)\right]$$

$$= \sum_n \frac{p(x_n, z_n = j)}{p(x_n)} \Sigma_j^{-1}(x_n-\mu_j)$$

posterior probability

$r_{nj} = p(z_{nj}=1 \mid x_n)$

what Bishop calls γ_{nj}

Multiply both sides by Σ_j and rearrange

$$\Sigma_j \vec{0} = \sum_n r_{nj} \cancel{\Sigma_j \Sigma_j^{-1}} (x_n - \mu_j)$$

$$\vec{0} = \sum_n r_{nj}(x_n - \mu_j)$$

$$\hookrightarrow \mu_j = \frac{\sum_n r_{nj} x_n}{\sum_n r_{nj}}$$

At fixed point, $\mu_j \in \mathbb{R}^D$ each mean will be equal to weighted empirical mean of points assigned to cluster j

The Coordinate Ascent "E-M" Algorithm

So, we can work out optimal updates for all parameters given "responsibilities" $r_n \in \Delta^K$

where $r_{nj} \triangleq p(z_{nj}=1 \mid x_n)$

$$\mu_j = \frac{\sum_n r_{nj} x_n}{\sum_n r_{nj}}$$

Bishop PRML (9.17)

$$\Sigma_j = \frac{1}{\sum_n r_{nj}} \sum_n r_{nj} (x_n - \mu_j)(x_n - \mu_j)^T \quad (9.19)$$

may need to add small diagonal term to ensure positive definite

$$\pi_j = \frac{\sum_n r_{nj}}{N} \quad (9.22)$$

Leads to a coordinate ascent algorithm known as "Expectation-Maximization" (similar to k-means)

init: Some valid $\pi \in \Delta^K$, $\mu \in \mathbb{R}^{K \times D}$, $\Sigma : \{\Sigma_k\}_{k=1}^K$ each valid covariance

while not converged:

for each n in 1...N:

for each j in 1, 2, ... K:

$r_{nj} = \pi_j \, \text{NormPDF}(x_n \mid \mu_j, \Sigma_j)$

$r_n = \text{make\_sum\_to\_one}(r_{n1}, \dots r_{nK})$

"E" step

for each j in 1...K:

Update μ_j with 9.17 given $\{r_n\}$

Σ_j with 9.19 "

π_j with 9.22 "

"M" step

Advantages: No step sizes to select.