SPR Day 18

Gaussian Mixture Models

Reading: Bishop PRML Sec. 9.2-9.2.1
Define GMM
Covers Parts 2-5 below

Sec. 2.3.9
basic motivation for mixture models

Outline:
(1) Why mixture models
(2) Gaussian mixture model
Two views: Without + with hidden assignments
(3) Estimating posterior over assignment indicators for one example
(4) Estimating parameters of GMM given dataset
(5) Problems with ML estimation

Why Mixture Models?

Common data analysis problem:

We observe many examples of a diverse population

Examples: many patients in hospital have vital signs taken

many DNA sequences obtained from cells found near a tumor

We have good domain knowledge that examples come from some "types" or "groups" or clusters"

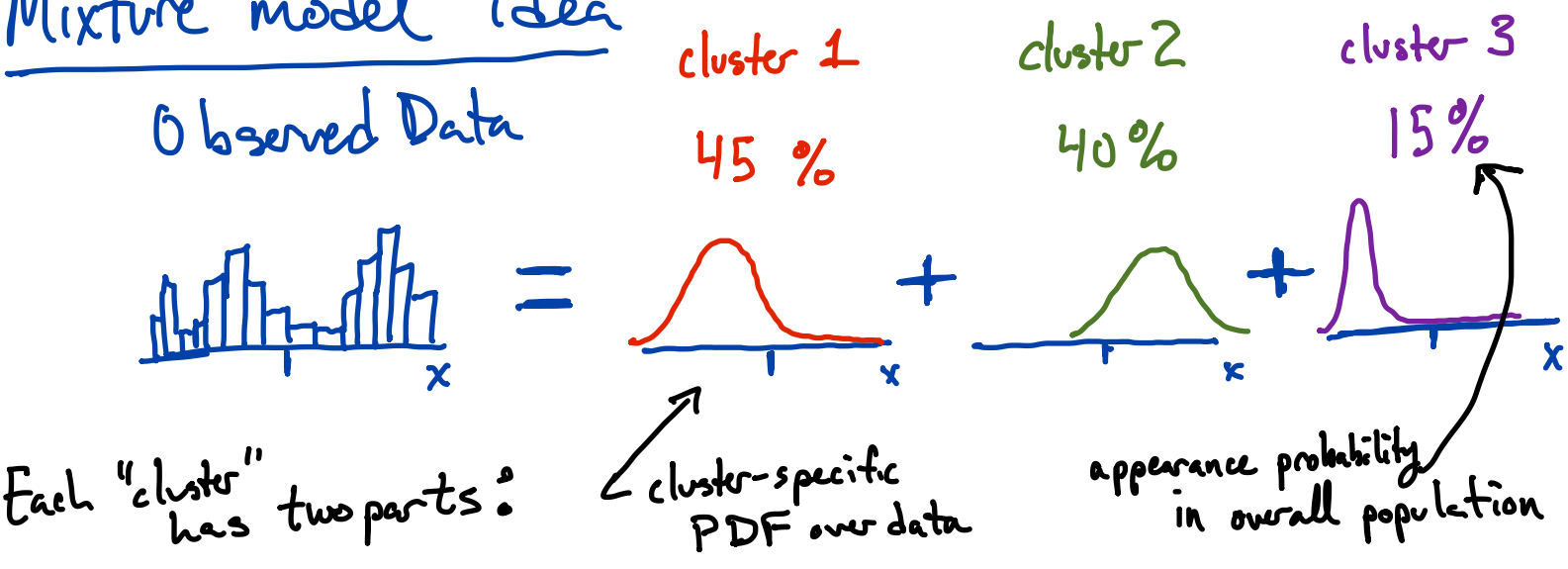
Examples: patients: disease A vs. disease B vs ... disease Z

cells: healthy vs. cancer vs. cancer w/ extra mutation

But, we just observe raw data, <u>not</u> cluster indicators.

Goal: Represent many clusters to get <u>flexible</u> model that fits data well.

Mixture model idea



Gaussian Mixture Model (GMM)

Defines a distribution over real valued vectors via a mixture of K separate Gaussian distributions

Random variable: $x \in \mathbb{R}^D$ D = # of dimensions in vector x

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

Parameters: $\pi_{1:K} = [\pi_1, \pi_2, \dots \pi_K]$ "mixture weights" aka "assignment probabilities"

$\pi_{1:K} \in \Delta^K$ vector of non-negative entries, sum to one

$\mu_{1:K}$ where $\mu_k \in \mathbb{R}^D$ "cluster locations" or "means"

$\Sigma_{1:K}$ where Σ_k is DxD symmetric pos. definite "cluster covariances"

PDF

$$GMMPDF(x \mid \pi_{1:K}, \mu_{1:K}, \Sigma_{1:K}) = \sum_{k=1}^{K} \underbrace{\pi_k}_{\text{proba. that cluster k generates this data}} \underbrace{MVNormPDF(x \mid \mu_k, \Sigma_k)}_{\text{PDF at x using cluster k's data-generating distribution}}$$

Is this a valid PDF? Yes! Meets two requirements:

- $GMMPDF(x) \geq 0 \quad \forall x$
- $\int GMMPDF(x)\,dx = 1$

Another View of GMMs: Hidden Assignment Variables

Consider adding another random variable

$$z : [z_1, z_2, \dots z_K]$$ one-hot vector size K ("assignment indicator")

$$z_k = \begin{cases} 1 & \text{if we assign cluster } k \text{ to generate our data } x \\ 0 & \text{otherwise} \end{cases}$$

Our model for z is as a Categorical r.v.

$$z \sim Cat(\pi_{1:K})$$ assign to cluster k with probability π_k

Now, our GMM in this view defines a JOINT distribution

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

$$= CatPMF(z|\pi_{1:K}) \cdot \prod_{k=1}^{K} MVNormPDF(x|\mu_k, \Sigma_k)^{z_k}$$

Key Idea: This expanded view is useful, but ultimately it's the same model as previous page.

Why? Marginal of x is: $p(x) = \sum_{k=1}^{K} p(x, z = e_k)$ (e_k = one hot vec indicating k)

This is equal to GMMPDF(x)

$$= \sum_{k=1}^{K} \pi_k\, MVNormPDF(x|\mu_k, \Sigma_k)$$ same as before!

Computing Posterior over Assignments

Given a GMM with known parameters π, μ, Σ and a single data example $x \in \mathbb{R}^D$, which of the K clusters generated this x?

Can write this as a posterior probability

$$\underbrace{p(z_k=1 \mid x)}_{\text{Prob. that cluster } k \text{ generated data example } x} = \frac{\overbrace{p(z_k=1)}^{\text{prior}} \overbrace{p(x \mid z_k=1)}^{\text{likelihood}}}{p(x)} \quad \text{by Bayes rule}$$

$$= \frac{\pi_k \cdot \text{MVNormPDF}(x \mid \mu_k, \Sigma_k)}{\underbrace{\sum_{l=1}^{K} \pi_l \, \text{MVNormPDF}(x \mid \mu_l, \Sigma_l)}_{\text{GMMPDF}(x)}} \quad \text{substitute in definitions}$$

Thus, can write posterior as a categorical distribution

$$p(z \mid x) = \text{CatPMF}(r_1, r_2, \dots r_K)$$

where $r_k = \frac{1}{\text{GMMPDF}(x)} \pi_k \, \text{MVNormPDF}(x \mid \mu_k, \Sigma_k)$

ML Estimation of Parameters

Given: N data examples $X_{1:N}$ s.t. $X_n \in \mathbb{R}^D$

K : number of assumed clusters

Goal: Estimate values of all GMM parameters

total size		
K	π	: K-length vector with non-negative entries that sum to one
$K \times D$	$\mu_{1:K}$	: μ_k is D-length vector of reals
$K \times D \times D$	$\Sigma_{1:K}$	: Σ_k is $D \times D$ sym, pos. definite matrix

Method: Maximize likelihood of N observed examples $X_{1:N}$

Optimization Problem:

$$\pi^*, \mu_{1:K}^*, \Sigma_{1:K}^* = \underset{\substack{\pi \in \Delta^K \\ \mu_{1:K} : \mu_k \in \mathbb{R}^D \\ \Sigma_{1:K} : \Sigma_k \in \text{D×D sym. pos. def.}}}{\operatorname{argmax}} \sum_{n=1}^{N} \log \text{GMMPDF}(x_n | \pi, \mu_{1:K}, \Sigma_{1:K})$$

$$= \log \sum_k \pi_k \, \text{MVNormPDF}(x_n | \mu_k, \Sigma_k)$$

$$= \text{logsumexp}\Big([\log \pi_1, \ldots \log \pi_K] + [\log \text{MVNPDF}(x_n | \mu_1, \Sigma_K), \ldots \log \text{MVNPDF}(x_n, \mu_K, \Sigma_K)] \Big)$$

Methods: - grad. ascent
- coord. ascent (next class)

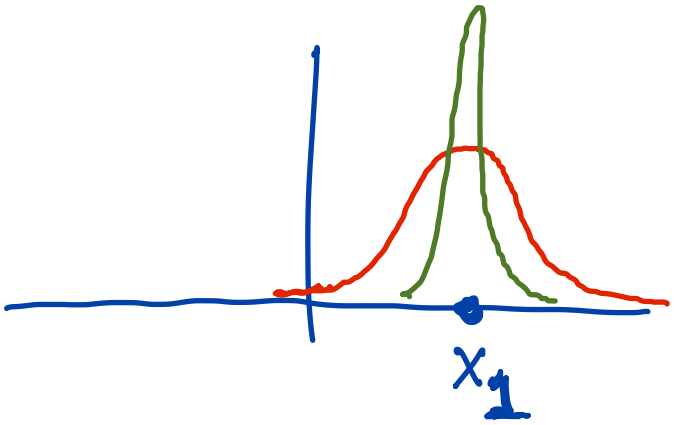
use logsumexp trick to avoid underflow

Problems with ML Estimation

As we've seen before, ML estimation can have problems with poor generalization performance especially when data set size is small.

Consider K=1 GMM (basically, a single Gaussian)

for N=1 example with D=1 dimension. $\Sigma_k = \sigma_k^2$ scalar variance



ML goal:

$$\max_{\mu_1 \in \mathbb{R}, \sigma_1 > 0} \log \text{NormPDF}(x_1 | \mu_1, \sigma_1^2)$$

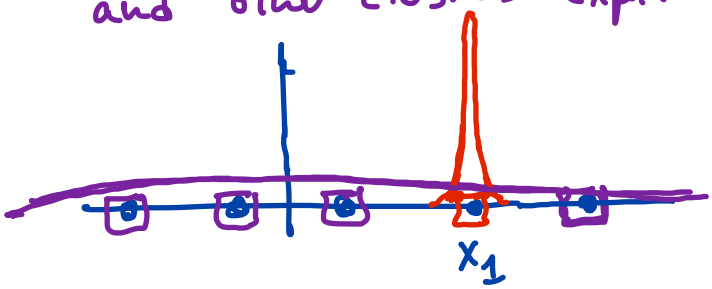
Two solutions shown

- $\mu_1 = x_1, \sigma_1 = 1$
- $\mu_1 = x_1, \sigma_1 = \frac{1}{100}$

Problem: We can make this point's PDF value $\to +\infty$ as $\sigma_1 \to 0^+$

Consider more general GMMs:

Problems arise when $N >> 1$ and $K \geq 2$,
when one cluster with $\sigma_k \to 0$ lands on top of one point,
and other clusters explain other examples



Problem summary: ML can yield
- zero variance
- $+\infty$ PDF

- Want all clusters to have non-zero variance to have hope of generalizing well.
- Want finite PDF values even on training set for sensible model selection

Ways to Avoid this Problem

(1) Constrain variances away from zero

$$\sigma^2 > \varepsilon \quad \text{not} \quad \sigma^2 > 0$$

for some $\varepsilon >> 0$, say 0.01

(2) Penalized ML

$$\max \log p(x_{1:N}) - \lambda \, \text{penalty}(\sigma)$$

Add penalty term to objective with weight $\lambda > 0$

Penalty term gives high "cost" to problematic σ values

(3) Do MAP instead of ML

Pick a prior distribution $p(\sigma)$ that favors variances far from zero

$$\max \log p(x_{1:N}) + \log p(\sigma)$$