11. Mixture Models and EM

Mixture Models:

Recap K-means clustering

$\underline{x} \in \mathbb{R}^N$

data space

M clusters: $q_1, q_2, \ldots q_M$

M prototypes (centroids): $\underline{w}_q \in \mathbb{R}^N$

assignment variables for each point

$$m_q^{(\alpha)} := \begin{cases} 1 & \text{point assigned to cluster } q \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{q=1}^{M} m_q^{(\alpha)} = 1$$ a point has to be assigned to something.

slide #4

measure "size" of a cluster post-hoc

radius of the circles

cluster "size" does not tell us how "important" a cluster is

high density vs. low density clusters

more ← important → less

Whatever is generating the data generates more points in cluster q_1 than q_2

We can measure this density post-hoc but what about an algorithm that does this as it's clustering?

This algorithm should be able to capture densities

Remember density estimation?

Mixture Models

Combine (A) density estimation
with (B) clustering

Many ways of describing the same thing:

- cluster the data and estimate the density of each cluster.

- increase the resolution of your density estimation such that you end up with a mode around each cluster.

- assume M sources are generating points in the data find a measure for:

* 1. the probability of source $q \in \{1, \dots, M\}$ generating any point. Are you contributing to the data?

** 2. the probability of a point coming from source q_1 vs. q_2 vs q_M
Who generated this point?

$$\underline{x} \in \mathbb{R}^N \sim P(\underline{x})$$

observations joint unknown distribution of the data

Mixture models express $P(\underline{x})$ as a liner combination of components, ~~and~~

$$P(\underline{x}) = \sum_{q=1}^{M} \underbrace{P(\underline{x} \mid q)}_{**} \underbrace{P(q)}_{*}$$

slide #6 Example of Gaussian mixture models

choice of basis function

What happens now?

$P(\underline{x})$ is unknown, we want to estimate it.

Pick a model class: $\underbrace{\text{Gaussian}}_{\text{parametric}}$ Mixture model

Recap from density estimation:

slide #9 objective: maximize likelihood $P(x^{(\alpha)}) \overset{!}{=} \max$

corresponds to minimizing neg. log likelihood

maximizing, minimizing
no real difference
depends on the cost funct.
any max can be turned into a min.

easier to optimize
sums are easier to handle
than products.

$$E^T = -\ln P =$$

$$= -\sum_{\alpha} \ln \sum_{q} P(x^{(\alpha)}|q) P(q) \overset{!}{=} \min$$

slide #10 We've actually seen a special case
of this before: soft-clustering

slides #15, #16, #17

Optimization

expressions for the different partial derivatives of all the parameters of the individual Gaussian components we are trying to find.

Lagrang multiplier method required because of constraint $\sum_{q}^{M} P(q) = 1$

ensures we don't want useless bystanding components.
each component has to contribute something to the dataset.
Not necessarily to every point.

side effect

degenerate solutions "collapse"
smaller M actually better. ← Actually we just want to avoid a component being only useful in explaining a very very small number of points.

slide #19

The EM algorithm

Expectation Maximization

An iterative method that alternates between two steps for approximately maximizing the likelihood function

Recap MLE

Problem: estimate distribution parameters (e.g. μ, Σ of Gaussian mixture models)

Approach:

likelihood function $\hat{P}(\{\underline{x}^{(\alpha)}\} \mid \underline{w}) = \prod \hat{P}(\underline{x}^{(\alpha)} \mid \underline{w})$

observations

$\underline{w}$

measures that the points come from this estimated distribution.

The true distribution $P(\underline{x})$ is unknown.

All we can do is find a $\hat{P}(\underline{x} \mid \underline{w})$ that produces non-zero probabilities for every observation $\underline{x}^{(\alpha)}$ and assigns more mass to more frequently occuring sample regions.

slide #9

Optimization

$$\ln \hat{P}(\{\underline{x}^{(\alpha)}\} \mid \underline{w}) = \sum_{\alpha=1}^{p} \ln \hat{P}(\underline{x}^{(\alpha)} \mid \underline{w}) \overset{!}{=} \max_{\underline{w}}$$

use the log to maximize sum instead of product.

OR minimize negative log likelihood

$$-\ln P(\{\underline{x}^{(\alpha)}\} \mid \underline{w}) = -\sum_{\alpha=1}^{p} \ln \hat{P}(\underline{x}^{(\alpha)} \mid \underline{w})$$

A scenario to justify approximation of MLE via EM

MLE with latent variable models is too costly.

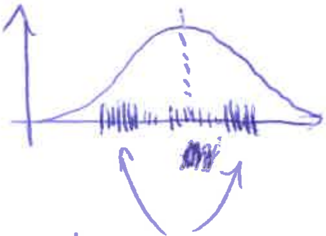
2 → see example (e.g. density estimation for clustered data)

assignment variables are unknown.

Example scenario for latent variable models:

What is the height distribution of students ~~in a school~~ at the university?

Approach 1: parametric fit ⇒ fit a Gaussian



latent variables

two groups

maybe gender / a lot of basketball players

There are hidden causes in these observations.

We want to
1. increase the resolution of our density estimation
2. estimate density around each unknown group

we need clustering

mixture models

$\hat{P}(\underline{x}^{(\alpha)} | \underline{w})$ = see slide #25

slide #25

So what is the EM algorithm?

A solution to the chicken-an-egg-problem

We need to know who is assigned to which group to estimate the density of that group

But we don't know the assignments.

~~To know the assignments we need~~

But to get the assignment we need the density of that group.....

show GIF

What do we do:

Initialization

- randomly assign samples ~~equally~~ evenly to groups
- estimate the density of each group

Step 1 — use the density to "re-" assign points to groups

Step 2 — use the current assignments to update the densities

repeat

Sounds a lot like K-means.

K-mean is a special case of EM.

Now we need reassurance that alternating between the steps improves both our assignments AND our density estimates.

slide #125

$\underline{x}$	$\underline{m}$	$\underline{w}$
visible variables	latent (hidden) variables	parameters we want to estimate

slide #126

$$+\ln P\left(\{\underline{x}^{(\alpha)}\}, \{\underline{m}^{(\alpha)}\} \,\middle|\, \underline{w}\right) \stackrel{!}{=} \max$$

But $\underline{w}$ depends on $\underline{m}$!

an $\underline{m}$ is unknown

How do ~~the~~ $\underline{x}$, $\underline{m}$ and $\underline{w}$ interact with one another?

$$P(\underline{x} \mid \underline{w}) = \prod_\alpha P(\underline{x}^{(\alpha)} \mid \underline{w}) =$$

$$= \prod_\alpha \sum_{\underline{m}} P(\underline{x}^{(\alpha)}, \underline{m} \mid \underline{w})$$

marginalizing over $\underline{m}$ (summing out $\underline{m}$) recovers the likelihood.

from the product rule: $P(a,b) = P(a \mid b) P(b)$

$$P(\underline{x}, \underline{m} \mid \underline{w}) = P(\underline{m} \mid \underline{x}, \underline{w}) \cdot P(\underline{x} \mid \underline{w})$$

$$P(\underline{x}, \underline{m} \mid \underline{w}) = P(\underline{x} \mid \underline{m}, \underline{w}) \cdot P(\underline{m} \mid \underline{w})$$

$$P(\underline{x} \mid \underline{w}) = \sum_{\underline{m}} P(\underline{x}, \underline{m} \mid \underline{w})$$

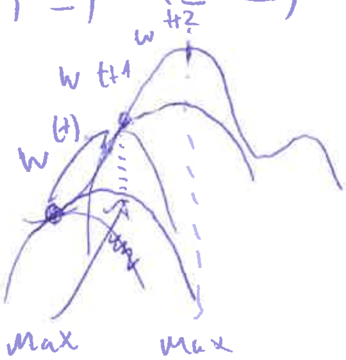
$$= \sum_{\underline{m}} P(\underline{m} \mid \underline{w}) P(\underline{x} \mid \underline{m}, \underline{w})$$

posterior:

$$P(\underline{m} \mid \underline{x}, \underline{w}) = \frac{P(\underline{x}, \underline{m} \mid \underline{w})}{P(\underline{x} \mid \underline{w})}$$

posterior of hidden $\underline{m}$ given observed $\underline{x}$

$$P(\underline{m}|\underline{x},\underline{w})\,P(\underline{x}|\underline{w}) = P(\underline{x}|\underline{m},\underline{w})\,P(\underline{m}|\underline{w})$$



show convergence plot

$$P(\underline{x}|\underline{w}) = \frac{P(\underline{x}|\underline{m},\underline{w})\,P(\underline{m}|\underline{w})}{P(\underline{m}|\underline{x},\underline{w})}$$

Define a distribution $P(\underline{m})$ and insert it above:

$$P(\underline{x}|\underline{w}) = \frac{P(\underline{x}|\underline{m},\underline{w})\,P(\underline{m}|\underline{w})}{P(\underline{m})} \cdot \frac{P(\underline{m})}{P(\underline{m}|\underline{x},\underline{w})}$$

take logarithm on both sides:

$$\ln P(\underline{x}|\underline{w}) = \ln\Big(\qquad\qquad\Big) + \ln\Big(\qquad\qquad\Big)$$

Compute the expectation w.r.t. $P(\underline{m})$:

$$\mathbb{E}[\ln P(\underline{x}|\underline{w})] = \int P(\underline{m}) \ln\left(\frac{P(\underline{x}|\underline{m},\underline{w})\,P(\underline{m}|\underline{w})}{P(\underline{m})}\right) d\underline{m}$$

$$+ \int P(\underline{m}) \ln\left(\frac{P(\underline{m})}{P(\underline{m}|\underline{x},\underline{w})}\right) d\underline{m}$$

$D_{KL}(P(\underline{m}) \| P(\underline{m}|\underline{x},\underline{w})) \geq 0$

Jensen's inequality

$\cup$ For convex functions $E[f(x)] \geq f(E[x])$

$\cap$ for concave $E[f(x)] \leq f(E[x])$

The logarithm function is concave

1st derivative of $\ln(a) = \frac{1}{a}$ strictly decreases when x increases

$$E\left[\ln P(\underline{x}\mid\underline{w})\right] = E\left[\mathcal{L}(P(\underline{m}), \underline{w})\right] + D_{KL}\left(P(\underline{m}) \,\|\, P(\underline{m}\mid\underline{x},\underline{w})\right)$$

$\ln P(\underline{X}\mid\underline{W})$ has a lower bound that is

$$\mathcal{L}(P(\underline{m}), \underline{w}) = \int P(\underline{m}) \ln\left(\frac{P(\underline{x}\mid\underline{m},\underline{w})\,P(\underline{m}\mid\underline{w})}{P(\underline{m})}\right) d\underline{m}$$

E - step using $\underline{w}^{old}$:

maximize $\mathcal{L}(P(\underline{m}), \underline{w}^{old})$ w.r.t $P(\underline{m})$

keeping $\underline{w}^{old}$ fixed

solution:

$\mathcal{L}(P(\underline{m}), \underline{w}^{old}) \overset{!}{=} \max$

when $D_{KL} \to 0$

this happens when $P(\underline{m}) = P(\underline{m}\mid\underline{x}, \underline{w}^{old})$

M-step

keep $P(\underline{m})$ fixed

maximize $\mathcal{L}(P(\underline{m}), \underline{w})$ w.r.t $\underline{w} \longrightarrow \underline{w}^{new}$

We either have a higher lower bound $\mathcal{L}$ or the same because we already reached the maximum.

~~This~~ increasing $\mathcal{L}$ $\to$ increase of likelihood yay!

but not by the same amount.

the difference is explained by the D_{KL} term. ≥ 0

increase in likelihood is actually larger.

if we are keeping $P(\underline{m})$ fixed between E and M steps we are actually measuring $P(\underline{m})$ as a function of $\underline{w}^{old}$.

Therefore:

$$P(\underline{m}) := P(\underline{m} \mid \underline{x}, \underline{w}^{old})$$

From this follows:

$$\mathcal{L}\left(P(\underline{m} \mid \underline{x}, \underline{w}), \underline{w}^{old}\right) =$$

no old for this one.

$$= \sum_{\underline{m}} P(\underline{m} \mid \underline{x}, \underline{w}^{old}) \ln P(\underline{x} \mid \underline{m}, \underline{w})$$

$$\underbrace{- \sum_{\underline{m}} P(\underline{m} \mid \underline{x}, \underline{w}^{old}) \ln P(\underline{m} \mid \underline{x}, \underline{w}^{old})}_{\text{entropy of } P(\underline{m} \mid \underline{x}, \underline{w}^{old}) \text{ independent of } \underline{w} \,\longrightarrow\, \text{constant.}}$$

$$) := Q(\underline{w}, \underline{w}^{old})$$

the