CISC 372
Text Analytic IIII
Expectation Maximization (em)

https://giphy.com/gifs/kid-costume-transformer-R9zXHWAHyTjnq

Last lectures

- Text Analysis
  - Vanilla RNN
  - Gated Recurrent Unit
  - LSTM
  - Attention Mechanism
    - Memory
      - things we memorized
    - Context
      - based on the context, on which part of memory should we focus
  - Multi-head attention
  - Attention – Explainability
  - Padding (sorting the sequence)
  - RNN for different sequence problems

Last lectures

- Language Model
- Neural Network for Language Model
- RNN for Language Modeling
- Transformer

Whenever we cluster data, we are assuming, implicitly, that the clusters reflect some real, underlying process.

reason + variation

So what if we assume that the data has been generated from a set of underlying distributions?

The problem of clustering becomes the problem of inferring the structure of these distributions from the evidence of the data.

Two levels of assumption:

What shape do the distributions have?

(Normal, uniform, Poisson)

What are their parameters?

(Normal distribution determined completely by mean and standard deviation)

The standard statistical machinery is designed to answer the question: how likely is it that the data I see comes from this particular model (distribution).

We want to ask a harder question: which distribution if the most likely one for this data to have come from?

Answer: maximum likelihood estimation

Suppose our dataset is n observations, D, from a distribution $f(x \mid \theta)$, where θ is the set of parameters of the distribution.

The likelihood function

$$L(\theta \mid D)$$

is the probability that the data came from the distribution f as a function of θ.

$$L(\theta \mid D) = p (x_1, x_2, \ldots, x_n \mid \theta)$$

$$= \prod f(x_i \mid \theta)$$

Because log is a monotonic function, it's convenient to take the log-likelihood

$$l(\theta \mid D) = \log L(\theta \mid D)$$

so the product turns into a sum.

The value of θ for which the data has the highest probability of having arisen is the maximum likelihood estimator (MLE).

If the form of f is tractable, we can find it by differentiating the expression for l and finding its maximum analytically.

This works for some simple distributions and mixtures of distributions, but is hard in general.

In many realistic situations, the data we see comes from a set of distributions, each modelling some subset of the objects.

In this case, the distribution function has the form

$$f(x) = \sum \pi_k f_k (x \mid \theta_k)$$

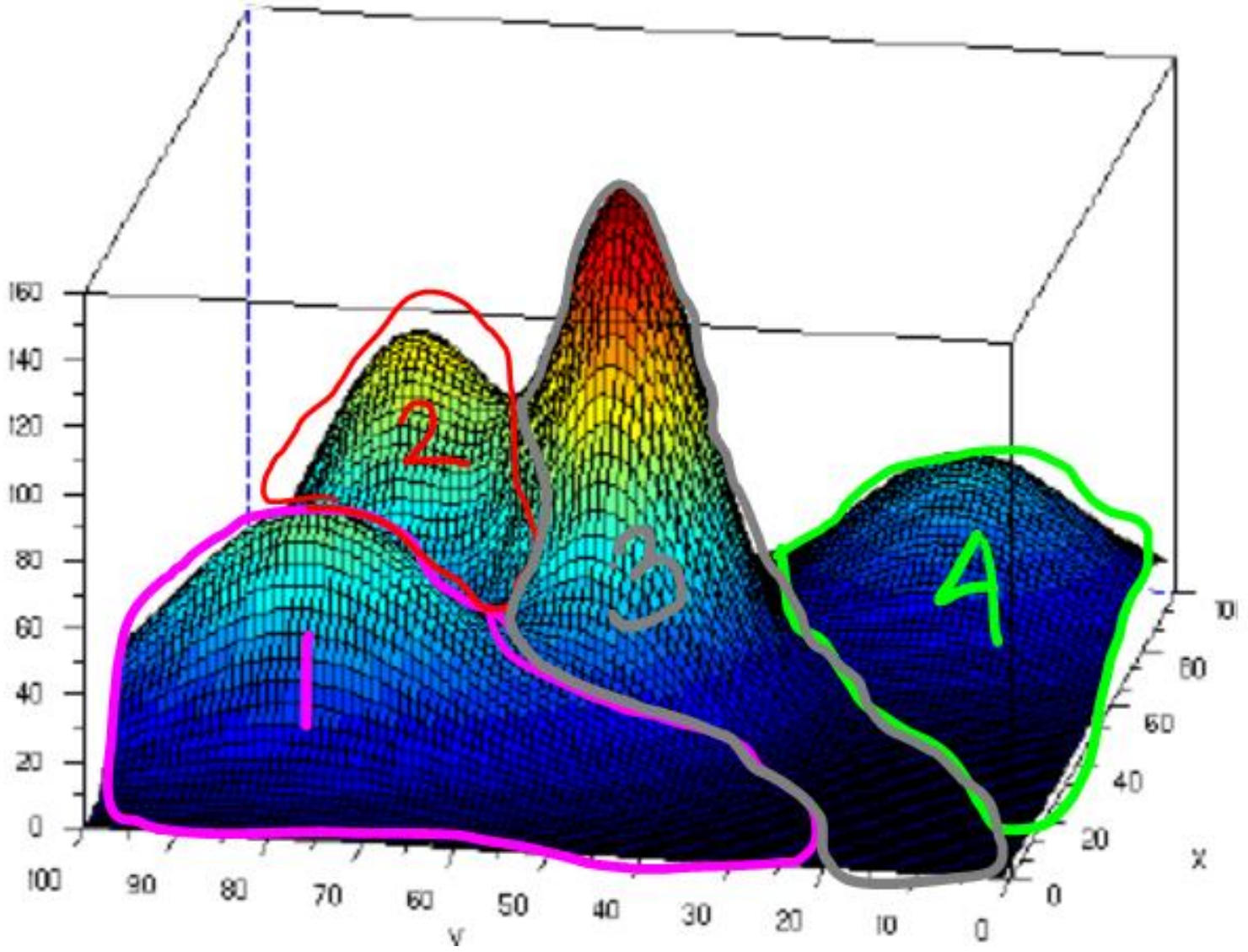
where π_k is the probability that an object will come from the *k* th distribution.

Sadly the optimization problem is much harder.

Expectation-Maximization (EM) is an iterative algorithm to maximize a likelihood function.

Suppose that the **different distributions represent different classes**. The reason the optimization problem is hard is that we are simultaneously trying to allocate objects to classes, and discover the properties of the classes!

$$f(x) = \sum \pi_k f_k (x \mid \theta_k)$$



https://ardianumam.files.wordpress.com/2017/11/em2.png

EM goes in two phases:

E: guess the parameter values for each distribution, and calculate the probability that each object came from one of the distributions.

M: given these probabilistic memberships in classes, calculate new parameter values for the distributions. (Often this can be solved analytically.)

Repeat until likelihood converges.

This seems like a heuristic, but the likelihood provably increases in each round, so it must converge to at least a local maximum.

A good idea to start from several sets of randomly chosen places.

The algorithm usually converges in 5-20 iterations.

Time complexity per step $O(k\ m^2\ n)$ for k distributions.

Notice that the class labels for the hypothetical clusters (distributions) could be thought of as missing values.

The EM algorithm `fills in' these missing values as well as the parameters of the distribution.

This idea can be extended to other missing values – just treat them as extra data that the EM algorithm has to infer.

This property makes EM a very strong technique for clustering in the presence of missing values.

And it's probably the technique of choice when enough is known about the data to make plausible guesses about

how many distributions

what their shape should be