AIMLC

Anshul Mittal

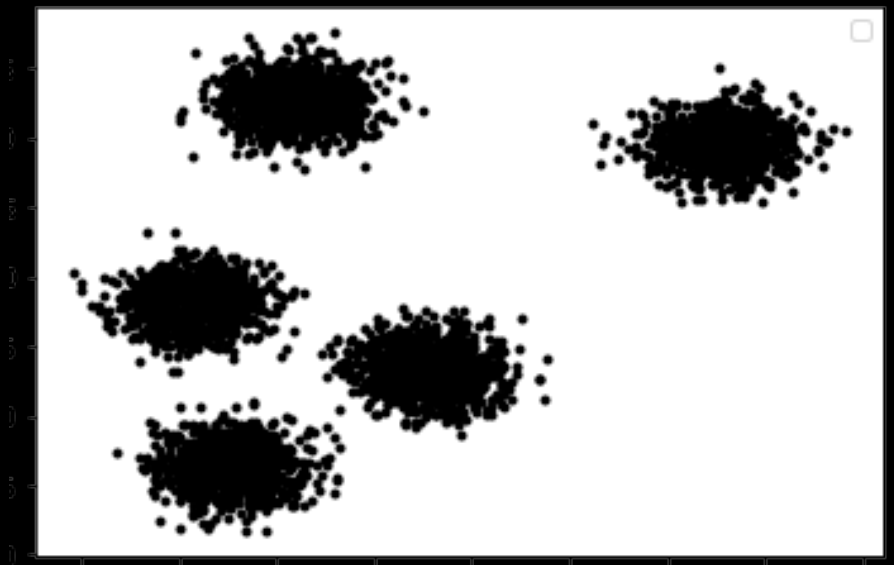
Google PhD Fellow, IIT Delhi

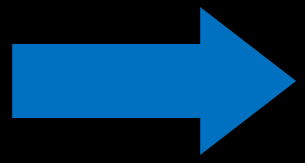
Today’s topics

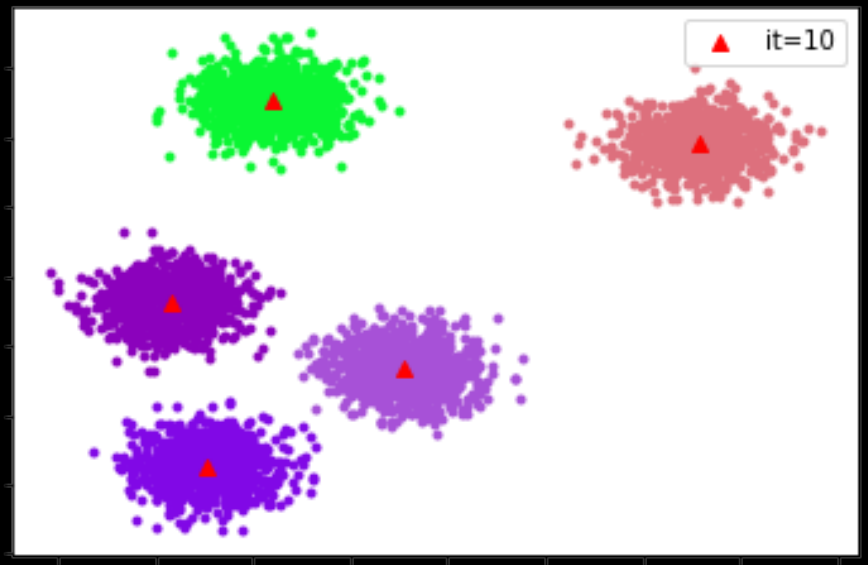
- K-Means
  - What is it?
  - How does it work?
  - Demo
- Expectation Maximization Technique
  - What is expectation Maximization
  - Deriving K-Means using expectation maximization
- Existing problems in K-Means and Possible Solution
  - HINT: GDA: Gaussian Discriminant Analysis
  - Demo

K-Means – What is it?

- Unsupervised Generative algorithm to cluster the feature space
- Problem setting:
  - N points are spread across the feature space which is D-Dimensional and task is to partition the feature space in order to have maximum likelihood of the dataset.







K-Means – How does it work?

- Iteratively clusters the feature space
  1. Assumes cluster centers μ_k, ϕ_k
  2. Assigns items to each cluster based on "some" similarity measure

$$\text{e.g. } p\left(x \middle| \mu_k\right) = \frac{e^{-0.5\left(X-\mu_k\right)^T (X-\mu_k)}}{\sqrt{2\pi}}$$

  3. Recompute cluster centers $\mu_k = Mean(X \in cluster\ k)$
  4. Repeat step 2,and 3 till dataset likelihood ($L(D)$) converges or maximum number of iteration are reached

$$L(D) = \prod_{i=1}^{N} P\left(x_i \middle| \mu_{ci}\right) * \phi_{ci}$$

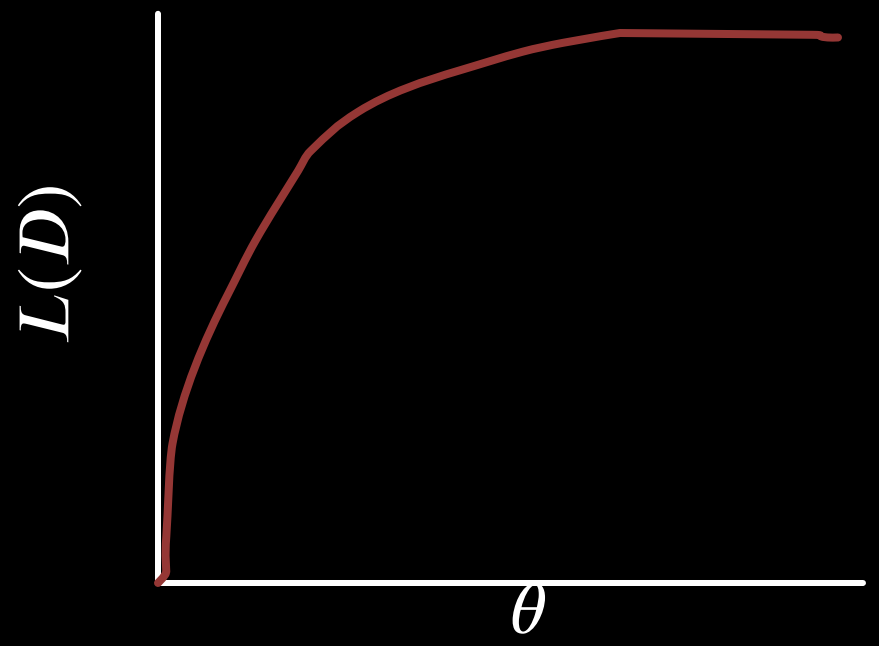
$$\text{s.t.} \quad \sum \phi_i = 1$$

ci – Cluster id of the document

ϕ_i – Probability of the document assigned to the cluster

Expectation Maximization Technique

- Likelihood is locally concave
- Start with some assumed parameters
- Maximize expectation w.r.t to the parameters
- Re-compute the parameters



Derivation K-Means using EM

- On the white board

GDA – How does it work?

- Iteratively clusters the feature space
  1. Assumes cluster centers $\mu_k, \phi_k, \sum_k$
  2. Assigns items to each cluster based on "some" similarity measure

     e.g. $p(x \mid \mu_k) = \frac{e^{-0.5(X-\mu_k)^T \sum^{-1} (X-\mu_k)}}{\sqrt[n]{2\pi} \sum^{0.5}}$
  3. Recompute cluster centers $\mu_k, \phi_k, \sum_k$
  4. Repeat step 2,and 3 till dataset likelihood ($L(D)$) converges or maximum number of iteration are reached

$$L(D) = \prod_{i=1}^{N} P(x_i \mid \mu_{ci}) * \phi_{ci}$$

s.t.

$$\sum \phi_i = 1$$

ci – Cluster id of the document

ϕ_i – Probability of the document assigned to the cluster