04/27/15



# UNSUPERVISED LEARNING

"Make sense of unlabelled data"

1. K-Means: "Show me different types of input in my dataset"

Given: unlabelled data $D = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ to some fixed integer $k$

Goal: Find k "clusters" or groupings of data

Variables:
- Cluster Centers: $\vec{\mu}_1, \vec{\mu}_2, \ldots, \vec{\mu}_K \in \mathbb{R}^d$
- Cluster-Assignment Ids: $c_1, c_2, \ldots c_N \in \{1, 2, \ldots, K\}$
- Cluster-Assignment: $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_K \in \{0,1\}^K$ 1-of-K encoding

"assignment" or "association" vector for pt $i$

$$\vec{a}_i = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \leftarrow 4^{th} \text{ cluster}$$

$a_{ij} = 1 \iff$ datapoint $i$ belongs to cluster $j$

$a_{ij} \in \{0,1\}$ $\quad \sum_{j=1}^{k} a_{ij} = 1$

K-Means Algorithm:

→ Assignment Step:

$$c_i \leftarrow \underset{j}{\operatorname{argmin}} \|\vec{x}_i - \vec{\mu}_j\|_2^2$$

Assign each data-pt to closest cluster center; closest in $L_2$-sense

→ Recenter Step

$$\mu_j \leftarrow \underbrace{\frac{1}{\#\text{pts-assigned-to-}j}}_{\sum_{i=1}^{N} a_{ij}} \sum_{a_{ij}=1} \vec{x}_i$$

$\iff$ $\mu_j$ is centroid/mean of $\vec{x}_i$ assigned to $j$

---

② K-means as an Optimization problem/alg.

→ So far, we have a "procedural" view of K-means

→ But what is k-means doing?

→ Does it ever converge?

→ What does it all mean???

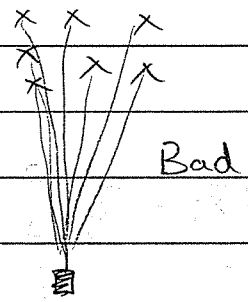
Objective:

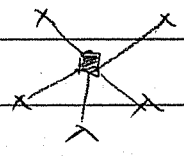
$$F(\vec{\mu}_1, \ldots, \vec{\mu}_K, c_1, \ldots, c_N) = \sum_{i=1}^{N} \|\vec{x}_i - \vec{\mu}_{c_i}\|_2^2$$

$$F(\vec{\mu}_1, \ldots, \vec{\mu}_K, \vec{a}_1, \ldots, \vec{a}_N) = \sum_{i=1}^{N} \sum_{j=1}^{K} a_{ij} \|x_i - \mu_j\|_2^2$$

"Distortion Function": sum of squared distances between pts & cluster-centers-assigned-to-them

Picture:



Under this view / objective, the optimal clustering:

$$\min_{\vec{u}_1,\dots\vec{u}_K} \min_{\vec{a}_1\dots\vec{a}_N} F(\vec{u},\vec{a})$$

Mixed Continuous / Discrete Optimization
Very very hard!

Approximate optimization Alg: Co-ordinate Descent

→ Fix $\{\vec{u}_j\}$, Optimize $\{\vec{a}_i\}$ (or $\{c_i\}$)

$$\min_{c_1\dots c_N} \sum_{i=1}^{N} \|\vec{x}_i - \vec{u}_{c_i}\|^2$$

$$= \sum_{i=1}^{N} \min_{c_i} \|\vec{x}_i - \vec{u}_{c_i}\|^2 \equiv \text{Assignment Step}$$

Set $c_i = j^*$ that minimizes distance (closest center center)

→ Fix $\{c_i\}$ (or $\{\vec{a}_i\}$), min $\{\vec{u}_j\}$

$$\min_{\vec{u}_1\dots\vec{u}_K} \sum_{i=1}^{N}\sum_{j=1}^{K} a_{ij} \|\vec{x}_i - \vec{u}_j\|^2$$

(swap)

$$= \min_{\vec{u}_1\dots\vec{u}_K} \sum_{j=1}^{K}\sum_{i=1}^{N} a_{ij} \|\vec{x}_i - \vec{u}_j\|^2$$

$$= \sum_{j=1}^{K} \min_{\vec{u}_j} \sum_{i=1}^{N} a_{ij} \|\vec{x}_i - \vec{u}_j\|^2$$

$$= \sum_{j=1}^{K} \boxed{\min_{u_j} \sum_{i|a_{ij}=1} \|u_j - x_i\|^2} \quad \text{Centroid Problem} = \text{Recenter Step}$$

Set $u_j$ = mean of $x_i$ st $a_{ij}=1$

## ③ Gaussian Mixture Model (GMM)

"Tell me how likely or rare each data point is"

Given: Data $D = \{\vec{x}_1, \ldots \vec{x}_N\}$ $\vec{x}_i \in \mathbb{R}^d$

Goal: $\vec{X} \sim P(\vec{X})$

"Fit" $P(\vec{X})$ to dataset

Note: $P(\vec{X})$ is not Gaussian (multimodal data)

But all I know are Gaussians :(

Maybe it's a "collection of Gaussians?"

Formally, we introduce a "Latent" or "Hidden" variable $Z = z \in \{1, 2, \ldots, k\}$ that indexes k Gaussians

So GMM:

① $Z \sim Cat(\vec{\pi}) \iff P(Z=c) = \pi_c$ $\quad \pi_c \geq 0$, $\sum_{c=1}^{k} \pi_c = 1$

② $\vec{X} \mid Z=c \sim N(\vec{\mu}_c, \Sigma_c)$ [ie $\vec{X}$ given $Z=c$ is a Gaussian w/ mean vector $\vec{\mu}_c \in \mathbb{R}^d$, covariance matrix $\Sigma_c \in \mathbb{R}^{d \times d}$]

$$P(\vec{X}=\vec{x} \mid Z=c) = \frac{1}{(2\pi)^{d/2} |\Sigma_c|^{1/2}} e^{\frac{-(\vec{x}-\vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x}-\vec{\mu}_c)}{2}}$$

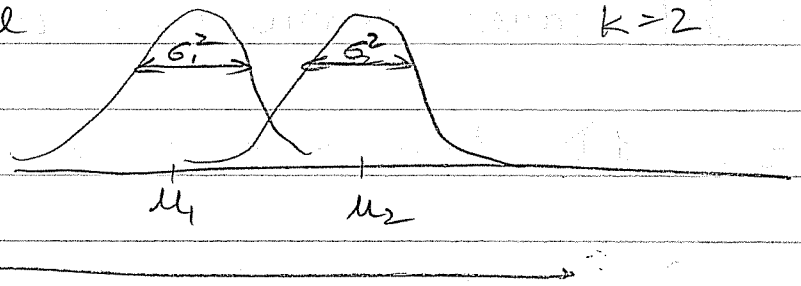
So $P(\vec{X}=\vec{x}) = \sum_{c=1}^{k} P(\vec{X}=\vec{x}, Z=c)$

$$= \sum_{c=1}^{k} P(\vec{X}=\vec{x} \mid Z=c) P(Z=c)$$

mixture $\leftarrow$ $= \sum_{c=1}^{k} \pi_c N(\vec{\mu}_c, \Sigma_c)$ $\leftarrow$ Gaussian

1D example

$d = 1$
$k = 2$




## (4) Parameter Learning in GMMs

Given: $D$

Estimate Parameters: $\{\pi_1, \ldots, \pi_K, \vec{\mu}_1, \ldots, \vec{\mu}_K, \Sigma_1, \ldots, \Sigma_K\} \equiv \vec{\theta}$

How? How else? Max-likelihood! (Max-Marginal-likelihood actually)

$$\vec{\theta}^* = \arg\max_{\theta} \prod_{i=1}^{N} P(\vec{X} = \vec{x}_i \mid \vec{\theta})$$

$$\equiv \arg\max_{\theta} \sum_{i=1}^{N} \log P(\vec{X} = \vec{x}_i \mid \theta)$$

Problem! $Z$ not available at training time

Must marginalize $\Leftrightarrow$ sum inside log

$\Leftrightarrow$ all parameters coupled

$$\vec{\theta}^* = \arg\max \sum_{i=1}^{N} \log \sum_{c=1}^{K} P(\vec{X} = \vec{x}_i, Z = c \mid \vec{\theta})$$

$$= \arg\max \sum_{i=1}^{N} \log \sum_{c=1}^{K} P(\vec{X} = \vec{x}_i \mid Z = c, \vec{\theta}) P(Z = c \mid \vec{\theta})$$

↑ Doesn't "factorize"

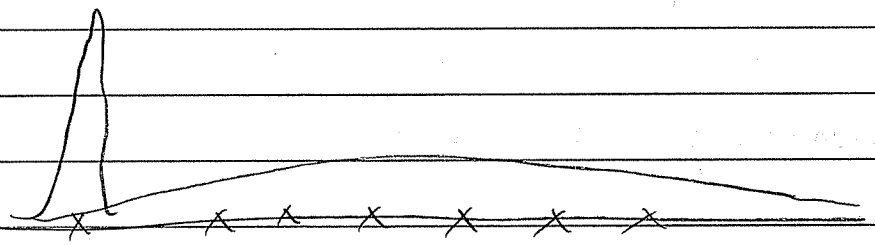
Compare to supervised learning

$$D = \{(\vec{x}_1, z_1), (\vec{x}_2, \vec{z}_2) \ldots (\vec{x}_N, z_N)\}$$

$$\vec{\theta}^* = \arg\max \sum_{i=1}^{N} \Big[ \underbrace{\log P(\vec{X} = \vec{x}_i \mid Z = z_i, \vec{\theta})}_{\text{Separate}} + \underbrace{\log P(Z = c \mid \vec{\theta})}_{\text{Separate}} \Big]$$

## (5) Hidden data causes singularities in marginal likelihood

Consider 1D data with $K=2$



What if I "fit" 1 very very peaky Gaussian to 1 point (say $x_1$) to another Gaussian to everything else.

So $\vec{\mu}_1 = \vec{x}_1$

$$P(X=x_1 \mid Z=1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{-(x_1-\mu_1)^2}{2\sigma_1^2}}$$

$$= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{0}$$

So if $\sigma_1 \to 0$ $\quad P(x_1 \mid Z=1, \vec{\theta}) \to \infty$

$\Rightarrow$ likelihood $\to \infty$

## (6) Relating Max-(Marginal) Likelihood Estimation M(M)LE in GMMs to K-Means

Assumption (1) $\pi_c = \frac{1}{K}$ All Gaussians/Clusters equally likely a priori

Assumption (2) Same Spherical Gaussian $\forall c$

$$\Sigma_c = \Sigma = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}_{d \times d} = \sigma^2 I_{d \times d}$$

$$\Leftrightarrow P(\vec{X} = \vec{x} \mid Z = c) \propto e^{-\frac{1}{2}(\vec{x} - \mu_c)^T \frac{I}{\sigma^2} (\vec{x} - \mu_c)}$$

$$= e^{-\frac{\|\vec{x} - \mu_c\|^2}{2\sigma^2}}$$

Assumption (3) "Hard Assignment": Each $\vec{x}_i$ "belongs to" exactly 1 cluster/Gaussian

$$P(\vec{X} = \vec{x}, Z = c) \propto \begin{cases} 0 & \text{if } z = 1 \\ 0 & \text{if } z = 2 \\ P(\vec{X} = \vec{x} \mid Z = c) & \\ \vdots & \\ 0 & \end{cases}$$

Assumptions (1), (2), (3)

$$\Rightarrow \text{Marginal likelihood} = \sum_{i=1}^{N} \log P(\vec{X} = \vec{x}_i \mid \vec{\theta}) = \sum_{i=1}^{N} \log \sum_c P(\vec{x}_i, c \mid \vec{\theta})$$

$$= -\sum_{i=1}^{N} \frac{\|x_i - \mu_{c_i}\|^2}{2\sigma^2} + \text{const}$$

$$\Rightarrow \boxed{\text{M(M)LE} \equiv \text{K-Means!}}$$