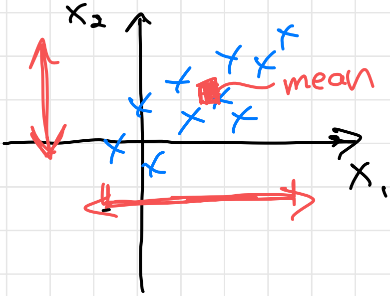
3/28/2024: Gaussian Mixture Model (GMM)

1. How to describe non-spherical cluster?
2. What is a GMM?
3. Inference – How to assign datapoint to cluster?
4. Learning – Decide shapes & locations of clusters AND assign points to clusters



$$\text{Mean} = \mu = \mathbb{E}[x] = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

Average squared distance from the mean

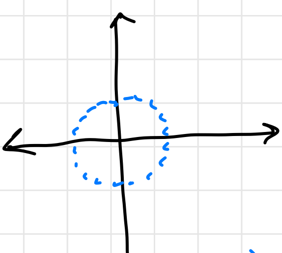
$$\text{Variance of } x_j = \mathbb{E}\left[\underbrace{(x_j - \mu_j)^2}\right] = \frac{1}{n}\sum_{i=1}^{n}(x_j^{(i)} - \mu_j)^2$$

covariance between x_1 & x_2

$$= \mathbb{E}[(x_1 - \mu_1)(x_2 - \mu_2)] = \frac{1}{n}\sum_{i=1}^{n}(x_1^{(i)} - \mu_1)(x_2^{(i)} - \mu_2)$$

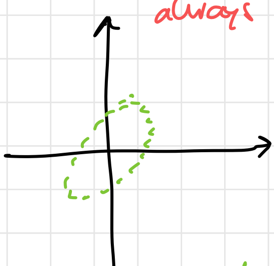
correlation between x_1 & x_2 $= \dfrac{\text{Covariance}(x_1, x_2)}{\boxed{\sqrt{\text{Var}(x_1)\,\text{Var}(x_2)}}}$

always positive



$\text{Var}(x_1) = \text{Var}(x_2) = 1$

$\text{cov}(x_1, x_2) = 0$



$\text{Var}(x_1) = \text{Var}(x_2) = 1$

$\text{cov}(x_1, x_2) > 0$

Covariance Matrix

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{pmatrix}$$

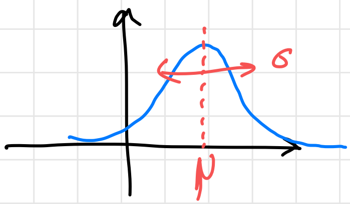
Captures
- Variance (spread) in each dimension
- Covariances (correlation) between every 2 dimensions
- Overall, says a lot about "shape" of data

Formula for

$$\Sigma = \mathbb{E}\left[(X-\mu)(X-\mu)^T \right] = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

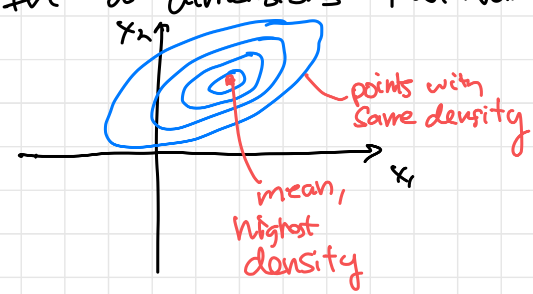
$$\underbrace{\begin{bmatrix} x_1^{(i)}-\mu_1 \\ \vdots \\ x_d^{(i)}-\mu_d \end{bmatrix}}_{d \times 1} \underbrace{\begin{bmatrix} x_1^{(i)}-\mu_1, \dots, x_d^{(i)}-\mu_d \end{bmatrix}}_{1 \times d} = \underbrace{\begin{bmatrix} (x_1^{(i)}-\mu_1)^2 & & \\ & \ddots & \\ (x_j^{(i)}-\mu_j)(x_k^{(i)}-\mu_k) & & \ddots \end{bmatrix}}_{d \times d}$$

In 1-D: Univariate Gaussian



μ = "location"
σ = "shape"

In d-dimensions: Multivariate Gaussian



mean $\mu \in \mathbb{R}^d$ "location"

and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ "shape"

In 1D:

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\sigma^2}} \cdot \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

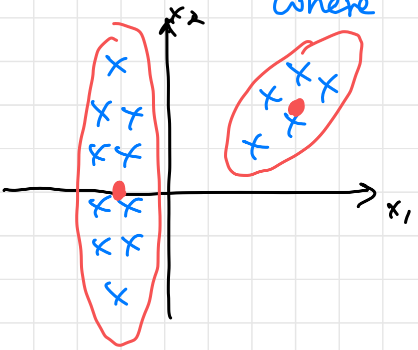
Multivariate:

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}} \cdot \frac{1}{\sqrt{\det(\Sigma)}} \cdot \exp\left(-\frac{1}{2}\underbrace{(x-\mu)^T}_{1\times d}\underbrace{\Sigma^{-1}}_{d\times d}\underbrace{(x-\mu)}_{d\times 1}\right)$$

Result: 1×1

What is a GMM?

Goal: Given dataset $\{x^{(1)},\dots,x^{(n)}\}$ produce clustering of points where clusters have custom shapes



How? For each cluster j, we will learn:

params of GMM

- μ_j : Center of each cluster "location"
- Σ_j : Covariance matrix of each cluster "shape"
- π_j : size (% of points) of each cluster "size/density"

For this dataset, we want to learn:

$$\mu_1 = \begin{bmatrix}-1\\0\end{bmatrix}, \quad \mu_2 = \begin{bmatrix}4\\2\end{bmatrix}$$

$$\Sigma_1 = \begin{pmatrix}1&0\\0&9\end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix}4&2\\2&4\end{pmatrix}$$

$$\pi_1 = \frac{2}{3}, \quad \pi_2 = \frac{1}{3}$$

Plan:
- Probabilistic Story
- Convert to loss function
- Minimize loss

Probabilistic Story:

For $i = 1,\dots,n$: — # of examples

① Randomly sample cluster z_i for point #i

$$p(z_i = j) = \pi_j$$

② Randomly sample x_i from a Multivariate Gaussian w/ mean μ_{z_i}, Σ_{z_i}

we only observe the $x^{(i)}$'s

For each i, there's:

unobserved AKA "latent"

- Random variable Z_i : denotes true cluster for i
- Random variable X_i : denotes value of example i

We observe $X_i = x^{(i)}$

Inference: Inferring the probability distribution of a latent random variable conditioned on observations

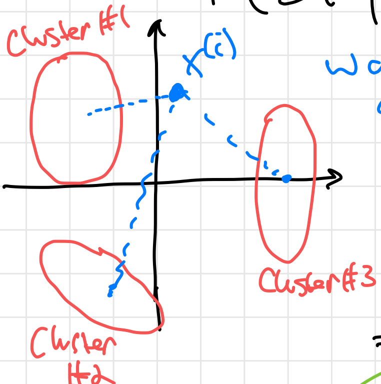
For GMMs: Given:

- Observed value $x^{(i)}$ for each X_i
- Best guess of parameters $\mu_j, \Sigma_j, \pi_j \quad \forall j$

Infer likely value of all Z_i's

i.e.:

$$P(Z_i \mid X_i = x^{(i)}; \pi_{1:k}, \mu_{1:k}, \Sigma_{1:k})$$



Was $x^{(i)}$ most likely generated from Cluster #1, 2, or 3?

How? Bayes Rule

$$P(Z_i = j \mid X_i = x^{(i)})$$

$$= \frac{P(Z_i = j) \cdot P(X_i = x^{(i)} \mid Z_i = j)}{\sum_{b=1}^{K} P(Z_i = b)\, P(X_i = x^{(i)} \mid Z_i = b)}$$

$P(Z_i = j) = \pi_j$

$= P(X_i = x^{(i)})$

$P(x^{(i)}; \mu_j, \Sigma_j)$

under multivariate Gaussian

Result: For each $x^{(i)}$, we get

$$P(Z_i = 1 \mid X_i = x^{(i)}) = 0.6$$
$$P(Z_i = 2 \mid X_i = x^{(i)}) = 0.1$$
$$P(Z_i = 3 \mid X_i = x^{(i)}) = 0.3$$

GMM vs Naive Bayes

① Unsupervised
↳ Have to keep alternating between guessing the clusters ("labels") and estimating parameters

Supervised learning
↳ First estimate all params
↳ Then predict

② We're trying to fit covariance Σ_j for each j, so x_1 & x_2 could be correlated conditioned on cluster ID

Assumes $P(x_1 \mid y)$ independent of $P(x_2 \mid y)$, etc...