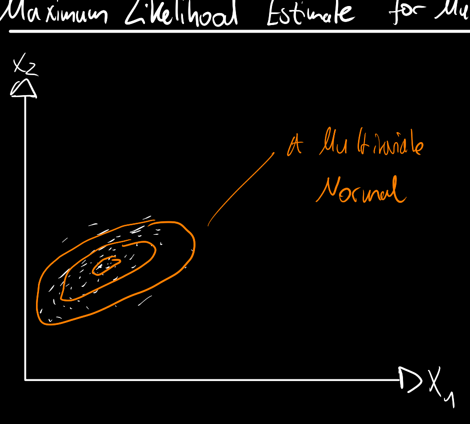
Maximum Likelihood Estimate for Multivariate Normal



$D = \{ \underbrace{(0.3, 1.7)}_{x^{[0]}}, \underbrace{(0.2, 1.6)}_{x^{[1]}} \dots \}$

$\Downarrow$

μ & Σ

$\Rightarrow$ MLE

$X_i \sim \mathcal{N}(X_i; \mu, \Sigma)$

defined in $\mathbb{R}^K$

not use D to avoid confusion with dataset

μ: mean (vector) $\in \mathbb{R}^K$

Σ: covariance (matrix) $\in \mathbb{R}^{K \times K}$

$\Sigma > 0$ $\to$ sym. positive definite

$(\Sigma = \Sigma^T)$

e.g. $K=2$ $\quad \Sigma = \begin{bmatrix} 1.7 & 0.3 \\ 0.3 & 1.5 \end{bmatrix}$ covariances, variances

$$\mathcal{N}(X_i; \mu, \Sigma) = \frac{1}{\sqrt{\det(\Sigma) \cdot (2\pi)^K}} \exp\left(-\frac{1}{2}(X_i - \mu)^T \Sigma^{-1} (X_i - \mu)\right)$$

Likelihood:

$$\mathcal{L}(D; \mu, \Sigma) \xrightarrow{\text{i.i.d.}} \prod_{i=0}^{N-1} p(X_i = x^{[i]})$$

$$= \prod_{i=0}^{N-1} \mathcal{N}(X_i = x^{[i]}; \mu, \Sigma)$$

$$= \prod_{i=0}^{N-1} \frac{1}{\sqrt{\det(\Sigma) \cdot (2\pi)^K}} \exp\left(-\frac{1}{2}(x^{[i]} - \mu)^T \Sigma^{-1} (x^{[i]} - \mu)\right)$$

$$= \det(\Sigma)^{-\frac{N}{2}} \cdot (2\pi)^{-\frac{NK}{2}} \exp\left(-\frac{1}{2}\sum_{i=0}^{N-1} (x^{[i]} - \mu)^T \Sigma^{-1} (x^{[i]} - \mu)\right)$$

don't confuse

Log-Likelihood

$$\ell(D; \mu, \Sigma) = \log(\mathcal{L}(D; \mu, \Sigma))$$

$$= -\frac{N}{2}\log\det(\Sigma) - \frac{NK}{2}\log(2\pi) - \frac{1}{2}\sum_{i=0}^{N-1}(x^{[i]} - \mu)^T \Sigma^{-1} (x^{[i]} - \mu)$$

Maximum Likelihood Estimate

$$\mu^*, \Sigma^* = \underset{\substack{\mu \in \mathbb{R}^K \\ \Sigma \in \mathbb{R}^{K \times K} \\ \Sigma > 0}}{\arg\max} \left(\ell(D; \mu, \Sigma)\right)$$

$\Sigma > 0$ — arises naturally *

Take derivative & set to zero

* More detail on this (was not 100% in the video)

Our MLE does not guarantee positive definiteness, although it guarantees symmetry.
In order to be on the safe side one can use sth called a "shrunk" covariance.

Essentially, this is: $\Sigma_{\text{shrunk}} = (1-\alpha)\Sigma_{MLE} + \alpha \frac{tr(\Sigma_{MLE})}{K} \cdot I$

which generates more diagonal dominance

μ

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{2}\sum_{i=0}^{N-1}\left(\Sigma^{-1}(x^{[i]} - \mu) + (x^{[i]} - \mu)^T \Sigma^{-1}\right) \stackrel{!}{=} 0$$

$(x^{[i]} - \mu)^T \Sigma^{-1} = \Sigma^{-1}(x^{[i]} - \mu)$

because $\Sigma = \Sigma^T$

$\hookrightarrow \Sigma^{-1} = \Sigma^{-T}$

$$= -\frac{1}{2}\sum_{i=0}^{N-1} 2 \cdot \Sigma^{-1}(x^{[i]} - \mu) \stackrel{!}{=} 0 \quad | \cdot (-1)$$

$$= \Sigma^{-1}\sum_{i=0}^{N-1}(x^{[i]} - \mu) \stackrel{!}{=} 0 \quad | \Sigma$$

$$\hookrightarrow \sum_{i=0}^{N-1}(x^{[i]} - \mu) \stackrel{!}{=} 0$$

$$\sum_{i=0}^{N-1} x^{[i]} - \underbrace{\sum_{i=0}^{N-1}\mu}_{N\mu} \stackrel{!}{=} 0$$

$$\Leftrightarrow \boxed{\mu = \frac{1}{N}\sum_{i=0}^{N-1} x^{[i]}}$$

Σ

$$\ell \doteq -\frac{N}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=0}^{N-1}\underbrace{(x^{[i]} - \mu)^T \Sigma^{-1} (x^{[i]} - \mu)}_{\text{Scalar}}$$

Σ^{-1} does not commute

$\hookrightarrow$ commutes inside trace

$tr(AB) = tr(BA)$

in general: $AB \neq BA$

$$= -\frac{N}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=0}^{N-1} tr\left((x^{[i]} - \mu)^T \Sigma^{-1} (x^{[i]} - \mu)\right)$$

$$= -\frac{N}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=0}^{N-1} tr\left(\underbrace{(x^{[i]} - \mu)(x^{[i]} - \mu)^T}_{\text{outer product}} \Sigma^{-1}\right)$$

$S = (x^{[i]} - \mu)(x^{[i]} - \mu)^T$

$= [\dots] \in \mathbb{R}^{K \times K}$

$\hookrightarrow$ Matrix Cookbook

$$\frac{\partial \det(A)}{\partial A} = \det(A) \cdot A^{-T} \qquad \frac{\partial tr(AB^{-1})}{\partial B} = -(B^{-1}AB^{-1})^T$$

$$\frac{\partial \ell}{\partial \Sigma} = -\frac{N}{2}\frac{1}{\cancel{\det(\Sigma)}} \cdot \cancel{\det(\Sigma)}\, \Sigma^{-T} + \frac{1}{2}\sum_{i=0}^{N-1}(\Sigma^{-1} S \Sigma^{-1})^T \stackrel{!}{=} 0$$

$= \Sigma^{-1}$

$$= -\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\sum_{i=0}^{N-1}\Sigma^{-T} S^T \Sigma^{-T} \stackrel{!}{=} 0 \quad | \cdot 2$$

$$= -N\Sigma^{-1} + \sum_{i=0}^{N-1}\Sigma^{-1}\left((x^{[i]} - \mu)(x^{[i]} - \mu)^T\right)^T \Sigma^{-1} \stackrel{!}{=} 0$$

$$= -N\Sigma^{-1} + \Sigma^{-1}\sum_{i=0}^{N-1}\left((x^{[i]} - \mu)(x^{[i]} - \mu)^T\right)\Sigma^{-1} \stackrel{!}{=} 0$$

$\Sigma \,|\, \Sigma$

$$= -N\Sigma + \sum_{i=0}^{N-1}(x^{[i]} - \mu)(x^{[i]} - \mu)^T \stackrel{!}{=} 0$$

$$\boxed{\Sigma = \frac{1}{N}\sum_{i=0}^{N-1}(x^{[i]} - \mu)(x^{[i]} - \mu)^T}$$

MLE

$[\,]^{[i]} \; [\dots]^{[i]}$

$= [\dots]^{[i]} \in \mathbb{R}^{K \times K}$

data matrix

$$X = \begin{bmatrix} \text{—} & x^{[0]} & \text{—} \\ \text{—} & x^{[1]} & \text{—} \\ & \vdots & \\ \text{—} & x^{[N-1]} & \text{—} \end{bmatrix} \in \mathbb{R}^{N \times K}$$

different way to express the dataset

n_samples

K dim ($\hat{=}$ features)

Mean alongside rows $\mu = [\dots] \in \mathbb{R}^K$

Centered data matrix

$$\tilde{X} = X - \underbrace{\begin{bmatrix} \text{—} & \mu & \text{—} \\ \text{—} & \mu & \text{—} \\ & \vdots & \\ \text{—} & \mu & \text{—} \end{bmatrix}}_{\in \mathbb{R}^{N \times K}}$$

$$\tilde{X} = X - 1_N \mu^T$$

$1_N = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

$$\Sigma_{MLE} = \frac{1}{N}\tilde{X}^T \tilde{X} \in \mathbb{R}^{K \times K}$$

$= \frac{1}{N} \cdot K$ $\cancel{N}$ $\cancel{N}$ K

$= \frac{1}{N} \cdot K$ K