Lect 2: Gaussian Distribution
and 3.

Gaussian Distribution: Extremely important.

Univariate Gaussian:

Toss a fair coin (head / tail $p=0.5$)
Define the random variable

$$X = \begin{cases} 1 & \text{if head} \\ 0 & \text{if tail.} \end{cases}$$

We will toss the coin N times
and count the # of heads, m.
calculate m/N.

The probability density function is

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

mean variance.

$\theta = (\mu, \sigma^2)$

In MLE:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log P(D|\theta)$$

$$= \underbrace{\arg\max_{\theta} \sum_{i=1}^{N} \log P(x^{(i)} \mid \theta)}.$$

$$\Downarrow$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(x^{(i)}-\mu)^2/2\sigma^2\right).$$

① Find μ_{MLE} first.

$$\boxed{\arg\max_{\mu} -\sum_{i=1}^{N} (x^{(i)}-\mu)^2.}$$

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} (x^{(i)}-\mu)^2 = 0 \Rightarrow \mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

② Find σ_{MLE}.

$$\arg\max \sum_{i=1}^{N} \left[-\log \sigma - \frac{1}{2\sigma^2} (x^{(i)}-\mu)^2\right]$$

$$\frac{\partial}{\partial \sigma} \left[\sum_{i=1}^{N} \log \sigma + \frac{1}{2\sigma^2} (x^{(i)}-\mu)^2\right] = 0$$

$$\Rightarrow \sigma^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)}-\mu)^2.$$

③ μ_{MLE} is unbiased

$$E_D[\mu_{MLE}] = E_D\left[\frac{1}{N}\sum_{i=1}^{N} x^{(i)}\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E_D[x^{(i)}] = \mu.$$

④ σ_{MLE} is biased. if we replace μ by μ_{MLE}.

$$E_D[\sigma^2_{MLE}] = E_D\left[\frac{1}{N}\sum_{i=1}^{N}(x^{(i)}-\mu_{MLE})^2\right]$$

$$= E_D\left[\frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)2} - 2x^{(i)}\mu_{MLE} + \mu^2_{MLE}\right)\right]$$

$$= E_D\left[\frac{1}{N}\sum_{i=1}^{N} x^{(i)2}\right] - E_D[\mu^2_{MLE}]$$

$$= E_D\left[\frac{1}{N}\sum_{i=1}^{N} x^{(i)2} - \mu^2\right] - E_D[\mu^2_{MLE} - \mu^2]$$

$$= \sigma^2 - \left(E_D[\mu^2_{MLE}] - E_D^2[\mu_{MLE}]\right)$$

$$= \sigma^2 - \mathrm{var}[\mu_{MLE}] = \sigma^2 - \mathrm{var}\left[\frac{1}{N}\sum_{i=1}^{N} x^{(i)}\right]$$

$$= \sigma^2 - \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{var}[x^{(i)}] = \sigma^2 - \frac{1}{N}\sigma^2$$

$$= \frac{N-1}{N}\sigma^2$$

The unbiased estimator is

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x^{(i)}-\mu_{MLE})^2$$

Multivariate Gaussian Distribution.

pdf is

$$p(\vec{x};\vec{\mu},\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(x-\mu)\right)$$

$\vec{x}, \vec{\mu} \in \mathbb{R}^D$ and $\Sigma \in \mathbb{R}^{D\times D}$.

Σ: positive semi-definite matrix (PSD)

i.e. for any $\vec{v} \in \mathbb{R}^D$, $\vec{v}^T\Sigma\vec{v} \geq 0$

Here we consider Σ as PD for now.

$|\Sigma|$ is the determinant of Σ.

eigenvalue decomposition:

$$\Sigma = U\Lambda U^T = [\vec{u}_1|\vec{u}_2|\cdots|\vec{u}_D]\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_D\end{bmatrix}[\vec{u}_1|\vec{u}_2|\cdots|\vec{u}_D]^T$$

$$= \sum_{i=1}^{D}\vec{u}_i\lambda_i\vec{u}_i^T$$

so $\Sigma^{-1} = \underset{\substack{\uparrow \\ \text{orthogonal matrix i.e. } UU^T = I.}}{U}\Lambda^{-1}U^T = \sum_{i=1}^{D}\vec{u}_i\frac{1}{\lambda_i}\vec{u}_i^T$

$$\Delta = (\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})$$

$$= \sum_{i=1}^{D} \underbrace{(\vec{x}-\vec{\mu})^T \vec{u}_i}_{\triangleq\, y_i} \frac{1}{\lambda_i} \underline{\vec{u}_i^T (\vec{x}-\vec{\mu})} = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

y_i is the projected vector of $(\vec{x}-\vec{\mu})$ on the eigenvector $\vec{u}_i$

equation of ellipsoid

$D_M(\vec{x})$ = Mahalanobis distance.

$$= \sqrt{(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}$$

MLE for θ:

$$\theta_{MLE} = \arg\max_{\theta} \sum_{i=1}^{N} \log P(x^{(i)} | \mu, \Sigma)$$

$$= -\frac{Nd}{2} \log(2\pi) + \frac{N}{2} \log(|\Lambda|)$$

$$- \frac{1}{2} \sum_{i=1}^{N} \left[(x^{(i)} - \vec{\mu})^T \Lambda (x^{(i)} - \mu) \right] \triangleq l(\mu, \Sigma)$$

where

$\Lambda = \Sigma^{-1}$ is the $\boxed{\text{precision matrix}}$

Fact:

① Trace is invariant under cyclic permutation of Matrix Product.

$$tr(ABC) = tr(BCA) = tr(CAB)$$

② $\frac{\partial}{\partial A} tr(AB) = \frac{\partial}{\partial B} tr(BA) = B^T$

③ $\frac{\partial}{\partial A} \log|A| = A^{-T}$ ④ $|A| = \frac{1}{|A^{-1}|}$

$$\frac{\partial \ell(\mu, \Sigma)}{\partial \mu} = \sum_{i=1}^{N} \Sigma^{-1}(x^{(i)} - \mu) = 0$$

$$\Leftrightarrow \Sigma^{-1} \sum_{i=1}^{N} (x^{(i)} - \mu) = 0$$ assume Σ is PD

$$\Rightarrow \mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} = \text{empirical mean.}$$

$$\ell(\mu, \Sigma) = -\frac{Nd}{2}\log(2\pi) + \frac{N}{2}\log(|\Lambda|) - \frac{1}{2}\sum_{i=1}^{N} tr\left((x^{(i)} - \mu)(x^{(i)} - \mu)^T \Lambda\right)$$

By trace trick.

$$= -\frac{Nd}{2}\log(2\pi) + \frac{N}{2}\log(|\Lambda|)$$

$$-\frac{1}{2}\operatorname{tr}(S_{\bar{x}}\Lambda).$$

where $S_{\bar{x}} = \sum_{i=1}^{N}(x^{(i)}-\mu)(x^{(i)}-\mu)^T.$

$$= \sum_{i=1}^{N} x^{(i)}(x^{(i)})^T - N\mu\mu^T$$

$\boxed{\text{scatter matrix}}$ centered on μ

Rewrite scatter matrix into a compact form.

$$S_{\bar{x}} = \tilde{X}^T\tilde{X} = X^T C_N^T C_N X = X^T C_N X$$

where $C_N = I_N - \frac{1}{N}1_N 1_N^T$

is the centering matrix

$$\frac{\partial l}{\partial \Lambda} = -\frac{N}{2}\Sigma + \frac{1}{2}S_{\bar{x}} = 0$$

Then $\Sigma_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)}-\mu)(x^{(i)}-\mu)^T.$

$\Downarrow$

① $E[\Sigma_{MLE}] = \frac{N-1}{N}\Sigma.$

② degree of freedom of θ is.

$$\underset{\mu}{\underset{\downarrow}{D}} + \underset{\Sigma}{\underset{\downarrow}{\frac{D(D+1)}{2}}} \sim O(D^2).$$

↳ overfitting.

possible way: assume Σ is a diagonal matrix.

Common theorems.

Denote $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix} \in R^D = \begin{bmatrix} \vec{x}_a \\ \vec{x}_b \end{bmatrix}$ $\quad \vec{x}_a \in R^m$, $\vec{x}_b \in R^n$

$(m+n=D)$.

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

$\vec{\mu} = (\vec{\mu}_a, \vec{\mu}_b)^T$ and $\vec{x} \sim N(\vec{\mu}, \Sigma)$.

① If $\vec{x} \sim N(\vec{\mu}, \Sigma)$ and $\vec{y} = Ax + b$

then $\vec{y} \sim N(A\mu + b, A\Sigma A^T)$

proof:

$$E[y] = E[Ax+b] = AE[x]+b = A\mu+b$$

$$var[y] = var[Ax+b] = var[Ax] = A \cdot var[x] \cdot A^T$$

② Find $P(x_a)$, $P(x_b)$, $P(x_a|x_b)$, $P(x_b|x_a)$

$$\vec{x_a} = \begin{bmatrix} I_{m\times m} & O_{m\times n} \end{bmatrix} \begin{bmatrix} \vec{x_a} \\ \vec{x_b} \end{bmatrix}$$

$\Downarrow$

$$E[\vec{x_a}] = [I, O] \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} = \mu_a.$$

$$var[\vec{x_a}] = [I, O] \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \begin{bmatrix} I \\ O \end{bmatrix}$$

$$= \Sigma_{aa}$$

so $\vec{x_a} \sim N(\mu_a, \Sigma_{aa})$

similarly: $\vec{x_b} \sim N(\mu_b, \Sigma_{bb})$

Define: ① $\vec{x}_{b|a} = \vec{x}_b - \Sigma_{ba}\Sigma_{aa}^{-1}\vec{x}_a$

② $\vec{\mu}_{b|a} = \vec{\mu}_b - \Sigma_{ba}\Sigma_{aa}^{-1}\vec{\mu}_a$

③ $\Sigma_{b|a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}$

Schur Complementary of Σ_{bb}.

$$x_{b|a} = \left[-\Sigma_{ba}\Sigma_{aa}^{-1},\ I_{n\times n}\right]\begin{bmatrix} x_a \\ x_b \end{bmatrix}$$

So $E[x_{b|a}]$

$$= \left[-\Sigma_{ba}\Sigma_{aa}^{-1},\ I_{n\times n}\right]\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} = \vec{\mu}_{b|a}$$

$\mathrm{var}[\vec{x}_{b|a}]$

$$= \left[-\Sigma_{ba}\Sigma_{aa}^{-1},\ I_{n\times n}\right]\begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\begin{bmatrix} -\Sigma_{aa}^{-1}\Sigma_{ba}^{T} \\ I_{n\times n} \end{bmatrix}$$

$$= \Sigma_{b|a}$$

Then $\vec{x}_b = \vec{x}_{b|a} + \Sigma_{ba}\Sigma_{aa}^{-1}\vec{x}_a$.

step 1: $\operatorname{cov}(\vec{X}_{b|a}, \vec{X}_a)$

$$= \operatorname{cov}(\vec{X}_b - \Sigma_{ba}\Sigma_{aa}^{-1}\vec{X}_a, \vec{X}_a)$$

$$= \Sigma_{ba} - \Sigma_{ba}\Sigma_{aa}^{-1} \cdot \Sigma_{aa} = 0$$

$\Rightarrow \vec{X}_{b|a}$ and $\vec{X}_a$ are unrelated.

They are jointly normal so they are independent

step 2:

$E[\vec{X}_b | \vec{X}_a]$

$$= E[\vec{X}_{b|a} + \Sigma_{ba}\Sigma_{aa}^{-1}\vec{X}_a \mid \vec{X}_a]$$

$$= E[\vec{X}_{b|a}] + \Sigma_{ba}\Sigma_{aa}^{-1}\vec{X}_a = \vec{\mu}_{b|a} + \Sigma_{ba}\Sigma_{aa}^{-1}\vec{X}_a$$

$$= \vec{\mu}_b + \Sigma_{ba}\Sigma_{aa}^{-1}(\vec{X}_a - \vec{\mu}_a).$$

step 3:

$\operatorname{var}[\vec{X}_b | \vec{X}_a]$

$$= \operatorname{var}[\vec{X}_{b|a} + \Sigma_{ba}\Sigma_{aa}^{-1}\vec{X}_a | \vec{X}_a]$$

$$= \operatorname{var}[\vec{X}_{b|a}] = \Sigma_{b|a}.$$