Likelihood

$D = \{x_1, x_2, x_3 \dots, x_n\}$

Categorical → $P(X = x)$

Continuous → $p(x)$

Let x be continuous

$p(x_1)$ → pdf value at x_1

$p(x_2)$ → pdf ... x_2

$p(x_n)$ → ... .

$p(x_1, x_2, x_3, \dots, x_n) = p(x_1)\, p(x_2) \dots p(x_n)$

↳ assuming independence

i.i.d

↑

independent & identically distributed

$$= \prod_{i=1}^{n} p(x_i)$$

$$p(D) = \prod_{i=1}^{n} p(x_i)$$

↑ likelihood.

$$\log p(D) = \sum_{i=1}^{n} \log(p(x_i))$$

log-likelihood

$D = \{1, 4, 16, 64\}$

What is likelihood of D given h.

$$\underline{\ell(D|h)} = \prod_{i=1}^{4} p(x_i|h)$$

$$= p(1|h)\, p(4|h)\, p(16|h)\, p(64|h)$$

$\underline{h \rightarrow \text{all even numbers.}}$

$p(1 \mid \text{even numbers}) = 0$  $\ell(D|h) = 0$

$\underline{h \rightarrow \text{odd numbers.}}$ ... $\ell(D|h) = 0$

$\underline{h \rightarrow \text{squares}}$

$1, 4, 9, 16, 25, \ldots\ldots, 100$

10 possible "squares"

$p(1 \mid \text{squares}) = \frac{1}{10}$

$p(4 \mid \text{squares}) = \frac{1}{10}$

$p(16 \mid \text{squares}) = \frac{1}{10}$

$p(64 \mid \text{squares}) = \frac{1}{10}$

$\ell(D \mid \text{squares}) = \frac{1}{10} * \frac{1}{10} * \frac{1}{10} * \frac{1}{10} = 10^{-4}$

$\underline{h \rightarrow \text{Powers of 4}}$ $\underline{1, 4, 16, 64}$

$p(1 \mid \text{powers of 4}) = \frac{1}{4}$

$p(64 \mid \text{powers of 4}) = \frac{1}{4}$

$\ell(D \mid \text{powers of 4}) = \frac{1}{4} * \frac{1}{4} * \frac{1}{4} * \frac{1}{4} = \underline{3.906 * 10^{-3}}$

$h \rightarrow$ all numbers between 1 and 100

$p(1 \mid \text{all numbers}) = \frac{1}{100}$

$\vdots$

$l(D \mid \text{all numbers}) = \frac{1}{100} * \frac{1}{100} * \frac{1}{100} * \frac{1}{100} = 10^{-8}$

Prior $p(h)$

Posterior $p(h \mid D) = \dfrac{p(D \mid h)\, p(h)}{\sum_{\text{all } h'} p(D \mid h')\, p(h')}$

$p(\text{powers of 4} \mid D) = \dfrac{3.906 * 10^{-3} * 0.075}{0 * 0.075 + 0 * 0.3 + \cdots + \cdots}$

Inference D – Training data. $p(h)$ – prior.

new data example x^*

What is the prob that x^* was also generated by the same hypothesis as D

Option 0. $p(x^* \mid h_{\text{prior}})$

$x^* = 4$

$p(4 \mid h_{\cdot}) = \frac{1}{100}$

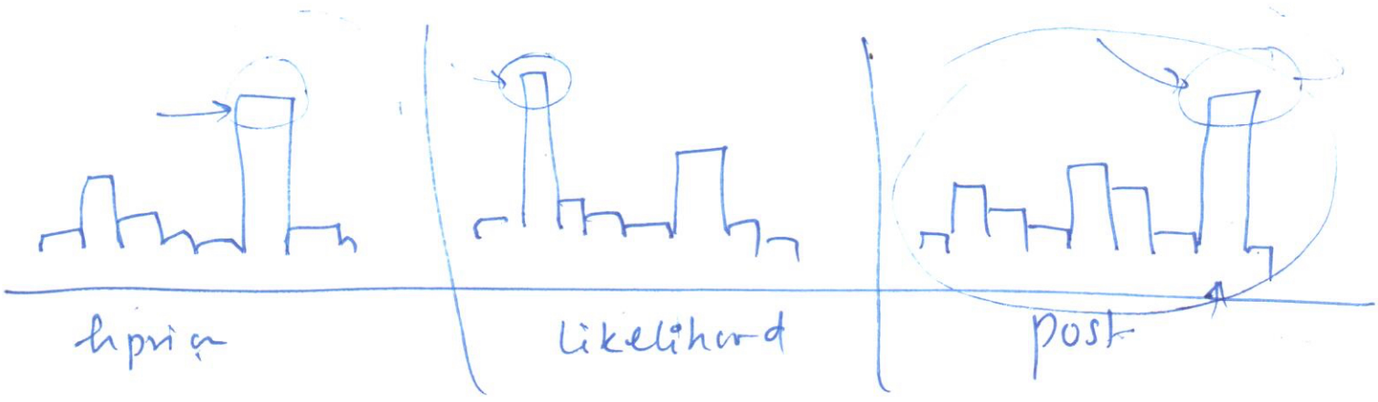
example

let $P(h \rightarrow$ numbers between 1 & 100 be the highest prior.

Option 1 $p(x^* | h_{MLE})$ Let $h \rightarrow$ powers of 4 have the high likelihood

$$p(x^* | D) = p(x^* | h_{MLE}) = \frac{1}{4}$$

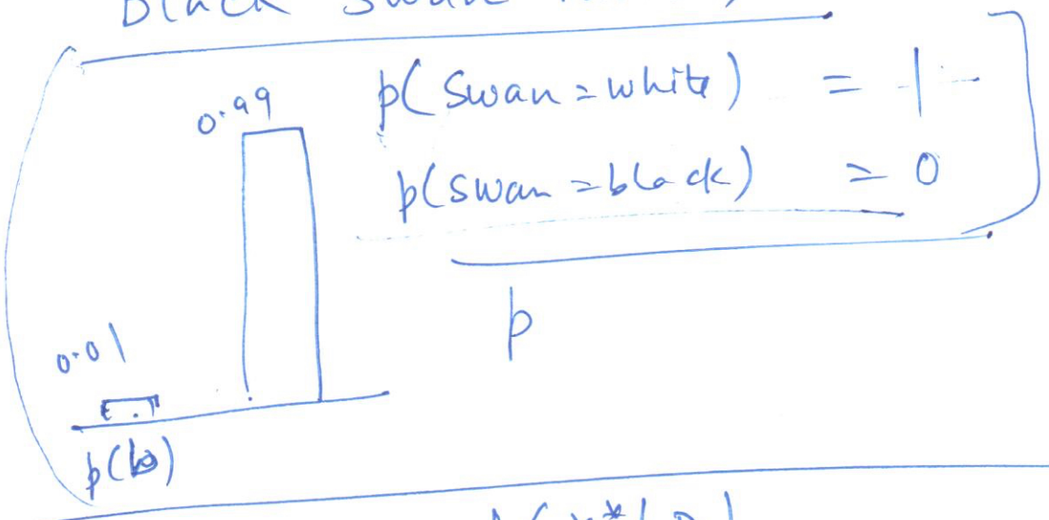
Option 2 $p(x^* | h_{MAP})$



Black Swan Paradox

$p(\text{Swan} = \text{white}) = 1$

$p(\text{swan} = \text{black}) \geq 0$



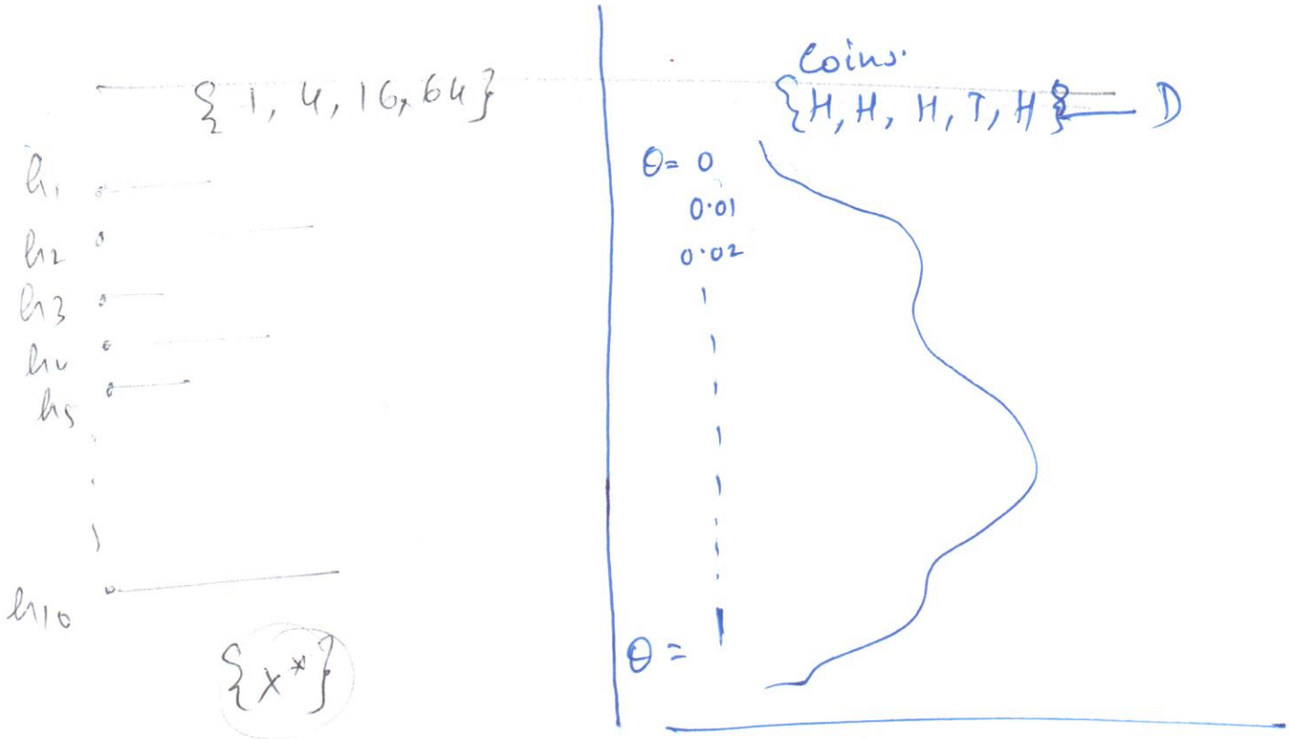
X

Option 3. $p(x^* | D)$

Bayesian Averaging

Daniel Sarlo. (F)

Dan Goings. (MF)



D = {H, H, H, T, T, H}

θ → Prob. of x to be H

likelihood of D = $P(X_1|\theta)\ P(X_2|\theta) \cdots$

$= \theta * \theta * \theta * (1-\theta) * (1-\theta) * \theta$

$= \theta^4 (1-\theta)^2$

if D has N_1 heads and N_0 tails.

$l(D|\theta) = \theta^{N_1} (1-\theta)^{N_0}$

Taking logs:

$$ll(D|\theta) = N_1 \log \theta + N_0 \log (1-\theta)$$

Find θ that maximizes $ll(D|\theta)$

θ_{max} will be when : $\frac{d}{d\theta} ll(D|\theta) = 0$

$$\frac{N_1}{\theta} + \frac{N_0}{(1-\theta)}(-1)$$

$$= \frac{N_1}{\theta} - \frac{N_0}{(1-\theta)}$$

Setting to 0

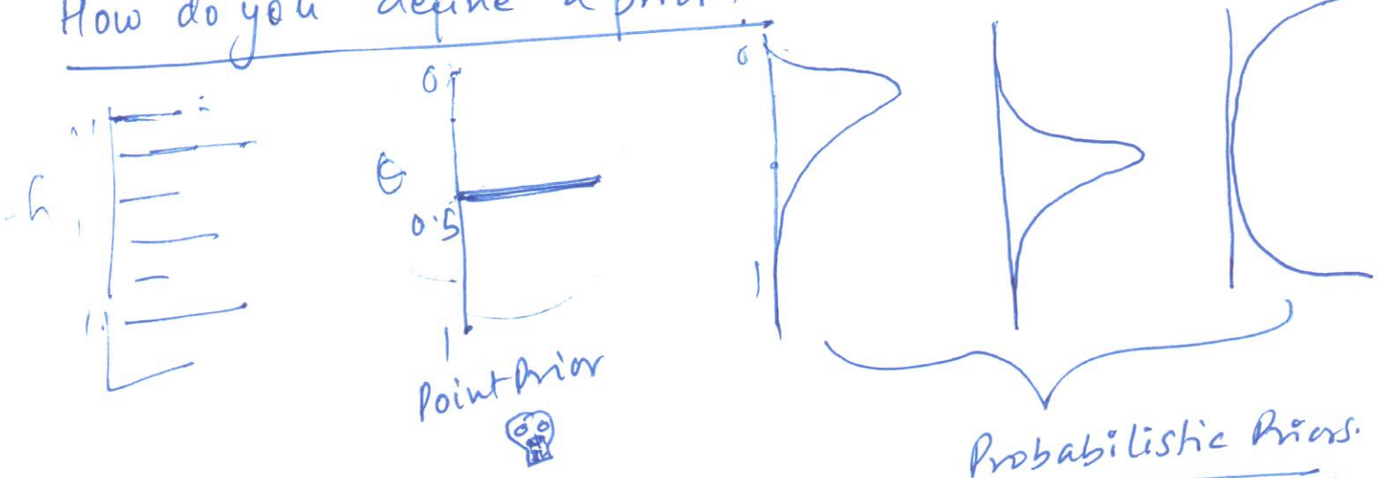
$$\frac{N_1}{\theta} = \frac{N_0}{1-\theta} \Rightarrow N_1 - N_1\theta = N_0\theta$$

$$\theta = \frac{N_1}{N_0+N_1} \rightarrow \text{MLE for } \theta$$

N = ~~Total~~ Size of D : ~~$N_0+N_1=N$~~

$P(x^* = \text{heads} \mid D) = \theta_{MLE} = \frac{N_1}{N_0+N_1}$

How do you define a prior?



Use a probability distribution to represent prior.

$\theta \rightarrow$ also a random variable

$\theta \in (0,1)$

Beta Distribution

a , b.

$p(\theta = 0.7 \mid a, b) = \theta^{a-1}(1-\theta)^{b-1}$

If $\theta = 0.7 \quad a = 3 \quad b = 2$

$p(0.7 \mid 3, 2) = (0.7)^2 (1-0.7)^1$

$= \_\_\_\_\_$

$E[\theta] = \int \theta \, p(\theta \mid a, b) \, d\theta$

$= \int \theta \; \theta^{a-1}(1-\theta)^{b-1} \, d\theta$

skipping Calculus

$= \frac{a}{a+b}$

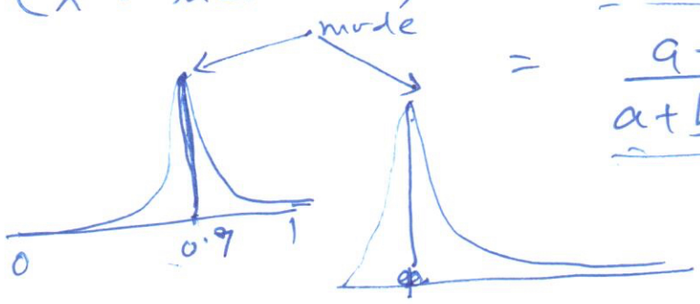
$Var[\theta] = E[(\theta - E(\theta))^2] = \frac{ab}{(a+b)^2(a+b+1)}$

If prior for θ is a Beta(a,b)

Ignoring D:

$P(x^* = \text{heads}) =$ θ that maximize pdf.

$= \frac{a-1}{a+b-2}$ [Same as the expected value of θ]



We have likelihood $p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$

prior $p(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$

We need posterior $p(\theta|D) = \dfrac{p(D|\theta)\,p(\theta)}{\int_{\theta\in 0,1} p(D|\theta)\,p(\theta)\,d\theta}$

Bayes Rule.

for some pairs of $p(D|\theta)$ (lik.) and $p(\theta)$ (prior)

computing $\int p(D|\theta)\,p(\theta)\,d\theta$ is easy.

→ Conjugate pairs.

$\int e^x dx = e^x$

$\int x^2 dx = \frac{1}{3}x^3$

$\int f(x) dx$

If $p(\theta)$ and $p(D|\theta)$ are conjugates then $p(\theta|D)$ will have the same form as $p(\theta)$

$p(\theta|D) := \text{Beta}(a+N_1-1,\ b+N_0-1)$

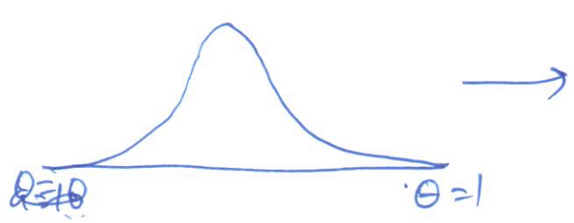
($p(\theta|D)$: posterior)

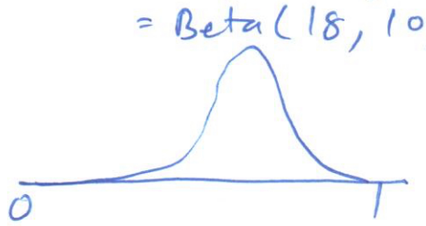
$N_1 = 9 \quad N_0 = 1$

let $a = 10 \quad b = 10$

$\text{Beta}(10+9-1,\ 10+1-1)$

$= \text{Beta}(18, 10)$





Learn parameters (θ) of a Bernoulli distribution, given some samples drawn from it.

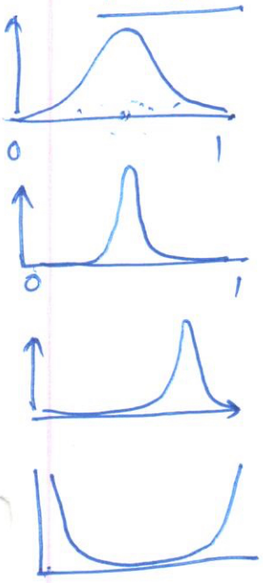
$0 \le \theta \le 1$

$D = \{H, H, H, T, T, H\}$

N_0 – # Tails in D

N_1 – # Heads in D

Prior.



Beta(a, b)

$\hat{\theta}_{Prior} = \max_{\theta} pdf(\theta | a, b)$

$= \frac{a-1}{a+b-2}$

(0)

MLE

$\hat{\theta}_{MLE} = \frac{N_1}{N_0 + N_1}$

(1)

Posterior

$p(\theta | D) = \frac{p(D|\theta)\, p(\theta)}{\int p(D|\theta)\, p(\theta)\, d\theta}$

posterior

Conjugate pairs.

$p(\theta | D)$ is also a Beta distribution (a', b')

$a' = a + N_1$

$b' = b + N_0$

$\max_{\theta} pdf(\theta | a', b')$

$\hat{\theta}_{MAP} = \frac{a + N_1 - 1}{a + b + N_0 + N_1 - 2}$

Given a new instance x^* what is $P(x^* = 1 | D)$

- $\hat{\theta}_{Prior}$
- $\hat{\theta}_{MLE}$
- $\hat{\theta}_{MAP}$ → Frequentist.
- Bayesian Averaging. → Bayesian.

Side discussion

$$\text{Beta}(\theta \mid a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta \mid D) = \frac{\theta^{N_1}(1-\theta)^{N_0}\,\theta^{a-1}(1-\theta)^{b-1}}{\int_\theta \theta^{N_1}(1-\theta)^{N_0}\,\theta^{a-1}(1-\theta)^{b-1}\,d\theta}$$

$$= \frac{\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}}{\int \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}\,d\theta}$$

$$= \frac{1}{Z} * \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

$Z \rightarrow$ constant

$$\boxed{p(\theta \mid D) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}}$$

Gaussian Distribution

$$x \sim \mathcal{N}(x \mid \mu, \Sigma)$$

$d \times 1$ $\quad d \times 1$ $\quad d \times d$

Let $d = 1$

$$x \sim N(x \mid \mu, \sigma^2)$$

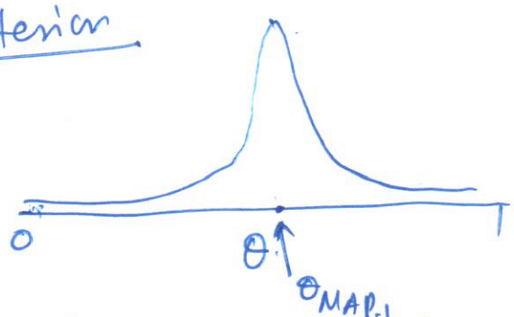
μ: scalar $\quad \sigma^2$: scalar

$D = (x_1, x_2, x_3, \ldots x_N) \leftarrow$ data.

Likelihood of D, $\mathcal{L}(D) = p(x_1 \mid \mu, \sigma^2) * p(x_2 \mid \mu, \sigma^2) \ldots$

log-lik: $ll(D) = \sum_{i=1}^{N} \log p(x_i \mid \mu, \sigma^2)$

Posterior



x^*

θ	
0.1	$0.1 * p(\theta = 0.1 \mid D)$
0.2	$0.2 * p(\theta = 0.2 \mid D)$
0.3	$0.3 * p(\theta = 0.3 \mid D)$
0.4	1
0.5	1
0.6	1
0.7	1
0.8	1
0.9	1
1.0	

$$\sum \theta \; p(\theta \mid D)$$

This is the same expression as $\mathbb{E}[\theta]$ posterior.

$$\int \theta \; p(\theta \mid D) \, d\theta$$

$$p(x^* = 1 \mid D) = \frac{a + N_1}{a + b + N_0 + N_1}$$

$$ll(D) = \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\right]$$

$$= \sum_{i=1}^{N}\left[\log\frac{1}{\sqrt{2\pi}\sigma} + \left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\right]$$

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

$$\frac{\partial ll(D)}{\partial \mu} = \sum_{i=1}^{N}\left[\frac{\partial}{\partial\mu}\left(-\frac{1}{2}\log(2\pi\sigma^2)\right) - \frac{1}{2\sigma^2}\frac{\partial}{\partial\mu}(x_i-\mu)^2\right]$$

$$= \sum_{i=1}^{N} \frac{1}{2\sigma^2} 2(x_i-\mu)(-1)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu)$$

$$\frac{\partial ll(D)}{\partial \mu} = 0 \Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0$$

$$\Rightarrow \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

MLE — Sample mean

MLE for Gaussian Distributions.

$x \in \mathbb{R}$ $\quad d=1$

$$pdf(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$D = \{x_1, x_2 \cdots x_N\}$

$$\ell(D|\mu,\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$ll(D|\mu,\sigma^2) = \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]$$

To get MLE for μ, σ:

$$\frac{\partial}{\partial \mu} ll(D|\mu,\sigma^2) = 0 \quad \Rightarrow \quad \boxed{\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

$$\frac{\partial}{\partial \sigma} ll(D|\mu,\sigma^2) = \sum_{i=1}^{N} \frac{\partial}{\partial \sigma}\left[\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \sum_{i=1}^{N} \left[-\frac{\partial}{\partial \sigma}(\log \sqrt{2\pi}\sigma) - \frac{\partial}{\partial \sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]\right]$$

$$= \sum_{i=1}^{N} \left[\frac{1}{\sigma} - (x_i-\mu)^2 (-2\sigma^{-3})\right]$$

$\log(\sqrt{2\pi}\sigma)$
$= \log\sqrt{2\pi} + \log\sigma$

Setting this to 0

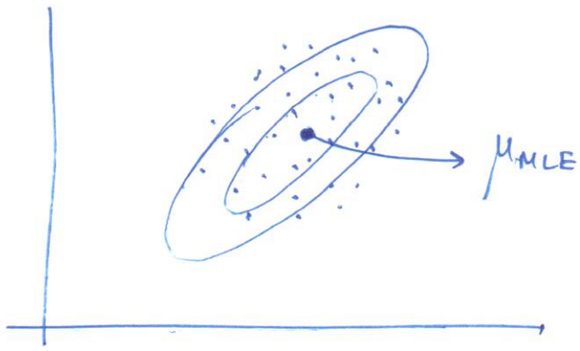
$$\boxed{\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{MLE})^2}$$

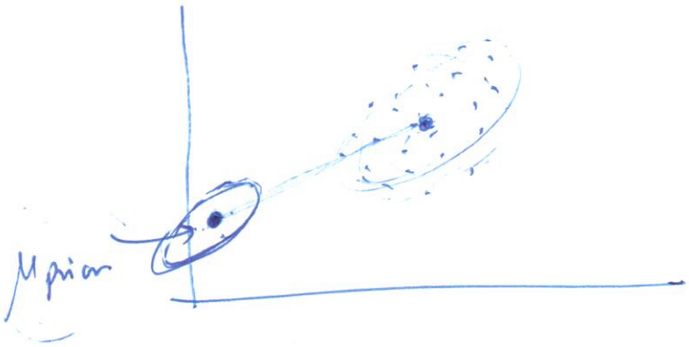
If $d > 1$ $\quad pdf(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \Sigma_{MLE} = \frac{1}{N}\sum_{i=1}^{N} \underbrace{(x_i-\mu_{MLE})}_{D\times 1} \underbrace{(x_i-\mu_{MLE})^T}_{1\times D}$$

← Matrix Cookbook Handouts.

$d = 2$





$$D = \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix}\begin{bmatrix} 3, 2 \\ 1, 3 \\ 2, 1 \end{bmatrix} \qquad \mu = \frac{1}{N}\sum x_i$$

$$\mu = \frac{3+1+2}{3}, \frac{2+3+1}{3}$$

$$= [2, 2]^T$$

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

$$x_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \quad x_2 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \mu = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$\Sigma_{MLE} = \frac{1}{3}\left[(x_1-\mu)(x_1-\mu)^T + (x_2-\mu)(x_2-\mu)^T + (x_3-\mu)(x_3-\mu)^T\right]$$

$$x_1-\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad (x_1-\mu)^T = [1 \quad 0]$$

$$(x_1-\mu)(x_1-\mu)^T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$(x_2-\mu)(x_2-\mu)^T = \begin{bmatrix} -1 \\ 1 \end{bmatrix}[-1 \quad 1] = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

$$(x_3-\mu)(x_3-\mu)^T = \begin{bmatrix} 0 \\ -1 \end{bmatrix}[0 \quad -1] = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma_{MLE} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 2/3 & -1/3 \\ 0 & 2/3 \end{bmatrix}$$

$X = N \times D$

np.mean(X)

np.mean(X, axis=0)

np.cov(X) — $N \times N$

np.cov(X.T) ✓