Lecture 9

*) MLE → You may not have likelihood parameters.

$p(x|\omega_1) \sim N(\mu_1, \sigma_1^2)$ μ_1, σ_1^2 are not known.

N iids. x_i's $\forall\ i = 1, \text{to}, N$

likelihood function $F(\theta) = \ln \prod_{i=1}^{N} p(x_i|\theta)$

$\nabla_\theta F(\theta) = 0$

θ_{MLE}

*) $\mu_{ML} = \frac{1}{N} \sum_i x_i$ $\sigma^2_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{ML})^2$

*) MLE for multivariate case,

→ Bernoulli distribution.

*) MAP / BCA.

Multivariat case Gaussian.

Given: N i.i.d. X_i $i = 1, 2, \cdots N$ $X_i \in R^d$

$$f(\theta) = \ln \prod_{i=1}^{N} \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{\left\{ -\frac{1}{2}(X_i - \mu)^T \Sigma^{-1} (X_i - \mu) \right\}}$$

$\mu \in R^d$, $\Sigma \in R^{d \times d}$

$$= \sum_{i=1}^{N} \ln p(X_i | \mu, \Sigma)$$

$$= \sum_{i=1}^{N} -\frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) \}$$

μ is unknown & Σ is known.

$\mu_{ML} = ?$

$$\frac{\partial}{\partial \mu} F(\theta) = 0$$

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

$$= -\frac{1}{2} \sum_{i=1}^{N} x_i^T \Sigma^{-1} x_i - 2\mu^T \Sigma^{-1} x_i + \mu^T \Sigma^{-1} \mu$$

$$= -\frac{1}{2} \sum_{i=1}^{N} \mu^T \Sigma^{-1} x_i - \frac{1}{2} \mu^T \Sigma^{-1} \mu$$

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

$$\mu^T \Sigma^{-1} x_i = (\mu^T \Sigma^{-1} x_i)^T$$

$$= x_i^T (\Sigma^{-1})^T \mu$$

$$= x_i^T \Sigma^{-1} \mu$$

$$= (\Sigma^{-1} x_i)^T \mu$$

$$\mu^T \Sigma^{-1} x_i$$

$$= \mathrm{tr}(\mu^T \Sigma^{-1} x_i)$$

$\frac{\partial (b^T y)}{\partial y} = b$ b & y are vectors. $\frac{\partial f(y)}{\partial y}$

$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$

$b^T y = b_1 y_1 + b_2 y_2$

$\frac{\partial}{\partial y} b^T y = \begin{bmatrix} \frac{\partial b^T y}{\partial y_1} \\ \frac{\partial b^T y}{\partial y_2} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$

$\mu^T \Sigma^{-1} \mu$ $\boxed{\mu^T A \mu}$

$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ $A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$

$\frac{\partial \mu^T A \mu}{\partial \mu} = \frac{\partial (2\mu_1^2 + \mu_2^2)}{\partial \mu}$

$= \begin{bmatrix} 2 \cdot 2\mu_1 \\ 2 \cdot \mu_2 \end{bmatrix}$

$= 2A\mu$

$\frac{\partial F(\theta)}{\partial \mu} = \sum_{i=1}^{N} \Sigma^{-1} x_i - \frac{1}{2} 2\Sigma^{-1}\mu$

$= \cancel{2}\,\cancel{\Sigma^{-1}} (\Sigma_i x_i - \mu) = 0$

$\sum_i x_i = N\mu$

$\mu_{ML} = \frac{1}{N}\sum_i x_i$

$y \Rightarrow$ d-dim vectors

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix}$$

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{bmatrix}$$

$b^T y$

x) ~~$E(\mu_{ML}) \rightarrow$ Biased?~~

$\mu_{ML} \rightarrow$ Biased?

$E(\mu_{ML}) = \mu \rightarrow$ unbiased.

$E(\mu_{ML}) = \frac{1}{N}\sum_i E(x_i) = \mu$

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

d^2

How many mult. do you need to obtain Σ_{ML}?

$\Sigma_{ML} = Nd^2$

$N = 10^3, d = 10^3$

- Example of a specific case: unknown μ
  - $P(x_i \mid \mu) \sim N(\mu, \Sigma)$
    (Samples are drawn from a multivariate normal population)

$$\ln P(x_k \mid \mu) = -\frac{1}{2}\ln\left[(2\pi)^d |\Sigma|\right] - \frac{1}{2}(x_k - \mu)^t \Sigma^{-1}(x_k - \mu)$$

$$\text{and} \quad \nabla_{\theta\mu} \ln P(x_k \mid \mu) = \Sigma^{-1}(x_k - \mu)$$

$\theta = \mu$ therefore:

- The ML estimate for μ must satisfy:

$$\sum_{k=1}^{k=n} \Sigma^{-1}(x_k - \hat{\mu}) = 0$$

- Multiplying by Σ and rearranging, we obtain:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{k=n} \mathbf{x}_k$$

Just the arithmetic average of the samples of the training samples!

Conclusion:

If $P(x_k \mid \omega_j)$ (j = 1, 2, …, c) is supposed to be Gaussian in a *d*-dimensional feature space; then we can estimate the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_c)^t$ and perform an optimal classification!

Bernoulli experiment

- Let's conduct N independent trials of the following Bernoulli experiment: We will ask each student whether they will take SML or ML.
- Let p be the probability that an individual will vote SML. In this example, each observation xi is a scalar. So it's better to represent it by xi . For each i, the value of xi is either SML (1) or ML (0).

$$x_i = \begin{cases} 1 & \text{if SML} \quad \text{with} \quad p = \theta \\ 0 & \text{if ML} \quad , \quad 1-p = 1-\theta \end{cases}$$

$$P(x_i \mid \theta) = \theta^{x_i} (1-\theta)^{1-x_i}$$

$$\theta_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$F(\theta) = \ln \prod_{i=1}^{N} \theta^{x_i} (1-\theta)^{1-x_i}$$

$$= \sum_{i=1}^{N} x_i \ln\theta + (1-x_i)\ln(1-\theta)$$

$$= \sum_{i=1}^{N} \frac{x_i}{\theta} - \frac{(1-x_i)}{1-\theta} = 0 \Rightarrow \theta_{MLE} =$$

*) Multivaiate Bernoulli θ_1 θ_2 θ_3

Student 1 $X_1 = \{ x_{1SML} = 1, x_{1GPA} = 1, x_{1Post} = 0 \}$

Student 2 $X_2 = \{ x_{2SML} = 0, x_{2GPA} = 0, x_{2Post} = 0 \}$

... $X_N = \{ x_{NSML} = 0, x_{NGPA} = 1, x_{NPost} = 1 \}$

$\theta_{ij} = Pr(x_{ij} = 1 | \theta_j)$

$\theta_j = P_r(x_j = 1 | \theta_j)$

'j' → denoting the dim.

θ_1 → dim. to SML

θ_2 → dim to GA

θ_3 → " " Poetry

$$P(x | \mu, \sigma^2) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{\left\{ -\frac{1}{2} (x-\mu)^2 \right\}}$$

μ, Σ

$x_i \leftarrow N(\mu, \Sigma)$

N samples.

$$P_\theta(x_j | \theta_j) = \theta_j^{x_j} (1-\theta_j)^{1-x_j}$$

$$P_\theta(x | \theta) = \prod_{j=1}^{3} \theta_j^{x_j} (1-\theta_j)^{1-x_j}$$

all dim. ~~decrease~~ ind.

$\theta = \{\theta_1, \theta_2, \theta_3\}$

$X = \{x_1, x_{SML}, x_{GA}, x_{Poet}\}$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}$$

$$X_1 = \begin{bmatrix} x_{1d_1} & x_{1d_2} \end{bmatrix} \leftarrow \mu_1, \Sigma_1$$

$$X_2 = \begin{bmatrix} x_{2d_1} & x_{2d_2} \end{bmatrix} \leftarrow \mu_2, \Sigma_2$$

$$\vdots$$

$$X_N = \begin{bmatrix} x_{Nd_1} & x_{Nd_2} \end{bmatrix} \leftarrow (\mu_{2N}, \Sigma_N)$$

*) N iids X_i's $\quad X_i \in \mathbb{R}^d$

$$\ln \prod_{i=1}^{N} p(x_i|\theta) = \ln \prod_{i=1}^{N} \prod_{j=1}^{d} p(x_{ij}|\theta_j)$$

$$= \ln \prod_{i=1}^{N} \prod_{j=1}^{d} \theta_j^{x_{ij}} (1-\theta_j)^{1-x_{ij}}$$

$$F(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{d} x_{ij} \ln \theta_j + (1-x_{ij}) \ln(1-\theta_j)$$

$$\frac{\partial}{\partial \theta_j} F(\theta) = \sum_{i=1}^{N} \frac{x_{ij}}{\theta_j} - \frac{1-x_{ij}}{1-\theta_j} = 0 \Rightarrow \theta_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij}$$

$$\theta_1 = \frac{1}{N}\sum_{i=1}^{N} x_{i1} = \frac{1}{4}, \quad \theta_2 = \frac{1}{4}, \quad \theta_3 = \frac{1}{9}$$

$$p(x|\theta) = \theta_1^{x_1}(1-\theta_1)^{1-x_1} \cdot \theta_2^{x_2}(1-\theta_2)^{1-x_2}\,\theta_3^{x_3}(1-\theta_3)^{1-x_3}$$

$x_1, x_2, x_3 : \{1, 0, 1\}$

ML	GA	Poetry
1	0	1
0	0	0
0	1	0
0	0	1

$$\frac{\partial}{\partial \theta_j} \sum_{i=1}^{N} \{x_{i1} \ln\theta_1 + x_{i2} \ln\theta_2 + x_{i3} \ln\theta_3 + x_{i4} \ln\theta_4\}$$

$$\frac{\partial}{\partial \theta_1}$$

Example: Bernoulli distribution

If $x_1, \cdots, x_n$ are a set of Bernoulli observations,

with $f(1 \mid p) = p$ and $f(0 \mid p) = 1 - p$, *i.e.* $f(x_i \mid p) = p^{x_i}(1-p)^{1-x_i}$

The likelihood function for observations $x_1, \cdots, x_n$ is

$$L(p \mid x_1, \cdots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i},$$

and the m.l.e $\hat{p}$ is the value that maximizes this.

The log-likelihood is $\ln(L) = \ln(p)\sum x_i + (n - \sum x_i)\ln(1-p)$

Set: $\dfrac{d\ln(L)}{dp} = \dfrac{\sum x_i}{p} - \dfrac{n - \sum x_i}{1-p} = 0 \overset{\text{solve for p}}{\Longrightarrow} \hat{p} = \dfrac{\sum x_i}{n}$

Drawback of MLE?

$$\theta_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i = 1 \approx \frac{1}{2}$$

$$x = \{H, T\}$$
$$1, 0$$

$$\theta = p_\theta(x = H \mid \theta)$$

$x_{1:3}$, x_1 x_2 x_3 x_4 ..... x_{1000},

1 1 1 1

Biased Estimate

/ Practice.

$\mathrm{Var}(\mu_{ML}) = ?$. Summation of ind. rvs

Suppose MLE for mean for a dataset is taken as the first sample.
Is the estimator unbiased?

MAP

→ Max. a Posteriori

- $\hat{\theta}_{MLE} = argmax_{\theta} P(D|\theta)$
- $\hat{\theta}_{MAP} = argmax_{\theta} P(\theta|D) \equiv argmax_{\theta} P(D|\theta)P(\theta)$

$f(\theta)$

Likelihood Prior.

Bernoulli Beta.

Posterior will be Beta distribution.

likelihood → Gaussian, Prior Gaussian.

Post. Gaussia.

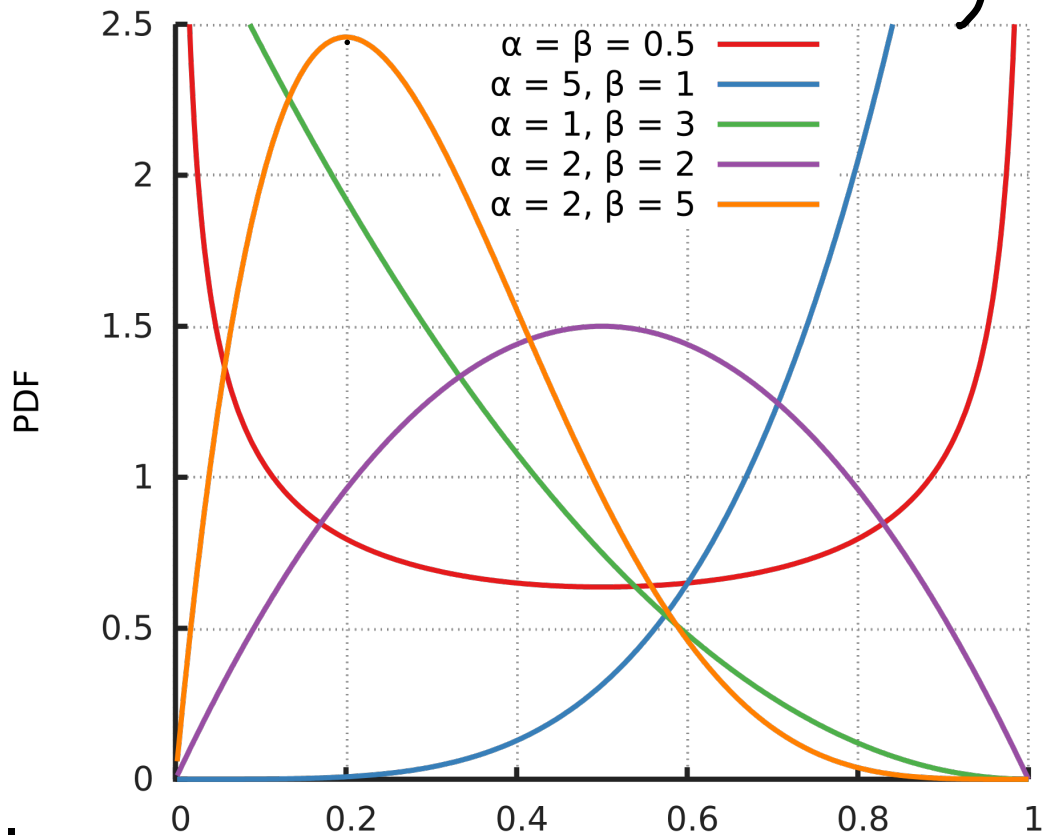
Conjugate prior - Beta distribution

- Conjugate prior of likelihood gives the posterior distribution from the same family of prior

$$f(p;\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$$

Beta distribution.

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$



Posterior and prior are conj. distributions

MAP estimate:

$$\max_{\theta} \left[\sum_{i=1}^{N} \ln p(x|\theta) \right] + \ln p(\theta)$$

$$\max_{\theta} \left[\sum_{i=1}^{N} \{ x_i \ln\theta + (1-x_i) \ln(1-\theta) \} \right] + (\alpha-1)\ln\theta + (\beta-1)\ln(1-\theta)$$

$$p(\theta) \prod_{i=1}^{N} p(x|\theta)$$

$$p(\theta)\, p(\mathcal{D}|\theta)$$

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N}$$