ORF 418 : OPTIMAL LEARNING

LECTURE 8 : September 29, 2025

BAYES FORMULA

1) Recall from Probability Theory
2) One discrete variable
3) Gaussian Case

I. TOOLS FROM PROBABILITY THEORY

Let X be a real-valued random variable (r.v.).

Its probability density function, pdf, $f_X(\cdot)$ satisfies:

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)\, dx,$$

for any continuous bounded function h. Formally,

$$f_X(x)\, dx \approx \mathbb{P}(X \in [x, x+dx]).$$

Given another r.v. Y, the joint probability distribution $f_{X,Y}(\cdot,\cdot)$ satisfies:

$$\mathbb{E}[H(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x,y) f_{X,Y}(x,y)\, dx\, dy.$$

Formally

$$f_{X,Y}(x,y)\, dx\, dy = \mathbb{P}(X \in [x, x+dx] \text{ and } Y \in [y, y+dy])$$

For given rv's X, Y, the conditional pdf of X given $\{Y=y\}$, denoted by $f_{X|Y}(x|y)$ satisfies:

$$\mathbb{E}[h(X) \mid Y=y] = \int_{-\infty}^{\infty} h(x) f_{X|Y}(x|y)\, dx.$$

As above formally,

$$f_{X|Y}(x|y)\, dx \approx \mathbb{P}(X \in [x, x+dx] \mid Y=y).$$

Proceeding formally

$$f_{X|Y}(x|y)\,dx \approx \mathbb{P}(X \in [x, x+dx] \mid Y \in [y, y+dy])$$

$$= \frac{\mathbb{P}(X \in [x, x+dx] \text{ and } Y \in [y, y+dy])}{\mathbb{P}(Y \in [y, y+dy])}$$

$$= \frac{f_{X,Y}(x,y)\,dx\,dy}{f_Y(y)\,dy}.$$

Hence we have

$$\boxed{f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.}$$

By symmetry we have

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)\, f_Y(y)$$

$$= f_{Y|X}(y|x)\, f_X(x).$$

This can be rewritten in the form of a Bayes formula:

$$\boxed{f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)\, f_X(x)}{f_Y(y)}}$$

This is the Bayes formula for continuous random variables.

II. Mixed Case.

We now consider the cases either X or Y is discrete valued. Note that the case X and Y are continuous valued is treated in Section I. When X and Y are both discrete valued, this is infact included in the Bayes formula for partition. Indeed, if Y takes values $\{y_1, \dots, y_M\}$ and X takes values $\{x_1, \dots, x_N\}$, then

$$\mathbb{P}(X=x_i \mid Y=y_k) = \frac{\mathbb{P}(Y=y_k \mid X=x_i)\,\mathbb{P}(X=x_i)}{\sum_{j=1}^{N} \mathbb{P}(Y=y_k \mid X=x_j)\,\mathbb{P}(X=x_j)}$$

for any $i=1,\dots,N$, $j=1,\dots,M$. See Exercise 3.4 in the lecture Notes.

Example. Consider a coin and let P be the probability of getting head once it is flipped. Initially, we have no knowledge about P and our prior distribution is uniform. Suppose that we flip this coin 100 times and 52 of them are heads. With this information

we would like to update our distribution of P.

Note that this is not estimation - that we will consider next lecture.

So mathematically

$$f_P(p) = 1 \quad , \quad p \in [0,1] \qquad \text{(uniform prior)}$$

Y = # of heads in 100 flips.

The conditional probabilities $\mathbb{P}(Y=k \mid P=p)$ are known:

$$\mathbb{P}(Y=k \mid P=p) = \binom{100}{k} p^k (1-p)^{100-k}.$$

for $k=0,\dots,100$. We would like to compute

$$f_{P|Y}(p \mid 52).$$

For any subset $A \subset [0,1]$ we have

$$\mathbb{P}(P \in A \mid Y=52) = \frac{\mathbb{P}(P \in A \text{ and } Y=52)}{\mathbb{P}(Y=52)}.$$

Also,

$$\mathbb{P}(Y=52 \mid P=p') = \binom{100}{52} (p')^{52} (1-p')^{48}.$$

Since $f_P(p)=1$ we have

$$\mathbb{P}(Y=52) = \int_0^1 \binom{100}{52} (p')^{52} (1-p')^{48} dp' =: c_{52}$$

Now formally take $A = [p, p+dp]$. Then

$$\mathbb{P}_{P|Y}(p|52)\,dp \simeq \mathbb{P}(P \in A \mid Y=52)$$

$$= \frac{\mathbb{P}(P \in [p,p+dp],\ Y=52)}{c_{52}}.$$

$$= \frac{\binom{100}{52}}{c_{52}} p^{52}(1-p)^{48}.$$

The above is the Beta distribution with parameters $a=53$ and $b=49$. Its mean is $a/(a+b) \simeq 0.5196$.

Now consider the case where X is a continuous r.v., and Y takes values in $\{y_1, \dots y_n, \dots\}$. Then,

$$f_{X|Y}(x|y_k) = \frac{\mathbb{P}(Y=y_k \mid X=x)}{\mathbb{P}(Y=y_k)},$$

and $$\mathbb{P}(Y=y_k) = \int \mathbb{P}(Y=y_k \mid X=x')\, f_X(x')\, dx'.$$

III. Gaussian case.

We recall that if $Z \in \mathbb{R}^m$ is Gaussian, then its pdf is given by,

$$f_Z(z) = (2\pi \det(\Sigma_Z))^{-m/2} \exp\left(-\tfrac{1}{2}(z-\mu_Z)^T \Sigma_Z^{-1}(z-\mu_Z)\right), \quad z \in \mathbb{R}^m,$$

where $\mu_Z \in \mathbb{R}^M$ is the mean and the $(m \times m)$ symmetric positive semi-definite matrix Σ_Z is the covariance matrix of Z.

Suppose that $X \in \mathbb{R}^k$, $Y \in \mathbb{R}^d$ are jointly Gaussian, corresponding to the case $Z = (X, Y) \in \mathbb{R}^M$ with $m = k+d$. But we seperate mean and the covariance matrix into its X and Y components:

$$\mu_Z = (\mu_X, \mu_Y) \in \mathbb{R}^k \times \mathbb{R}^d$$

$$\Sigma_Z = \begin{bmatrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_{Y,X} & \Sigma_Y \end{bmatrix}$$

where Σ_X is the $(k \times k)$ covariance matrix of X, Σ_Y is the $(d \times d)$ covariance matrix of Y, $\Sigma_{X,Y}$ is the $k \times d$ covariance matrix of X, Y and $\Sigma_{Y,X} = (\Sigma_{X,Y})^T$, i.e.,

$$(\Sigma_X)_{i,n} = \mathbb{E}\left[(X - \mathbb{E}(X))_i \, (X - \mathbb{E}(X))_n\right], \quad i,n = 1, \ldots, k$$

$$(\Sigma_Y)_{\ell,j} = \mathbb{E}\left[(Y - \mathbb{E}(Y))_\ell \, (Y - \mathbb{E}(Y))_j\right], \quad \ell, j = 1, \ldots, d$$

$$(\Sigma_{X,Y})_{ij} = \mathbb{E}\left[(X-\mathbb{E}(X))_i\,(Y-\mathbb{E}(Y))_j\right] \quad \begin{matrix} i=1,\dots,k \\ j=1,\dots,d \end{matrix}.$$

THEOREM 3.1.1. The conditional density of X given Y is also Gaussian with conditional mean

$$\mu_{X|Y} = \mathbb{E}[X|Y] = \mathbb{E}[X] + \Sigma_{X,Y}\Sigma_Y^{-1}(Y-\mathbb{E}[Y])$$

or equivalently

$$\boxed{\mu_{X|Y=y} = \mu_X + \Sigma_{X,Y}\Sigma_Y^{-1}(y-\mu_Y),}$$

and conditional covariance

$$\boxed{\Sigma_{X|Y} = \Sigma_X - \Sigma_{X,Y}\Sigma_Y^{-2}(\Sigma_{X,Y})^T}$$

is independent of the observation $Y=y$.

The conditional density is then given by

$$\boxed{f_{X|Y}(x|y) = (2\pi\Sigma_{X|Y})^{-\frac{k}{2}} \exp\left(-\tfrac{1}{2}(x-\mu_{X|Y=y})^T\Sigma_{X|Y}^{-1}(x-\mu_{X|Y=y})\right)}$$

for $x\in\mathbb{R}^k$, $y\in\mathbb{R}^d$.

example. Consider $k=d=1$, $\mu_X=1$, $\mu_Y=2$,

$$\Sigma := \Sigma_{X,Y} = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} \Rightarrow \Sigma^{-1} = \frac{1}{3}\begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix}.$$

Then applying the general formula $X|Y=2$ has the pdf.

$$f_{X|Y}(x|2) = \frac{f_{X,Y}(x,2)}{f_Y(2)},$$

and since $\mu_Y = 2$ and $\sigma_Y^2 = 2$,

$$f_Y(y) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{8}(y-2)^2\right), \quad y \in \mathbb{R}.$$

Also,

$$f_{X,Y}(x,y) = \underbrace{(2\pi|\det(\Sigma)|)^{-1}}_{=(2\pi\sqrt{3})^{-1}} \exp\left(\underbrace{-\frac{1}{2}(x-1,y-2)^T \Sigma^{-1} (x-1,y-2)}_{=: Q(x,y)}\right)$$

and

$$Q(x,y) = \frac{1}{6}\left[4(x-1)^2 - 2(x-1)(y-2) + (y-2)^2\right].$$

Substitute all these into the formula to obtain,

$$f_{X|Y}(x|2) = \frac{(2\pi\sqrt{3})^{-1} \exp\left(-\frac{2}{3}(x-1)^2\right)}{(2\sqrt{2\pi})^{-1}}$$

$$= \frac{2}{\sqrt{6\pi}} \exp\left(-\frac{2}{3}(x-1)^2\right), \quad x \in \mathbb{R}.$$

Formulae from the Theorem are:

$$\mu_{X|Y=2} = 1 + 1 \cdot \frac{1}{4}(2-2) = 1,$$

$$\Sigma_{X|Y} = 1 - 1 \cdot \frac{1}{4} \cdot 1 = \frac{3}{4}.$$

Agrees with the above calculation.