DS 4400
Ehsan Elhamifar

Probability Review :

A random variable, e.g. X, is a variable whose possible values are numerical outcomes of a random phenomenon.

Discrete rvs : takes countable number of distinc values, e.g. 2, 3, 4, ... Ex: flipping a coin $H \longrightarrow X=1$, $T \longrightarrow X=0$

probability distribution : list of probabilities assigned to each possible value. $P(X=1)=p(H)=p_1$, $P(X=0)=p(T)=p_2$

probability distribution = probability mass function

Continuous rvs : takes infinite number of possible values, e.g., temperature of Boston $X \in (-\infty, +\infty)$

cont. rv is not defined at specific values, it is defined over an interval of values and represented by area under curve. $P(X=x)=0$, $p(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\,dx$

$p(x)$: density function $\begin{cases} p(x) \geq 0 \\ \int p(x)dx = 1 \end{cases}$

Joint distribution : For collection of random variables e.g. X, Y or $X_1\, X_2 \cdots X_N$

X: Boston temperature, Y: flipping a coin; $X_1\, X_2 \cdots X_N$: flipping a coin N times

$p(X,Y)$ or $p(X_1, X_2, \dots, X_n)$

Conditioning : $P(X=x \mid Y=y) = \dfrac{P(X=x, Y=y)}{P(Y=y)}$, $\sum_x P(X=x \mid Y=y) = 1$

but not necessarily $\sum_y P(X=x \mid Y=y) \neq 1$

Chain rule of probability : $P(X \mid Y) = \dfrac{P(X,Y)}{P(Y)} \longrightarrow P(X,Y) = P(X \mid Y)P(Y)$ or $P(Y \mid X)\,p(X)$

$$P(X_1, X_2, \dots, X_N) = p(X_1)\,p(X_2 \mid X_1)\,p(X_3 \mid X_1, X_2) \cdots p(X_N \mid X_1, \dots, X_{N-1})$$

Marginalization : Given joint p.d. $p(X,Y) \longrightarrow p(X) = \sum_y p(X, Y=y) = \sum_y P(Y=y \mid X)\,p(X)$
$= p(X) \times 1 = p(X)$.

$p(X,Y,Z) \longrightarrow p(Y) = \sum_x \sum_z p(X=x, Y, Z=z)$

Bayes Rule : $P(X \mid Y) = \dfrac{P(Y \mid X)\,P(X)}{P(Y)}$ e.g. $P(\theta \mid D) = \dfrac{P(D \mid \theta)\,P(\theta)}{P(D)}$ MAP

Independence : X and Y are independent if observing X does not affect information about Y

$X \perp\!\!\!\perp Y \Leftrightarrow P(Y \mid X) = P(Y) \equiv P(X,Y) = P(X)\,P(Y)$ e.g. $X = H, T$; Y = Boston temp.

Conditional Independence : X and Y are conditionally independent given Z iff $P(X,Y|Z) = P(X|Z)\,P(Y|Z)$

$X \perp\!\!\!\perp Y \mid Z$

e.g. Height and Vocabulary are not independent (small child has a limited vocab.)

but given age they become independent : Height $\perp\!\!\!\perp$ Vocabulary | Age

Expectation: Given X being a rv, any function $f(x)$ is also a rv.

$$E[f(x)] = \sum_{x} f(x)\,p(x)$$

$$E[f(x)] = \int f(x)\,p(x)\,dx$$

$$E[f(x) \mid Y=y] = \int f(x)\,p(x|y)\,dx \quad \text{or} \quad \sum_{x} p(x|y)\,f(x)$$

IID: independent and identically distributed rvs are mutually independent and all come from the same distribution: E.g. flip a coin N times.

Probability Model: probability distribution is used as models of observed data

Ex) We flip a coin N times and record $X_1, X_2, \dots, X_N$

We measure Boston city temperature on Jan 1, each year $X_1, \dots, X_N$

These prob. dist. are parameterized and the goal is to learn the parameters of the p.d.

Ex) Flip a coin N times : $X_1, X_2, \dots, X_N$: $P(X_i=1) = \theta$, $P(X_i=0) = 1-\theta \Rightarrow P(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$

Given Dataset : $P_\theta(x_1, x_2, \dots, x_N) \overset{iid}{=} \prod_{i=1}^{N} P_\theta(x_i)$

$X_1 = x_1, \dots, X_N = x_N$

$$= \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{N} x_i}(1-\theta)^{N-\sum_{i=1}^{N} x_i} = \text{likelihood function}$$

* Given $x_1, \dots, x_N$, determine parameters of the prob. model on data.

Here, determine θ ? $\max_\theta P_\theta(x_1, \dots, x_N)$ for what value of θ, the observations are most likely?

e.g. $X_1=1, X_2=1, \dots, X_N=1 \longrightarrow$ if $\theta = 0 \longrightarrow P(x_i=1) = 0 \longrightarrow$ we never observe H!

if $\theta = 1 \longrightarrow P(x_i=1) = 1 \longrightarrow$ we always observe H!

To determine parameters of the model, we use maximum likelihood:

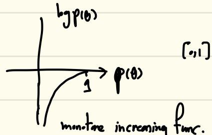
$$L(\theta) \triangleq p_\theta(x_1, \dots, x_N) \longrightarrow \max_\theta L(\theta) = \max_\theta p_\theta(x_1, \dots, x_N)$$

Generally, it is more convenient to take log (simplifies operations, increases dynamic range)

$$l(\theta) \triangleq \log p_\theta(x_1, \dots, x_N) \longrightarrow \max_\theta l(\theta) = \arg\max_\theta \log p_\theta(x_1, \dots, x_N)$$

$\hookrightarrow$ maximum log-likelihood

Same results for θ via $L(\theta)$ or $l(\theta)$



$[0,1] \longrightarrow [-\infty, 0]$ increased dynamic range!

$$l(\theta) = \sum_{i=1}^{N} x_i \log\theta + \left(N - \sum_{i=1}^{N} x_i\right) \log(1-\theta) \longrightarrow \text{is it concave?}$$

$$\frac{\partial l}{\partial \theta} = \sum_{i=1}^{N} x_i / \theta - \frac{N - \sum x_i}{1-\theta} = 0 \longrightarrow \hat{\theta}_{ML} = \sum_{i=1}^{N} x_i / N = \text{Sample mean}$$

Gaussian MLE: $\quad x_i \sim \mathcal{N}(x_i; \mu, \sigma^2) \Rightarrow p(x_i) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$

$$\theta = \{\mu, \sigma^2\} : \quad p_\theta(x_1, \dots, x_N) \overset{\text{iid}}{=} \prod_{i=1}^{N} p_\theta(x_i) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\sum_{i=1}^{N}(x_i-\mu)^2/2\sigma^2}$$

$$l(\theta) = -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} (x_i-\mu)^2/2\sigma^2$$

$$\frac{\partial l}{\partial \theta} = \begin{pmatrix} \frac{\partial l}{\partial \mu} \\ \frac{\partial l}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \longrightarrow \frac{\partial l}{\partial \mu} = \sum_{i=1}^{N} (\mu - x_i)/\sigma^2 = 0 \longrightarrow \hat{\mu}_{ML} = \sum_{i=1}^{N} x_i / N$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2} \times \frac{1}{\sigma^2} + \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^4} = 0 \longrightarrow \hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{ML})^2$$

empirical mean and variance $\longrightarrow \hat{\mu}_{ML}, \hat{\sigma}^2_{ML}$