Probability and Bayes' Review

Discretely...

Marginal Distributions $p(A), p(B)$

Joint Distributions $p(A, B) = p(A)p(B)$ if A, B are independent events

Conditional Distributions $p(A|B)$, $p(B|A)$

$$p(A) = \sum_B p(A,B), \quad (B) = \sum_A p(A)$$

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A,B)}{\sum_A p(A,B)}, \quad (B|A) = \frac{p(A,B)}{p(A)} = \frac{p(A,B)}{\sum_B p(A,B)}$$

Given $p(A|B)$, $p(B)$, we can also find $P(B|A)$.

$$p(A|B) = \frac{p(A,B)}{p(B)} \Rightarrow p(A,B) = p(A|B)p(B)$$

Bayes' Rule

$$\boxed{p(B|A) = \frac{p(A,B)}{p(A)} = \frac{p(A,B)}{\sum_B p(A,B)} = \frac{p(A|B)p(B)}{\sum_B p(A|B)p(B)}}$$

Continuously...

Joint Probability Density $p(x,y)$

Marginal Probability Density $p(x) = \int p(x,y)\,dy$, $p(y) = \int p(x,y)\,dx$

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x,y)}{\int p(x,y)\,dx}. \quad p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x,y)}{\int p(x,y)\,dy}$$

$$= \frac{p(y|x)p(x)}{\int p(y|x)p(x)\,dx} \quad\quad = \frac{p(x|y)p(y)}{\int p(x|y)p(y)\,dy}.$$

Bayes' Rule $\rightarrow$ we can find $p(x|y)$ given that we know $p(y|x)$ and $p(x)$!

Maximum Likelihood Estimation

or pdf for continuous case ← parameter that we are trying to solve for our distribution.

Let X be a discrete r.v. with pmf p depending on parameter θ.

$L(\theta|x) = p_\theta(x) = P_\theta(X=x)$ is the likelihood function, given the outcome x of the r.v. X.

For Bernoulli r.v. X,

$$L(\theta|x) = p_\theta(x) = \theta^x(1-\theta)^{1-x} \quad \text{s.t. } P_\theta(X=1) = \theta \quad ; \quad P_\theta(X=0) = (1-\theta) = 1 - P_\theta(X=1)$$

Suppose we have collected some data ... (bernoulli coin tosses) i.i.d.

$$\text{data} = \{x_1, x_2, x_3, \dots, x_N\}$$

Likelihood is : $L(\theta|\text{data}) = p_\theta(\text{data}) = \prod_{i=1}^{N} p_\theta(x_i) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$

Probability of observing the data that we observe assuming each coin toss was i.i.d

$$\text{E.g. } L(\theta|x_1=1, x_2=0, x_3=1) = \theta^1(1-\theta)^1\theta^1$$

We want to find the value of θ that maximizes $L(\theta|\text{data})$!

i.e. the value of θ that makes the data that we collect most probable.

To Maximize $L(\theta|\text{data})$, we want to take its derivative wrt. θ. s.t. $\frac{dL}{d\theta} = 0$

$$\hat{\theta} = \arg\max_\theta L(\theta).$$

Note that argmax is the operation that finds the argument that gives the max value from function $L(\theta)$.

Most of the time it is better to take the log of the likelihood before differentiating,

- and it usually leads to a simpler expression for the derivative that is easier to set to 0 and solve.
- since log is a monotonous function, i.e. whatever maximizes L also maximizes $\log L$.

$$\ell(\theta) = L(\theta) = \log \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} = \sum_{i=1}^{N} \{x_i \log\theta + (1-x_i)\log(1-\theta)\}.$$

Set $\frac{d\ell}{d\theta} = \frac{1}{\theta}\sum_{i=1}^{N} x_i - \frac{1}{1-\theta}\sum_{i=1}^{N}(1-x_i) = 0$

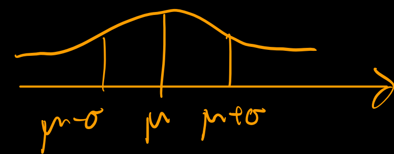
Solve for θ $\quad (1-\theta)\sum_{i=1}^{N} x_i = \theta\sum_{i=1}^{N}(1-x_i),$

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

Note that for Binomial dist.,

$$p(x=k) = \binom{N}{k}\theta^k(1-\theta)^{N-k}$$

For Gaussian (Normal) r.v.,



$$\text{data} = \{x_1, x_2, \ldots, x_N\} \quad \text{iid}$$

$$L(\theta \mid \text{data}) = \prod_{i=1}^{N} p_\theta(x_i), \quad \text{where } \theta = \{\mu, \sigma^2\}$$

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

Goal : Find parameters that best describe data collected.
These values will maximizes the likelihood

$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu, \sigma^2} L(\mu, \sigma^2 \mid \text{data})$$

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

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

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

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)\frac{1}{\sigma}$$

Set $\frac{\partial L}{\partial \mu} = 0$, $\quad \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0.$

$$\sum_{i=1}^{N} x_i - N\mu = 0.$$

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i = f(x_1, x_2, \ldots, x_N).$$

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

$$l(\mu, v) = -\frac{N}{2}\log(2\pi v) - \frac{1}{2}\frac{1}{v}\sum_{i=1}^{N}(x_i - \mu)^2, \text{ where } v = \sigma^2$$

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

Set $\frac{\partial l}{\partial v} = 0$, $N = \frac{1}{v}\sum_{i=1}^{N}(x_i - \mu)^2$

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

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 = g(x_1, x_2, \ldots, x_N).$$

Our MLE estimates are also random variables!

If they are random variables, then we can ask:

- what is their distribution and expectation?

$$\hat{\mu} \sim ? \quad E(\hat{\mu}) = ?$$
$$\hat{\sigma}^2 \sim ? \quad E(\hat{\sigma}^2) = ?$$

Can use characteristics functions to prove:

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

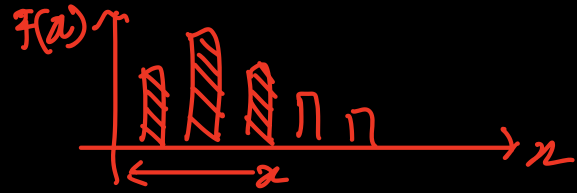
$$E(\hat{\mu}) = \mu$$

$E(\hat{\sigma}^2) \neq \sigma^2$; $E(\hat{\sigma}^2) = \frac{N-1}{N}\sigma^2 \to \sigma^2$ as $N \to \infty$.

See 'unbiased estimate of the covariance matrix' for full derivation

CDF and Percentiles

for Discrete rvs:



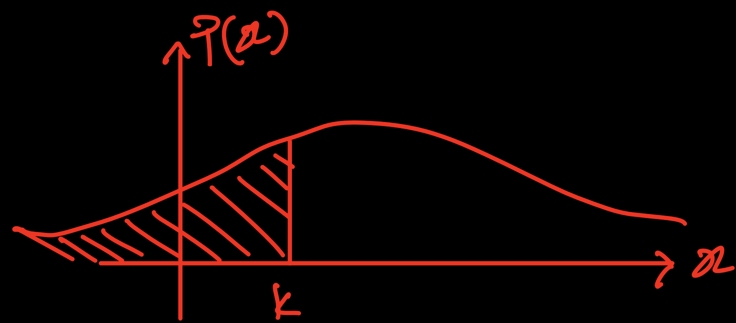
CDF: $F(x) = P(X \le x) = \sum_{k=-\infty}^{x} p(k)$, where $p(k) = \text{Prob}(X = k)$

for Continuous rvs:

CDF: $F(k) = P(X \le k) = \int_{-\infty}^{k} f(t)\,dt$

t is a dummy variable, and disappears after integration.

PDF: $f(x) = \frac{dF(x)}{dx}$



Inverse CDF (percentile function)

$F^{-1}(p)$ — What value of x would yield a CDF value of p?

E.g. Heights — $\mu = 170$ cm, 7cm

$$F^{-1}(0.95, \mu = 170, \sigma = 7) = 181.5$$

$$F(160, \mu = 170, \sigma = 7) = 0.08$$