Basic Probability

Axioms of Probability

- Non negativity: P(A) >=0
- Normalization: P(Ω) =1
- Additivity: if $A \cap B = \phi, then\ P(A \cup B) = P(A) + P(B)$

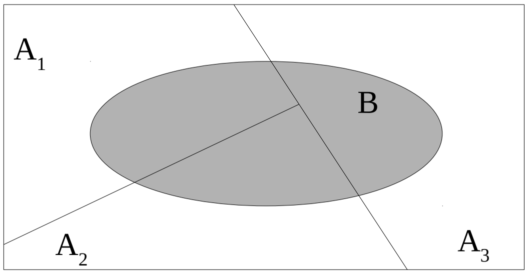
Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

P(A|B) is undefined if P(B)=0

Bayes's Rule

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i \cap B)}{P(B)} \\ &= \frac{P(A_i)P(B|A_i)}{P(B)} \\ &= \frac{P(A_i)P(B|A_i)}{\sum_j P(A_j)P(B|A_j)} \end{aligned}$$



Independence of Two events

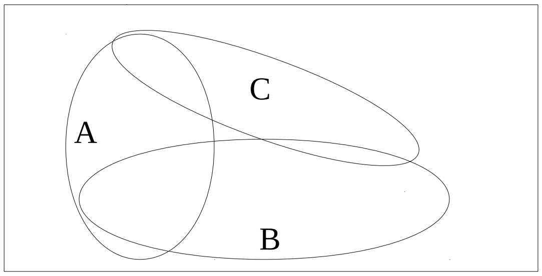
$$P(B \mid A) = P(B)$$

Occurrence of event A provide no information about B's occurrence.

$$P(A \cap B) = P(A)P(B|A) \Rightarrow P(A \cap B) = P(A)P(B)$$

Conditional Independence

$$P(A \cap B|C) = P(A|C)P(B|C)$$



Having Independence in original model does not imply independence in conditional model.
Example Here $P(A \cap B|C) = 0$ but we cannot say whether P(A|C)P(B|C)=0

Discrete Random Variables

An assignment of a variable (number) to every possible outcome.
Mathematically, A function from sample space Ω to the real numbers.

Probability Mass function(pmf)

$$\begin{aligned} p_X(x) &= P(X = x) \\ &= P(\{\omega \in \Omega\ st\ X(\omega) = x\}) \end{aligned}$$

$$p_X(x) \geq 0 \qquad \sum_x p_X(x) = 1$$

Expectation $$E[X]=\sum_x x\, p_X(x)$$

Interpretations:

- Center of gravity of PMF
- Average in large number of repetitions of the experiment

In general $E[g(x)]\neq g(E[X])$, if g(.) is a non-linear function then surely these expectation are not equal but if g(.) is a linear function then the equality holds.

Variance

$$\begin{aligned} var(X) &= E[(X-E[X])^2] \\ &= \sum_x (x-E[X])^2\, p_X(x) \\ &= E[X^2]-(E[X])^2 \end{aligned}$$

Properties:

- var(X) ≥ 0
- var(αX+β)= α^2var(X)

Joint PMF $$p_{XY}(x,y)=P(X=x\cap Y=y)$$

Properties:

$$\sum_x \sum_y p_{XY}(x,y)=1$$
$$p_X(x)=\sum_y p_{XY}(x,y)$$
$$p_{(X|Y)}(x|y)=P(X=x|Y=y)=\frac{P_{XY}(x,y)}{P_Y(y)}$$
$$\sum_x p_{(X|Y)}(x|y)=1$$

Bernoulli Distribution

$$p_X(X=x)=p^x(1-p)^{1-x} \qquad x\in\{0,1\}$$

Mean =p Variance =p(1-p)

Binomial Distribution

$$p_X(X=k)=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \qquad 0\leq k\leq n$$

Mean = np Variance = np(1-p)

Geometric Distribution

$$p_X(X=k)=(1-p)^k p \quad k\geq 0$$

Mean = $\frac{1}{p}$ Variance = $\frac{1-p}{p^2}$

Poisson Distribution

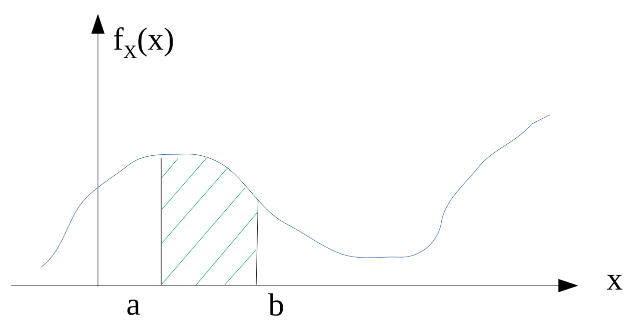
$$p_x(X=k)=\frac{\lambda^k e^{-\lambda}}{k!} \qquad k\in\mathbb{N}$$

Mean = λ Variance = λ

Continuous Random Variable

It is describe by a probability density function f_X.

$$P(a\leq X\leq b)=\int_a^b f_X(x)\,dx$$



Properties:

- P(X = a) = 0
- $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$
- $E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$
- $var(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)\,dx$

Cumulative density function

for continuous case $$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(x)\,dx$$

for discrete case $$F_X(x) = P(X \le x) = \sum_{k \le x} p_X(k)$$

Gaussian Distribution

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R}$$

Mean = μ Variance = σ^2

Exponential Distribution

$$f_X(x) = \lambda e^{-\lambda x} \quad x \ge 0$$

Mean = 1/λ Variance = $1/\lambda^2$

Gamma Distribution

$$f_X(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{\lambda x}}{\gamma(\alpha)} \quad x \ge 0$$

where $\gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\,dx$

Mean = $\frac{\alpha}{\lambda}$ Variance = $\frac{\alpha}{\lambda^2}$

Beta Distribution

$$f_X(x) = \frac{\gamma(\alpha+\beta) x^{\alpha-1} (1-x)^{\beta-1}}{\gamma(\alpha)\,\gamma(\beta)}$$

Mean = $\frac{\alpha}{\alpha+\beta}$ Variance = $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Sum of 2 independent random variable

Discrete case W=X+Y

$$p_W(w) = \sum_x p_X(x)\,p_Y(w-x)$$

Continuous case W=X+Y

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w-x)$$

Example

X ~ $\mathcal{N}(\mu_x, \sigma_x)$ and Y ~ $\mathcal{N}(\mu_y, \sigma_y)$ [Independent]

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

$$= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left\{ \frac{-(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2} \right\}$$

PDF is constant on the ellipse where $\frac{(x-\mu_x)^2}{2\sigma_x^2}+\frac{(y-\mu_y)^2}{2\sigma_y^2}$ is constant. We get contours in the form of ellipse (or circle when $\sigma_x = \sigma_y$)

Covariance

$$cov(X,Y)=E\left[(X-E[X])(Y-E[Y])\right]$$
$$=E[XY]-E[X]E[Y]$$

Independent implies cov(X,Y) = 0 whereas converse is not true

$$var\left(\sum_{i=1}^{n} X_i\right)=\sum_{i=1}^{n} var(X_i)+\sum_{(i,j):i\neq j} cov(X_i,X_j)$$

Correlation coefficient

$$\rho=E\left[\frac{(X-E[X])}{\sigma_x}.\frac{(Y-E[Y])}{\sigma_y}\right]$$
$$=\frac{cov(X,Y)}{\sigma_x\sigma_y}$$

- $-1 \leq \rho \leq 1$
- $|\rho| = 1$ implies (X-E[X]) = c(Y-E[Y]) *//linearly related*
- Independent implies $\rho = 0$

Conditional Expectation

$$E[X|Y=y]=\sum_x x p_{(X|Y)}(x|y)$$

Conditional Expectation is itself a random variable.

Law of Iterated expectation

$$E[X]=E\left[E[X|Y]\right]$$

Proof:

$$Suppose\ E[X|Y]=g(Y)$$
$$E[X|Y=y]=g(y)$$
$$E\left[E[X|Y]\right]=E[g(Y)]$$
$$=\sum_y g(y)p_Y(y)$$
$$=E[X|Y=y]p_Y(y)$$
$$=E[X]$$

Conditional Variance

$$var(X|Y=y)=E\left[(X-E[X|Y=y])^2 \middle| Y=y\right]$$

Var(X|Y) is also a random variable

Law of Total Variance

$$var(X)=E[var(X|Y)]+var(E[X|Y])$$

Proof: Recall $var(X)=E[X^2]-(E[X])^2$

Similarly, we can write $var(X|Y)=E[X^2|Y]-(E[X|Y])^2$ since Var(X|Y) is a random variable.

Taking expectation of var(X|Y), we get,

$$E[var(X|Y)]=E[X^2]-E\left[(E[X|Y])^2\right] \qquad ...(1)$$

E[X|Y] is a random variable, we calculate variance of this and get.

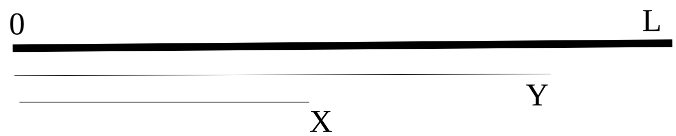
$$var(E[X\mid Y]) = E\left[(E[X\mid Y])^2\right] - (E[X])^2 \qquad \text{...(2)}$$

On adding (1) and (2), we get,

$$E[var(X\mid Y)] + var(E[X\mid Y]) = E[X^2] - (E[X])^2 \qquad \blacksquare$$

<u>Example</u> *Variance of Stick breaking problem*

First we break the stick at Y, then we break it at X; where X and Y are uniformly distributed.



P(Y = y) = 1/L such that $y \in [0, L]$

P(X = x | Y=y) = 1/y such that $x \in [0, y]$

$$E[Y] = \frac{L}{2}$$

$$var(Y) = \frac{L^2}{12}$$

X ~ U[0 , y] , then $E[X\mid Y] = \frac{Y}{2}$ and $var(X\mid Y) = \frac{Y^2}{12}$

Using Law of iterative expectation, we get

$$\begin{aligned} E[X] &= E\left[E[X\mid Y]\right] \\ &= E[\frac{Y}{2}] \\ &= \frac{L}{4} \end{aligned}$$

$$We\ know,\ var(X) = E[var(X\mid Y)] + var(E[X\mid Y])$$

$$\begin{aligned} E[var(X\mid Y)] &= E\left[\frac{Y^2}{12}\right] = \frac{1}{12}E[Y^2] \\ &= \frac{1}{12}\left(var(Y) + (E[Y])^2\right) \\ &= \frac{1}{12}\left(\frac{L^2}{12} + \frac{L^2}{4}\right) \\ &= \frac{L^2}{36} \end{aligned}$$

$$\begin{aligned} var(E[X\mid Y]) &= var(\frac{Y}{2}) \\ &= \frac{1}{4}var(Y) = \frac{L^2}{48} \end{aligned}$$

Hence var(X) = $\frac{7L^2}{144}$

Bernoulli Process (Discrete Memoryless)

A sequence of **independent** Bernoulli trials.

At each trial 'i' : P(success) = P(X_i = 1) = p and P(X_i = 0) = 1-p

Example : Sequence of lottery wins/loss.

So we have a sequence of random Variable X_1, X_2, … X_t …

$$E[X_t] = p \qquad var(X_t) = p(1-p)$$

But we are more interested in the joint probability of the distribution (for inference)

Interarrival Times:

T_1 : number of trials until first success

$$\underbrace{0\quad 0\quad 0\quad 1}_{T_1}\quad \underbrace{0\quad 0\quad 1}_{T_2}\quad \underbrace{0\quad 1}_{T_3}\quad \underbrace{0\quad 0\quad 0\quad 0\quad 1}_{T_4}\quad \ldots.$$

$P(T_1 = t) = (1\text{-}p)^{t-1}p \qquad t=1, 2, 3, \ldots$

$E[T_1] = 1/p$

$var(T_1) = (1\text{-}p)/p^2$

Memoryless Property!!

Y_k : Number of trials until k successes.

$$Y_k = T_1 + T_2 + \ldots + T_k$$

All the T_i , i = {1, 2, …, k} are geometric random variable with parameter ‘p’

$$P(Y_k = t) = P(k-1 \; arrival \in [1,2,3,\ldots t-1] \cap an \; arrival \; at \; time \; 't')$$

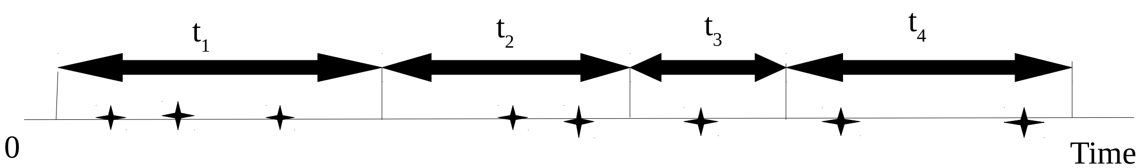
$$= \frac{(t-1)!}{(k-1)!\,(t-k)!} p^{k-1}(1-p)^{t-k} p \qquad \text{for } t \geq k$$

$$E[Y_k] = \frac{k}{p}$$

$$var(Y_k) = \frac{k(1-p)}{p^2}$$

Poissons Process (Continuous Memoryless)

Time Homogeneity : P(k, τ) = Probability of ‘k’ arrivals in interval duration ‘τ’
Also the number of arrivals in disjoint time interval are **independent**



For a very small ‘δ’ :

$$P(k, \delta) = \begin{cases} 1 - \lambda\delta & \text{if } k=0 \\ \lambda\delta & \text{if } k=1 \\ 0 & \text{if } k>1 \end{cases}$$

λ = Arrival Rate = Expected number of arrivals per unit time

$$P(k, \; \tau) = \frac{(\lambda\tau)^k e^{-k\tau}}{k!} \qquad k = 0, \; 1, \; 2, \; \ldots$$

$$E[N_t] = \lambda t, \qquad var(N_t) = \lambda t$$

Interarrival Times

Y_k : Time of k_{th} arrival

$$f_{Y_k}(t)\,\delta = P(t \leq Y_k \leq t + \delta)$$

$$= P(k-1 \; arrivals \in [0, t] \cap 1 \; arrival \in \delta \; time \; interval)$$

$$= \frac{(\lambda t)^{k-1}}{(k-1)!} \; e^{-\lambda t} \; \lambda\delta$$

Erlang distribution

$$f_{Y_k}(t)=\lambda\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-\lambda t} \qquad , \ t \geq 0$$
$$=\frac{\lambda^k y^{k-1} e^{-\lambda t}}{(k-1)!}$$

Time of first arrival (k=1) : Exponential $f_{Y_1}(y)=\lambda e^{-\lambda y}, \quad y\geq 0$ and $E[T_1] = 1/\lambda$

Markov Chains

Finite state Markov Chain

X_n : state after ‘n’ transition. X_n belongs a finite set e.g {1, 2, 3, … m} and X_0 (Initial State) is either random or given.

Markov Property (Given current state the past does not matters)

$$p_{ij}=P(X_{n+1}=j|X_n=i)$$
$$=P(X_{n+1}=j|X_n=i, X_{n-1}, \ldots X_0)$$

n-step transition probabilities

$$r_{ij}(n)=P(X_n=j|X_0=i)$$

Key Recursion

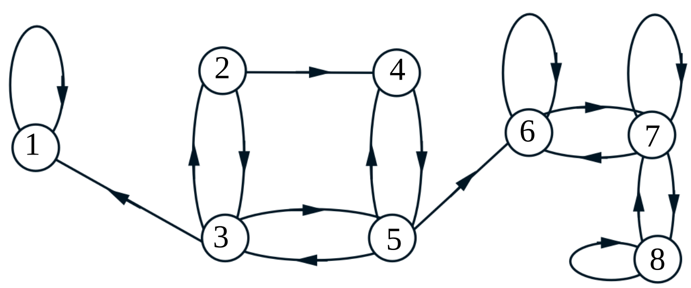
$$r_{ij}(n)=\sum_{k=1}^{m} r_{ij}(n-1)\, p_{kj}$$

With random Initial state

$$P(X_n=j)=\sum_{i=1}^{m} P(X_0=i) r_{ij}(n)$$

State ‘i’ is **recurrent** if starting from ‘i’, from wherever you go, there is way of returning to ‘i’. And if state ‘i’ is not recurrent then it is **transient**

e.g. In the diagram, we see the state 1, 6, 7, 8 are recurrent cause if you start from any of these states its possible to get to state where you started. And the state 2, 3, 4, 5 are transient, cause once we move out from these 4 states and its not possible going back. #Transient $P(X_n = i) \rightarrow 0$, ‘i’ visited finite number of times.



The State in recurrent class are periodic if
they can be grouped into d > 1 groups so that all transitions from one group lead to next group. OR
A state in a Markov chain is periodic if the chain can return to the state only at multiples of some integer larger than 1.

Steady state Probabilities

Question *Do $r_{ij}(n)$ converges to* π_j *?*

Yes if :

- Recurrent states are all in single class
- Single recurrent class is not periodic

Question *How do we calculate* π_j ?

Assuming the above conditions, start from key recursion:

$$r_{ij}(n)=\sum_{k} r_{ik}(n-1)\, p_{kj} \qquad \textit{for all } j$$

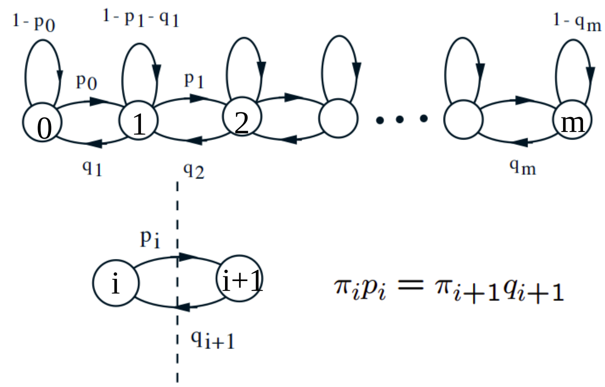
Take the limit as $n \to \infty$

$$\pi_j = \sum_k \pi_k p_{kj} \quad for\ all\ j$$

Additional Equation

$$\sum_j \pi_j = 1$$

Birth Death Process



Apart from the condition

$$\pi_i\, p_i = \pi_{i+1} q_{i+1}$$

We also have to use the normalization condition:

$$\sum_j \pi_j = 1$$

Special case when $p_i = p$ and $q_i = q$

$\rho = p/q$ = load factor, and $\pi_{i+1} = \pi_i \rho$

$\pi_i = \pi_0 \rho^i$, $\quad$ i =0, 1, … m

When $\rho = 1$ then $\pi_i = 1/(m+1)$, for all i

Assume $p < q$ and $m \approx \infty$, then

$\pi_0 = 1 - \rho$ and $\quad E[X_n] = \dfrac{\rho}{1-\rho}$

Example A phone company problem

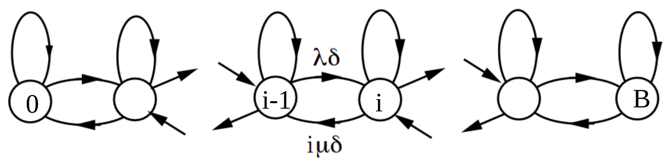
#Calls are generated as Poissons process, rate λ

Each call duration is exponentially distributed (parameter μ)

Number of lines needed ??(One call per line)

We assume discrete time intervals of (small) length δ

Let number of states be equal to the number of calls going on



Probability of going upward = Poisson Process recording an arrival in time interval δ = λδ

Suppose there is only one call happening so probability of call drop = μδ. And if ‘i’ calls are happening collective probability of one call drop = iμδ. (We are assuming probability of two calls happening and dropping to be zero cause of the $O(\delta^2)$ terms)

So the chain has BD process.

$$\lambda \delta \pi_{i-1} = i \mu \delta \pi_i$$
$$\lambda \pi_{i-1} = i \mu \pi_i$$
$$Hence, \pi_i = \pi_0 \frac{\lambda_i}{\mu^i i!}$$
$$\pi_0 = \frac{1}{\sum_{i=0}^{B} \frac{\lambda_i}{\mu^i i!}}$$

Now Probability of all lines to be busy = π_B , we set this value to a lower number to calculate the value of ‘B’

Limit Theorems

Markov Inequality If $X \geq 0$, we know $\quad E[X] = \sum_x x p_X(x)$

$$
\begin{aligned}
E[X] &\geq \sum_{x \geq a} x p_X(x) \\
&\geq \sum_{x \geq a} a p_X(x) \\
&= aP(X \geq a)
\end{aligned}
$$

Markov Inequality relates probability to the Expectation. So if the Expected value is small then the probability of X being is also small.

Since $var(X) = E[(X-\mu)^2]$, We do the same calculations as above to get,

$$E[(X-\mu)^2] \geq a^2 P((X-\mu)^2 \geq a^2)$$

$$var(X) \geq a^2 P(|X-\mu| \geq a)$$

This relates variance of X to the probability. If the variance is small then probability of being far away from the mean is also small.

$$P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

#Convergence a_n converges to a

$$\lim_{n \to \infty} a_n = a$$

"a_n eventually gets and stay (arbitrary) close to a"

For every $\epsilon > 0$, there exist n_0, such that every $n \geq n_0$, we have $|a_n - a| \leq \epsilon$

Convergence in Probability Sequence of random variable Y_n converges in probability to a number 'a' "(almost all) of the PMF/PDF of Y_n , eventually gets concentrated (arbitrarily) close to a".

For every $\epsilon > 0$,

$$\lim_{n \to \infty} P(|Y_n - a| \geq \epsilon) = 0$$

Central Limit Theorem $X_1, X_2, \ldots X_n$ are iid, with finite variance σ^2

"Standardized" $S_n = X_1 + X_2 + \ldots + X_n$:

$$Z_n = \frac{S_n - E[S_n]}{\sigma_{S_n}} = \frac{S_n - nE[X]}{\sigma\sqrt{n}}$$

$E[Z_n] = 0$, $\qquad$ $var(Z_n) = 1$

Let Z be an standard normal random variable (zero mean, unit variance), then for every c :

$$P(Z_n \leq c) \to P(Z \leq c)$$

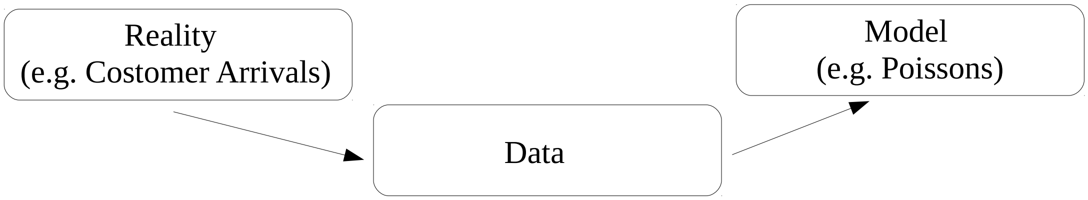
$P(Z \leq c)$ is the standard normal CDF, $\Phi(c)$ available from the normal tables.

Usefulness :

- Universal, only mean and variance matter
- Accurate computational shortcut
- Justification of normal models

Inference

We have a real phenomenon and we have to model it, all we have is the data from reality. So we have to use that data to come up with a model and its parameter. Then we predict about reality or tell certain hidden aspects of reality that we can not infer directly.



Types of Inference models/approaches:

- Model building versus inferring unknown variables. e.g., assume X = aS + W [‘S’ is the signal ; ‘a’ is the magnitude by which it is amplified ; ‘W’ is random noise ; ‘X’ is the observed sample]
  Model building : know signal ‘S’, observe ‘X’, infer ‘a’
  Estimation in the presence of noise : know ‘a’, observe ‘X’, estimate ‘S’
- Hypothesis testing: unknown takes one of few possible values; aim at small probability of incorrect decision
- Estimation: aim at a small estimation error

Bayesian Statistical Inference

Use Bayes Rule

$$P_{(\Theta|X)}(\theta \mid x)=\frac{P_{\Theta}(\theta)P_{(X\mid\Theta)}(x\mid\theta)}{P_{x}(x)}$$

Assume a prior on Θ , to estimate probability of (Θ | X)

Since output is PMF/PDF , if we are interested in single answer, then take the value:

- Maximum aposteriori probability(MAP)

$$P_{(\Theta|X)}(\hat{\theta}|x)=max_{\theta}P_{(\Theta|X)}(\theta|x)$$

- Conditional Expectation

$$E[\Theta|X=x]=\int \theta f_{(\Theta|X)}(\theta|x)\,dx$$

LMS Estimation $\hat{\Theta}=E[\Theta \mid X]$ minimizes $E[(\Theta-g(X))^2]$ over all estimators g(.) ; for any x, $\hat{\theta}=E[\Theta|X=x]$ minimizes $E[(\Theta-\hat{\theta})^2|X=x]$ over all estimator of $\hat{\theta}$

Classical Statistical Inference

Maximum Likelihood Estimation Pick θ, “that makes data most likely”

$$\hat{\theta}_{ML}=arg\,max_{\theta}\,p_X(x;\theta)$$

Desirable Properties of estimators:

- Unbiased $E[\hat{\Theta}_n]=\theta$
- Consistent $\hat{\Theta}_n \to \theta$ (in probability)
- “Small” mean Squared Error

$$E_{\theta}[(\hat{\Theta}_n-\theta)^2]=var_{\theta}(\hat{\Theta}-\theta)+(E_{\theta}[\hat{\Theta}-\theta])^2$$
$$=var_{\theta}(\hat{\Theta})+(bias_{\theta})^2$$

Confidence Interval (An estimate $\hat{\Theta}_n$ may not be informative enough.)

An (1–α) confidence interval [Θ^-_n, Θ^+_n] such that

$$P(\Theta^-_n \le \theta \le \Theta^+_n) \ge 1-\alpha \qquad \forall\,\theta$$

often α = 0.05 or 0.01

Classical Statistics

Linear Regression

Data: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ Model : $\theta_0 + \theta_1 x$

$$min_{\theta_0,\theta_1}\sum_{i=1}^{n}(y_i-\theta_0-\theta_1 x_i)^2$$

Can also be thought as $Y_i=\theta_0+\theta_1 x_i+W_i$, where $W_i \sim N(0, \sigma^2)$

Solution

$$\bar{x}=\frac{x_1+x_2+\ldots+x_n}{n} \qquad \bar{y}=\frac{y_1+y_2+\ldots+y_n}{n}$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

INTERPRETATION : Assume the model $Y = \theta_0 + \theta_1 X + W$, W is independent of X with mean zero

$$E[Y] = \theta_0 + \theta_1 E[X]$$

$$\theta_0 = E[Y] + \theta_1 E[X]$$

Since we don’t have E[X] and E[Y] , we replace them by their estimated value, also we don’t know θ_1 , but we have an estimate $\hat{\theta}_1$, we can predict $\hat{\theta}_0$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

Foe estimating θ_1 , assume E[X]=0 and E[W]=0

$$YX = \theta_0 X + \theta_1 X^2 + WX$$

taking expectation both side

$$cov(X, Y) = \theta_1 var(x)$$

Since we don't have cov(X,Y) and var(X), we estimate them.
After estimating we got the same formula as above.
Some common concerns:

- Heteroskedasticity
- Multicollinearity