* Probability Distributions :-

* Binomial (Discrete)

$PMF = {}^nC_x \, p^x (1-p)^{n-x}$

n, p are the parameters of distribution

Mean and Variance :-

$\mu = E(X) = np$

$\sigma^2 = V(x) = np(1-p)$

* Poisson Dist :-

Let λ be the rate at which the event occurs.

t be the length of time interval

X be total no. of events in that time interval.

$\mu = \lambda t$

$P(X=x) = e^{-\mu} \dfrac{\mu^x}{x!}$

μ is parameter of the distribution

$E(X) = \mu$

$V(X) = \mu$

Continuous dist :-

* Uniform dist :-

If Random variable is defined by

$f(x) = \dfrac{1}{b-a} \quad , -\infty < a \leq x \leq b < \infty$

$E(X) = \dfrac{a+b}{2}$

$V(X) = \dfrac{(b-a)^2}{2}$

* Normal Dist :-

$f(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$

$E(X) = \mu, \quad V(X) = \sigma^2$

* Exponential Dist :-

$f(x) = \lambda e^{-\lambda x}, \; \forall x \geq 0$

λ - Rate (Instantanious failure Rate)

$E(X) = \frac{1}{\lambda} \quad , V(X) = (1/\lambda)^2$

$P\{X \leq x\} = 1 - e^{-\lambda x}$

$P\{X > x\} = e^{-\lambda x}$

$P\{x_1 < X \leq x_2\} = e^{-\lambda x_1} - e^{-\lambda x_2}$

MLE of Poission Dist:-

Given $x_1 --- x_n$ are independent random Poission variables with rate (mean) λ and,

$$PMF\ (x_i = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

for $k = 0, 1, 2 ---$

$$L(\lambda \mid x_1, x_2 --- x_n) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$$l(\lambda \mid x_1, x_2 --- x_n) =$$

$$\sum_1^n \log\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right)$$

$$= \sum_1^n \left[-\lambda + x_i \log\lambda - \log(x_i!)\right]$$

To find maximum likelihood we diff wrt λ:-

$$\frac{d\,l(\lambda)}{d\lambda} = -n + \sum_1^n \frac{x_i}{\lambda} = 0$$

$$\hat{\lambda} = \frac{1}{n}\sum_1^n x_i$$

MLE of normal dist:-

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$L(\mu, \sigma) x_1 --- x_n) = \prod_1^n \left[\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)\right]$$

$$l(\mu, \sigma \mid x_1 --- x_n) =$$

$$\sum_1^n \left[-\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]$$

$$= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \sum_1^n\left(\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

For μ we diff wrt μ and for σ we diff wrt σ.

$$\hat{\mu} = \frac{\sum_1^n x_i}{n} \quad \Big| \quad \sigma = \sqrt{\frac{\sum_1^n (x_i-\mu)^2}{n}}$$

MLE for Bernoulli:-

$$f(x) = p^x (1-p)^{1-x}$$

$$L(p) = f(x_1, --- x_n \mid p) = \prod_1^n p^{x_i} (1-p)^{x_i}$$

$$l(p) = \sum_1^n \left[x_i \log p + x_i \log(1-p)\right]$$

Diff wrt p:-

$$= \frac{\sum_1^n x_i}{p} + \frac{\sum_1^n x_i}{1-p}$$

$$\hat{p} = \frac{\sum_1^n x_i}{n}$$

*MLE for Beta dist:-

It is a continuous dist. on [0,1] parameterised by 2 +ve shape parameters denoted as α and β which represent prior knowledge about dist.

$$L(\alpha, \beta \mid x_1, --- x_n) = \prod_1^n \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1} (1-x_i)^{\beta-1}$$

Γ is gama function

$$l(\alpha, \beta) = \sum_1^n (\alpha-1)\log(x_i) + (\beta-1)\log(1-x_i)$$

$$-\log(\Gamma(\alpha)) - \log(\Gamma(\beta))$$

$$+\log(\Gamma(\alpha+\beta))$$

We will diff and will get:-

$$\hat{\alpha} = \frac{\sum_1^n x_i (\alpha-1) + \sum_1^n (1-x_i)(\beta-1)}{\alpha+\beta-2n}$$

$$\hat{\beta} = \frac{n(\alpha+\beta)}{\alpha+\beta-2n}$$

Bayes Thm :-

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$

* Maximum A-Posteriori Estimation :-

From Bayes :-

$$P(\theta|D) = \frac{\overbrace{P(D|\theta)}^{\text{Likelihood}}\,\overbrace{P(\theta)}^{\text{Prior}}}{P(D)}$$

$$\theta_{ML} = \arg\max_{\theta} \sum_{i}^{n} \log P(x_i|\theta)$$

$$\theta_{MAP} = \arg\max_{\theta} \left[\sum_{i}^{n} \log P(x_i|\theta) + \sum_{i}^{n} \log P(\theta) \right]$$

* Linear Regression :-

$$\hat{y} = w_0 + w_1 x_i$$
$$= x^T w$$

Learn a function which passes through as many lines as possible : Minimize Least square Error :-

$$E(w) = \frac{1}{2} \sum_{i}^{n} (y_i - \hat{y})^2$$

diff wrt w

$$\nabla E(w) = \nabla \left[\frac{1}{2} \sum_{i}^{n} \| (y_i - x^T w) \|^2 \right]$$

$$= xy - xx^T w = 0$$

$$w_{ML} = (xx^T)^{-1} xy$$

* While calculating linear regression :-

$$m = \frac{(n\Sigma xy - \Sigma y \Sigma x)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$c = \frac{(\Sigma y - m \Sigma x)}{n}$$

* Polynomial Regression :-

$$y(x, w) = w_0 + w_1 x + w_2 x^2 + \dots + w_{M-1} x^{M-1}$$
$$= \sum_{j=0}^{M-1} w_j x^j$$

Coefficients can be learned by minimizing Error function :-

$$E(w) = \frac{1}{2} \sum_{n=1}^{n} (y(x, w) - t_n)^2$$

t_n - actual data-point

$$\phi = \begin{bmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_{M-1}(x_2) \\ \vdots & \vdots & & \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_{M-1}(x_N) \end{bmatrix}_{N \times M}$$

$$\boxed{w = (\phi^T \phi)^{-1} \phi^T t}$$

-> Adding Regularization coefficient to Error function in order to minimise over-fitting :-

$$\hat{E}(w) = \frac{1}{2} \sum_{n=1}^{N} \{ y(x, w) - t_n \}^2 + \frac{\lambda}{2} \|w\|^2$$

Regularized Least Squares :-

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \{ t_n - w^T \phi(x_n) \}^2 + \frac{\lambda}{2} \sum_{j=1}^{M} |w_j|^q$$

If $q = 2$ -> Ridge Regression

If $q = 1$ -> Lasso Regression

Generally the family of regression containing this regularized least squares error function with different values of q -> Elastic net regularization

The regularized least squares solution is :-

$$w = (\lambda I + \phi^T \phi)^{-1} \phi^T t$$

* Least Squares = MLE

$$\arg\min_{w} E = \arg\max_{w} L$$

* Regularized Least Squares = MAP

* **Bias Variance decomposition:-**

$$E_D[\{y(x;D) - h(x)\}^2]$$

$$= \underbrace{\{E_D[y(x;D)] - h(x)\}^2}_{\text{bias}^2} + \underbrace{E_D[\{y(x;D) - E_D[y(x;D)]\}^2]}_{\text{Variance}}$$

- Overfitting - High variance
- Underfitting - High bias

* **Linear Discriminant Analysis:-**

$$m_1 = \frac{1}{N}\sum_{i \in C_1} x_i \;, \; m_2 = \frac{1}{N}\sum_{i \in C_2} x_i$$

Maximise $m_2 - m_1 = w^T(m_2 - m_1)$

Fishers Discriminant Analysis:-

Improvement of LDA.

Maximise a function that will give large separation b/w projected classes while giving small variance within each class.

$\therefore$ find w that maximise

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$

$$m_1 = \frac{\sum_t w^T x^t x^t}{\sum_t x^t} \;, \; s_1^2 = \sum_t (w^T x^t - m_1)^2 x^t$$

FLD solution $\rightarrow$ $w = c \cdot S_w^{-1}(m_1 - m_2)$

$$S_w = S_1 + S_2$$

* WLS:-

$$Y = \beta x + \epsilon$$

$\beta \to$ Params

$\epsilon \to$ error

$x \to$ dependent term

Here,
ϵ assumes to be (multi-variate) normally distributed with mean vector 0 and non-constant variance-covariance matrix.

$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & \sigma_n^2 \end{bmatrix}$$

If we define the reciprocal of each variance σ_i^2, as the weight $w_i = 1/\sigma_i^2$, then let matrix W be a diagonal matrix containing these weights:-

$$W = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & w_n \end{bmatrix}$$

The weighted least squares estimated is then:-

$$\hat{\beta}_{WLS} = \arg\min_{\beta} \sum_{i=1}^{n} (\epsilon_i)^2$$

$$= (X^T W X)^{-1} X^T W Y$$

In OLS:-

$$\text{Least squares} = \sum_{i=1}^{n} (y_i - x_i \beta)^2$$

$$WLS = \sum_{i=1}^{n} w_i (y_i - x_i \beta)^2$$

*Polynomial Regression:-

$$y = b_0 + b_1 x + b_2 x^2$$

$$\begin{bmatrix} n & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \\ \Sigma x_i^2 y_i \end{bmatrix}$$

---

* Mid Sem revision:-

* Supervised Learning - Regression and Classification

  - Liner regression
  - Polynomial "
  - Logistic regression
  - Poission Regression
  - Naive Bayes
  - KNN
  - Decision Trees
  - Ensemble methods
  - Model Selection and regularization
  - Bias and Variance

* Unsupervised Learning:-

  - Dimensionality Reduction
    - LDA, FDA

-> Rain prediction:-

Assuming Bernoulli dist:-

$$L(P|x) = (P)^{N_0} (1-P)^{N_1}$$

$$l(P|x) = N_0 \log P + N_1 \log (1-P)$$

$$MLE = \frac{d}{dP}(l(P|x)) = \frac{N_0}{P} - \frac{N_1}{1-P} = 0$$

$$\therefore P = \frac{N_0}{N_0 + N_1}$$

Assuming a Beta prior:-

$$MAP = \frac{N_0 + \alpha - 1}{N_0 + N_1 + \alpha + \beta - 2}$$