Pre-req:

Matrices $\longrightarrow A_{m \times n}$

Determinants.

Defn: If $x \in \mathbb{R}^n$ is a vector such that, $Ax = \lambda x$ (λ is a scalar), then

i) $\lambda \rightarrow$ eigenvalue of A.

ii) $x \rightarrow$ eigen vector corr'to A

$\det(A - \lambda I) = 0$ (Caley-Hamilton equation)

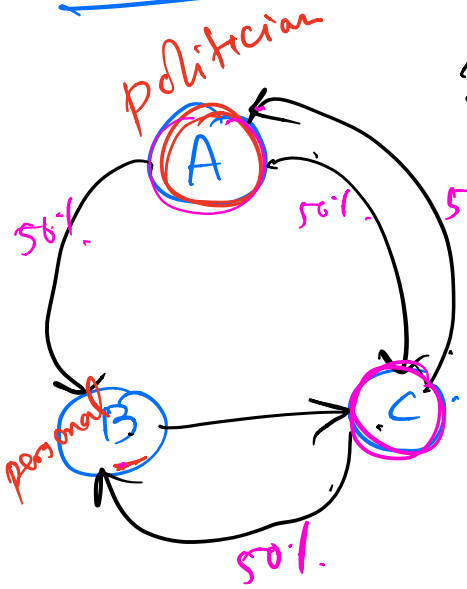
$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{bmatrix}$$

$$\begin{vmatrix} 1-\lambda & 2 & 0 \\ 0 & -\lambda & 1 \\ 0 & 0 & -1-\lambda \end{vmatrix} = 0$$

$\Rightarrow \lambda =$ 3 values

$n \times n \rightarrow$ n eigen values ($\in \mathbb{R} / \mathbb{C}$)

Page-Rank (Page & Sergei B.)



1 Billion webpages

How to rank these pages!

Ranks are recursive.

Ref: [Markov chains]

If a node has m. neighbours, each of them get $\frac{1}{m}$ traff

R_A is higher

search.

$$R_A = R_C \times \frac{1}{2}$$

$$R_B = R_A \times \frac{1}{2} + R_C \times \frac{1}{2}$$

$$R_C = R_B \times 1 + R_A \times \frac{1}{2}$$

} Flow Balance

$$x = \begin{pmatrix} R_A \\ R_B \\ R_C \end{pmatrix}$$

$$A \to \begin{bmatrix} 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 1 & 0 \end{bmatrix} \begin{pmatrix} R_A \\ R_B \\ R_C \end{pmatrix} = 1 \begin{pmatrix} R_A \\ R_B \\ R_C \end{pmatrix}$$

all column sums are 1

$Ax = x \Rightarrow x$ is an eigen vector of A

Network analytics, stochastics

$Ax = a$

$(A - I)x = 0$ → finding exact solution is time taking if A is large.

If A has size say $100k \times 100k$

① start with arbitrary v.

② $v, Av, A^2v, A^3v, A^4v, \cdots$

$A^n v \Rightarrow x$ [PF thm. in MC.]

$A^{n+1}v = A \cdot (A^n v)$, $A^{n+1}v \approx x$, $A^n v \approx x$

Iteratively getting approx sol.

$\Rightarrow Ax \approx x$ [approximation]

(Machine learning)

Thm 1:
If A is a real, symmetric matrix, then the eigen-vectors of A, form an ortho-gonal basis. $\#\{v_1, v_2, \ldots, v_n\}$ → Basis for C(A)

$Av_i = \lambda_i v_i$

$v_i \perp v_j$

Thm 2 (Spectral / Eigen decomposition)
If A is symmetric, then you can express A as $A = Q \Lambda Q^T$, where

$$Q = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{bmatrix}, \quad \Lambda^n = \begin{pmatrix} \lambda_1^n & 0 & \\ 0 & \lambda_2^n & \\ & & \ddots \end{pmatrix}$$

$A = Q \Lambda Q^T$ → orthogonal

$QQ^T = I$, $Q^TQ = I$

$A^n v$

A^{100} A is 10000×10000

$A^n = Q \Lambda^n Q^T$

$A = Q \Lambda Q^T$

$A^2 = Q \Lambda Q^T Q \Lambda Q^T$

$= Q \Lambda^2 Q^T$

SVD (Singular Value Decomposition)

What if A is a general matrix?

Rectangular

$A = U \Sigma V^t$

U is orthogonal

V is orthog

σ_i's singular values

$$\Sigma = \left[\begin{array}{c|c} \sigma_1 \sigma_2 \cdots \sigma_r & 0 \\ \hline 0 & 0 \end{array}\right]$$

How to find the SVD?

$A = U \Sigma V^t$

$A^t = V \Sigma U^t$

$AA^t = U \Sigma \underbrace{V^t V}_{I} \Sigma U^t$

$= U \Sigma^2 U^t$

$\Rightarrow (AA^t) U = U \Sigma^2 \underbrace{U^t U}_{I}$

$$= U \left[\begin{array}{c|c} \sigma_1^2 \sigma_2^2 \cdots \sigma_r^2 & 0 \\ \hline 0 & 0 \end{array}\right].$$

$(AA^t) u_i = \sigma_i^2 u_i$

σ_i^2's are the eigenvalues of AA^t

$U = [u_1 \; u_2 \; \cdots]$. u_i's are eigen-vectors

$$A = \begin{bmatrix} 2 & 3 & 1 \\ -1 & 2 & 0 \end{bmatrix} \qquad AA^T$$

QR Decomposition

$A = QR \longrightarrow$ upper triangular.

Q — orthogonal

Regression

$y = X\beta + \varepsilon$, $\quad \hat{\beta} = (X^T X)^{-1} X^T y$

$X = QR$, $\quad \hat{\beta} = (R^T R)^{-1} X^T y$

$X_{n \times p}$ — p: dimension; n: data size

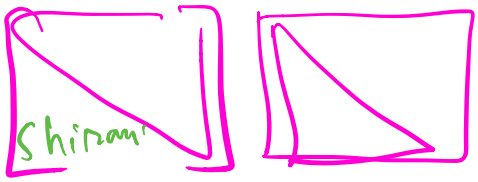
$X_{n \times 1000}$

$X^T X = R^T Q^T Q R = R^T R$

Big Data: $p = 1$ million

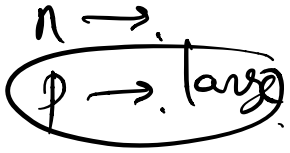
$X^T X \longrightarrow$ inversion of large matrix is hard

Ref: Elements of stat learning Tib Shirani



Sparse

Principal Comp. Analysis



Assume X to be centred.

[otw, do $X \leftarrow (I - \frac{1}{n}J)X$]

Genetics (Health AI).

$p = 100k$ features.

$n << p$.

$n = 100$ patients

Traditional Stat.

$n > p$

many would be correlated

Var-Cov. Matrix. (sample)

$$\underset{(p \times p)}{\hat{\Sigma}} = \frac{1}{n-1} \sum (x_i - \bar{x})(x_i - \bar{x})^t$$

$p = 100k$, $\Sigma =$ $\Sigma_{ij} = \text{cov}(i^{th}\text{col}, j^{th}\text{col})$, symmetric

SVD (Dimension Red.)

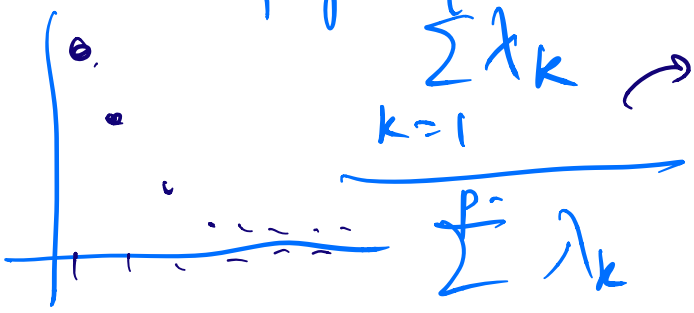
$\Sigma = Q \Lambda Q^t$

$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix}$

Thumb Rule: Reduce Size $\Rightarrow$ Do SVD

usually, a few top λ's are very large compared to others

screeplot



If you take i components, $\frac{\sum_{k=1}^{i} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$ → % of explained variance

often

e.g. top 10% λ_i's ∧ explain 90% variance

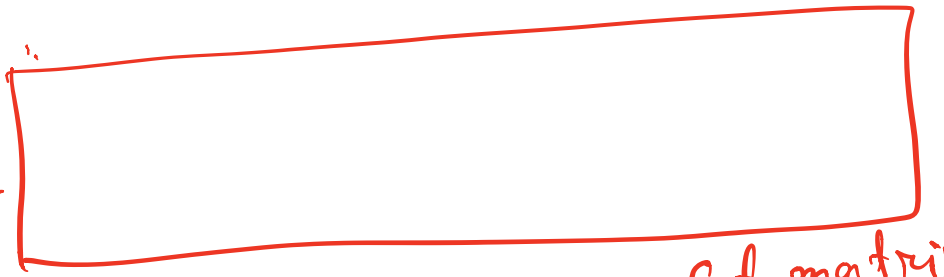
$X_{n \times p}$

$Q_{p \times p}$

$Q = [\, q_1 \; \cdots \; q_k \; \cdots \; q_p \,]$

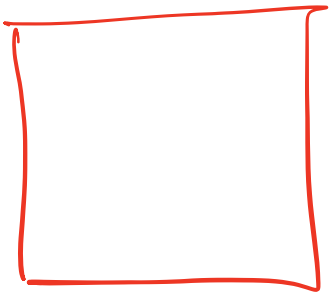
$Y = X Q_k$ linear dim. reduction $Q_k \to p \times k$.

$n \times k$.

$X =$ 

$n \times 1$ long.

fat matrix.

$Y =$ 

skinny matrix

$n \times k$

Choose k so the 90% variance is explained

i.e. $\dfrac{\sum_{1}^{k} \lambda_i}{\sum_{1}^{p} \lambda_i} \geq 0.9$.

Σ is always psd.

(positive-semi-definite).

Σ is symmetric

$\Sigma_{ij} = \text{cov}(i^{th}\text{ rv}, j^{th}\text{ rv.})$

$= \Sigma_{ji}$

Σ is symmetric (Thm 2).

$\rightarrow \Sigma = Q \Lambda Q^T$

$= Q \Lambda Q^{-1}$

(so, diagonalizable)

Manifold learning.
Spectral embedding
t-SNE

(dimension reduction)

References are in several online resources.

Probability Distributions

* There are more complex stuff, but I'm going to ignore them for now.

1) Continuous Distn.
   1) Gaussian.
   2) Cauchy.
   3) exponential
   4) Gamma
   5) Beta

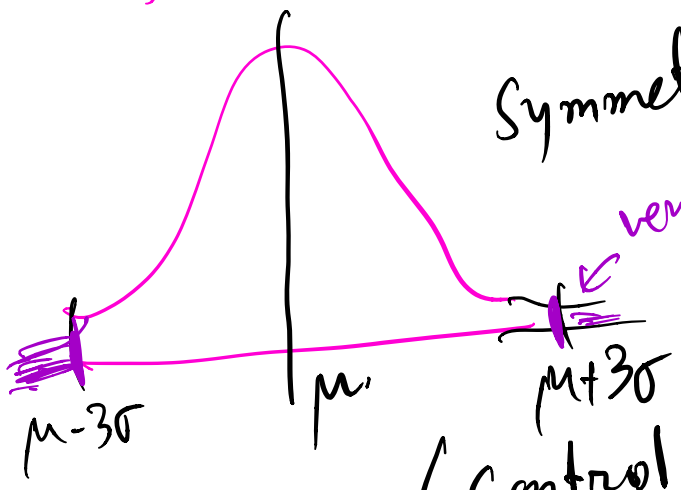
2) Discrete
   1) Binomial
   2) Bernoulli
   3) Poisson
   4) Geometric.

Continuous [$f(x) \rightarrow$ PDF].

$$P(a<X<b) = \int_a^b f(x)\,dx \quad \& \quad f \geq 0.$$

① normal (μ, σ^2).

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$



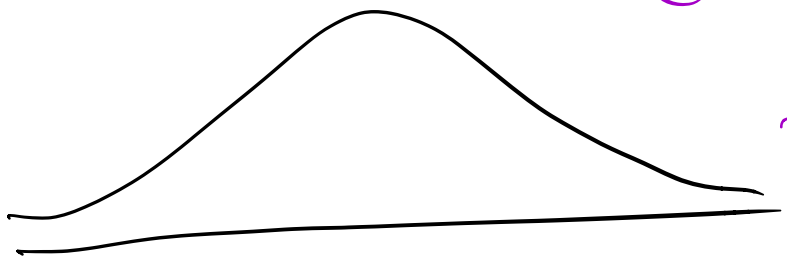
Symmetric.

$P(X \in \mu \pm 3\sigma) \simeq 99\%$.

Range $(X) = \mathbb{R}$

(control limits)

Cauchy has long tail.



$$f(x) = \frac{1}{\pi(1+x^2)}$$

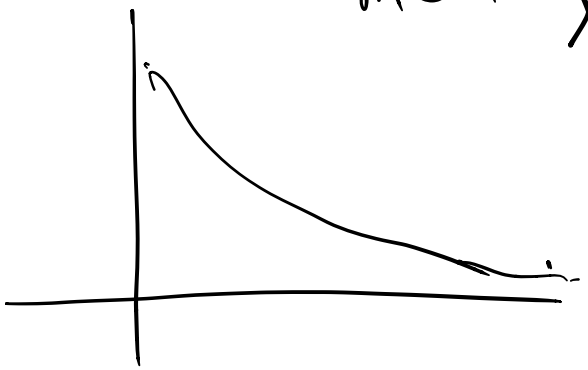
$$f(x) = \frac{1}{\pi(1+(x-\mu)^2)}$$

Cauchy's mean is undefined

[Feller 1- -].

exp : $f(x) = \lambda e^{-\lambda x}$ $\quad \lambda \to$ rate

$\text{mean} = \frac{1}{\lambda} =$



Gamma : $f(x) = e^{-\lambda x} x^{\alpha-1} \frac{\lambda^{\alpha}}{\Gamma(\alpha)}$

(Poisson process) $\quad \alpha = 1 \to$ exp.

Binomial $\to$ n trials, $p = P(\text{succ.})$

$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

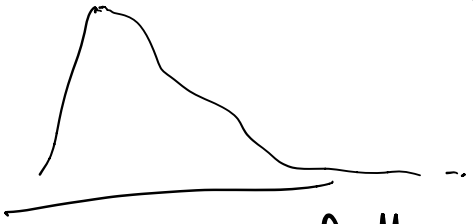
Bernouli $\to$ $n = 1$, $\text{Bin}(1, p)$

$P(X = 1) = p, \quad P(X = 0) = 1 - p$

$X \in \{0, 1\}$ $\quad$ H/T

Poisson → $P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$

$x \geq 0, \quad x \in \mathbb{Z}$



If falls very very quickly.

Modelling Rare events (accidents on Hanz Khas)

Naive Bayes

p features (discrete / cont.)

$X = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}_{n \times p}$ training data

$X = \begin{bmatrix} \text{gender} & \text{Race} & \text{Nat} \end{bmatrix}$

$\{C_1, \ldots C_k\}$

Posterior

$$P(C_i|x) = \frac{P(x|C_i)\,P(C_i)}{\sum P(x|C_i)\,P(C_i)}$$

$\pi_i \simeq \frac{n_i}{n}$ [JEE days]

C_1, C_2, C_3

Test sample

$x = (x_1, x_2, \ldots, x_p)$

$P(C_1|x), P(C_2|x)$

$P(C_3|x)$

Naive assumption: Given C_i, $x_1, x_2 \ldots x_p$ are indep.

UCI, IRIS

$$P(x|C_i) = \prod_{j=1}^{p} P(x_j|C_i)$$

[Works fantastic in many cases]

$\hat{P}(x_j|C_i)$

$P(C_1|x) = 42\%$

$P(C_2|x) = 27\%$

$P(C_3|x) = 31\%$

$x_j \rightarrow$ nationality $\rightarrow$ Indian, Chinese, Srilankan

assign x to C_1

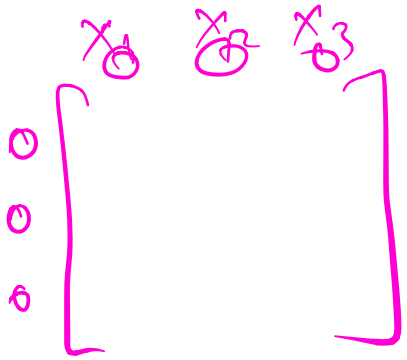
$\hat{P}(x_j = \text{Srilanka} \mid C_i) \Rightarrow$ training

$$\frac{\#\{\text{Srilankans in } C_i \text{ in training data}\}}{\#\{C_i\}}.$$

Assign x to k, where

$$k = \underset{i}{\operatorname{argmax}}\ \hat{P}(C_i \mid x).$$

Some answers to questions/doubts :



$$\Sigma = \begin{bmatrix} \mathrm{cov}(X_1, X_1) & & \\ & \mathrm{var}(X_2) & \\ & & \mathrm{var}(X_3) \end{bmatrix}$$

$(i, j) \rightarrow \mathrm{cov}(X_i, X_j)$

orthogonal $\{v_1 \ldots v_n\}$

$v_i \perp v_j \quad \forall i \neq j$

orthonormal : orthogonal + $\|v_i\| = 1 \quad \forall i$

References:

1. Tib Shirani, ESL (Stanford)
2. Wiki pages on Naive Bayes
3. Bhima-Sankaran & Rao, Linear algebra.
4. Gilbert Strang, (MIT), Linear algebra

---

errata: (1) In the page ranking example I'd referred x as eigen-value. I meant x to be the eigenvector.

(2) In the naive bayes case, ..., I was,

During the presentation, trying to mean. $P(C_1|x) \propto a$, $P(C_2|x) \propto b$, $P(C_3|x) \propto c$. and so on. [ignoring the common Denominator]. I mistakenly referred to them as exact equalities.

The actual probabilities would be.

$\frac{a}{a+b+c}$, $\frac{b}{a+b+c}$, $\frac{c}{a+b+c}$ resp.

Feel free to get in touch if you want to discuss anything about ML.