INTRODUCTION TO DATA SCIENCE

1. PROBABILITY: (lecture 1)

- DEFINITIONS:
  - Experiment is an activity that produces distinct well defined possibilities called outcomes
  - Ω = {all outcomes} := sample space
  - Trial: doing an experiment once and getting an outcome.
  - Events: subsets of Ω.
  - $\omega \in E \Rightarrow E \subset \Omega$ occured

- "THE LONG TERM RELATIVE FREQUENCY": $N(A,n) = \frac{1}{n} \cdot (\#\text{ of times } A \text{ happens})$

  1. $N(\Omega,n) = 1 \quad \wedge \quad N(A,n) \in [0,1]$
  2. $A \cap B = \emptyset \rightarrow N(A \cup B, n) = N(A,n) + N(B,n)$
     - 2.1 $A_1, \dots, A_j, \dots \; A_i \cap A_k = \emptyset \Longrightarrow N(\bigcup_{k=1}^{\infty} A_k, n) = \sum_{k=1}^{\infty} N(A_k, n)$
  3. Each experiment is independent of eachother

- DEF: $\mathcal{F}$ a collection of subsets of Ω is σ-algebra if:
  (1) $\Omega \in \mathcal{F}$
  (2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
  (3) $A_1, A_2, \dots \in \mathcal{F} \Longrightarrow \bigcup A_i \in \mathcal{F}$ "

- DEF: $\mathcal{F}$ σ-algebra. $\mathbb{P}: \mathcal{F} \longmapsto [0,1]$ is a probability measure if:
  (1) $\mathbb{P}(\Omega) = 1$
  (2) $A \cap B = \emptyset \Longrightarrow P(A \cup B) = P(A) + P(B)$

  $(\Omega, \mathcal{F}, \mathbb{P})$:= "probability triple"

- DEF: $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{P(A)}$:= conditional prob (B given A)

- EX: Ω = {"free, spam", "no free, spam", "free, not spam", "no free, not spam"}
  Probability that the text is spam given that it contains "free"?
  B = {"free, spam", "free, not spam"}
  A = {is spam} = {"free, spam"; "no free, spam"}
  $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{\text{free spam}\})}{P(B)}$

- FORMULAS:
  - $P(A^c) = 1 - P(A)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - $P(A \cup B) \leq P(A) + P(B)$
  - $A \perp B$ INDEED $\Rightarrow P(A \cap B) = P(A)P(B)$
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$

2. RANDOM VARIABLES

- DEF: $(\Omega, \mathcal{F}, \mathbb{P})$ prob. triple. A random variable $X: \Omega \longmapsto \mathbb{R}$ is a function such that $\forall x \in \mathbb{R}, \quad X^{-1}((-\infty, x]) := \{\omega : X(\omega) \leq x\} \in \mathcal{F}$.

  We assign probability to the RV X as follows:

$$\mathbb{P}(X \leq x) = \mathbb{P}(X^{-1}((-\infty, x])) = \mathbb{P}(\{\omega : X(\omega) \leq x\})$$

- EX: $\Omega = \{H, T\}$. $X(\omega) = \begin{cases} 1, & \text{if } \omega = H \\ 0, & \text{if } \omega = T \end{cases}$

  $X^{-1}((-\infty, 0]) = \{T\}$

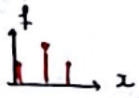
  $X^{-1}((-\infty, 1]) = \{H, T\} = X^{-1}((-\infty, 2])$

- DEF: X is a RV DISCRETE if it takes discrete values: $0, 1, 2, 3, \ldots$

- DEF: Let X be a $\mathbb{R}$-valued discret RV. $f: \mathbb{R} \longmapsto [0,1]$ where

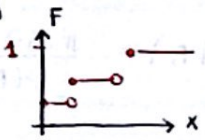
$$f(x) := P(X = x) = \mathbb{P}(\{\omega : X(\omega) = x\}) = \begin{cases} p_i, & x = x_i \in X \\ 0, & \text{otherwise} \end{cases}$$

X takes $x_1, x_2, \ldots$



"PROBABILITY MASS FUNCTION" PMF

- DEF: $F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega : X(\omega) \leq x\})$:= "DISTRIBUTION FUNCTION" (or comulative DF)
  - non decreasing (gorakor): $x_1 < x_2 \rightarrow F(x_1) \leq F(x_2)$
  - continuous (right)
  - X has DF $F(x) \Longrightarrow X \sim F$



- FORMULAS:
  - $F_X(x) = P(X \leq x) = P(\{X = x_1\} \cup \{X = x_2\} \cup \ldots \cup \{X = x_k\} \mid x_{k+1} > x\})$

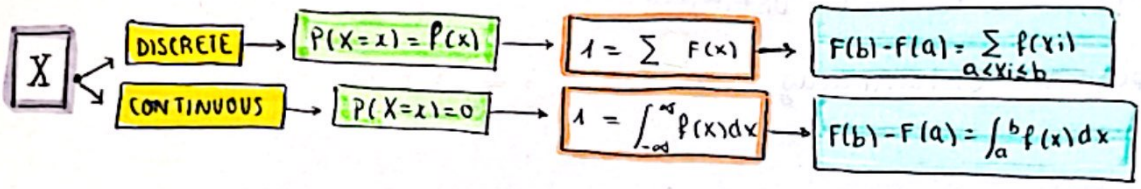
    $= \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} f_X(x_i)$
  - $F_X(b) - F_X(a) = \sum_{a < x_i \leq b} f_X(x_i)$
  - $\sum_{x_i} f_X(x_i) = 1$
  - $\lim_{x \to -\infty} F(x) = 0$
  - $\lim_{x \to +\infty} F(x) = 1$

- DEF: $\mathbb{E}[X] = \sum_x x f(x)$:= expectation / mean

- COMMON EXPECTATIONS:
  - $V[x] = \mathbb{E}\left[(x - \mathbb{E}(X))^2\right] = \sum_{x_i} (x_i - \mathbb{E}(X))^2 f_x(x_i)$
  - $\mathbb{E}\left[(X - E(X))^K\right]$:= K-th central moment
  - $\sigma(X) = \sqrt{V(X)}$:= Standard Deviation
  - $\mathbb{E}\left[\left(\frac{X - \mathbb{E}(X)}{\sqrt{V(X)}}\right)^K\right]$:= K-th standardized moment < K=3: Skewness; K=4: Kurtosis

- DEF: $\exists f_x : \mathbb{R} \longmapsto [0, \infty)$ s.t $F_x(x) = \int_{-\infty}^{x} f_x(s)ds = \mathbb{P}(X \leq x) \Longrightarrow X$ RV CONTINUOUS

⚠ • f_x := "probability density"
• $P(X = x) = 0$



- DEF: $F_X(x) = \mathbb{P}\left(\bigcap_{i=1}^{m} (X_i \leq x_i)\right) = \mathbb{P}(X_1 \leq x_1, \dots, X_m \leq x_m) =$

$= \mathbb{P}(\{\omega : X_1(\omega) \leq x_1, \dots, X_m(\omega) \leq x_m\})$:= Joint Distribution Function

where $X : \Omega \longmapsto \mathbb{R}^m$: $X = (X_1, \dots, X_m)$

- DEF: $F_X(X, Y)$:= JDF OF $\mathcal{X} = (x, y)$. We call "Marginal Distribution" to:

$$F_X(X) = F_{X,Y}((x, \infty)) = \mathbb{P}(X \leq x, Y \leq \infty) = \mathbb{P}(X \leq x)$$

- DEF: $\mathcal{Z} = (X, Y)$
$\forall (x,y) \in \mathbb{R}^2 \quad F_{\mathcal{Z}}(x,y) = F_X(x) F_Y(y)$ $\Longrightarrow$ X and Y INDEPENDENT

- DEF: $\mathcal{Z} = (X, Y)$. $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$:= Conditional PMF / Density

▸ IF Y is discrete: $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{P(X \leq x, Y = y)}{P(Y = y)}$

(X,Y) $\mathbb{R}^2$ RV
$A \subset \mathbb{R}$: $\mathbb{P}(Y \in A) > 0$

$\Rightarrow$ $F_{X|Y}(x|A) := \dfrac{\mathbb{P}(X \le x,\ Y \in A)}{\mathbb{P}(Y \in A)}$

" KASU ORORORRA "

- DEF: $g(X)$ function of a RV X.

$$\mathbb{E}(g(X)) = \begin{cases} \sum g(x) f(x), & \text{DISCRETE} \\ \int_{-\infty}^{\infty} g(x) f(x)\,dx, & \text{CONTINUOUS} \end{cases}$$

«SEQUENCES»

- It's a random vector: $X = (X_1, \dots, X_n)$ For a fixed n, where $\underline{X} = (X_1, \dots, X_n)$.

. $F(X \le x) = F_1(X_1 \le x_1) \cdots F_n(X_n \le x_n) \Longrightarrow$ Independent Sequence

. $F_{X_i} = F_{X_j} \Longrightarrow$ Identically Distributed

. Independent $\oplus$ Identically Distributed $:=$ "I.I.D"

- PROPERTIES OF EXPECTATION:

(1) $\alpha \in \mathbb{R}$: $\mathbb{E}(\alpha X) = \alpha\, \mathbb{E}(X)$

(2) $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

(3) X & Y INDEP $\Longrightarrow$ $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$

(4) $\mathbb{E}[X \mid Y=y] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\,dx$

(5) TOWER PROPERTY: $\mathbb{E}[\mathbb{E}(X|Y)] = \mathbb{E}[X]$

⚠ $g(y) = \mathbb{E}(X \mid Y=y)$. Define: $g(\underline{Y}) = \mathbb{E}(X|Y)$

so, $\mathbb{E}[X|Y]$ is another random variable

3. CONCENTRATION

(1) LEARNING FROM DATA:

- Experiment: Randomly pick a swedish person and washing them
- RV: X represents the weight of the random individual. Assume that the weight is from 0 to 300: $\mathbb{P}(0 \leq X \leq 300) = 1$.
- We want to learn: $\mathbb{E}[X]$
- n-product experiment: check n people $\vec{X} = (X_1, \dots, X_n)$ where X_i is the weight of person i.
- Estimator: $\frac{1}{n} \sum^{n} X_i \approx \mathbb{E}[X]$ empirical mean

"How much is the empirical mean concentrated around $\mathbb{E}(X)$?"

- THEOREM: "Chebychev's inequality"

$$\forall \text{ RV } X, \forall \varepsilon > 0, \quad \mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}$$

$\bar{X}_n$:= empirical mean

Choose $\varepsilon = 10$, $X_i \leq 300 \longrightarrow \mathbb{E}(\bar{X}_n) \leq 300$

$\mathbb{P}(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| \geq 10) \leq V(\bar{X}_n)/100$

} (...) Lecture 4

(2) CONFIDENCE INTERVAL:

- $\mathbb{P}(|\bar{X}_n - \mathbb{E}(X)| \geq \varepsilon) \leq \delta$
- $\mathbb{P}(|\bar{X}_n - \mathbb{E}(X)| < \varepsilon) \geq 1 - \delta$
- $\mathbb{P}(\bar{X}_n - \varepsilon < \mathbb{E}(X) < \bar{X}_n + \varepsilon) \geq 1 - \delta$

} $\Rightarrow (\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon)$

$$I = (\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon) \Rightarrow \mathbb{P}(\mathbb{E}(X) \in I) \geq 1 - \delta$$

"Can We Do Better?":

- THEOREM: "Hoeffdings inequality"

$X_1, \dots, X_n \overset{IID}{\sim} F$

$\mathbb{P}(X_i \in [a,b]) = 1$

$\forall \varepsilon > 0$

} $\Rightarrow \mathbb{P}(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| \geq \varepsilon) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$

$\varepsilon = \sqrt{\frac{(b-a)^2}{2n} \ln\left(\frac{\alpha}{2}\right)}$ → Confidence Interval: $I = (\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon)$

$P(\mathbb{E}[\bar{X}_n] \text{ is in } I) \geq 1 - \alpha$

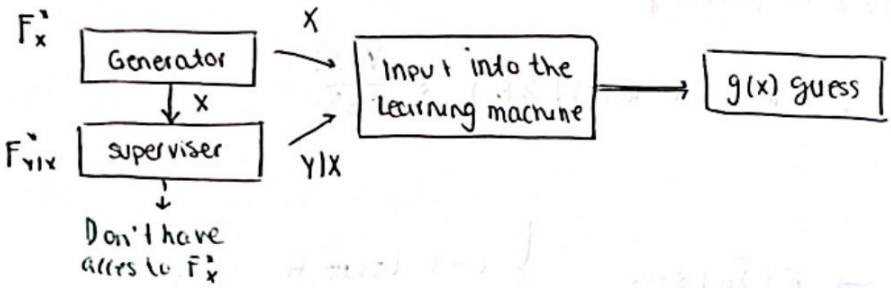
Ex: $X_1, \dots, X_n$ i.i.d Bernoulli(p) → $P(\bar{X}_n - p \geq \varepsilon) \leq e^{-2n\varepsilon^2}$

- $P(\bar{X}_n - \varepsilon < p) \geq 1 - e^{-2n\varepsilon^2}$ → $P(\bar{X}_n - \varepsilon < p < \bar{X}_n + \varepsilon) \geq 1 - 2e^{-2n\varepsilon^2}$
- $\varepsilon = \sqrt{\frac{-1}{2n} \ln\left(\frac{\alpha}{2}\right)}$

▸ BENNETS INEQUALITY: $P(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| \geq \varepsilon) \leq 2e^{\left(-\frac{na^2}{b^2} h\left(\frac{b\varepsilon}{a^2}\right)\right)}$

where $h(u) = (1+u)\log(1+u) - u$

«SUPERVISED LEARNING»



- DEF: $L: \mathbb{R}^2 \to \mathbb{R}$ loss function. Let $(X,Y) \sim F_{XY}^*$. The RISK of a function $g: X \to Y$ is $R[g] := \mathbb{E}[L(g(X), Y)]$

- EXAMPLES:

(1) Quadratic loss: $L(a,b) = (a-b)^2$ → $R(g) = \mathbb{E}[(g(x)-Y)^2]$

↳ care about LARGE errors

(2) Absolute loss: $L(a,b) = |a-b|$ → $R(g) = \mathbb{E}[|g(x)-Y|]$

↳ care about SMALL errors

(3) 0-1 Loss: $Y = \{0,1\}$ $L(a,b) = \begin{cases} 1, & a \neq b \text{ (guess + expected)} \\ 0, & \text{otherwise} \end{cases}$ → $R(g) = P(g(x) \neq Y)$

↳ $1 - R(g) :=$ "accuracy"

- DEF: "Training Error": $\hat{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} L(g(x_i), y_i)$ empirical: $(x_1,y_1), \dots, (x_n,y_n)$

"Testing Error": $\hat{R}_m(g) = \frac{1}{m} \sum_{j=1}^{m} L(g(x_{n+j}), y_{n+j})$: $(x_{n+1}, y_{n+1}), \dots, (x_{n+m}, y_{n+m})$

<<DATA >>

- D_{Tr}: $((x_1,y_1),\dots,(x_n,y_n))$ ⟶ Training Data
- D_{Te}: $((x_{n+1},y_{n+1}),\dots,(x_{n+m},y_{n+m}))$ ⟶ Testing Data

- The LM minimizes the empirical risk:

$$\hat{g} = \arg\min_{g \in \mathcal{M}} \left(\frac{1}{n} \sum_{i=1}^{n} L(g(x_i), y_i) \right)$$

We test it on D_{Te} ⟶

$$\frac{1}{n} \sum_{j=1}^{m} L(\hat{g}(x_{n+j}), y_{n+j}) \approx \mathbb{E}(R(\hat{g}) \mid \hat{g}) \neq \mathbb{E}(R(\hat{g}) \mid D_{Tr})$$

- TRUE RISK: $R(g) = \mathbb{E}[L(g(x), Y)]$

<<ESTINATION >>

possible values of T

- STATISTIC T: something we can calculate based on data $T: x_n \longrightarrow$ (T)
  - EXAMPLE: $\hat{g}$ is a statistic,
  - EXAMPLE: Testing Error: $\frac{1}{m} \sum_{i=1}^{m} L(\hat{g}(x_{n+i}), y_{n+i}) = T(D_{TE})$
- DEF: $\text{bias}(T) = \mathbb{E}[T] - \theta$ where T. is the estimator of θ
  - guess $> \theta$: positive bias
  - guess $< \theta$: negative bias
  - $\text{bias}(T) = 0$ ⟶ unbiased

↑ samples ⇒ ↓ bias

- DEF: $Se(T) = \sqrt{V(X)}$ "standard error"
- EXAMPLES:
  - $T = \frac{1}{n} \sum x_i$ ⟶ $Se(T) = \frac{\sqrt{V(x_i)}}{\sqrt{n}}$
  - $T = x_i$ ⟶ $Se(T) = \sqrt{V(x_i)}$
- DEF: "Mean Squared Error":

$$MSE(T) = \mathbb{E}[(T-\theta)^2] = \mathbb{E}[(T-E(T))^2] + [\mathbb{E}[T]-\theta]^2 = [Se(T)]^2 + [\text{bias}(T)]^2$$

↳ The risk of the estimator w.r.t the quadratic loss can be decomposed as above.

<<REGRESSION >>

1 Find the regression function:

- $(X,Y) \sim F_{XY}$ (unknown)
- $L(a,b) = (a-b)^2$ LOSS FUNCTION
- Guess: f
- $R(f) = \mathbb{E}[(Y-f(x))^2]$ RISK
- $r(x) := \mathbb{E}[Y \mid X = x]$ (x Fixed)

$$R(f) = \mathbb{E}[((Y-r(x)) + (r(x)-f(x)))^2] = \dots =$$

$$= \underbrace{\mathbb{E}[(Y-r(x))^2]}_{\text{Measures the Noise (CAN'T REDUCE IT)}} + \underbrace{\mathbb{E}[(r(x)-f(x))^2]}_{\text{BIAS (CAN REDUCE IT)}}$$

Best Guess ⟶ $R(r)$ smallest

[2] FIND h:

- Assume $\mathbb{E}[(Y-r(X))^2] = 0$

- "PATTERN RECOGNITION":
  - $Y = \{0,1\}$ class
  - $L(a,b) = \begin{cases} 1, & a \neq b \\ 0, & a = b \end{cases}$
  - $r(x) = \mathbb{E}(Y \mid X=x) = P(Y=1 \mid X=x)$ .Regression Function
  - $h^*(x) = \begin{cases} 1, & r(x) > 1/2 \\ 0, & r(x) \leq 1/2 \end{cases}$:= Bayes Classifying Rule
    - ↳ $R(h^*) \leq R(h)$ for any guessing function h

- "MAXIMUM LIKELIHOOD":
  - $\mathcal{S} = \{P_\alpha(x) : \alpha \in \mathbb{R}^q\}$ parametric model
  - $L(x, \underset{\uparrow}{\alpha}) = -\ln P_\alpha(x)$ log-loss
    - ↑ parameter
  - $X \sim P_{\alpha^*} \in \mathcal{S}$: $R(\alpha) = \mathbb{E}[L(X,\alpha)] = \mathbb{E}[-\ln P_\alpha(X)] = -\int \ln P_\alpha(x)\, P_{\alpha^*}(x)\, dx$
  - $X_1, \dots, X_n \overset{IND}{\sim} P_{\alpha^*}$: $\hat{R}(\alpha) = \frac{1}{n}\sum^{n} -\ln P_\alpha(x_i)$ empirical Risk

EXAMPLES:

(1) Linear Regression:

- $P_\alpha(y \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y-(ax+b))^2}{\sigma^2}}$ $\Big\} \Rightarrow \hat{R}(\alpha) = \ln|\sigma| + \frac{1}{n}\sum^{n} \frac{(y_i-(ax_i+b))^2}{\sigma^2} + \ln\sqrt{2\pi}$
- $\alpha = (a,b,\sigma)$

(2) Logistic Regression

- $P_\alpha(y \mid x) = \text{Bernoulli}(p) = p^y(1-p)^{1-y}$
- $G(x) = \frac{1}{1+e^{-x}} \in (0,1)$ Logistic Function

$L(z,\alpha) := -\ln P_\alpha(z)$ where P_α proposed density for our data

- $\hat{R}(\alpha) = \frac{1}{n}\sum -\ln(P_\alpha(x_i))$
- $R(\alpha) = E(\ln(P_\alpha(x_i)))$

« GENERATING RANDOM VARIABLES » (Markov Chain)

1 PRNG : PSEUDO RANDOM NUMBER GENERATOR

- DEF: A sequence $u_0, u_1, \dots$ is pseudorandom if for any $a \in \mathcal{U} = \{0,1,\dots,M-1\}$

$$\frac{N_n(a)}{n} \longrightarrow \frac{1}{M}.$$

- EX: $0,1,0,1\dots$ is pseudorandom on $\{0,1\}$
- DEF: A congruential generator with parameters (a,b,M):
  - $D(x) = (ax+b) \bmod M$
  - start at some u_0 - random seed
  - $u_i = D(u_{i-1})$
  - Period: of D starting at u_0 is the smallest T_0 such that $u_i + T_0 = u_i \ \forall i$
  - Full period: on $\{0,1,\dots,M-1\}$ is M (you get all the numbers)
    - $u_0 = 0 \xrightarrow[\text{steps}]{M}$ get 0 : $\frac{N_{K\cdot M}(0)}{K\cdot M} = \frac{K}{KM} = \frac{1}{M}$

↳ So: $u_0, u_1, \dots$ is pseudorandom

IMPROVE IT :

- Large M
- Assume full period : $u_0, \dots, u_n$
- Fix K s.t $K \mid M$
- Construct : $V_i = \lfloor u_i \frac{K}{M} \rfloor$ (floor in python)
  - V_i pseudorandom on $\{0,1,\dots,K-1\}$
  - Period: M

⚠ To create $\text{Unif}([0,1])$ we take : $V_i = \frac{u_i}{M}$

« SAMPLING »

1 INVERSION SAMPLING

- We want to sample from F
  1. Calculate F^{-1} (if we can)
  2. Generate $u_{\text{unif}}([0,1])$
  3. Consider $X = F^{-1}(u) \longrightarrow X \sim F$

u uniform
$F_u(u) = u$

$$F_X(x) = P(X \le x) = P(F^{-1}(u) \le x) = P(u \le F(x)) = F(x)$$

• EX: $F(x) = 1 - e^{-\lambda x}$ exponential

$y = 1 - e^{-\lambda x} \longrightarrow e^{-\lambda x} = 1 - y \longrightarrow -\lambda x = \ln(1-y) \longrightarrow x = -\frac{1}{\lambda}\ln(1-y) = F^{-1}(y)$

2 ACCEPT - REJECT METHOD

- Input target density: f
- Sampling density: g

} we need a sampling density that dominates f: $f \leq Ng(x)$ for all x ($N < 1$)

1. Draw $X \sim g$ and calculate $r(X) = \frac{f(X)}{Ng(X)}$
2. Draw a $U \sim \text{Unif}[0,1]$
3. { accept X, $U \leq r(X)$
   reject x, otherwise + try again

↳ $I = \begin{cases} 1, & U \leq r(X) \\ 0, & \text{otherwise} \end{cases}$

↳ IF $I = 1$ (ACCEPT):

$P(X \leq x \mid I = 1) = F(x)$

- Bayes theorem: $f_{X|I}(x|1) = \frac{f_{I|X}(1|x)\, f_X(x)}{f_I(1)}$
- $f_{I|X}(1|x) = P(U \leq r(Y) \mid X = x) = r(x)$
- $f_I(1) = \int f_{I|X}(1|x)\, f_X(x)\, dx = \int r(x) g(x)\, dx = \int \frac{f(x)}{Ng(x)} g(x)\, dx =$
  $= \frac{1}{N}\int f(x)\, dx = \frac{1}{N}$
- $f_X(x) := g(x)$

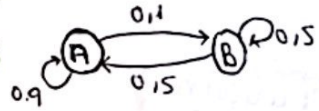
• DEF: A stochastic process is indexed set of random variables $\{X_i\}_{i \in I}$.
If $I = \{1,2,3,\dots\}$, we say it is discrete.

• DEF: A Markov process $\{X_i\}_{i \in \mathbb{N}}$, $X_i \in X = \{S_1, \dots, S_N\}$:

$P(X_t = y \mid x_1, \dots, x_{t-1}) = P(X_t = y \mid X_{t-1})$:= Markov Property

↳ Only cares about the last one (previous don't affect)

• EX:



$P = \begin{pmatrix} 0,9 & 0,1 \\ 0,5 & 0,5 \end{pmatrix}$ A = "dry", B = "wet"

▲ $P_{t-1} = [P(X_{t-1} = \text{dry}),\ P(X_{t-1} = \text{wet})]$ we know

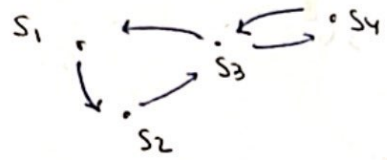
- $P(X_t = \text{dry}) = P(X_t = \text{dry} \mid X_{t-1} = \text{dry})\, P(X_{t-1} = \text{dry}) +$
  $+ P(X_t = \text{dry} \mid X_{t-1} = \text{wet})\, P(X_{t-1} = \text{wet})$

▲ $P_t = P_{t-1} \cdot P = P_0 \cdot P^t$

- DEF: $X = \{S_1, \dots, S_N\}$. We say that $S_i \longrightarrow S_j$ "state S_i communicates with S_j" if $P(X_t = S_j \mid X_1)$ for some t
- DEF: $S_i \longrightarrow S_j \ \wedge \ S_j \longrightarrow S_i \longrightarrow$ They intercommunicate
- DEF:
  - Everything communicates $\longrightarrow$ Irreducible
  - If not $\longrightarrow$ Reducible
- DEF: Return times: How many steps to go back (get stuck):

$$T(x) = \{t \in \mathbb{N} \cdot (P^t)_{xx} > 0\}$$

↳ EXAMPLE:



$T(S_1) = \{3, 5, 7 \dots\}$
$T(S_3) = \{2, 3, 5, \dots\}$

- DEF: $\gcd(T(x))$:= "Period of the state x"
- DEF: $\gcd(T(S_i)) = 1$ for all states $\longrightarrow$ The chain is APERIODIC
- THEOREM:
  - Irreducible
  - Aperiodic

$\Big\}$ $P_0 P^t \longrightarrow \pi_{mixing}$ (UNIQUE)

$\pi = \pi P \longrightarrow$ STATIONARY DISTRIBUTION

$\pi^T = (\pi P)^T = P^T \pi^T \longrightarrow$ Eigen vector of P^T with eigenvalue 1

« LINEAR CLASSIFIER »

- $x \in \mathbb{R}^d$, $y \in \{-1, 1\}$, $w \in \mathbb{R}^d$, $c \in \mathbb{R}$
  - $g_w(x) = \operatorname{sgn}(\langle w, x \rangle + c)$
  - $L(y, g_w(x)) = -\dfrac{\operatorname{sgn}(\langle w, x \rangle + c)\, y + 1}{2} = \begin{cases} 1, & y \neq g_w(x) \\ 0, & y = g_w(x) \end{cases}$:= 0-1 LOSS
  - $\langle \tilde{w}, \tilde{x} \rangle = \langle w, x \rangle + c$
    - $\tilde{w} = (w_1, \dots, w_d, c)$
    - $\tilde{x} = (x_1, \dots, x_d, 1)$

Two classes

$\langle \tilde{w}, \tilde{x} \rangle = 0$:= Plane from the origin

How do we find a linear classifier?

[1] PERCEPTION ALGORITHM:

(1) Set $w = (0, \dots, 0)$

(2) Points (x_i, y_i) $i = 1, \dots, n$

▲ If $\exists (x_i, y_i)$ s.t $\langle w, x_i \rangle y_i < 0 \implies$ misclassified

$$w \leftarrow w + \underbrace{x_i}_{\in \mathbb{R}^{d+1}} \underbrace{y_i}_{\in \{-1, 1\}}$$

(3) Repeat until all satisfy: $\langle w, x_i \rangle y_i > 0$

[2] SUPPORT VECTOR MACHINE:

We want to solve the problem of Uniqueness

(1) Take $\|w\| = 1$. $\langle w, x \rangle = 1$

$x - \tilde{x} \;/\!/\; w$: $|(x - \tilde{x}) \cdot w| = \|x - \tilde{x}\| \|w\| = 1$

⚠ If $\|w\| \neq 1 \implies \|x - \tilde{x}\| = 1/\|w\|$

(2) Consider the function:

$$\text{Hinge loss} := \max\{0, 1 - \langle w \cdot x \rangle y\} = \begin{cases} 1 - \langle wx \rangle y & ; \ \langle wx \rangle y < 1 \\ 0 & ; \ \text{otherwise} \end{cases}$$

(3) Minimize: $\|w\|$ under the constraint: "$\langle w \cdot x_i \rangle y_i > 1$"

"SOFT MARGIN SVM":

$$\min_w \|w\|^2 + \frac{c}{n} \sum_{j=1}^{n} \max\{0, 1 - \langle w \cdot x_i \rangle y_i\}$$

[3] THE KERNEL TRICK:

▲ Perception update rule: $w \leftarrow w + x_i y_i$ $\qquad \langle \phi(x_i), \phi(x_k) \rangle$

we will find w of the form: $w = \sum c_i x_i$

▲ $\langle w, x_k \rangle = \sum c_i \langle x_i, x_k \rangle = \sum c_i K_{ik} \xrightarrow{z_i = \phi(x_i)} \langle \tilde{w}, z_k \rangle = \sum \tilde{c}_i \tilde{K}_{ik}$

"TRICK": If we have a function $K(a,b)$ such that $\tilde{K}_{ik} = K(x_i, x_k)$ (we forget about ϕ).

$K :=$ Kernel Function

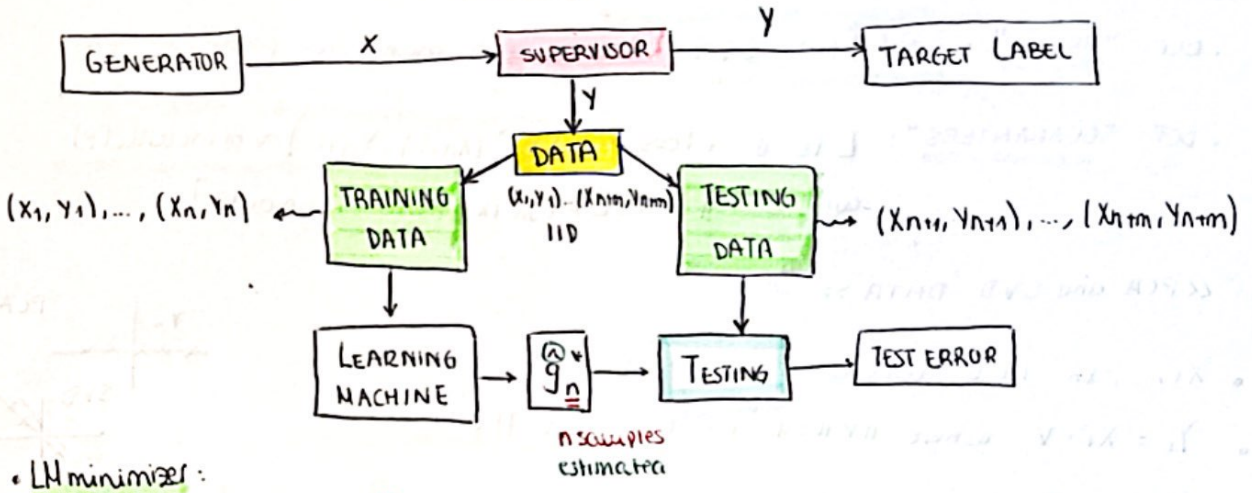
EXAMPLES:

- $K(a,b) = \langle a, b \rangle \longrightarrow$ linear
- $K(a,b) = (\gamma \langle a, b \rangle + m)^r \longrightarrow$ polynomial
- $K(a,b) = e^{-\|a-b\|^2} \longrightarrow$ radical basis function

"RADICAL BASIS FUNCTION":

$$\langle w, x \rangle = \lambda = \sum c_i K(x_i, x)$$

« TRAIN - TEST PROCEDURE »



- LM minimizer:

$$\hat{g}_n^* = \arg\min_{g \in \mathcal{U}} \frac{1}{n} \sum_{i=1}^{n} L(g(X_i), Y_i)$$ only training data

- Test error:

$$\frac{1}{m} \sum_{i=1}^{m} L(\hat{g}_n^*(X_{n+i}, Y_{n+i}))$$

PYTHON: `lr.fit(X_train, Y_train)` := LM

$\hat{g}_n^*(x)$ = `lr.predict(x)`

`lr.score(X_test, Y_test)` := Testing Error

« TEST ERROR »

- Once trained, we consider $\hat{g}_n^*$ as fixed: $\mathbb{E}[L(\hat{g}_n^*(X), Y) \mid \text{Training data}]$ is the expected loss.

- BIAS VARIANCE:

$$\mathbb{E}[|Y - \hat{g}(x)|^2 \mid X=x] = \mathbb{E}[|Y - r(x) + r(x) - \mathbb{E}[\hat{g}(x) \mid X=x] + \mathbb{E}[\hat{g}(x) \mid X=x] - \hat{g}(x)|^2 \mid X=x]$$

$$= \mathbb{E}[|Y - r(x)|^2 \mid X=x] + \mathbb{E}[|r(x) - \mathbb{E}(\hat{g}(x) \mid X=x)|^2 \mid X=x] +$$

$$+ \mathbb{E}[|\mathbb{E}(\hat{g}(x) \mid X=x) - \hat{g}(x)|^2 \mid X=x]$$

- ☐ Noise
- ☐ Bias squared: Bias^2
- ☐ Variance

- DEF: "PRECISION": Given $Y \in \{0,1,\dots,K-1\}$ and $\hat{g}$, then the precision for class $\ell \in \{0,1,\dots,K-1\}$: $\mathbb{P}(Y=\ell \mid \hat{g}(X)=\ell)$

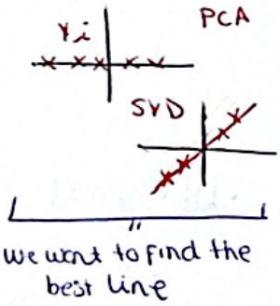
  "I have covid" — "test tells I have covid"

- DEF: "RECALL": $\mathbb{P}(\hat{g}(X)=\ell \mid Y=\ell)$ → How many sick people can detect

- DEF: "GUARANTEES": L is 0-1 loss. $L(\hat{g}_n^*(X_{n+1}), Y_{n+1}) \sim \text{Bernoulli}(p)$

  where $p = \mathbb{E}[L(\hat{g}_n^*(X_n), Y_n) \mid \text{Train data}]$

<< PCA and SVD DATA >>



- $X_1, \dots, X_n$ i.i.d samples
- $Y_i = X_i \cdot V$ where $\|V\|=1$ "coordinate on the line"
- $Y_i V$: "real coordinate"
- $\{Y_i\}_{i=1}^n$ assume $\bar{Y}_n = 0$ empirical mean
- $\frac{1}{n}\sum (Y_i - \bar{Y}_n)^2 = \frac{1}{n}\sum (X_i \cdot V)^2$ empirical variance

GOAL: select V s.t variance is maximize

- $V_1 = \underset{\|V\|=1}{\operatorname{argmax}} \left(\frac{1}{n}\sum (X_i \cdot V)^2 \right)$ → 1ST SINGULAR VECTOR
  - $\sigma_1 := \sqrt{\sum (X_i \cdot V_1)^2}$ → 1ST SINGULAR VALUE
- $V_2 = \underset{\substack{\|V\|=1 \\ V \perp V_1}}{\operatorname{argmax}} \left(\frac{1}{n}\sum (X_i \cdot V)^2 \right)$ → 2ND SINGULAR VECTOR
  - $\sigma_2 := \sqrt{\sum (X_i \cdot V_2)^2}$ → 2ND SINGULAR VALUE
- $V_3 = \underset{\substack{\|V\|=1 \\ V_2 \perp V \perp V_1}}{\operatorname{argmax}} \left(\frac{1}{n}\sum (X_i \cdot V)^2 \right)$ → 3RD SINGULAR VECTOR
  - $\sigma_3 := \sqrt{\sum (X_i \cdot V_3)^2}$ → 3RD SINGULAR VALUE

(...)

TWO STOPPING CONDITIONS

1. Reach the size of the space
2. There is no vector $V \perp V_1, \dots, V_r$ or $\sum (X_i \cdot V)^2 = 0$ for all vectors, $r \leq \min(m,n)$

- $A = \begin{pmatrix} - x_1 - \\ \vdots \\ - x_n - \end{pmatrix}_{n \times m}$:= Gram Matrix
- $Av = \begin{pmatrix} x_1 \cdot v \\ \vdots \\ x_n \cdot v \end{pmatrix} \longrightarrow \|Av\|^2 = \sum_{i=1}^{n} (x_i \cdot v)^2$
- $v_1 = \underset{\|v\|=1}{\arg\max} \left(\|Av\|^2 \right) \longrightarrow \sigma_1 = \|Av_1\|$
- Once $Av = 0$ for all $v \perp v_1, \dots, v \perp v_k$ choose a basis for the null space of A $v_{k+1}, \dots, v_r$ where $r = \min(m, n)$ so $\sigma_{k+1} = \dots = \sigma_r = 0$.
- THEOREM: $m \leq n$ (more data than dimension). We have $v_1, \dots, v_m$. Choose $k \leq m$:
  - $W_i = \sum_{j=1}^{k} (x_i \cdot v_j) v_j$ k-dimensional subspace (passes from 0) "The best fit subspace"
  - $\sum_{i=1}^{n} \|x_i - W_i\|^2 \leq \sum_{i=1}^{n} \|x_i - \text{Proj}_{H_k}(x_i)\|^2$ where H_k is any k-dim subspace (with any other the error will be bigger)
- $v_1, \dots, v_m$ basis in $\mathbb{R}^m \longrightarrow x_i = \sum_{j=1}^{m} (x_i \cdot v_j) v_j$

$$Av_j v_j^T = \begin{pmatrix} x_1 v_j \\ \vdots \\ x_n v_j \end{pmatrix} v_j^T = \begin{pmatrix} - (x_1 \cdot v_j) v_j - \\ \vdots \\ - (x_n \cdot v_j) v_j - \end{pmatrix}$$

$$\sum_{j=1}^{m} Av_j v_j^T = \begin{pmatrix} \sum_{j=1}^{m} (x_1 \cdot v_j) v_j \\ \vdots \\ \sum_{j=1}^{m} (x_n \cdot v_j) v_j \end{pmatrix} = A$$

$$u_j = \begin{cases} \frac{Av_j}{\sigma_j}, & \sigma_j > 0 \\ 0, & \text{otherwise} \end{cases} \longrightarrow \text{"Left Singular Vector"}$$

$$\sigma_j u_j = Av_j$$

- $A = \sum_{j=1}^{m} \sigma_j u_j v_j^T = \sum_{j=1}^{m} Av_j v_j^T$
- $U = (u_1 | u_2 | \cdots | u_m)_{n \times m}$
- $D = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_m \end{pmatrix}$
- $V = (v_1 | \cdots | v_m)_{m \times m}$

$\longrightarrow A = UDV^T \rightsquigarrow$ "SVD"

- "Low Rank Approximation": $A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T$ if $\sigma_j > 0$ $\rightsquigarrow \text{Rank}(A_k) = k$
- "Error": $\|A - A_k\|_{2 \text{ Frobenius}}^2 = \sum (a_{ij} - (a_k)_{ij})^2$
  - $A = \sum_{i=1}^{m} \sigma_i u_i v_i^T$
  - $A - A_k = \sum_{i=k+1}^{m} \sigma_i u_i v_i^T$

- $\|A - A_k\|_2^2 = \sum_{j=1}^{n} \|\tilde{x}_i\|_2^2 = \sum_{j=1}^{m}\sum_{i=1}^{n} |\tilde{x}_i \cdot v_j|^2 = \sum \underbrace{\|(A-A_k)v_i\|^2}_{\sum_{i=k+1}^{m} \sigma_i u_i v_i^T v_j} =$

$$= \sum_{j=k+1}^{m} \|\sigma_j u_j\|^2 = \sum_{j=k+1}^{m} \sigma_j^2$$

- SELECT K LARGE → $\sum_{j=k+1}^{m} \sigma_j^2$ small

- $\|Av\|^2 = \langle Av, Av \rangle = \langle A^T A v, v \rangle$
- $\tilde{A} = A^T A$: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m > 0$
- $w_1, \dots, w_m$ eigenvectors
- $v = \sum (v \cdot w_i) w_i$
- $\tilde{A} v = \sum_{i=1}^{m} \lambda_i (v \cdot w_i) w_i$
- $\langle \tilde{A} v, v \rangle = \sum_{i=1}^{m} \lambda_i (v \cdot w_i)^2$

$\Longrightarrow$ $\max \|Av\|^2 = \lambda_1$

⚠ Before logistic Regression: STANDARDIZE

⚠ CENTER THE DATA! ⊕ STANDARDIZE DATA

- CLEAN (quit outliers)
- TRANSFORM
- STANDARDIZE

$\Longrightarrow$ TRAINING DATA

⚠ PIPELINE (recomendado)

▸ CALIBRATOR ERROR: $c(f) = \sqrt{\mathbb{E}\left[\left|\mathbb{E}(Y|f(x)) - f(x)\right|^2\right]}$ (f(x): my prediction)

$Y \in \{0,1\}$ → $c(f) = \sqrt{\mathbb{E}\left[\left|P(Y=1|f(x)) - f(x)\right|^2\right]}$

If it is perfectly calibrated → $c(f) = 0$

▸ FEATURE ENGINEERING
- Construct features from data to make a model more powerful
- Transformation of features: replace x with log(x)
- Transform the target: Y price → $Y' = \log(y)$

« TEST DATA »

- Dictionary: enumeration of all possible words in your data.
- TF − IDF := Term Frequency − Inverse Document Frequency
  - $f_{t,d}$: count how many times t appears in d. document
  - $TF(t,d) = f_{t,d} / \sum_{s \in d} f_{s,d}$ → How many words in d
  - $IDF(t) = \log\left(\frac{N}{|\{d : t \in d\}|}\right)$ → N: # documents; denominator: in how many documents this word appears

$\Longrightarrow$ $TF \cdot IDF(t,d) = TF(t,d) \cdot IDF(t)$

▸ TF↑ : very often
▸ IDF↑ : not in every doc

PRODUCT ↑ : "interesting word"