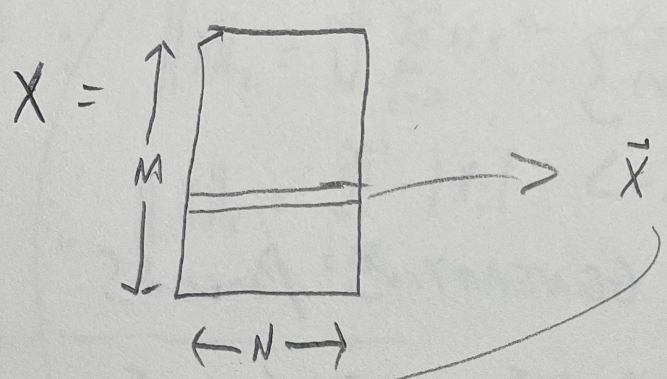
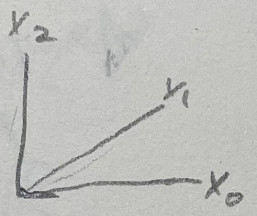
FEATURE REPRESENTATION

$X \in \mathbb{R}^{M \times N}$

M = Number of observations

N = Number of Features





↳ VECTOR SPACE:

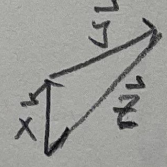
$X = [0.2, 0.5, 1.0]$

NORM: $f(\cdot) : \mathbb{R}^N \to \mathbb{R}$

$f(\cdot)$ MUST SATISFY:

① $f(\vec{x}) = 0$ iff $\vec{x} = \vec{\emptyset}$

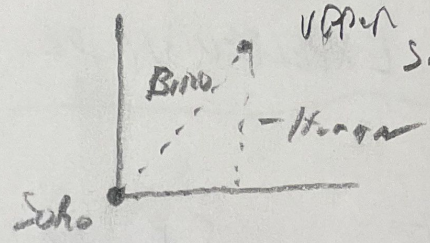
② $f(\vec{x} + \vec{y}) \leq f(\vec{x}) + f(\vec{y})$



③ $f(\alpha \vec{x}) = |\alpha| f(x)$ $\alpha \in \mathbb{R}$

L_p Norms

$$\|x\|_p = \left[\sum_{i=0}^{N-1} |x_i|^p \right]^{\frac{1}{p}}$$



L_2 : $\|x\|_2 = \sqrt{\sum_{i=0}^{N-1} x_i^2}$ <=> EUCLIDEAN

L_1 : $\|x\|_1 = \sum_{i=0}^{N-1} |x_i|$ <=> MANHATTAN

FROBENIUS NORM

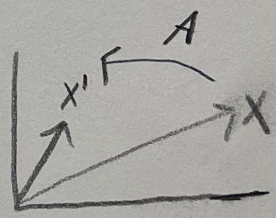
$A \in \mathbb{R}^{M \times N}$

$$\|A\|_F = \sqrt{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} A_{ij}^2} \in \mathbb{R}$$

INNER PRODUCT ("DOT" PRODUCT)

$$\langle X, Y \rangle = X \cdot Y = \sum_{i=0}^{N-1} x_i y_i \in \mathbb{R}$$

TENSOR (LINEAR TRANSFORMATION)



$x' = Ax$

$x, x' \in \mathbb{R}^N$

$A \in \mathbb{R}^{N \times N}$

$$\begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \\ A_{N1} & & A_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} A_{1i} x_i \\ \vdots \\ \sum_{i=1}^{N} A_{Ni} x_i \end{bmatrix}$$

Bag of Words (BOW)

$$X = [0, 0, 0, \ldots, \underbrace{1}_{N}, 0, 0, \ldots, 2, \ldots 0]$$

↳ D Sparse Representation

Singular Value Decomposition (SVD)

$X \in \mathbb{R}^{M \times N}$

$$X = U \Sigma V^T$$

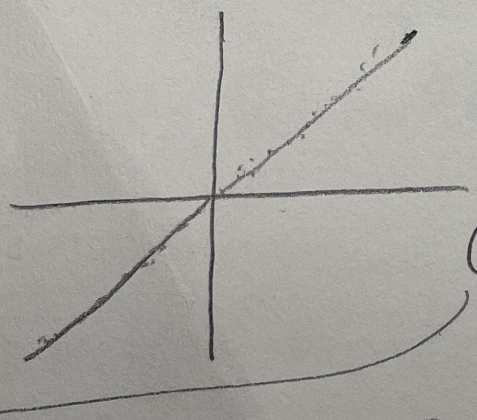
$M \times N \quad M \times M \quad M \times N \quad N \times N$

U, V are orthogonal

$U^{-1} = U^T$

$V^{-1} = V^T$

$$\Sigma = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_N \\ \hline & \text{Null Space} & \end{bmatrix}$$



① Rank(X) = Number of non-zero σ_i's := K

② Truncated SVD

$$X = U' \Sigma' V'^T$$

$M \times N \quad M \times K \quad K \times K \quad K \times N$

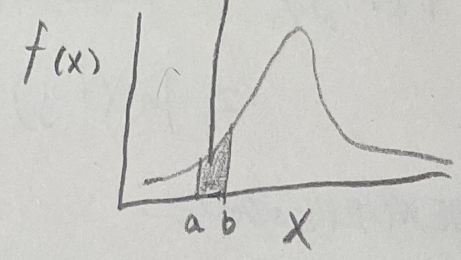
③ $q \in \mathbb{R}^N \Rightarrow \hat{q}' = qV'$

$1 \times K \quad 1 \times N \quad N \times K$

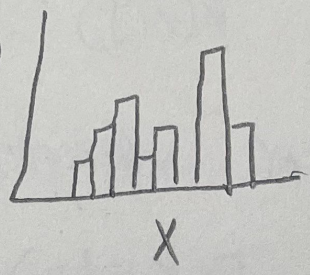
Random Variable (RV)

$P(a \le x \le b) = \int_a^b f(x)\,dx$

Continuous: PDF



Discrete: PMF



PMF must satisfy

① $\forall_{x \in X} \quad 0 \le x \le 1$

② $\sum_{x \in X} P(x) = 1$

$\sigma_{\text{softmax}}(x) = \frac{e^x}{\sum_{x' \in X} e^{x'}}$

Joint, Marginal, Conditional Dist.

Joint: $P(X, Y)$ = |X| × |Y|

Marginal: $P(x) = \sum_{y \in Y} P(x, y) \Rightarrow P(y) = \sum_{x \in X} P(y, x)$

Conditional: $P(y|x) = \frac{P(x, y)}{P(x)}$

PRODUCT RULE:

$$P(X,Y) = P(Y|X)P(X)$$
$$= P(X|Y)P(Y)$$

INDEPENDENCE:

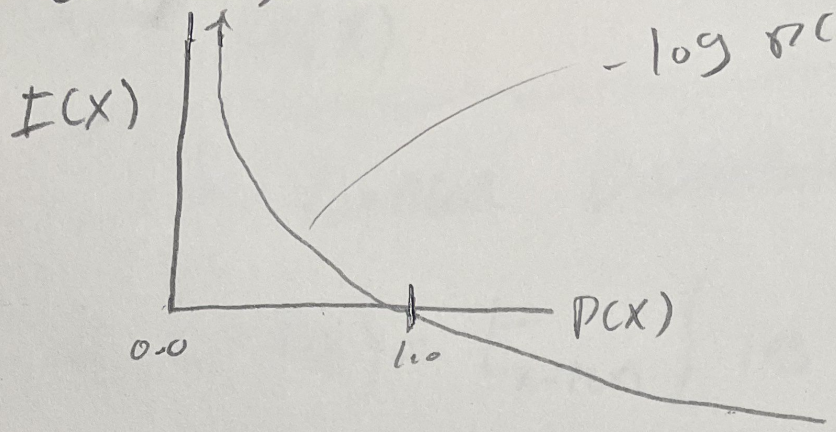
$$P(X,Y) = P(X)P(Y)$$

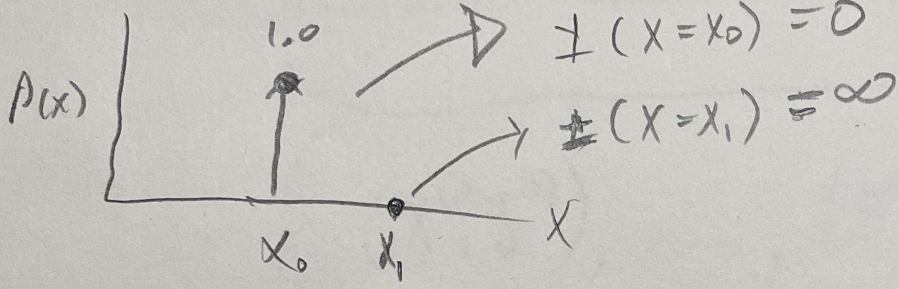
CONDITIONAL INDEPENDENCE:

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

INFORMATION THEORY

- Claude Shannon



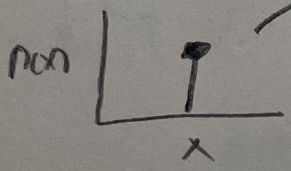


↳ **SELF INFORMATION: $I(x) = -\log P(x)$**

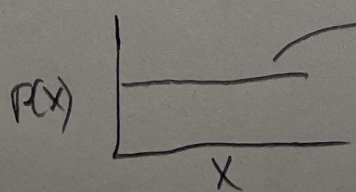
"The" → P(X = "The") is fairly High, I(x) is Low

"Russia" → P(X = "Russia") is Lower I(x) is High

ENTROPY:



→ H(P(x)) is zero



H(P(x)) is Large

$$H(P) = E_{X \sim P(x)}[I(x)]$$

$$= -\sum_{x \in X} P(x) \log P(x)$$

Divergence Between two two distributions

$P(x)$, $Q(x)$

Kullback-Leibler Divergence

$$D_{KL}(P \| Q) = E_{x \sim P(x)} \left[\log \frac{P(x)}{Q(x)} \right]$$

MAXIMUM LIKELIHOOD

COIN FLIP:

HEADS TAILS

R.V.: $y \in Y \quad Y = \{0, 1\}$

PMF: $P(y) = \begin{cases} \theta & y=0 \\ 1-\theta & y=1 \end{cases}$ "BERNOULLI"

EXPERIMENT CONSISTS OF M-TRIALS

$D = \{y^{(1)}, \ldots, y^{(M)}\}$

LIKELIHOOD IS A JOINT DISTRIBUTION OVER OUR DATA:

- $P(y^{(1)}, \ldots, y^{(M)}) \rightarrow$ JOINT DISTRIBUTION

↳ FOR EACH TRIAL, i, WE ARE DRAWING $y^{(i)}$ FROM $P(y; \hat{\theta})$

↳ EACH SAMPLE MUST BE DRAWN, I.E., EACH TRIAL OUTCOME, IS INDEPENDENT FROM ALL OTHER M-1 TRIALS

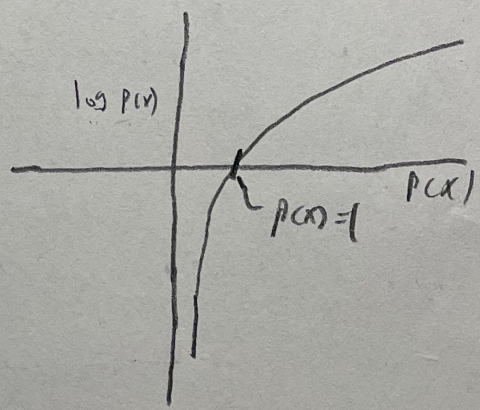
↳ THIS ALLOWS US TO REWRITE THE JOINT DIST. AS:

$$P(y^{(1)}, \ldots, y^{(M)}) = P(y^{(1)}; \theta) \cdot P(y^{(2)}; \theta); \ldots, P(y^{(M)}; \theta)$$

MLE continued...

$$P(y^{(1)}, \dots y^{(m)}) = \prod_{i=1}^{m} P(y^{(i)}; \theta)$$

- Learning problem is of course to find the "best" θ
  - $\hat{\theta}$ is the most likely θ given our set of observations in $D \rightarrow D = \{y^{(1)}, \dots, y^{(m)}\}$
  - $\hat{\theta} = \underset{\theta}{\text{ARGMAX}} \prod_{i=1}^{m} P(y^{(i)}; \theta)$
  - Numerical underflow:



  - $\hat{\theta} = \underset{\theta}{\text{ARGMAX}} \sum_{i=1}^{m} \log P(y^{(i)}; \theta)$
    - Equivalent to the above because of the fact that $P(x)$ and $\log P(x)$ change monotonically!
  - $\hat{\theta} = \underset{\theta}{\text{ARGMIN}} \sum_{i=1}^{m} -\log P(y^{(i)}; \theta)$

Bayes Rule:

↳ Product Rule of Probabilities

$$P(X,Y) = P(X|Y)P(Y)$$
$$= P(Y|X)P(X)$$

↳ Bayes Rule:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Evidence}}$$