Probabilistic Models and Machine Learning

No. 2

Math Background

Hui Jiang

Department of Electrical Engineering and Computer Science

Lassonde School of Engineering

York University, Toronto, Canada

Math Review

- Probability and Statistics
  - Random variables/vectors: discrete vs. continuous
  - Conditional probability & Bayes theorem: independence
  - Probability distribution of random variables:
  - Statistics: mean, variance, moments
  - Joint Probability distribution/marginal distribution
  - Some useful distributions: Multinomial, Gaussian, Uniform, etc.
- Information Theory:
  - entropy, mutual information, information channel, KL divergence
- Decision Trees:
  - CART (Classification and Regression Tree)
- Function Optimization
  - KKT conditions, Gradient descent, Newton's, etc.
- Linear Algebra:
  - Vector, matrix and tensor;
  - Matrix calculus
  - Applications: matrix factorization

Probability Definition

- **Sample Space:** Ω
  - **collection of all possible observed outcomes**
- **An Event *A*:** $A \subseteq \Omega$ **including null event** ϕ
- σ**-field: set of all possible events** $A \in F_{\Omega}$
- **Probability Function (Measurable)** $P : F_{\Omega} \rightarrow [0,1]$
  - **Meet three axioms:**

1. $P(\phi) = 0$ $P(\Omega) = 1$

2. If $A \subseteq B$ **then** $P(A) \leq P(B)$

3. If $A \cap B = \phi$ **then** $P(A \cup B) = P(A) + P(B)$

Some Examples

- Example I: experiment to toss a 6-face dice once:
  - Sample space: {1,2,3,4,5,6}
  - Events: X={even number}, Y={odd number}, Z={larger than 3}.
  - σ-field: set of all possible events
  - Probability Function (Measurable) ➔ relative frequency
- Example II:
  - Sample Space:

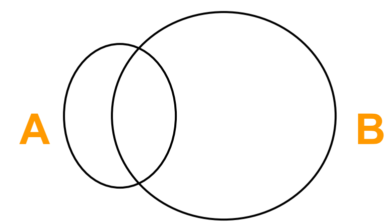
    Ω_c = {x: x is the height of a person on earth}
  - Events:
    - A={x: x>200cm}
    - B={x: 120cm<x<130cm}
  - σ-field: set of all possible events F_Ω
  - Probability Function (Measurable) $P : F_\Omega \to [0,1]$
  - measuring A, B:

$$\Pr(A) = \frac{\#\text{ of persons whose height over 200cm}}{\text{total }\#\text{ of persons in the earth}}$$

Conditional Events

- ***Prior Probability***
  - probability of an event before considering any additional knowledge or observing any other events (or samples): ***P(A)***
- **Joint probability of multiple events: probability of several events occurring concurrently, e.g.,** $P(A \cap B)$
- ***Conditional Probability:*** **probability of one event *(A)* after another event *(B)* has occurred, e.g., *P(A|B).***
  - **updated probability of an event given some knowledge about another event. Definition is:**

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



  - **Prove the *Addition Rule*:**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

  - **From *Multiplication Rule*, show *Chain Rule:***

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2 \mid A_1) \cdots P(A_n \mid \bigcap_{i=1}^{n-1} A_i)$$

Bayes' Theorem

- **Swapping dependency between events**
  - **calculate *P(B|A)* in terms of *P(A|B)* that is available and more relevant in some cases**

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

- **In some cases, not important to compute P(A)**

$$B^* = \arg\max_B P(B \mid A) = \arg\max_B \frac{P(A \mid B)P(B)}{P(A)} = \arg\max_B P(A \mid B)P(B)$$

- **Another Form of Bayes' Theorem**
  - **If a set B partitions A, i.e.**

$$A = \bigcup_{i=1}^{n} B_i \quad B_i \cap B_k = \phi$$

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{P(A)} = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(B_i)}$$

Quick Exercise: Monty Hall Problem

- **Three-Door Probability Puzzle**

Random Variable

- **A random variable (*R.V.*) is a variable which could take various values with different probabilities.**
- **A R.V. is said to be discrete if its set of possible values is a discrete set. The *probability mass function (p.m.f.)* is defined:**

$$f(x) = \Pr(X = x) \quad \text{for } x = x_1, x_2, \cdots \qquad \sum_{x_i} f(x_i) = 1$$

- **A univariate discrete R.V., one *p.m.f.* example:**

x	**1**	**2**	**3**	**4**
f(x)	**0.4**	**0.3**	**0.2**	**0.1**

- **A R.V. is said to be continuous if its set of possible values is an entire interval of numbers. Each continuous R.V. has a distribution function: for a *R.V. X*, its *cumulative distribution function (c.d.f.)* is defined as:**

$$F(t) = \Pr(X \le t) \qquad (-\infty < t < \infty)$$

$$\lim_{t \to -\infty} F(t) = 0 \qquad \lim_{t \to \infty} F(t) = 1$$

- **A *probability density function (p.d.f.)* of a continuous R.V. is a function that for any two number a, b (a<b),**

$$\Pr(a \le X \le b) = \int_a^b f(x)\,\mathrm{d}x \qquad F(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x \qquad \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$$

Random Variable

- **Expectation of random variables and its functions**

$$E(X)=\int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x \qquad \text{or} \qquad \sum_i x_i \cdot p(x_i)$$

$$E(q(X))=\int_{-\infty}^{\infty} q(x)\cdot f(x)dx \quad \text{or} \quad \sum_i q(x_i)\cdot p(x_i)$$

- **Mean and Variance**

$$\mathrm{Mean}(X)=\mathrm{E}(X) \quad \mathrm{Var}(X)=\mathrm{E}([X-\mathrm{E}(X)]^2)$$

- ***r*-th moment *(r=1,2,3,4,…)***

$$E(X^r)=\int_{-\infty}^{\infty} x^r \cdot f(x)\mathrm{d}x \qquad \text{or} \qquad \sum_i x_i^r \cdot p(x_i)$$

- **Random vector is a vector whose elements are all random variables.**

Joint and Marginal Distribution

- Joint Event and Product Space of two (or more) *R.V.'s* $\Omega_c \times \Omega_d$
  - e.g. E=(A,B)=(200cm<height, live in Canada)
- Joint p.m.f of two discrete random variables *X, Y*:

X \ Y	*0*	*1*	*2*
T	*0.03*	*0.24*	*0.17*
F	*0.23*	*0.11*	*0.22*

- Joint p.d.f. (c.d.f.) of two continuous random variables *X, Y*:

$$p(x,y) = \Pr(X \le x, Y \le y)$$

$$\Pr(a \le x \le b, c \le y \le d) = \int_a^b \int_c^d f(x,y)\,dy\,dx$$

- Marginal p.m.f. and p.d.f.:

$$p(x) = \sum_y p(x,y) \qquad f(x) = \int f(x,y)\,dy$$

Conditional Distribution of RVs

- Conditional p.m.f. or p.d.f. for discrete or continuous R.V.'s

$$f(x \mid y) = f(x,y) / f(y)$$

- Conditional Expectation

$$E(q(X) \mid Y = y_0) = \int_{-\infty}^{\infty} q(x) f(x \mid y_0) dx \quad \text{or} \quad \sum_i q(x_i) p(x_i \mid y_0)$$

- Conditional Mean:

$$\mathrm{E}(X \mid Y = y_0) = \int x \cdot f(x \mid y_0) dx$$

- Independence:

$$f(x,y) = f(x) f(y) \qquad f(x \mid y) = f(x)$$

- Covariance between two R.V.'s

$$\mathrm{Cov}(X,Y) = \mathrm{E}([X - \mathrm{E}(X)][Y - \mathrm{E}(Y)])$$

$$= \int_x \int_y (x - E(X))(y - E(Y)) \cdot f(x,y) \, \mathrm{d}x \, \mathrm{d}y$$

- Uncorrelated R.V.'s:

$$\mathrm{Cov(X, Y)} = \mathrm{E}([X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]) = 0$$

Some Useful Distributions (I)

- **Binomial Distribution: *B(R=r; n, p)***
  - **probability of *r* successes in *n* trials with a success rate *p***

$$B(r;n,p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \quad \text{where} \quad 0 \le r \le n$$

  - ***For binomial distribution:***

$$\sum_{r=0}^{n} B(r;n,p) = 1 \qquad \mathrm{E}_B(R) = \sum_{r=0}^{n} rB(r;n,p) = np \quad \mathrm{Var}_B(R) = np(1-p)$$

- **Multinomial Distribution**

$$M(r_1,\ldots,r_m;n,p_1,\ldots,p_m) = \frac{n!}{r_1!\cdots r_m!} \prod_{i=1}^{m} p_i^{r_i} \quad 0 \le r_i \quad \sum_{i=1}^{m} r_i = n$$

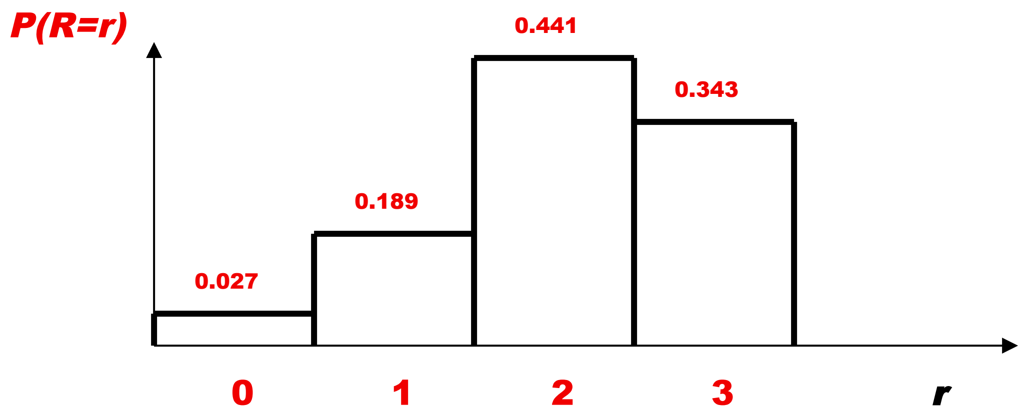
  - **For multinomial distribution**

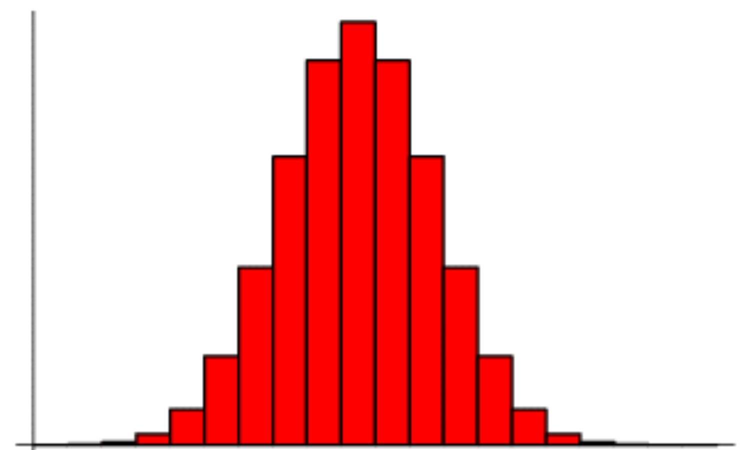
$$E(R_i) = np_i \quad Var(R_i) = np_i(1-p_i) \quad Cov(R_i, R_j) = -np_i p_j$$

Plot of Probability Mass Function

- **Binomial distribution: n=3, p=0.7**

$$B(r;n,p)=\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r} \quad \text{where} \quad 0\le r\le n$$





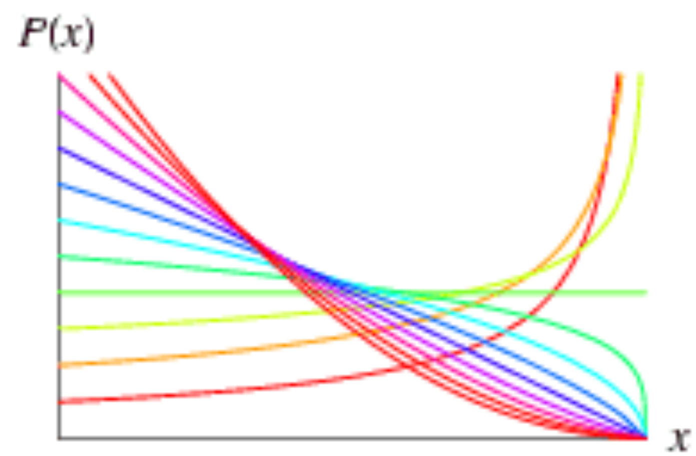
Some Useful Distributions (II)

- **Poisson Distribution with mean (and var) as** $\lambda\ (\lambda \geq 0)$

$$p(x \mid \lambda) = \begin{cases} \dfrac{e^{-\lambda} \cdot \lambda^{x}}{x!} & \text{for } x = 0,1,2,\cdots \\ 0 & \text{otherwise} \end{cases}$$

- **Beta distribution with parameters**

$$p(x \mid \alpha, \beta) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$



  - **For Beta distribution:**

$$\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta} \qquad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^{2}(\alpha+\beta+1)}$$

Some Useful Distributions (III)

- **Dirichlet distribution: a random vector *$(X_1,\dots,X_k)$* has a Dirichlet distribution with parameter vector *$(\alpha_1,\dots,\alpha_k)$ (for all $\alpha_k>0$) if***

$$p(X_1,\cdots,X_k \mid \alpha_1,\cdots,\alpha_k) = \frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} x_1^{\alpha_1-1}\cdots x_k^{\alpha_k-1}$$

$$\text{for all } x_i > 0 \; (i=1,2,\cdots,k) \;\text{ and } \sum_{i=1}^{k} x_i = 1.$$

  – **For Dirichlet distribution:**

$$\text{Denote } \alpha_0 = \sum_{i=1}^{k} \alpha_i$$

$$E(X_i) = \frac{\alpha_i}{\alpha_0} \quad Var(X_i) = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$$

$$Cov(X_i, X_j) = -\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}$$

Some Useful Distributions (IV)

- **Uniform Distribution: *U(X=x; a, b)***

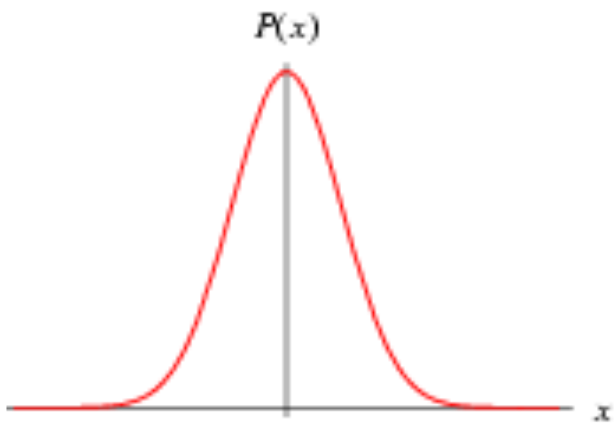
$$U(x;a,b)=\begin{cases}1/(b-a) & a\le x\le b\\ 0 & \text{otherwise}\end{cases} \quad \text{with} \quad a<b$$

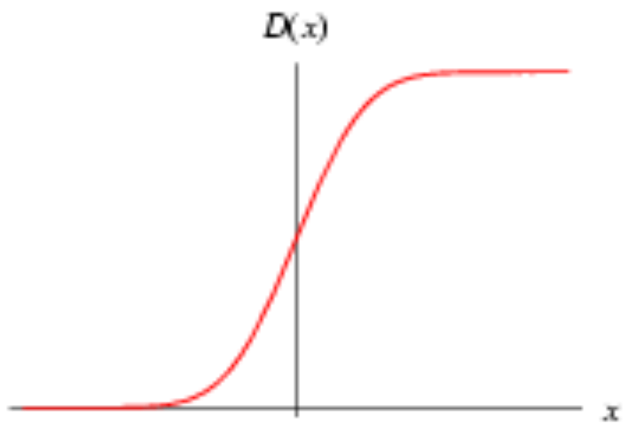
- ***Normal* (or *Gaussian*) Distribution: *Bell Curve***

$$N(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(x-\mu)^2/2\sigma^2} \quad -\infty<x<\infty \quad \sigma>0$$

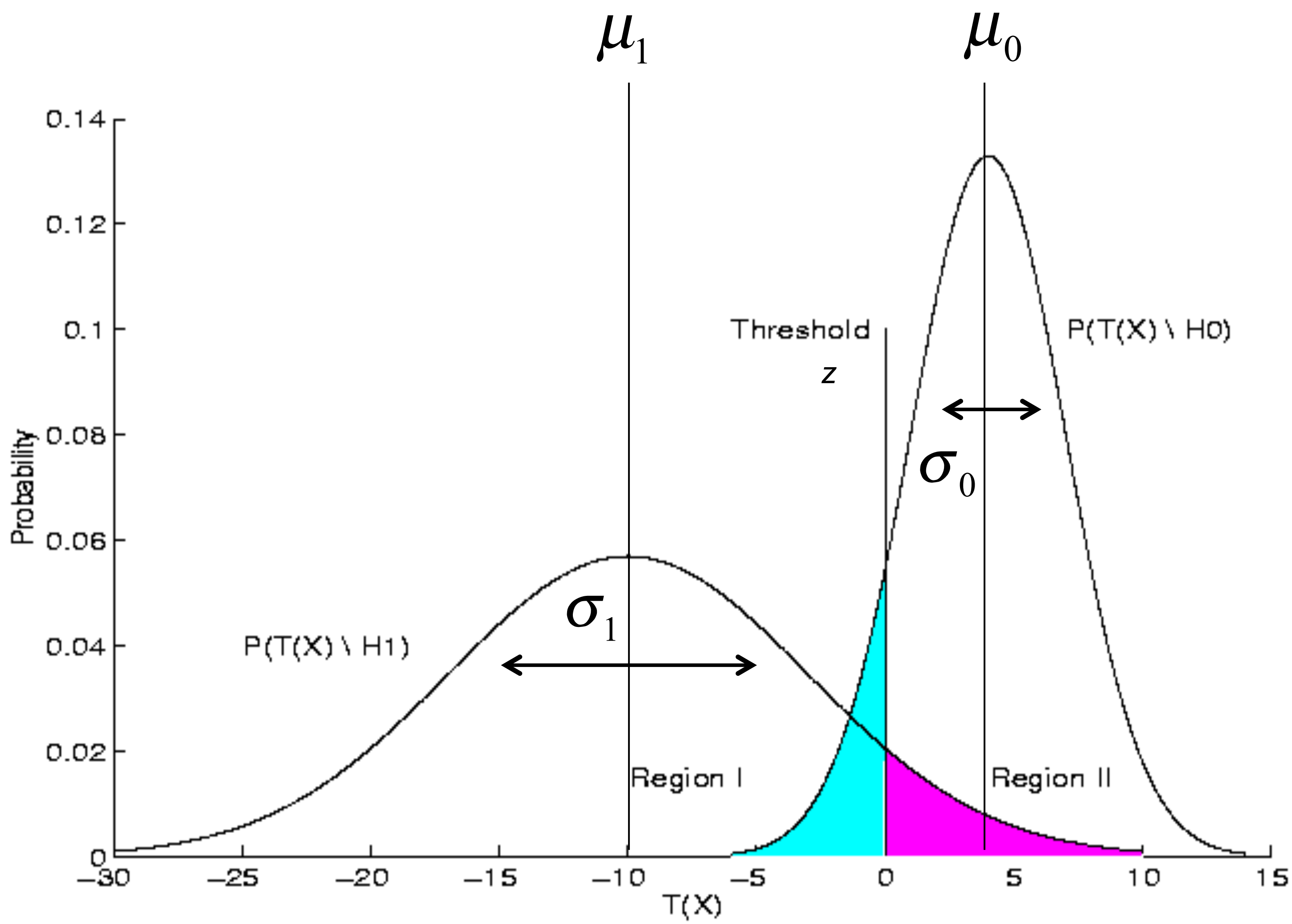
- **Show**

$$\mathrm{E}_U(X)=\frac{a+b}{2} \quad \text{and} \quad \mathrm{E}_N(X)=\mu \qquad \mathrm{VAR}_U(X)=\frac{(b-a)^2}{12} \quad \text{and} \quad \mathrm{VAR}_N(X)=\sigma^2$$





Typical Normal Distributions



Standard deviation (s.d. or spread): $\sigma_1 > \sigma_0$

Some Useful Distributions (V)

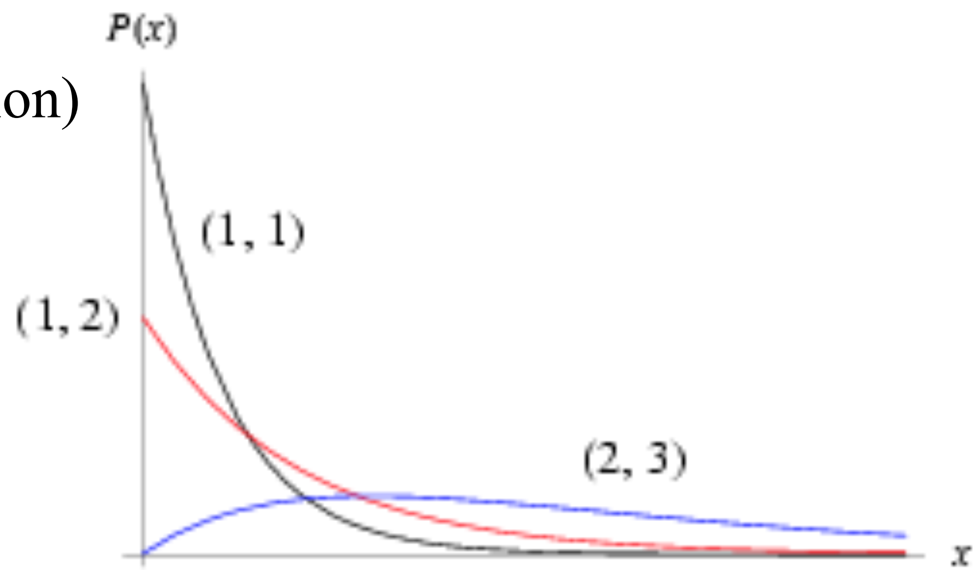
- **Gamma Distribution: a random variable X has a gamma distribution with parameters α and β (α>0, β>0) if**

$$p(x \mid \alpha, \beta) = \begin{cases} \dfrac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \cdot e^{-\beta x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

with

$$\Gamma(\alpha) = \int_0^{\infty} u^{\alpha-1} e^{-u} \mathrm{d}u \quad \text{(gamma function)}$$

$$\mathrm{E}(X) = \frac{\alpha}{\beta} \qquad \mathrm{Var}(X) = \frac{\alpha}{\beta^2}$$



Some Useful Distributions (VI)

- ***2-D Uniform*** **Distribution:**

$$U(x, y; a, b, c, d) = \begin{cases} 1/(b-a)(d-c) & a \le x \le b, c \le y \le d \\ 0 & \text{otherwise} \end{cases} \quad \text{with} \quad a < b, c < d$$

- ***Multivariate Normal*** **Distribution**

$$N(\mathbf{x}; \mu, \mathbf{C}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{C}|}} \mathrm{e}^{-(\mathbf{x}-\mu)^t \mathbf{C}^{-1} (\mathbf{x}-\mu)/2} \quad -\infty < \mathbf{x} < \infty$$

- **Show** $\mathrm{E}_N(\mathbf{X}) = \mu \quad \text{and} \quad \mathrm{VAR}_N(\mathbf{X}) = \mathbf{C}$

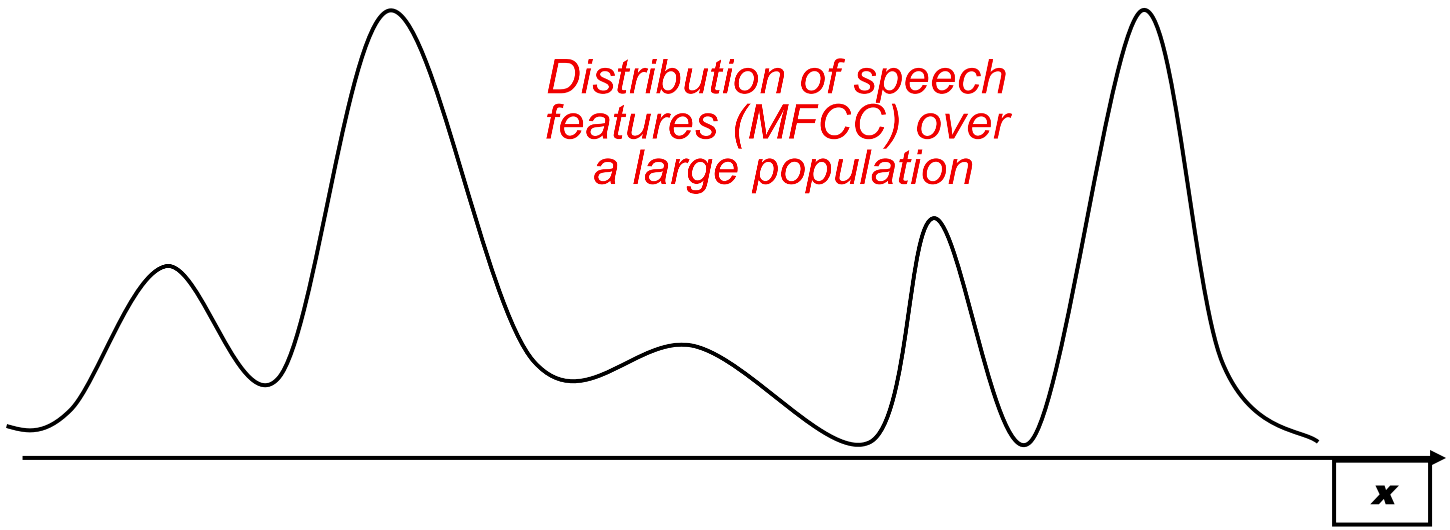
- **Can you write down the 2-D distribution form, compute Cov(*X,Y*), and derive the marginal and conditional densities, *f(y)* and *f(x|y)* ?**

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix} \qquad \mathbf{ì} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} \sigma_x^2 & r\sigma_x\sigma_y \\ r\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

Gaussian Mixture Distribution

- **Gaussian Mixture distribution:**

$$MG(x) = \sum_{m=1}^{M} \omega_m N(x; \mu_m, \sigma_m^2) \quad \text{with} \sum_{m=1}^{M} \omega_m = 1 \quad 0 \le \omega_m \le 1 \quad \sigma_m > 0$$



- **In theory, *MG(x)* matches *any probabilistic density* up to second order statistics (mean and variance)**
- **Approximating multi-modal densities which is more likely to describe real-world data.**

Multinomial Mixture Models

- The idea of mixture applies to other distributions.
- Multinomial Mixture model (MMM):

$$MMM(x) = \sum_{k=1}^{k} \omega_k \cdot M(r_1, \ldots, r_m; n, p_{k1}, \ldots, p_{km}) \quad \text{with } \sum_{k=1}^{K} \omega_k = 1 \quad 0 \le \omega_k \le 1$$

  – Useful for modeling complex discrete data, such as text, biological sequences, etc…

Statistical Distribution

- Non-Parametric Distribution
  - usually described by the data samples themselves
  - Sample distribution & histogram (pmf / bar chart): counting samples in equally-sized bins and plot them
- Parametric Distribution
  - r.v. described by a small number of parameters in pdf/pmf
  - e.g. Gaussian (2), Binomial (2), 2-d uniform (4)
  - many useful and known parametric distributions
  - Probability distribution of independently and identically distributed (i.i.d.) samples from such distributions can be easily derived.
- *Statistic*: Function of random samples
  - sample mean and variance, maximum/minimum, etc.
- *Sufficient Statistics*
  - minimum number of statistics to remember all samples
  - for Gaussian r.v. need count, sample mean and variance
  - for some r.v.'s, no sufficient statistics, need all samples

Function of Random Variables

- **Function of r.v.'s is also a r.v.**
  - **e.g. *X=U+V+W,* if we know *f(u,v,w)* how about *f(x)* ?**
  - **e.g. sum of dots on two dices**
- **Problem easier for known and popular r.v.'s …**

  If U and V are independent Gaussian, so is X=U+V

$$N(.\,|\,\mu_1,\sigma_1^2)+N(.\,|\,\mu_2,\sigma_2^2)=N(.\,|\,\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$$

  - **Sample mean of *n* independent samples of Gaussian r.v.'s is also Gaussian, show that:**

$$\mathrm{E}(\overline{X})=\mu \quad \mathrm{Var}(\overline{X})=\sigma^2/n$$

  - ***How about n samples and n is very large?***

    ➔ ***Law of large numbers* – asymptotic Normal p.d.f. !!**

- **If W and Z are independent uniform, is Y=W+Z uniform ??**

  ➔ Average of two independent samples of uniform r.v.'s form a triangular shape p.d.f.

Transformation of Random Variables

- **Given random vectors** $\vec{X} = (X_1, \cdots X_n)$ and $\vec{Y} = (Y_1, \cdots, Y_n)$
- **We know** $Y_1 = g_1(\vec{X}), \cdots, Y_n = g_n(\vec{X})$
- **Given p.d.f. of** $\vec{X}$, $p_X(\vec{X}) = p_X(X_1, \cdots X_n)$, **how to derive p.d.f. for** $\vec{Y}$ **?**
- **If the transformation is one-to-one mapping, we can derive an inverse transformation as:** $X_1 = h_1(\vec{Y}), \cdots, X_n = h_n(\vec{Y})$
- **We define the Jacobian matrix as:**

$$J(\vec{Y}) = \begin{bmatrix} \frac{\partial h_1}{\partial Y_1} & \cdots & \frac{\partial h_1}{\partial Y_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial h_n}{\partial Y_1} & \cdots & \frac{\partial h_n}{\partial Y_n} \end{bmatrix}$$

- **We have**

$$p_Y(\vec{Y}) = p_X(h_1(\vec{Y}), \cdots h_n(\vec{Y})) \cdot \left| J(\vec{Y}) \right|$$

Probability Theory Recap

- **Probability Theory Tools**
  - **fuzzy description of phenomena**
  - **statistical modeling of data for inference**
- **Statistical Inference Problems**
  - ***Classification*: choose one of the stochastic sources**
  - ***Hypothesis Testing*: comparing two stochastic assumptions and decide on how to accept one of them**
  - ***Estimation*: given random samples from an assumed distribution, find "good" guess for the parameters**
  - ***Prediction*: from past samples, predict next set of samples**
  - ***Regression* (*Modeling*): fit a model to a given set of samples**
- **Parametric vs. Non-parametric Distributions**
  - **Parsimonious or extensive description (model vs. data)**
  - **Sampling, data storage and sufficient statistics**
- **Real-World Data vs. Ideal Distributions**
  - **"there is no perfect goodness-of-fit"**
  - **ideal distributions are used for approximation**
  - **sum of random variables and Law of Large Numbers**

Information Theory & Shannon

- **Claude E. Shannon (1916-2001, from Bell Labs to MIT): Father of Information Theory, Modern Communication Theory …**
- **Information of an event:** $I(A) = \log_2 1/\Pr(A) = -\log_2 \Pr(A)$
- **<u>Entropy</u> (Self-Information) – in b*it,* amount of info in a r.v.**

$$H(X) = -\sum_{x \in \mathrm{X}} p(x)\log_2 p(x) = \mathrm{E}[\log_2 \frac{1}{p(X)}] \quad 0\log_2 0 = 0$$

  – **Entropy represents average amount of information in a r.v., in other words, the average uncertainty related to a r.v.**

- **Contributions of Shannon:**
  – **Study of English – Cryptography Theory, *Twenty Questions* game, Binary Tree and Entropy, etc.**
  – **Concept of Code – Digital Communication, Switching and Digital Computation (optimal Boolean function realization with digital relays and switches)**
  – **Channel Capacity – Source and Channel Encoding, Error-Free Transmission over Noisy Channel, etc.**
  – **C. E. Shannon, "A Mathematical Theory of Communication", Parts 1 & 2, *Bell System Technical Journal*, 1948.**
  – **He should have won a Nobel Prize for his contributions (1948 is also the year of the discovery of transistor at Bell Labs)**

Joint and Conditional Entropy

- **Joint entropy: average uncertainty about two r.v.'s; average amount of information provided by two r.v.'s.**

$$H(X,Y) = \mathrm{E}[\log_2 \frac{1}{p(X,Y)}] = -\sum_{x\in \mathrm{X}} \sum_{y\in \mathrm{Y}} p(x,y) \log_2 p(x,y)$$

- **Conditional entropy: average amount of information (uncertainty) of Y after X is known.**

$$H(Y \mid X) = -\sum_{x\in \mathrm{X}} p(x) H(Y \mid X = x) = \sum_{x\in \mathrm{X}} p(x)[-\sum_{y\in \mathrm{Y}} p(y \mid x) \log_2 p(y \mid x)]$$
$$= -\sum_{x\in \mathrm{X}} \sum_{y\in \mathrm{Y}} p(x,y) \log_2 p(y \mid x)$$

- **Chain Rule for Entropy :**

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1, \ldots, X_{n-1})$$

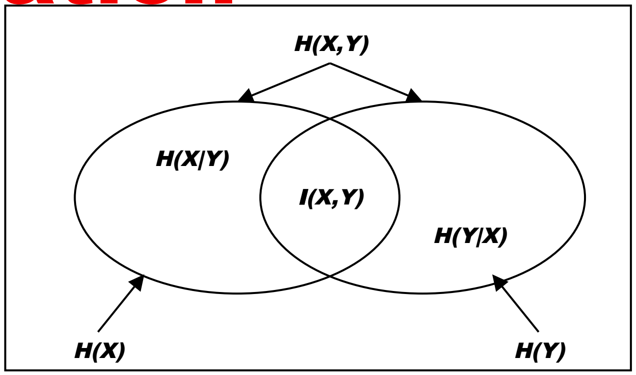
- **Independence:**

$$H(X,Y) = H(X) + H(Y) \quad \text{or} \quad H(Y \mid X) = H(Y)$$

Mutual Information

- **Definition :**

$$I(X,Y) = H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$
$$= H(X) + H(Y) - H(X,Y)$$



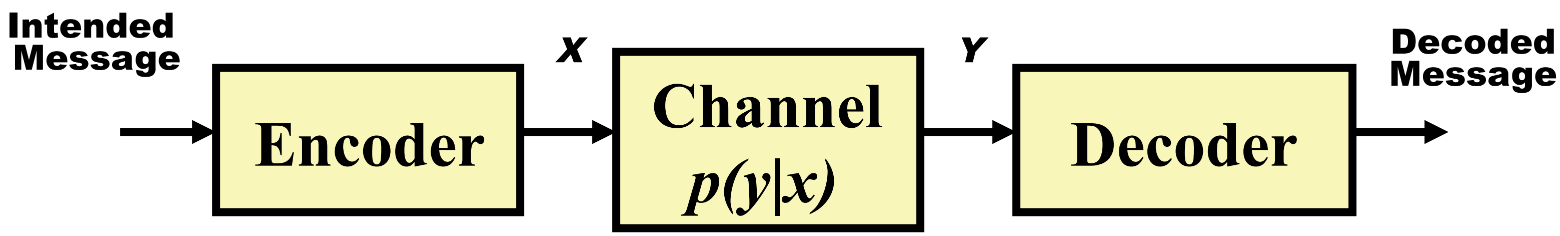
$$I(X,Y) = \sum_{x \in X} p(x)\log_2 \frac{1}{p(x)} + \sum_{y \in Y} p(y)\log_2 \frac{1}{p(y)} - \sum_{x \in X}\sum_{y \in Y} p(x,y)\log_2 \frac{1}{p(x,y)}$$

$$I(X,Y) = \sum_{x \in X}\sum_{y \in Y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \quad \text{or} \quad \int_x\int_y p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}\,\mathrm{d}x\mathrm{d}y$$

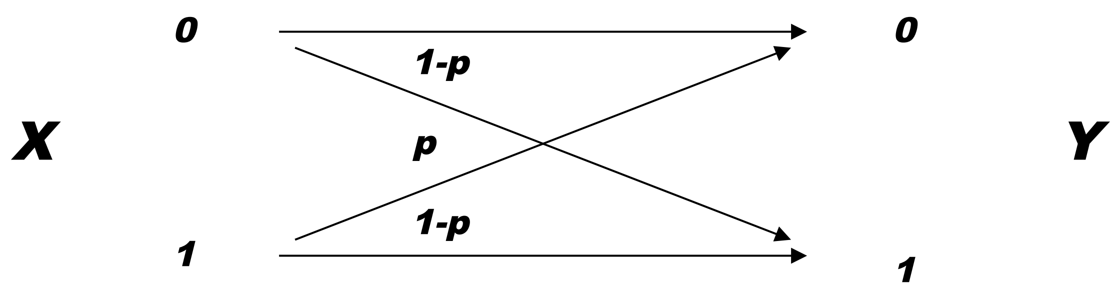
- **Intuitive meaning of mutual information: given two r.v.'s, *X and Y*, mutual information *I(X,Y)* represents average information about *Y* (or *X*) we can get from *X* (or *Y*).**
- **Maximization of *I(X,Y)* is equivalent to establishing a closer relationship between *X* and *Y*, i.e., obtaining a low-noise information channel between *X* and *Y*.**

Shannon Noisy Channel Model

- Shannon's Noisy Channel Model



- A Binary Symmetric Noisy Channel (Modem Application)



$$I(X,Y)$$

- Channel Capacity

$$C = \max_{p(X)} I(X,Y) = \max_{p(X)} [H(Y) - H(Y \mid X)]$$

$$C = 1 - H(p) \leq 1$$

- p(X) & p(Y|X) can be given by design or by nature.

Mutual Information: Example (I)

- In Shannon's noisy channel model: assume X={0,1} Y={0,1}

  X is equiprobable Pr(X=0)=Pr(X=1)=0.5 ➔ H(X) = 1 bit

  joint distribution p(X,Y)=p(X) p(Y|X)

  – Case I : *p=0.0 (noiseless)*

p(X,Y)	*0*	*1*
0	*0.5*	*0.0*
1	*0.0*	*0.5*

$$I(X,Y)=\sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$$

$$=0.5\cdot\log_2\frac{0.5}{0.5\cdot 0.5}+0.0+0.5\cdot\log_2\frac{0.5}{0.5\cdot 0.5}+0.0=1.0$$

  – *Case II: p=0.1 (weak noise)*

p(X,Y)	*0*	*1*
0	*0.45*	*0.05*
1	*0.05*	*0.45*

$$I(X,Y)=\sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$$

$$=2\cdot 0.45\cdot\log_2\frac{0.45}{0.5\cdot 0.5}+2\cdot 0.05\cdot\log_2\frac{0.05}{0.5\cdot 0.5}=0.533$$

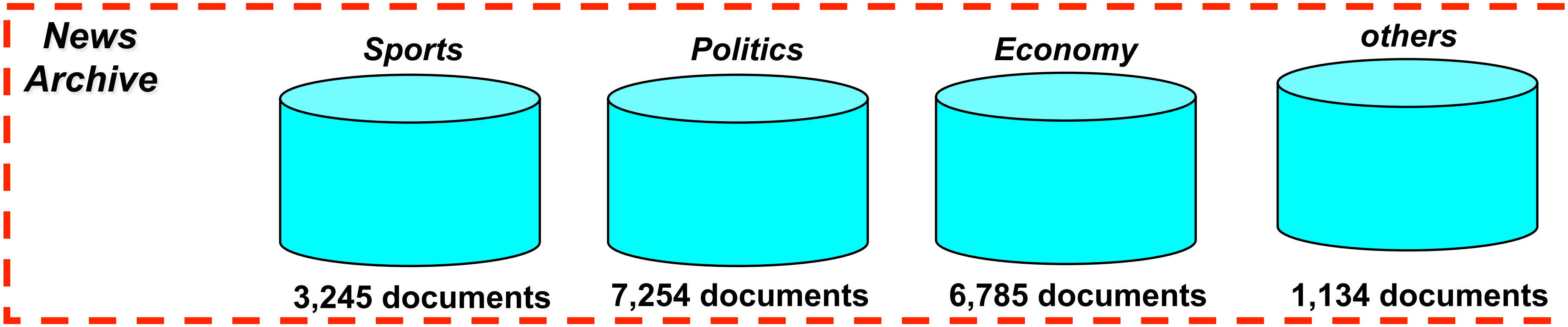
  – *Case III: p=0.4 (strong noise)*

p(X,Y)	*0*	*1*
0	*0.3*	*0.2*
1	*0.2*	*0.3*

$$I(X,Y)=\sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$$

$$=2\cdot 0.3\cdot\log_2\frac{0.3}{0.5\cdot 0.5}+2\cdot 0.2\cdot\log_2\frac{0.2}{0.5\cdot 0.5}=0.03$$

Mutual Information Example(II): Identifying keywords in Text Categorization



- All documents contain 10,345 distinct words in total (vocabulary)
- How to identify which words are more informative with respect to any one topic? (keywords of a topic)
- Use Mutual information as a criterion to calculate correlation of each word with any one topic.
- Example: word "*score*" vs. topic "*sports*"
  - Define two binary random variables:

    X: document topic is "sports" or not. *{0,1}*

    Y: document contains "score" or not. *{0,1}*
  - *I(X,Y)* ➔ relationship between word "*score*" vs. topic "*sports*"

Identifying keywords in Text Categorization

- **Count documents in archive to calculate *p(X,Y)***

$$p(X=1,Y=1)=\frac{\#\text{ of docs with topic "sports" and contains "score"}}{\text{total }\#\text{ of docs in the archive}}$$

$$p(X=1,Y=0)=\frac{\#\text{ of docs with topic "sports" and don't contains "score"}}{\text{total }\#\text{ of docs in the archive}}$$

Y→"score"

	p(X,Y)	*0*	*1*	
X	*0*	*0.802*	*0.022*	0.824
	1	*0.106*	*0.070*	0.176
		0.908	0.092	

$$I(X,Y)=\sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$$
$$=0.126$$

- **How about word "what" – topic "sports"**

Y→"what"

	p(X,Y)	*0*	*1*	
X	*0*	*0.709*	*0.115*	0.824
	1	*0.153*	*0.023*	0.176
		0.862	0.138	

$$I(X,Y)=\sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$$
$$=0.000070$$

- **"score" is a keyword for the topic "sports"; "what" is not;**

Identifying keywords in Text Categorization

- For topic T_i, choose its keywords (most relevant)
  - For each word W_j in vocabulary, calculate $I(W_j,T_i)$;
  - Sort all words based on $I(W_j,T_i)$;
  - Keywords w.r.t. topic T_i : top N words in the sorted list.

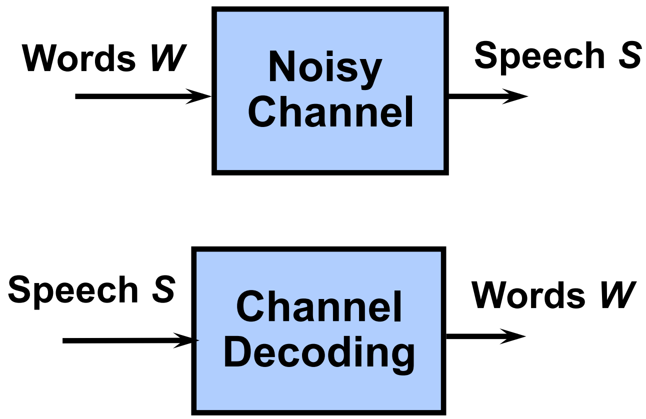
- Keywords for the whole text categorization task:
  - For each word W_j in vocabulary, calculate

$$I(W_j) = \frac{1}{|T|}\sum_{i=1}^{|T|} I(W_j, T_i) \;\text{ or }\; I'(W_j) = \max_{\mathrm{i}} I(W_j, T_i)$$

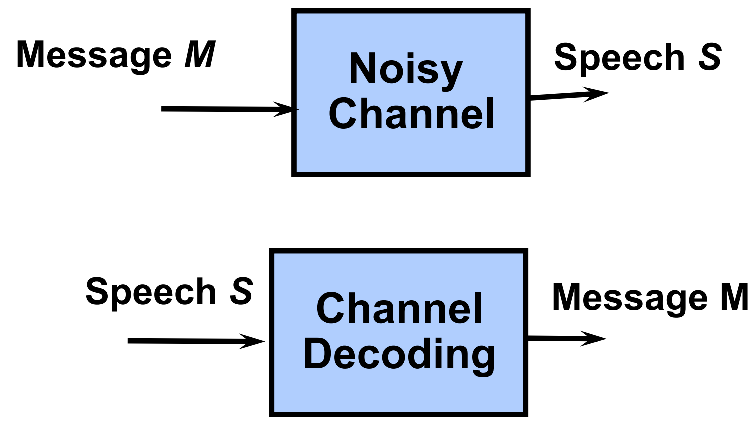
  - Sort all words based on $I(W_j)$ or $I'(W_j)$.
  - Top M words in the sorted list.

Channel Modeling and Decoding

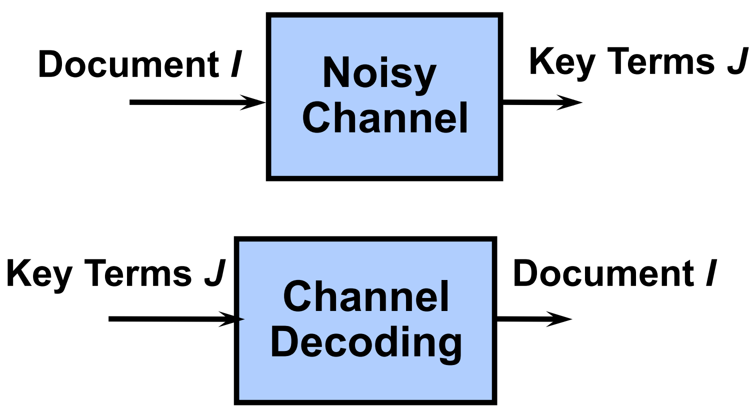
Speech Recognition



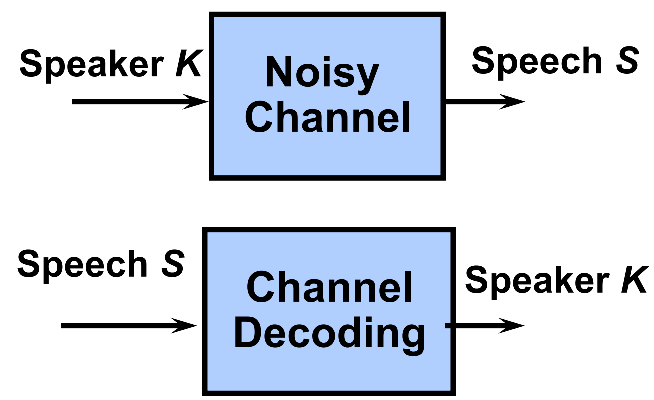
Speech Understanding



Information Retrieval



Speaker Identification



Bayes Theorem Applications

- **Bayes Theorem for Channel Decoding**

$$I^* = \arg\max_I P(I \mid \hat{O}) = \arg\max_I \frac{P(\hat{O} \mid I)P(I)}{P(\hat{O})} = \arg\max_I P(\hat{O} \mid I)P(I)$$

Application	Input	Output	p(I)	p(O|I)
Speech Recognition	Word Sequence	Speech Features	Language Model (LM)	Acoustic Model
Character Recognition	Actual Letters	Letter images	Letter LM	OCR Error Model
Machine Translation	Source Sentence	Target Sentence	Source LM	Translation (Alignment) Model
Text Understanding	Semantic Concept	Word Sequence	Concept LM	Semantic Model
Part-of-Speech Tagging	POS Tag Sequence	Word Sequence	POS Tag LM	Tagging Model

Kullback-Leibler (KL) Divergence

- **Distance measure between two p.m.f.'s (relative entropy)**

$$D(p \| q) = \mathrm{E}_p[\log_2 \frac{p(x)}{q(x)}] = \sum_{x \in \mathrm{X}} p(x) \log_2 \frac{p(x)}{q(x)}$$

  - ***D(p||q)>=0* and *D(p||q)=0* if only if *q=p***

- ***KL Divergence* is a measure of the average distance between two probability distributions.**

$$D(p(x,y) \| q(x,y)) = D(p(x) \| q(x)) + D(p(y \mid x) \| q(y \mid x))$$

- **Mutual information is a measure of independence**

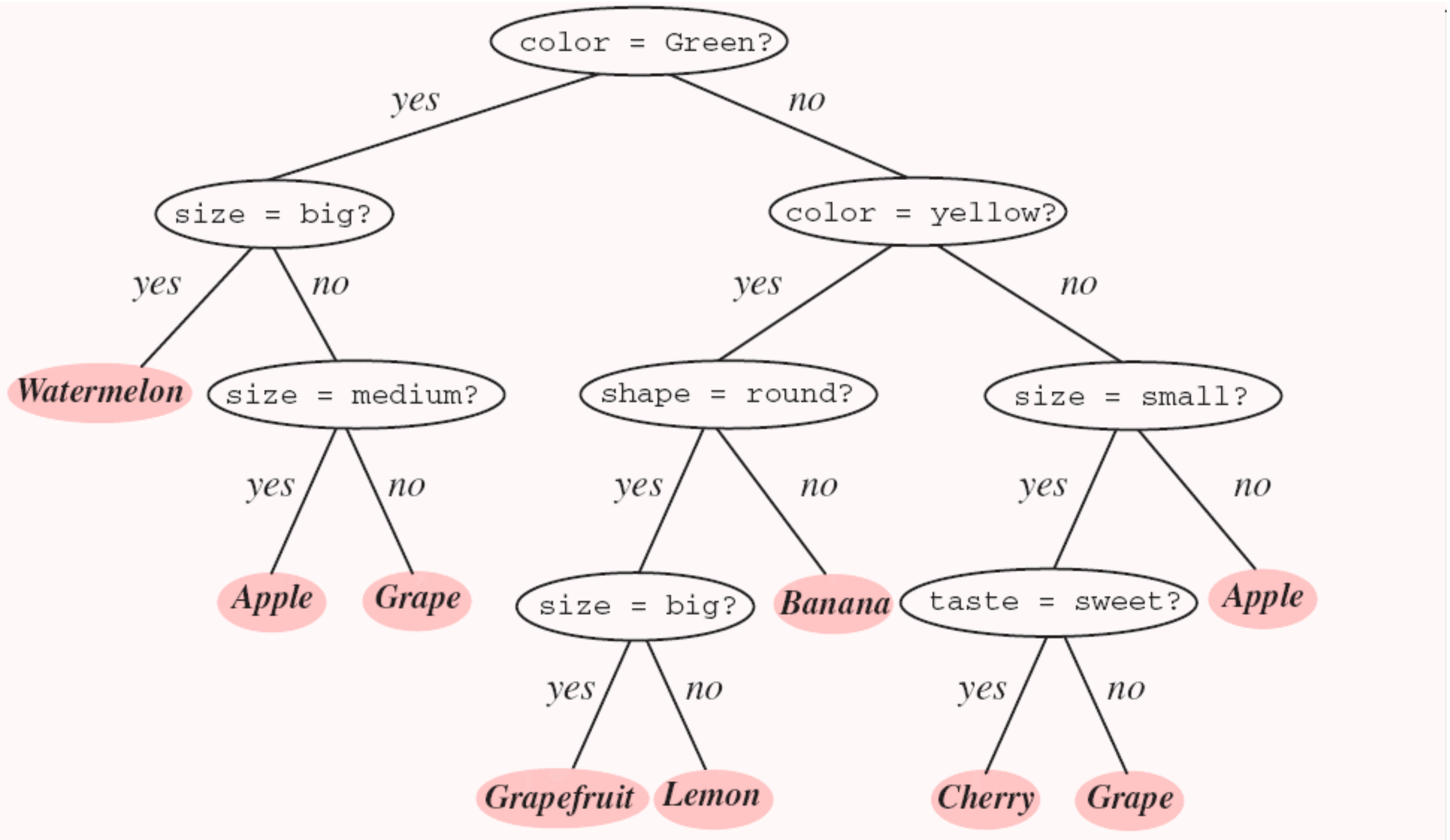
$$I(X,Y) = \sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} = D(p(x,y) \| p(x)p(y))$$

- **Conditional Relative Entropy**

$$D(p(y \mid x) \| q(y \mid x)) = \sum_{x \in \mathrm{X}} p(x) \sum_{y \in \mathrm{Y}} p(y \mid x) \log_2 \frac{p(y \mid x)}{q(y \mid x)}$$

Classification: Decision Trees

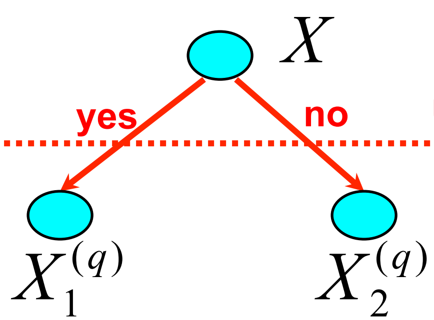
Decision Tree classification: interpretability
Example: fruits classification based on features



Classification and Regression Tree (CART)

- Binary tree for classification: each node is attached a YES/NO question; Traverse the tree based on the answers to questions; each leaf node represents a class.
- CART: how to automatically grow such a classification tree on a data-driven basis.
  - Prepare a finite set of all possible questions.
  - For each node, choose the best question to split the node. "best" is in sense of maximum entropy reduction between "before splitting" and "after splitting".
    - Entropy→ uncertainty or chaos in data;

      Small entropy → more homogeneous the data is; less impure



Use question q

$$H(X)$$

$$\frac{|X_1^{(q)}|}{|X|}H(X_1^{(q)}) + \frac{|X_2^{(q)}|}{|X|}H(X_2^{(q)})$$

Choose a q from the question set to maximize the difference

The CART algorithm

1) Question set: create a set of all possible YES/NO questions.
2) Initialization: initialize a tree with only one node which consists of all available training samples.
3) Splitting nodes: for each node in the tree, find the best splitting question which gives the greatest entropy reduction.
4) Go to step 3) to recursively split all its children nodes unless it meets certain stop criterion, e.g., entropy reduction is below a pre-set threshold OR data in the node is already too little.

CART method is widely used in machine learning and data mining:

1. Handle categorical data in data mining;
2. Acoustic modeling (allophone modeling) in speech recognition;
3. Letter-to-sound conversion;
4. Automatic rule generation
5. etc.

Optimization of objective function (I)

- **Optimization:**
  - **Set up an objective function *Q()* ;**
  - **Maximize or minimize the objective function with respect to the variable(s) in question.**
- **Maximization (minimization) of a function:**
  - **Differential calculus;**
    - **Unconstrained maximization/minimization**

$$Q = f(x) \Rightarrow \frac{\mathrm{d}\, f(x)}{\mathrm{d}x} = 0 \Rightarrow x = ?$$

$$Q = f(x_1, x_2, \cdots, x_N) \Rightarrow \frac{\partial\, f(x_1, x_2, \cdots, x_N)}{\partial\, x_i} = 0 \Rightarrow ??$$

  - **Lagrange Optimization:**
    - **Constrained maximization/minimization**

$$Q = f(x_1, x_2, \cdots, x_N) \quad \text{with constraint} \quad \mathrm{g}(x_1, x_2, \cdots, x_N) = 0$$

$$Q' = f(x_1, x_2, \cdots, x_N) + \lambda \cdot g(x_1, x_2, \cdots, x_N)$$

$$\frac{\partial\, Q'}{\partial\, x_1} = 0, \frac{\partial\, Q'}{\partial\, x_2} = 0, \cdots, \frac{\partial\, Q'}{\partial\, x_N} = 0, \frac{\partial\, Q'}{\partial\, \lambda} = 0$$

Karush–Kuhn–Tucker (KKT) conditions

- **A general optimization problem:**

$$\begin{aligned} &\min_{\mathrm{x}} \quad f(\mathrm{x}) \\ &\text{subject to} \\ &\mathrm{g}_i(\mathrm{x}) \le 0 \qquad (i = 1, \cdots, m) \\ &h_j(\mathrm{x}) = 0 \qquad (j = 1, \cdots n) \end{aligned}$$

- **Introduce KKT multipliers:**
  - **For each inequality constraint:** $\mu_i \quad (i = 1, \cdots, m)$
  - **For each equality constraint:** $\lambda_i \quad (i = 1, \cdots, m)$

Karush–Kuhn–Tucker (KKT) conditions

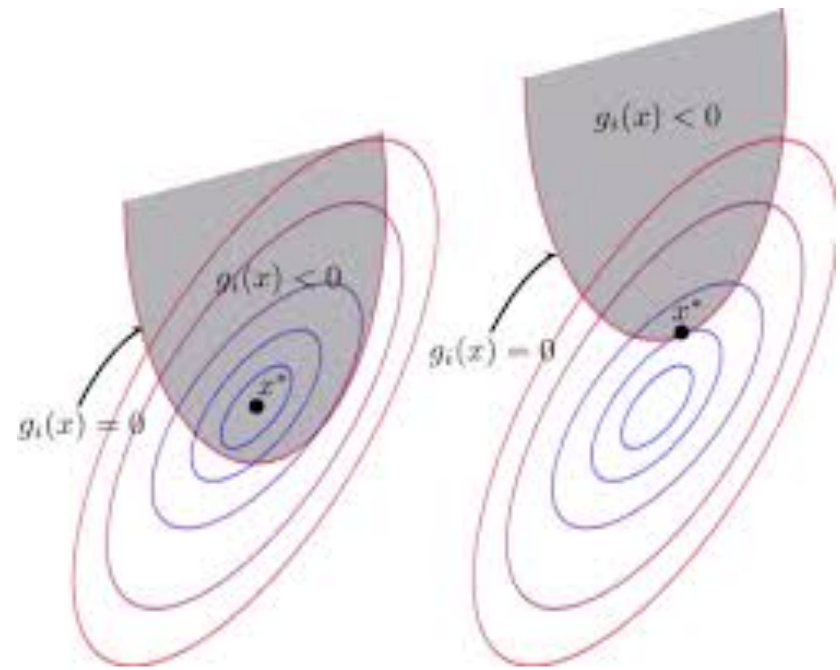
- **Necessary condition: If x* is local optimum of the primary problem, x* satisfies:**

$$\nabla f(\mathrm{x}^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(\mathrm{x}^*) + \sum_{j=1}^{l} \lambda_i \nabla h_j(\mathrm{x}^*) = 0$$

$$\mu_i \geq 0 \quad (i = 1, \cdots, m)$$

$$\mu_i g_i(\mathrm{x}^*) = 0 \quad (i = 1, \cdots, m)$$

- **Physical meaning of KKT multipliers:**



Karush–Kuhn–Tucker (KKT) conditions

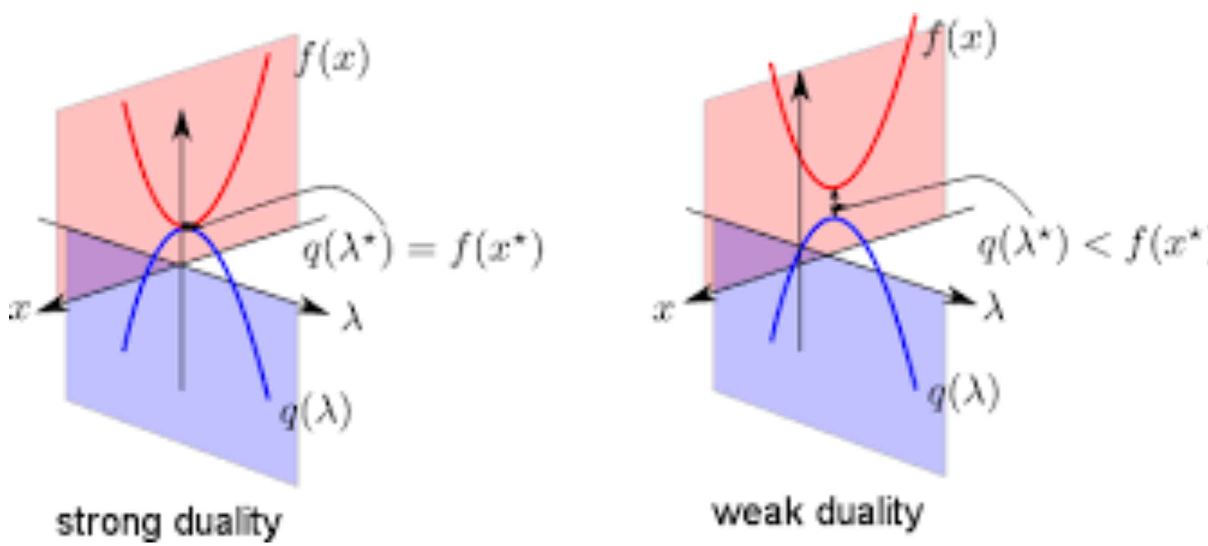
- **Prime problem**

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad i = 1, \ldots, m \end{array}$$

- **Dual problem:**

$$\begin{array}{ll} \underset{u}{\text{maximize}} & \inf_x \left(f(x) + \sum_{j=1}^{m} u_j g_j(x) \right) \\ \text{subject to} & u_i \geq 0, \quad i = 1, \ldots, m \end{array}$$

- **Strong duality vs. Weak duality**



Numerical Optimization (I): 1st order

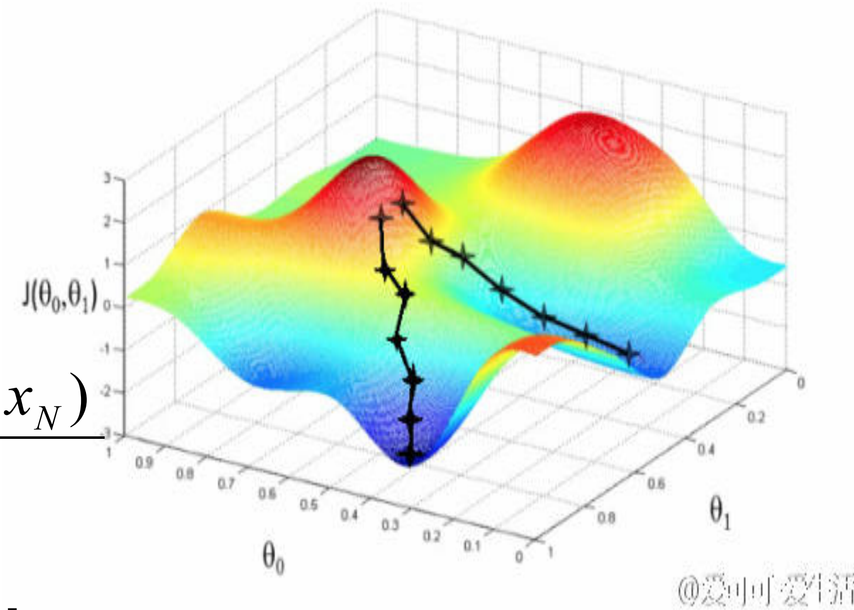
- **Gradient descent (ascent) method:**

$$Q = f(x_1, x_2, \cdots, x_N)$$

For any x_i, start from any initial value $x_i^{(0)}$

$$x_i^{(n+1)} = x_i^{(n)} \pm \varepsilon \cdot \nabla_{x_i} f(x_1, x_2, \cdots, x_N)\,|_{x_i = x_i^{(n)}}$$

$$\text{where } \nabla_{x_i} f(x_1, x_2, \cdots, x_N) = \frac{\partial f(x_1, x_2, \cdots, x_N)}{\partial x_i}$$



  - **Step size is hard to determine**
  - **Slow convergence**
- **Conjugate gradient descent (ascent)**
- **Stochastic gradient descent (SGD)**

Stochastic Gradient Descent (SGD)

- **The cost function in machine learning normally looks like:**

$$R_N(\theta) = \frac{1}{N}\sum_{n=1}^{N} Q(x_n, y_n, \theta)$$

- **Regular Gradient Decent (GD):**

$$\begin{aligned}\hat{\theta}_{t+1} &= \hat{\theta}_t - \lambda_t \cdot \nabla_\theta R_N(\hat{\theta}_t) \\ &= \hat{\theta}_t - \lambda_t \cdot \frac{1}{N}\sum_{n=1}^{N} \frac{\partial Q(x_n, \hat{\theta}_t)}{\partial \theta}\end{aligned}$$

- **Stochastic Gradient Descent (SGD):**

$$\bar{\theta}_{t+1} = \bar{\theta}_t - \lambda_t \cdot \frac{\partial Q(x_n, \bar{\theta}_t)}{\partial \theta}.$$

- **Mini-batach SGD**

- **SGD is extremely effective in optimizing a complex objective function but the reason remains unknown in theory.**

Numerical Optimization(II): 2nd order

- **Newton's method:**

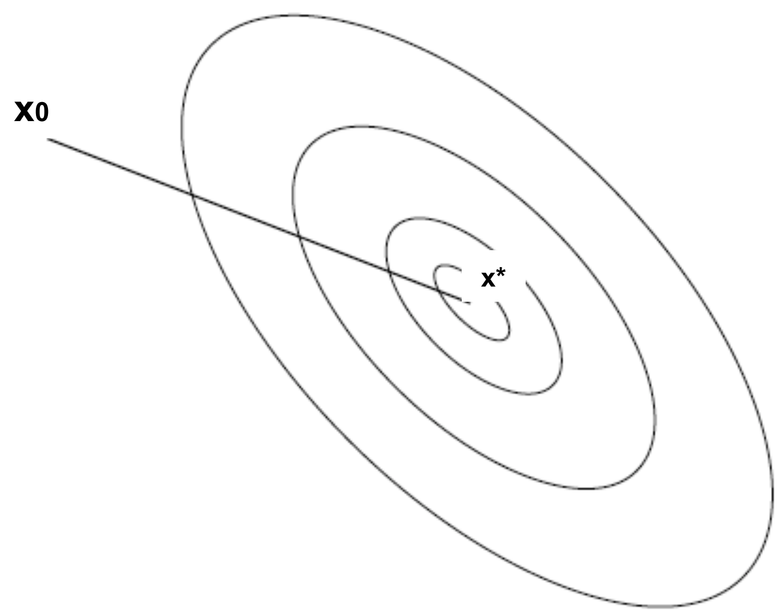
$$Q = f(\mathrm{x})$$

Given any initial value x_0

$$f(\mathrm{x}) \approx f(\mathrm{x}_0) + \nabla f(\mathrm{x}_0)(\mathrm{x} - \mathrm{x}_0)^t + \frac{1}{2}(\mathrm{x} - \mathrm{x}_0)^t H(\mathrm{x} - \mathrm{x}_0)$$

$$H = \left.\begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_N} \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_N} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_N} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_N} & \cdots & \frac{\partial^2 f(x)}{\partial x_N^2} \end{bmatrix}\right|_{\mathrm{x}=\mathrm{x}_0}$$

$$\mathrm{x}^* = \mathrm{x}_0 - H^{-1} \cdot \nabla f(\mathrm{x}_0)$$



  - **Hessian matrix is too big; hard to estimate**
- **Quasi-Newton's method: no need to compute Hessian matrix; quick update to approximate it.**
  - **Quickprop; R-Prop; BFGS; L-BFGS**

More Optimization Methods

- Convex optimization algorithms:
  - Linear Programming
  - Quadratic programming (nonlinear optimization)
  - Semi-definite Programming
- EM (Expectation-Maximization) algorithm
- Dual Coordinate Descent/Ascent
- Growth-Transformation method

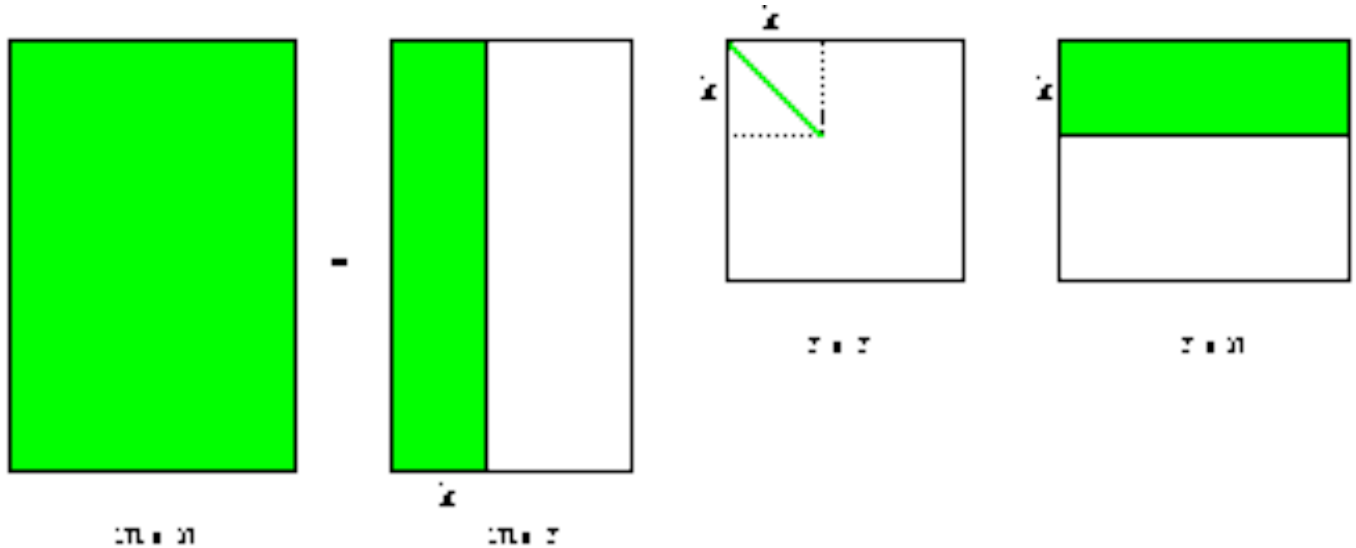
Vector, Matrix and Tensor

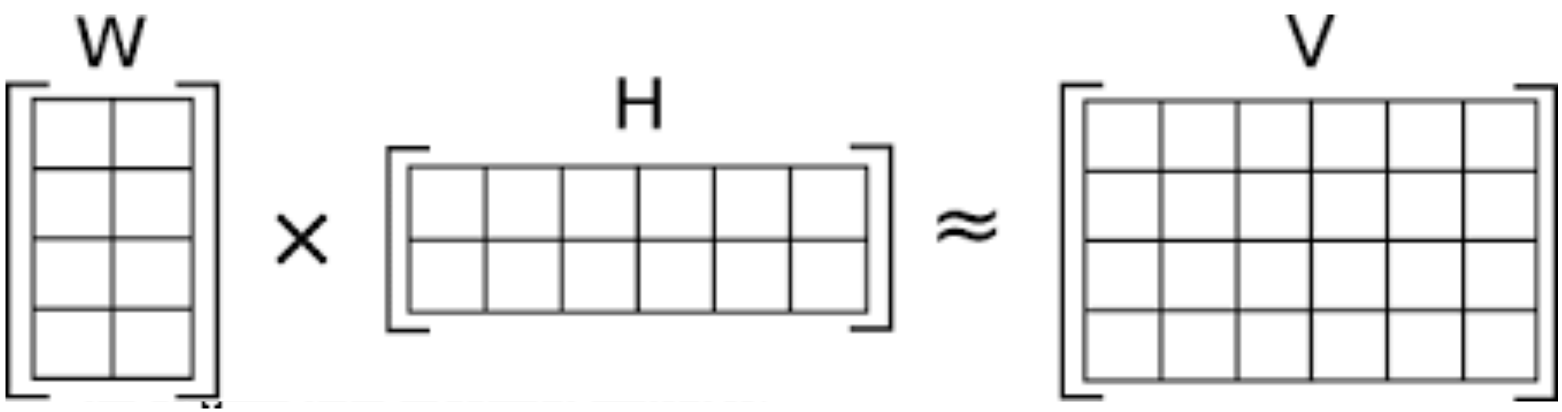
- Linear Algebra:
  - Vector, matrix, Tensor
  - Determinant and matrix inversion
  - Eigen-value and eigen-vector
  - Matrix Factorization
  - Derivatives of Matrices
  - etc.

- A good on-line matrix reference manual
  http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/
  http://www.psi.toronto.edu/matrix/matrix.html

Matrix Factorization

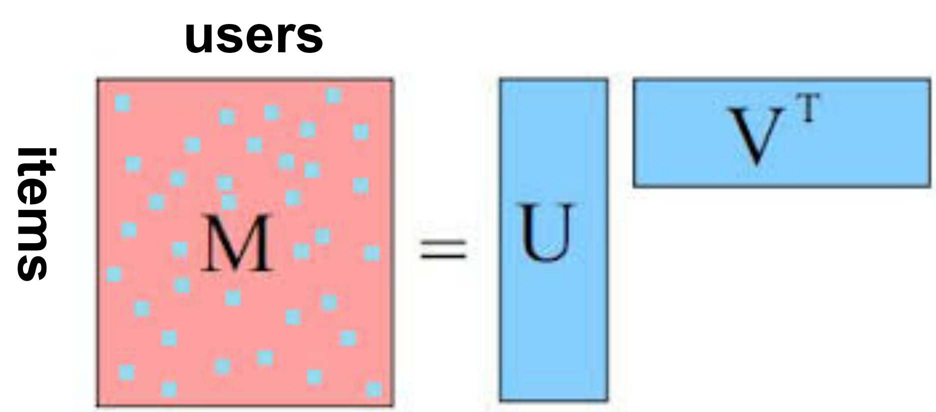
- Singular-Value Decomposition (SVD)
- Non-negative Matrix Factorization (NMF)

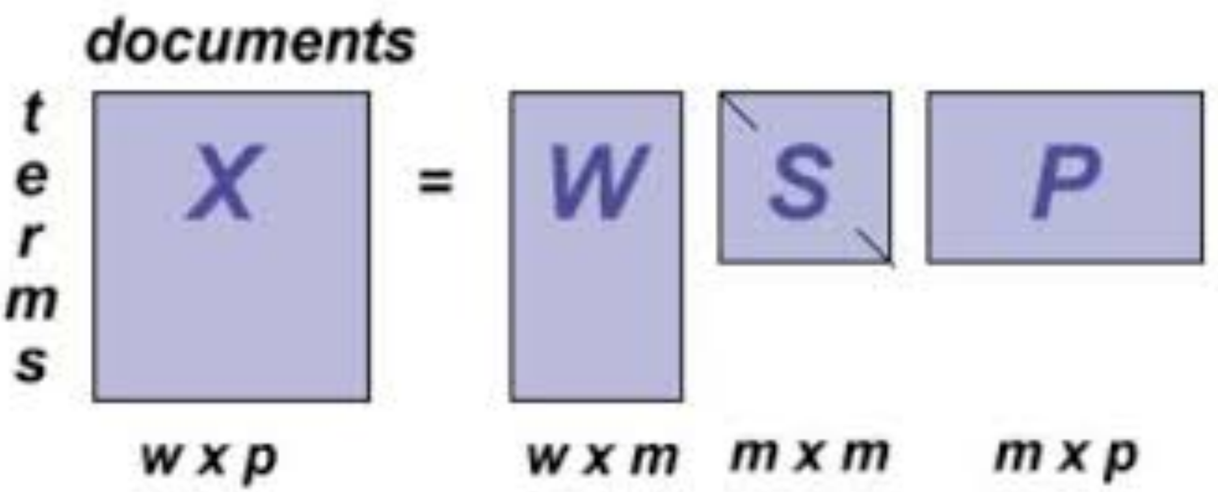




Matrix Factorization

- Popular recommending algorithm: collaborative filtering
- Popular NLP algorithm: Latent Semantic Analysis (LSA)





Matrix Calculus

- **Derivation w.r.t. a matrix or a vector**
- **Exercise: try to prove**

$\mathbf{y}$	$\dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}$
$\mathbf{x}^T\mathbf{x}$	$2\mathbf{x}$
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$\mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x}$