In The Name of Allah

Digital Media Laboratory
Sharif University of Technology

Statistical Pattern Recognition

A Brief Mathematical Review

Hamid R. Rabiee
Jafar Muhammadi, Ali Jalali, Alireza Ghasemi

Spring 2012
http://ce.sharif.edu/courses/90-91/2/ce725-1/

Agenda

- Probability theory
- Distribution Measures
- Gaussian distribution
- Central Limit Theorem
- Information Measure
- Distances
- Linear Algebra

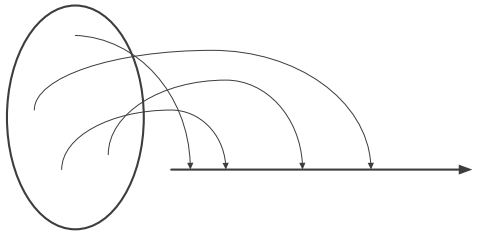
Probability Space

- **A triple of (Ω, F, P)**
  - **Ω: represents a nonempty set, whose elements are sometimes known as *outcomes* or *states of nature (Sample Space)***
  - **F: represents a set, whose elements are called *events*. The events are subsets of Ω. F should be a "Borel Field".**
    - If a field has the property that, if the sets A_1, A_2,...,A_n,... belong to it, then so do the sets $A_1+A_1+...+A_n+...$ and A1.A1.....An...., then the field is called a Borel field.
  - **P: represents the probability measure.**
- **Fact: $P(\Omega) = 1$**

Random Variables

- **Random variable is a *"function" ("mapping")* from a set of possible outcomes of the experiment to an interval of real (complex) numbers.**
- **In other words:**

$$\begin{cases} F \subseteq \Omega \\ I \subseteq R \end{cases} : \begin{cases} X : F \to I \\ X(\beta) = r \end{cases}$$



- **Examples:**
  - **Mapping faces of a dice to the first six natural numbers.**
  - **Mapping height of a man to the real interval (0,3] (meter or something else).**
  - **Mapping success in an exam to the discrete interval [0,20] by quantum 0.1**

Random Variables (Cont'd)

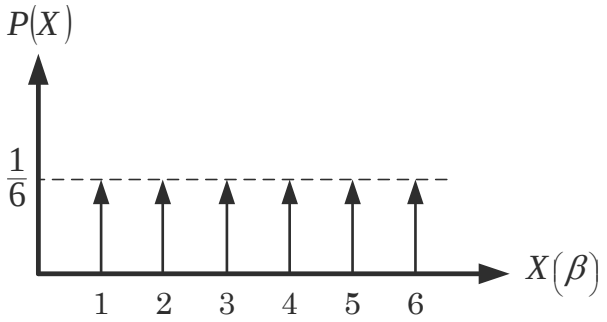
- **Random Variables**
  - **Discrete**
    - Dice, Coin, Grade of a course, etc.
  - **Continuous**
    - Temperature, Humidity, Length, etc.

- **Random Variables**
  - **Real**
  - **Complex**

Density/Distribution Functions

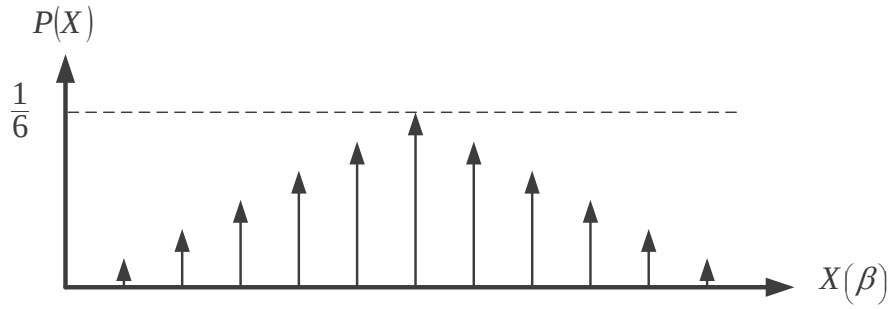
- **Probability Mass Function (PMF)**
  - **Discrete random variables**
  - **Summation of impulses**
  - **The magnitude of each impulse represents the probability of occurrence of the outcome**
  - **PMF values are probabilities.**

- **Example I:**
  - **Rolling a fair dice**



Density/Distribution Functions (Cont'd)

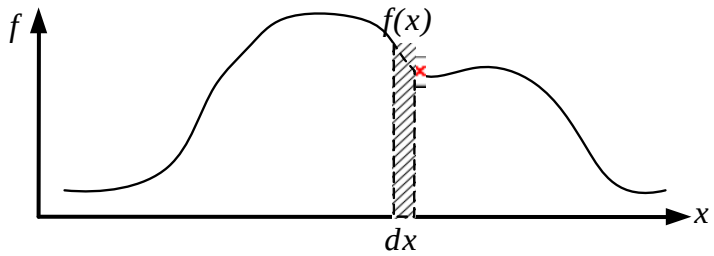
- **Example II:**
  - **Summation of two fair dices**



- **Note : Summation of all probabilities should be equal to ONE. (Why?)**

Density/Distribution Functions (Cont'd)

- **Probability Density Function (PDF)**
  - **Continuous random variables**
  - **The probability of occurrence of** $x_0 \in \left(x - \frac{dx}{2}, x + \frac{dx}{2}\right)$ **will be** $f(x).dx$

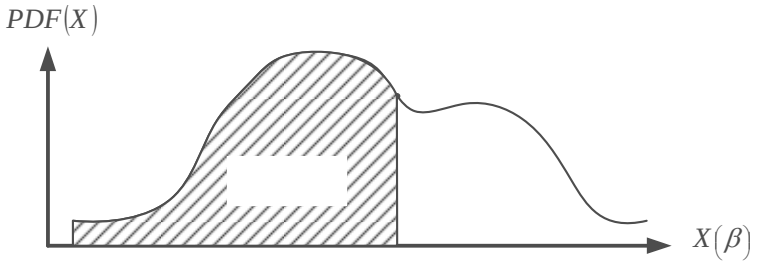


Density/Distribution Functions (Cont'd)

- **Cumulative Distribution Function (CDF)**
  - **Both Continuous and Discrete**
  - **Could be defined as the integration of PDF**

$$CDF(x) = F_X(x) = P(X \leq x)$$

$$F_X(x) = \int_{-\infty}^{x} f_X(x).dx$$

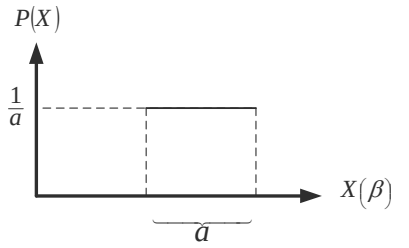


  - **Some CDF properties**
    - Non-decreasing
    - Right Continuous
    - F(-infinity) = 0
    - F(infinity) = 1

Famous Density Functions

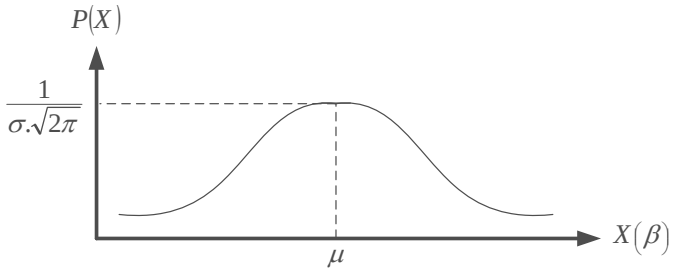
- **Some famous masses and densities**
  - **Uniform Density**

$$f(x) = \frac{1}{a}.\left(U(end) - U(begin)\right)$$



  - **Gaussian (Normal) Density**

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



  - **Exponential Density**

$$f(x) = \lambda e^{-\lambda x} U(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Joint/Conditional Distributions

✧ Joint Probability Functions

$$F_{X,Y}(x,y) = P(X \le x \ \ and \ \ Y \le y)$$
$$= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x,y)\,dydx$$

✧ Example I

- ✧ **In a rolling fair dice experiment represent the outcome as a 3-bit digital number "xyz".**

$$f_{X,Y}(x,y) = \begin{cases} 1/6 & x=0; y=0 \\ 1/3 & x=0; y=1 \\ 1/3 & x=1; y=0 \\ 1/6 & x=1; y=1 \\ 0 & OW. \end{cases}$$

$$\begin{array}{c} xyz \\ 1 \rightarrow 001 \\ 2 \rightarrow 010 \\ 3 \rightarrow 011 \\ 4 \rightarrow 100 \\ 5 \rightarrow 101 \\ 6 \rightarrow 110 \end{array}$$

Joint/Conditional Distributions (Cont'd)

- **Example II**
  - **Two normal random variables**

$$f_{X,Y}(x,y) = \frac{1}{2\pi.\sigma_x.\sigma_y.\sqrt{1-r^2}} e^{-\left(\frac{1}{2(1-r^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x{}^2}+\frac{(y-\mu_y)^2}{\sigma_y{}^2}-\frac{2r(x-\mu_x)(y-\mu_y)}{\sigma_x.\sigma_y}\right)\right)}$$

  - **What is "r" ?**

- **Independent Events (Strong Axiom)**

$$f_{X,Y}(x,y) = f_X(x).f_Y(y)$$

Joint/Conditional Distributions (Cont'd)

- **Obtaining one variable density functions**

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$$

- Distribution **functions can be obtained just from the density functions. (How?)**

Joint/Conditional Distributions (Cont'd)

✧ Conditional Density Function

- ✧ **Probability of occurrence of an event if another event is observed (we know what "Y" is).**

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

- ✧ **Bayes' Rule**

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x).f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|x).f_X(x)dx}$$

Joint/Conditional Distributions (Cont'd)

- **Example I**
  - **Rolling a fair dice**
    - X : the outcome is an even number
    - Y : the outcome is a prime number

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{1/6}{1/2} = \frac{1}{3}$$

- **Example II**
  - **Joint normal (Gaussian) random variables**

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi}.\sigma_x.\sqrt{1-r^2}} e^{-\left(\frac{1}{2(1-r^2)}\left(\frac{x-\mu_x}{\sigma_x} - r\times\frac{y-\mu_y}{\sigma_y}\right)^2\right)}$$

Joint/Conditional Distributions (Cont'd)

✧ Conditional Distribution Function

$$F_{X|Y}(x|y) = P(X \le x \ \ while \ \ Y = y)$$

$$= \int_{-\infty}^{x} f_{X|Y}(x|y)dx$$

$$= \frac{\int_{-\infty}^{x} f_{X,Y}(t,y)dt}{\int_{-\infty}^{\infty} f_{X,Y}(t,y)dt}$$

✧ **Note that "y" is a constant during the integration.**

Joint/Conditional Distributions (Cont'd)

✧ Independent Random Variables

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$
$$= \frac{f_X(x).f_Y(y)}{f_Y(y)}$$
$$= f_X(x)$$

Distribution Measures

- **Most basic type of descriptor for spatial distributions, include:**
  - **Mean Value**
  - **Variance & Standard Deviation**
  - **Covariance**
  - **Correlation Coefficient**
  - **Moments**

Expected Value

- Expected value (population mean value)

$$\mathrm{E}[g(X)] = \sum xf(x) = \int_{-\infty}^{\infty} xf(x)dx$$

- Properties of Expected Value
  - **The expected value of a constant is the constant itself. E[b]= b**
  - **If a and b are constants, then E[aX+ b]= a E[X]+ b**
  - **If X and Y are independent RVs, then E[XY]= E[X]* E[Y]**
  - **If X is RV with PDF f(X), g(X) is any function of X, then,**
    - if X is discrete

$$\mathrm{E}[g(X)] = \sum g(X)f(x)$$

    - if X is continuous

$$E[g(X)] = \int_{-\infty}^{\infty} g(X)f(x)dx$$

Variance & Standard Deviation

- **Let X be a RV with E(X)=u, the distribution, or spread, of the X values around the expected value can be measured by the variance (δ_x is the standard deviation of X).**

$$\begin{aligned}\text{var}(X) &= \delta_x^2 = E(X-\mu)^2 \\ &= \sum_x (X-\mu)^2 f(x) \\ &= E(X^2) - \mu^2 = E(X^2) - [E(X)]^2\end{aligned}$$

- **The Variance Properties**
  - **The variance of a constant is zero.**
  - **If a and b are constants, then** $\text{var}(aX+b) = a^2\,\text{var}(X)$
  - **If X and Y are independent RVs, then** $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$
    $\text{var}(X-Y) = \text{var}(X) + \text{var}(Y)$

Covariance

- **Covariance of two RVs, X and Y: Measurement of the nature of the association between the two.**

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = E(XY) - \mu_x\mu_y$$

- **Properties of Covariance:**
  - **If X, Y are two independent RVs, then** $Cov = E(XY) - \mu_x\mu_y = E(x)E(y) - \mu_x\mu_y = 0$
  - **If a, b, c and d are constants, then** $Cov(a+bX, c+dY) = bd\ \mathrm{cov}(X,Y)$
  - **If X is a RV, then** $Cov = E(X^2) - \mu_x{}^2 = Var(X)$
- **Covariance Value**
  - **Cov(X,Y) is positively big = Positively strong relationship between the two.**
  - **Cov(X,Y) is negatively big = Negatively strong relationship between the two.**
  - **Cov(X,Y)=0 = No relationship between the two.**

Variance of Correlated Variables

- $Var(X+Y) = Var(X)+Var(Y)+2Cov(X,Y)$
- $Var(X-Y) = Var(X)+Var(Y)-2Cov(X,Y)$
- $Var(X+Y+Z) = Var(X)+Var(Y)+Var(Z)+2Cov(X,Y)+2Cov(X,Z) + 2Cov(Z,Y)$

Covariance Matrices

✧ **If X is a n-Dim RV, then the covariance defined as:**

$$\Sigma = E[(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)^t]$$

whose ijth element σ_{ij} is the covariance of x_i and x_j:

$$Cov(x_i, x_j) = \sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)], \quad i, j = 1, \ldots, d.$$

then

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Covariance Matrices (cont'd)

✧ Properties of Covariance Matrix:

- ✧ **If the variables are statistically independent, the covariances are zero, and the covariance matrix is diagonal.**

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix} \Rightarrow \quad \Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_d^2 \end{bmatrix} \qquad \Rightarrow \left(\frac{x \ -\mu}{\sigma} \right)^2 = (\mathbf{x}-\boldsymbol{\mu})^t \, \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})$$

- ✧ **noting that the determinant of Σ is just the product of the variances, then we can write**

$$p(x) = \frac{1}{(2\pi)^{d/2} \left|\Sigma\right|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \, \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- ✧ **This is the general form of a multivariate normal density function, where the covariance matrix is no longer required to be diagonal.**

Correlation Coefficient

- **Correlation: Knowing about a random variable "X", how much information will we gain about the other random variable "Y" ?**
- **The population correlation coefficient is defined as**

$$\rho = \frac{\mathrm{cov}(X, Y)}{\delta_x \delta_y}$$

- **The Correlation Coefficient is a measure of linear association between two variables and lies between -1 and +1**
  - **-1 indicating perfect negative association**
  - **+1 indicating perfect positive association**

Moments

- **Moments**
  - **n^{th} order moment of a RV X is the expected value of X^n:**

$$M_n = E\left(X^n\right)$$

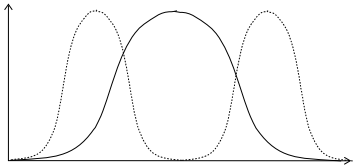
  - **Normalized form (Central Moment)**

$$M_n = E\left(\left(X - \mu_X\right)^n\right)$$

- **Mean is first moment**
- **Variance is second moment added by the square of the mean.**

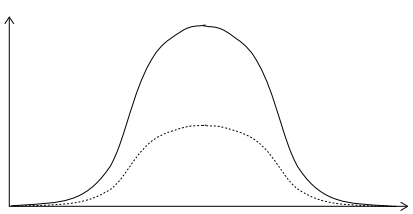
Moments (cont'd)

- **Third Moment**
  - **Measure of asymmetry**
  - **Often referred to as skewness** $s = \int_{-\infty}^{\infty} (x-\mu)^3 f(x)dx$

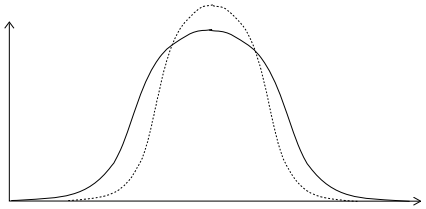


If symmetric s= 0

- **Fourth Moment**
  - **Measure of flatness**
  - **Often referred to as Kurtosis** $k = \int_{-\infty}^{\infty} (x-\mu)^4 f(x)dx$



small kurtosis



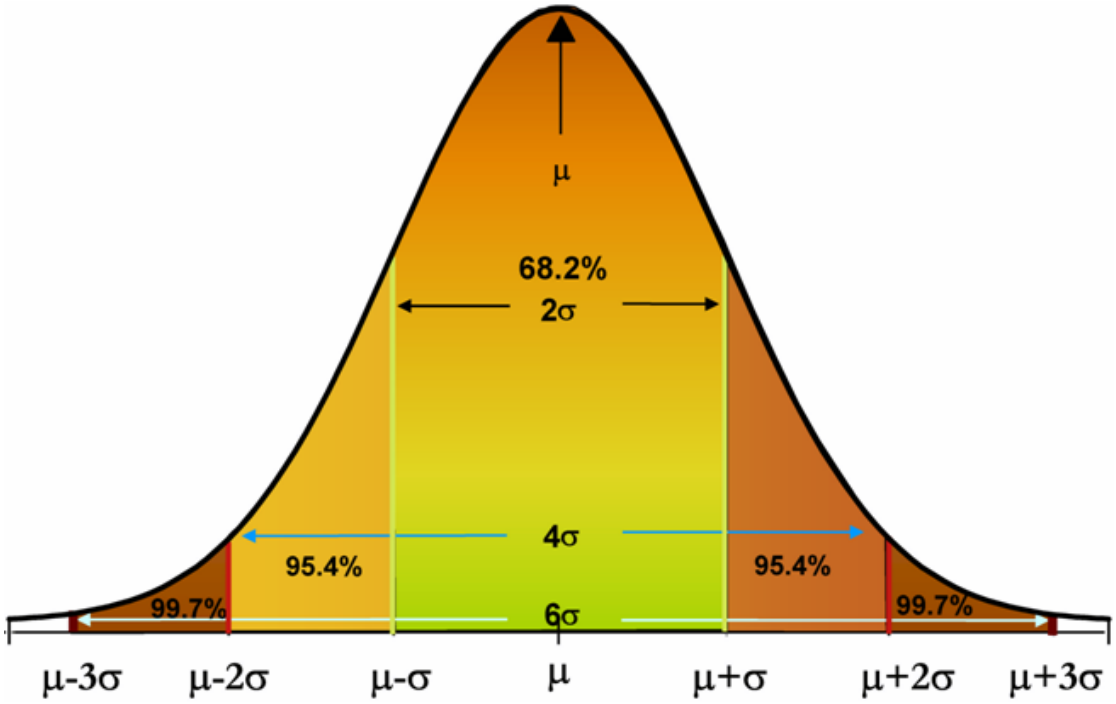
large kurtosis

Sample Measures

- **Sample: a random selection of items from a lot or population in order to evaluate the characteristics of the lot or population**
  - **Sample Mean:** $\bar{x} = \sum_{i=1}^{n} x_i$
  - **Sample Variance:** $S_x^2 = \sum_{i=1}^{n} \frac{(X - \overline{X})^2}{n-1}$
  - **Sample Covariance:** $Cov(X,Y) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$
  - **Sample Correlation:** $Corr = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})/(n-1)}{S_x S_y}$
  - **3th center moment** $\sum_{i=1}^{n} \frac{(X - \overline{X})^3}{n-1}$
  - **4th center moment** $\sum_{i=1}^{n} \frac{(X - \overline{X})^4}{n-1}$

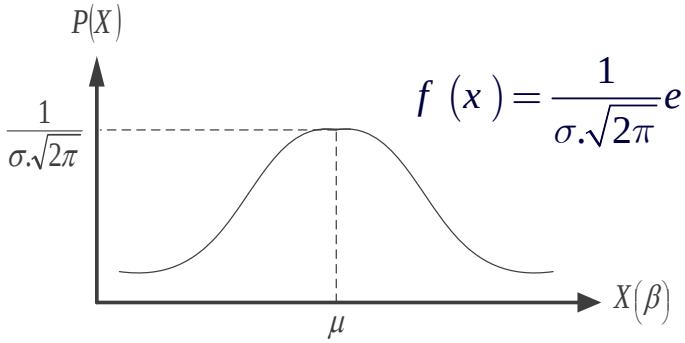
Gaussian distribution

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

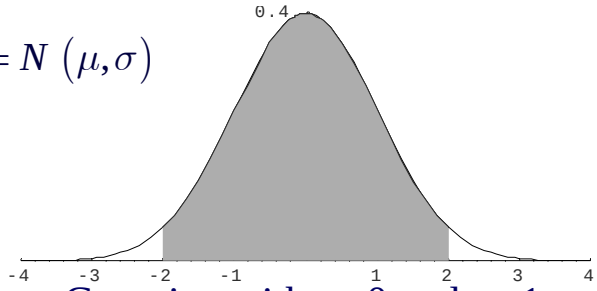


More on Gaussian Distribution

- **The Gaussian Distribution Function**
  - **Sometimes called "Normal" or "bell shaped"**
  - **Perhaps the most used distribution in all of science**
  - **Is fully defined by 2 parameters**
  - **95% of area is within 2σ**
  - **Normal distributions range from minus infinity to plus infinity**



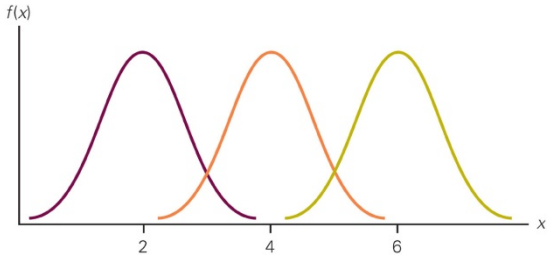
$$f(x) = \frac{1}{\sigma.\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2.\sigma^2}} = N(\mu, \sigma)$$



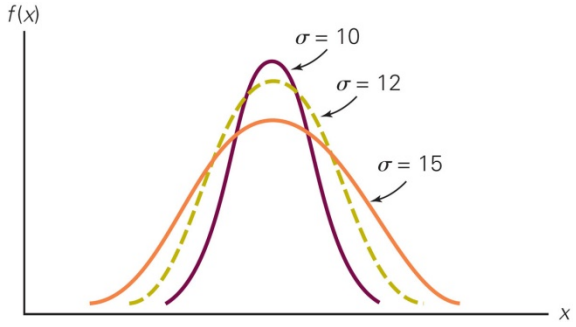
Gaussian with μ=0 and σ=1

More on Gaussian Distribution (Cont'd)

- **Normal Distributions with the Same Variance but Different Means**



- **Normal Distributions with the Same Means but Different Variances**



More on Gaussian Distribution

- **Standard Normal Distribution**
  - **A normal distribution whose mean is zero and standard deviation is one is called the standard normal distribution.**

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

  - **any normal distribution can be converted to a standard normal distribution with simple algebra. This makes calculations much easier.**

$$Z = \frac{X - \mu}{\sigma}$$

Gaussian Function Properties

- Gaussian has relatively simple analytical properties
- It is closed under linear transformation
- The Fourier transform of a Gaussian is Gaussian
- The product of two Gaussians is also Gaussian
- All marginal and conditional densities of a Gaussian are Gaussian
- Diagonalization of covariance matrix
  - rotated variables are independent
- Gaussian distribution is infinitely divisible

Why Gaussian Distribution?

- **Central Limit Theorem**
  - **Will be discussed later**
- **Binomial distribution**
  - **The last row of Pascal's triangle (the binomial distribution) approaches a sampled Gaussian function as the number of rows increases.**
- **Some distribution cab be estimated by Normal distribution for sufficiently large parameter values**
  - **Binomial distribution**
  - **Poisson distribution**

Covariance Matrix Properties

$$\Sigma = E[(X - E[X])(X - E[X])^T] \quad \mu = E[X]$$

$$\Sigma = E(XX^T) - \mu\mu^T$$

$$\text{var}(AX + a) = A\,\text{var}(X)A^T$$

$$\text{cov}(X, Y) = \text{cov}(Y, X)^T$$

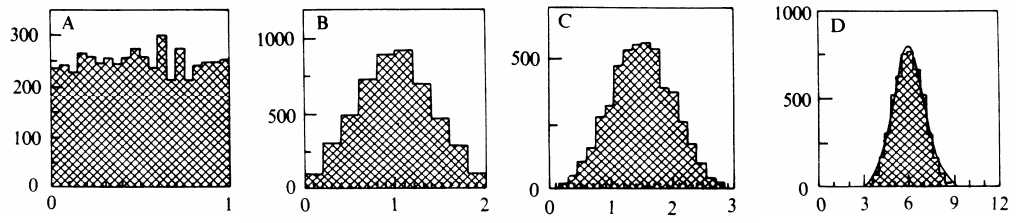
$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + \text{cov}(X, Y) + \text{cov}(Y, X)$$

$$\text{cov}(AX, BY) = A\,\text{cov}(X, Y)B^T$$

$$\text{cov}(X_1 + X_2, Y) = \text{cov}(X_1, Y) + \text{cov}(X_2, Y)$$

Central Limit Theorem

- Why is The Gaussian PDF is so applicable? ➔Central Limit Theorem
- Illustrating CLT
  - a) 5000 Random Numbers
  - b) 5000 Pairs (r1+r2) of Random Numbers
  - c) 5000 Triplets (r1+r2+r3) of Random Numbers
  - d) 5000 12-plets (r1+r2+…+r12) of Random Numbers



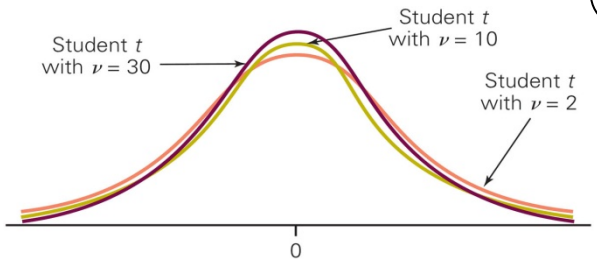
Central Limit Theorem

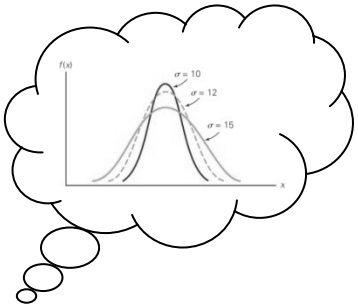
- **Central Limit Theorem says that**
  - **we can use the standard normal approximation of the sampling distribution regardless of our data. We don't need to know the underlying probability distribution of the data to make use of sample statistics.**
  - **This only holds in samples which are large enough to use the CLT's "large sample properties." So, how large is large enough?**
    - Some books will tell you 30 is large enough.
  - **Note: The CLT does not say: "in large samples the data is distributed normally."**

Two Normal-like distributions

✧ T-Student Distribution

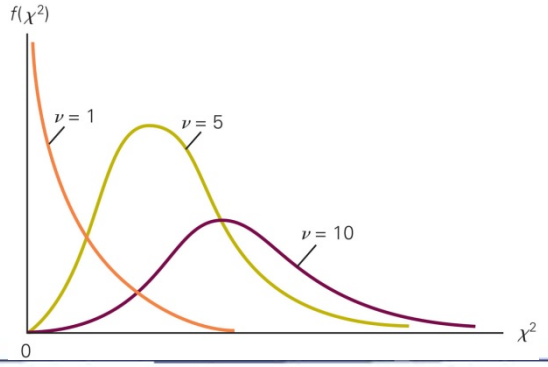
$$f(t) = \frac{\Gamma\left[(v+1)/2\right]}{\sqrt{v\pi}\,\Gamma(v/2)}\left[1+\frac{t^2}{v}\right]^{-(v+1)/2}$$





✧ Chi-Squared Distribution

$$f(\chi^2) = \frac{1}{\Gamma(v/2)}\frac{1}{2^{v/2}}(\chi^2)^{(v/2)-1}e^{-\chi^2/2}$$



Information Measure Criteria

- **Information gain**
  - Let p_i be the probability that a sample in D belongs to class C_i (estimated by $|C_i|/|D|$)
  - Expected information (entropy) needed to classify a sample in D:

$$\mathrm{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

  - Information needed to classify a sample in D, after using A to split D into v partitions :

$$\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \mathrm{Info}(D_j)$$

  - Information gained by attribute A

$$\mathrm{Gain}(A) = \mathrm{Info}(D) - \mathrm{Info}_A(D)$$

Information Measure Criteria

- **Gain Ratio**
  - Information gain measure is biased towards attributes with a large number of values
  - Gain ratio overcomes to this problem (normalization to information gain)

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

  - GainRatio(A) = Gain(A)/SplitInfo(A)

Information Measure Criteria

- **Gini index**
  - If a data set D contains examples from n classes, gini index, gini(D) is defined as

$$\text{gini}(D) = 1 - \sum_{j=1}^{n} p_j^2$$

  - If a dataset D is split on A into two subsets D1 and D2, the gini index gini(D) is defined as

$$\text{gini}_A(D) = \frac{|D_1|}{|D|}\text{gini}(D_1) + \frac{|D_2|}{|D|}\text{gini}(D_2)$$

  - Reduction in Impurity:

$$\Delta\text{gini}(A) = \text{gini}(D) - \text{gini}_A(D)$$

  - The attribute provides the largest reduction in impurity is the best

Information Measure Criteria

- **Comparison**
  - **Information gain:**
    - biased towards multivalued attributes
  - **Gain ratio:**
    - tends to prefer unbalanced splits in which one partition is much smaller than the others
  - **Gini index:**
    - biased to multivalued attributes
    - has difficulty when # of classes is large
    - Tends to favor tests that result in equal-sized partitions and purity in both partitions

Distances

- **Each clustering problem is based on some kind of "distance" between points.**
  - **An Euclidean space has some number of real-valued dimensions and some data points.**
    - A Euclidean **distance is based on the locations of points in such a space.**
  - **A Non-Euclidean distance is based on properties of points, but not their "location" in a space.**
- **Distance Matrices**
  - **Once a distance measure is defined, we can calculate the distance between objects. These objects could be individual observations, groups of observations (samples) or populations of observations.**
  - **For *N* objects, we then have a symmetric distance matrix D whose elements are the distances between objects *i* and *j*.**

Axioms of a Distance Measure

- ***d*** **is a distance measure if it is a function from pairs of points to reals such that:**
  1. **d(x,y) > 0.**
  2. **d(x,y) = 0 iff x = y.**
  3. **d(x,y) = d(y,x).**
  4. **d(x,y) < d(x,z) + d(z,y) (triangle inequality).**
- **For the purpose of clustering, sometimes the distance (similarity) is not required to be a metric**
  - **No Triangle Inequality**
  - **No Symmetry**

Distance Measures

Minkowski Distance

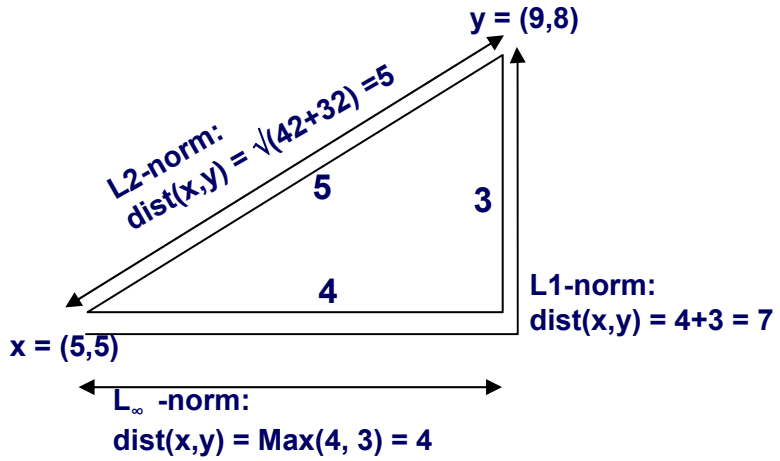
- The Minkowski distance of order *p* between two points is defined as:

$$\left(\sum_{k=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- **L_2 Norm (Euclidean Distance): p=2**
- **L_1 Norm (Manhattan Distance): p=1**
- **L_∞ Norm: p= ∞**

Euclidean Distance (L_1, L_∞ Norm)

- **L_1 Norm: sum of the differences in each dimension.**
  - **Manhattan distance = distance if you had to travel along coordinates only.**
- **L_∞ norm : d(x,y) = the maximum of the differences between x and y in any dimension.**



Mahalanobis Distance

- **Distances are computed based on means, variances and covariances for each of *g* samples (populations) based on *p* variables.**
- **Mahalanobis distance "weights" the contribution of each pair of variables by the inverse of their covariance.**

$$D_{ij}^2 = (\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)$$

  - μ_i , μ_j **are samples means.**
  - Σ **is the Covariance between samples.**

KL Divergence

- **Defined between two probability distributions**

$$D_{KL}(P||Q) = \sum_i P(i) \frac{\log P(i)}{\log Q(i)}$$

  - **Measures the expected number of extra bits in case of using Q rather than P**
  - **Can be used as a distance measure when feature vectors form distributions**

- **However, it is not a true metric**
  - **Triangle equality and symmetry do not hold**
  - **Symmetric variants have been proposed**

Linear Algebra – Review on Vectors

- **N-dimensional vectors**

$$v = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T$$

- **Unit Vector:**
  - **Any vector with magnitude equal to one**

$$||v|| = \sqrt{v^T v}$$

- **Inner product: Given two vectors $v=(x_1,x_2,x_3,\ldots,x_n)$ and $w=(y_1,y_2,y_3,\ldots,y_n)$**

$$v.w = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

More About Vectors

- **Geometrically-based definition of dot product**

$$v.w = ||v||.||w||.\cos(\theta) \quad (\theta \text{ is the smallest angle between v and w})$$

- **Orthonormal Vectors**
  - **A set of vectors $x_1, x_2, \ldots, x_n$ is called orthonormal if**

$$x_i x_j^T = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

  - **Any basis (x_1, x_2, . . . , x_n) can be converted to an orthonormal basis (o_1, o_2, . . . , o_n) using the Gram-Schmidt orthogonalization procedure.**

Vector Space

- **Given a set of objects W we see that W is a vector space if**

$$\lambda u + v \in W \text{ for all } \lambda \in F \text{ and } u,v \in W$$

  - **In general, F is the set of real numbers and W is a set of vectors**

- **A linear combination of vectors $v_1,\ldots v_k$ is defined as**

$$v = c_1 v_1 + c_2 v_2 + \cdots c_k v_k$$

  - **$c_1,\ldots c_k$ are scalars**

- **Spanning**

  - **We say that the set of vectors S=($v_1,\ldots v_k$) span a space W if every vector in W can be written as a linear combination of the vectors in S**

Linear Independence

- A set of vectors $v_1,\ldots,v_k$ is linearly independent if:
  - $c_1v_1 + c_2v_2 + \ldots + c_k v_k = 0$ implies $c_1 = c_2 = \ldots = c_k = 0$
- Geometric interpretation of linear independence
  - In R^2 or R^3, two vectors are linearly independent if they do not lie on the same line.
  - In R^3, there vectors are linearly independent if they do not lie in the same plane.

Basis

- A set of vectors $S=(v_1, ..., v_k)$ is said to be a basis for a vector space W if
  - (1) the v_js are linearly independent
  - (2) S spans W
  - Warning: The vectors forming a basis are not necessarily orthogonal !
- Theorem I: If V is an n-dimensional vector space, and if S is a set in V with exactly n vectors, then S is a basis for V if either S spans V or S is linearly independent.

Matrices

- **Matrix transpose**

$$A=\begin{bmatrix} a_{11} & a_{12} & . & . & a_{1n} \\ a_{21} & a_{22} & . & . & a_{2n} \\ . & . & . & . & . \\ . & . & . & & . \\ a_{m1} & . & . & . & a_{mn} \end{bmatrix}, A^T=\begin{bmatrix} a_{11} & a_{21} & . & . & a_{m1} \\ a_{12} & a_{22} & . & . & a_{m2} \\ . & . & . & . & . \\ . & . & . & & . \\ a_{1n} & a_{2n} & . & . & a_{mn} \end{bmatrix}$$

  - $(AB)^T=B^TA^T$

- **Symmetric matrix : $A^T=A$**

Linear Algebra – Determinant

- **Determinant**

$$A=[a_{ij}]_{n\times n};\qquad \det(A)=\sum_{j=1}^{n} a_{ij}A_{ij};\quad i=1,....n;\qquad A_{ij}=(-1)^{i+j}\det(M_{ij})$$

  - **det(AB)= det(A)det(B)**

- **Eigenvectors and Eigenvalues**

$$Ae_j=\lambda_j e_j,\ j=1,...,n;\quad ||e_j||=1$$

  - **Characteristic equation**

$$\det[A-\lambda I_n]=0$$

  - **Determinant & Eigen values**

$$\det[A]=\prod_{j=1}^{n}\lambda_j$$

Matrix Inverse

- **Matrix inverse (matrix must be square)**
  - **The inverse A^{-1} of matrix A has the property: $AA^{-1}=A^{-1}A=I$**
  - **A^{-1} exists only if $\det(A)\neq 0$**
    - Singular: the inverse of A does not exist
    - ill-conditioned: A is nonsingular but close to being singular
- **Some properties of the inverse:**
  - $\det(A^{-1}) = \det(A)^{-1}$
  - $(AB)^{-1}=B^{-1}A^{-1}$
  - $(A^T)^{-1} = (A^{-1})^T$

Rank of a Matrix

- **It is equal to the dimension of the largest square submatrix of A that has a non-zero determinant.**

$$A = \begin{pmatrix} 4 & 5 & 2 & 14 \\ 3 & 9 & 6 & 21 \\ 8 & 10 & 7 & 28 \\ 1 & 2 & 9 & 5 \end{pmatrix} \text{ has rank 3}$$

- **Alternatively, it is the maximum number of linearly independent columns or rows of A.**

Matrix Properties Based on Rank

- If A is m×n, rank(A) ≤ min m, n
- If A is n×n, rank(A) = n iff A is nonsingular (i.e., invertible).
- If A is n×n, rank(A) = n iff det(A)≠0 (full rank).
- If A is n×n, rank(A) < n iff A is singular

Any Question

End of Lecture 2

Thank you!

Spring 2012

http://ce.sharif.edu/courses/90-91/2/ce725-1/