Lecture 2: Review of Probability and Statistics

- **Probability**
  - Definition of probability
  - Axioms and properties
  - Conditional probability
  - Bayes Theorem

- **Random Variables**
  - Definition of a Random Variable
  - Cumulative Distribution Function
  - Probability Density Function
  - Statistical characterization of Random Variables

- **Random Vectors**
  - Mean vector
  - Covariance matrix

- **The Gaussian random variable**
Basic probability concepts

- **Definitions (informal)**
  - Probabilities are numbers assigned to events that indicate “*how likely*” it is that the event will occur when a random experiment is performed.
  - A *probability law* for a random experiment is a rule that assigns probabilities to the events in the experiment.
  - The *sample space* $S$ of a random experiment is the set of all possible outcomes.

- **Axioms of probability**
  - Axiom I: $0 \leq P[A_i]$
  - Axiom II: $P[S] = 1$
  - Axiom III: if $A_i \cap A_j = \emptyset$, then $P[A_i \cup A_j] = P[A_i] + P[A_j]$
More properties of probability

PROPERTY 1: \[ P[A^C] = 1 - P[A] \]

PROPERTY 2: \[ P[A] \leq 1 \]

PROPERTY 3: \[ P[\emptyset] = 0 \]

PROPERTY 4: given \(\{A_1, A_2, \ldots, A_N\}\), if \(A_i \cap A_j = \emptyset \ \forall i, j\) then \[ P[\bigcup_{k=1}^{N} A_k] = \sum_{k=1}^{N} P[A_k] \]

PROPERTY 5: \[ P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2] \]

PROPERTY 6: \[ P[\bigcup_{k=1}^{N} A_k] = \sum_{k=1}^{N} P[A_k] - \sum_{j<k}^{N} P[A_j \cap A_k] + \ldots + (-1)^{N+1} P[A_1 \cap A_2 \cap \ldots \cap A_N] \]

PROPERTY 7: if \(A_1 \subset A_2\), then \(P[A_1] \leq P[A_2]\)
**Conditional probability**

- If A and B are two events, the probability of event A when we already know that event B has occurred is defined by the relation

\[
P[A \mid B] = \frac{P[A \cap B]}{P[B]} \quad \text{for} \quad P[B] > 0
\]

- This conditional probability $P[A|B]$ is read:
  - the “conditional probability of A conditioned on B”, or simply
  - the “probability of A given B”

**Interpretation**

- The new evidence “B has occurred” has the following effects
  - The original sample space $S$ (the whole square) becomes $B$ (the rightmost circle)
  - The event $A$ becomes $A \cap B$
- $P[B]$ simply re-normalizes the probability of events that occur jointly with $B$
Theorem of total probability

- Let $B_1$, $B_2$, $\ldots$, $B_N$ be mutually exclusive events whose union equals the sample space $S$. We refer to these sets as a partition of $S$.

- An event $A$ can be represented as:

$$ A = A \cap S = A \cap (B_1 \cup B_2 \cup \ldots \cup B_N) = (A \cap B_1) \cup (A \cap B_2) \cup \ldots (A \cap B_N) $$

- Since $B_1$, $B_2$, $\ldots$, $B_N$ are mutually exclusive, then

$$ P[A] = P[A \cap B_1] + P[A \cap B_2] + \ldots + P[A \cap B_N] $$

- and therefore

$$ P[A] = P[A | B_1]P[B_1] + \ldots P[A | B_N]P[B_N] = \sum_{k=1}^{N} P[A | B_k]P[B_k] $$
**Bayes Theorem**

- Given $B_1, B_2, \ldots, B_N$, a partition of the sample space $S$. Suppose that event $A$ occurs; what is the probability of event $B_j$?
  - Using the definition of conditional probability and the Theorem of total probability we obtain
    \[
    P[B_j \mid A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A \mid B_j] \cdot P[B_j]}{\sum_{k=1}^{N} P[A \mid B_k] \cdot P[B_k]}
    \]

- This is known as Bayes Theorem or Bayes Rule, and is (one of) the most useful relations in probability and statistics
  - Bayes Theorem is definitely the fundamental relation in Statistical Pattern Recognition

---

Rev. Thomas Bayer (1702-1761)
Bayes Theorem and Statistical Pattern Recognition

- For the purposes of pattern recognition, Bayes Theorem can be expressed as

\[
P[\omega_j \mid x] = \frac{P[x \mid \omega_j] \cdot P[\omega_j]}{\sum_{k=1}^{N} P[x \mid \omega_k] \cdot P[\omega_k]} = \frac{P[x \mid \omega_j] \cdot P[\omega_j]}{P[x]}
\]

- where \( \omega_j \) is the \( i^{th} \) class and \( x \) is the feature vector

- A typical decision rule (class assignment) is to choose the class \( \omega_i \) with the highest \( P[\omega_i \mid x] \)
  - Intuitively, we will choose the class that is more “likely” given feature vector \( x \)

- Each term in the Bayes Theorem has a special name, which you should be familiar with
  - \( P[\omega_j] \)  **Prior probability** (of class \( \omega_i \))
  - \( P[\omega_j \mid x] \) **Posterior Probability** (of class \( \omega_i \) given the observation \( x \))
  - \( P[x \mid \omega_j] \) **Likelihood** (conditional probability of observation \( x \) given class \( \omega_i \))
  - \( P[x] \) A normalization constant that does not affect the decision
Random variables

- When we perform a random experiment we are usually interested in some measurement or numerical attribute of the outcome
  - When we sample a population we may be interested in their weights
  - When rating the performance of two computers we may be interested in the execution time of a benchmark
  - When trying to recognize an intruder aircraft, we may want to measure parameters that characterize its shape

- These examples lead to the concept of random variable
  - A random variable \( X \) is a function that assigns a real number \( X(\zeta) \) to each outcome \( \zeta \) in the sample space of a random experiment
    - This function \( X(\zeta) \) is performing a mapping from all the possible elements in the sample space onto the real line (real numbers)
  - The function that assigns values to each outcome is fixed and deterministic
    - as in the rule “count the number of heads in three coin tosses”
    - the randomness the observed values is due to the underlying randomness of the argument of the function \( X \), namely the outcome \( \zeta \) of the experiment
  - Random variables can be
    - Discrete: the resulting number after rolling a dice
    - Continuous: the weight of a sampled individual
Cumulative distribution function (cdf)

- The cumulative distribution function $F_X(x)$ of a random variable $X$ is defined as the probability of the event $\{X \leq x\}$

$$F_X(x) = P[X \leq x] \text{ for } -\infty < x < +\infty$$

- Intuitively, $F_X(b)$ is the long-term proportion of times in which $X(\zeta) \leq b$

- Properties of the cdf

$$0 \leq F_X(x) \leq 1$$

$$\lim_{x \to \infty} F_X(x) = 1$$

$$\lim_{x \to -\infty} F_X(x) = 0$$

$F_X(a) \leq F_X(b)$ if $a \leq b$

$$F_X(b) = \lim_{h \to 0} F_X(b + h) = F_X(b^+)$$
**Probability density function (pdf)**

- The **probability density function** of a continuous random variable $X$, if it exists, is defined as the derivative of $F_X(x)$

  \[ f_X(x) = \frac{dF_X(x)}{dx} \]

- For discrete random variables, the equivalent to the pdf is the **probability mass function**:

  \[ f_X(x) = \frac{\Delta F_X(x)}{\Delta x} \]

- **Properties**

  - $f_X(x) > 0$
  - $P[a < x < b] = \int_a^b f_X(x)dx$
  - $F_X(x) = \int_{-\infty}^x f_X(x)dx$
  - $1 = \int_{-\infty}^{+\infty} f_X(x)dx$
  - $f_X(x \mid A) = \frac{d}{dx} F_X(x \mid A)$ where $F_X(x \mid A) = \frac{P\{X < x\} \cap A}{P[A]}$ if $P[A] > 0$
Probability density function Vs. Probability

- **What is the probability of somebody weighting 200 lb?**
  - According to the pdf, this is about 0.62
  - This number seems reasonable, right?

- **Now, what is the probability of somebody weighting 124.876 lb?**
  - According to the pdf, this is about 0.43
  - But, intuitively, we know that the probability should be zero (or very, very small)

- **How do we explain this paradox?**
  - The pdf DOES NOT define a probability, but a probability DENSITY!
  - To obtain the actual probability we must integrate the pdf in an interval
  - So we should have asked the question: what is the probability of somebody weighting 124.876 lb plus or minus 2 lb?

- **The probability mass function is a ‘true’ probability (the reason why we call it ‘mass’ as opposed to ‘density’)**
  - The pmf is indicating that the probability of any number when rolling a fair dice is the same for all numbers, and equal to 1/6, a very legitimate answer
  - The pmf DOES NOT need to be integrated to obtain the probability (it cannot be integrated in the first place)
The cdf or the pdf are SUFFICIENT to fully characterize a random variable, however, a random variable can be PARTIALLY characterized with other measures:

- **Expectation**
  \[ E[X] = \mu = \int_{-\infty}^{+\infty} x f_X(x) \, dx \]
  - The expectation represents the center of mass of a density.

- **Variance**
  \[ \text{VAR}[X] = E[(X - E[X])^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x) \, dx \]
  - The variance represents the spread about the mean.

- **Standard deviation**
  \[ \text{STD}[X] = \text{VAR}[X]^{1/2} \]
  - The square root of the variance. It has the same units as the random variable.

- **Nth moment**
  \[ E[X^N] = \int_{-\infty}^{+\infty} x^N f_X(x) \, dx \]
The notion of a random vector is an extension to that of a random variable
- A vector random variable $X$ is a function that assigns a vector of real numbers to each outcome $\zeta$ in the sample space $S$
- We will always denote a random vector by a column vector

The notions of cdf and pdf are replaced by ‘joint cdf’ and ‘joint pdf’
- Given random vector, $X = [x_1, x_2, \ldots, x_N]^T$ we define
  - **Joint Cumulative Distribution Function** as:
    \[
    F_X(x) = P_X\left(\{X_1 < x_1\} \cap \{X_2 < x_2\} \cap \ldots \cap \{X_N < x_N\}\right)
    \]
  - **Joint Probability Density Function** as:
    \[
    f_X(x) = \frac{\partial^N F_X(x)}{\partial x_1 \partial x_2 \ldots \partial x_N}
    \]

The term **marginal pdf** is used to represent the pdf of a subset of all the random vector dimensions
- A marginal pdf is obtained by integrating out the variables that are not of interest
- As an example, for a two-dimensional problem with random vector $X = [x_1, x_2]^T$, the marginal pdf for $x_1$, given the joint pdf $f_{X_1X_2}(x_1, x_2)$, is
  \[
  f_{X_1}(x_1) = \int_{x_2=-\infty}^{x_2=+\infty} f_{X_1X_2}(x_1, x_2) \, dx_2
  \]
A random vector is also fully characterized by its joint cdf or joint pdf.

Alternatively, we can (partially) describe a random vector with measures similar to those defined for scalar random variables

- **Mean vector**
  \[
  \mathbb{E}[\mathbf{X}] = [\mathbb{E}[X_1] \mathbb{E}[X_2] \ldots \mathbb{E}[X_N]]^T = [\mu_1 \mu_2 \ldots \mu_N] = \mu
  \]

- **Covariance matrix**
  \[
  \text{COV}[\mathbf{X}] = \sum = \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] =
  \begin{bmatrix}
  \mathbb{E}[(x_1 - \mu_1)(x_1 - \mu_1)] & \ldots & \mathbb{E}[(x_1 - \mu_1)(x_N - \mu_N)] \\
  \ldots & \ldots & \ldots \\
  \mathbb{E}[(x_N - \mu_N)(x_1 - \mu_1)] & \ldots & \mathbb{E}[(x_N - \mu_N)(x_N - \mu_N)]
  \end{bmatrix}
  =
  \begin{bmatrix}
  \sigma_1^2 & \ldots & c_{1N} \\
  \sigma_2^2 & \ldots & c_{1N} \\
  \ldots & \ldots & \sigma_N^2
  \end{bmatrix}
  \]
Covariance matrix (1)

- The covariance matrix indicates the tendency of each pair of features (dimensions in a random vector) to vary together, i.e., to co-vary*

- The covariance has several important properties
  - If $x_i$ and $x_k$ tend to increase together, then $c_{ik} > 0$
  - If $x_i$ tends to decrease when $x_k$ increases, then $c_{ik} < 0$
  - If $x_i$ and $x_k$ are uncorrelated, then $c_{ik} = 0$
  - $|c_{ik}| \leq \sigma_i \sigma_k$, where $\sigma_i$ is the standard deviation of $x_i$
  - $c_{ij} = \sigma_i^2 = \text{VAR}(x_i)$

- The covariance terms can be expressed as
  \[ c_{ii} = \sigma_i^2 \text{ and } c_{ik} = \rho_{ik} \sigma_i \sigma_k \]
  - where $\rho_{ik}$ is called the correlation coefficient

*extracted from http://www.engr.sjsu.edu/~knapp/HCIRODPR/PR_home.htm
Covariance matrix (2)

- The covariance matrix can be reformulated as*
  \[ \Sigma = E[(X - \mu)(X - \mu)^T] = E[XX^T] - \mu\mu^T = S - \mu\mu^T \]
  
  with \( S = E[XX^T] = \begin{bmatrix} E[x_1x_1] & \cdots & E[x_1x_N] \\ \vdots & \ddots & \vdots \\ E[x_Nx_1] & \cdots & E[x_Nx_N] \end{bmatrix} \)
  
  - S is called the autocorrelation matrix, and contains the same amount of information as the covariance matrix

- The covariance matrix can also be expressed as
  \[ \Sigma = \Gamma R \Gamma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{12} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1N} & \vdots & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N \end{bmatrix} \]
  
  - A convenient formulation since \( \Gamma \) contains the scales of the features and \( R \) retains the essential information of the relationship between the features.
  - \( R \) is the correlation matrix

- Correlation Vs. Independence
  - Two random variables \( x_i \) and \( x_k \) are **uncorrelated** if \( E[x_i x_k] = E[x_i]E[x_k] \)
    - Uncorrelated variables are also called **linearly independent**
  - Two random variables \( x_i \) and \( x_k \) are **independent** if \( P[x_i x_k] = P[x_i]P[x_k] \)

*extracted from Fukunaga
The Normal or Gaussian distribution

- The multivariate Normal or Gaussian distribution $N(\mu, \Sigma)$ is defined as
  \[ f_x(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \]

- For a single dimension, this expression is reduced to
  \[ f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \]

- Gaussian distributions are very popular since
  - The parameters $(\mu, \Sigma)$ are **sufficient** to uniquely characterize the normal distribution
  - If the $x_i$'s are mutually **uncorrelated** ($c_{ik}=0$), then they are also **independent**
    - The covariance matrix becomes a diagonal matrix, with the individual variances in the main diagonal
  - **Central Limit Theorem**
  - The **marginal and conditional densities** are also Gaussian
  - Any **linear transformation** of any $N$ jointly Gaussian rv’s results in $N$ rv’s that are also Gaussian
    - For $X=[X_1 \ X_2 \ldots \ X_N]^T$ jointly Gaussian, and $A$ an $N \times N$ invertible matrix, then $Y=AX$ is also jointly Gaussian
      \[ f_y(y) = f_x(A^{-1}y) \frac{1}{|A|} \]
Central Limit Theorem

The central limit theorem states that given a distribution with a mean $\mu$ and variance $\sigma^2$, the sampling distribution of the mean approaches a normal distribution with a mean ($\mu$) and a variance $\sigma_i^2/N$ as $N$, the sample size, increases.

- No matter what the shape of the original distribution is, the sampling distribution of the mean approaches a normal distribution.
- Keep in mind that $N$ is the sample size for each mean and not the number of samples.

A uniform distribution is used to illustrate the idea behind the Central Limit Theorem.

- Five hundred experiments were performed using an uniform distribution.
  - For $N=1$, one sample was drawn from the distribution and its mean was recorded (for each of the 500 experiments).
    - Obviously, the histogram shown a uniform density.
  - For $N=4$, 4 samples were drawn from the distribution and the mean of these 4 samples was recorded (for each of the 500 experiments).
    - The histogram starts to show a Gaussian shape.
  - And so on for $N=7$ and $N=10$.
  - As $N$ grows, the shape of the histograms resembles a Normal distribution more closely.

![Histograms showing Central Limit Theorem](image)