Probability Theory Review

• Reading Assignments


"Everything I need to know about Probability" (on-line).
Probability Theory Review

• Definitions

Random experiment: an experiment whose result is not certain in advance (e.g., throwing a die)

Outcome: the result of a random experiment

Sample space: the set of all possible outcomes (e.g., \{1,2,3,4,5,6\})

Event: a subset of the sample space (e.g., obtain an odd number in the experiment of throwing a die = \{1,3,5\})

• Axioms of Probability

(1) $0 \leq P(A) \leq 1$

(2) $P(S) = 1$ (S is the sample space)

(3) If $A_1, A_2, ..., A_n$ are mutually exclusive events (i.e., $P(A_i \cap A_j) = 0$), then:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

*Note:* we will denote $P(A \cap C)$ as $P(A, B))$

• Other laws of probability

$$P(A) = 1 - P(\bar{A})$$

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

$$P(A) = P(A, B) + P(A, \bar{B}) \text{ (law of total probability)}$$
• Prior or Unconditional Probability

- It is the probability of an event prior to arrival of any evidence.

\[ P(\text{Cavity}) = 0.1 \text{ means that in the absence of any other information, there is a 10\% chance that the patient is having a cavity.} \]

• Posterior or Conditional Probability

- It is the probability of an event given some evidence.

\[ P(\text{Cavity/Toothache}) = 0.8 \text{ means that there is an 80\% chance that the patient is having a cavity given that he is having a toothache.} \]

- Conditional probabilities can be defined in terms of unconditional probabilities:

\[ P(A/B) = \frac{P(A, B)}{P(B)} = \frac{P(A, B)}{P(B)} \]

- The following formulas can be derived (chain rule):

\[ P(A, B) = P(A/B)P(B) = P(B/A)P(A) \]

- Using the above formula, we can rewrite the law of total probability as follows:

\[ P(A) = P(A, B) + P(A, \bar{B}) = P(A/B)P(B) + P(A/\bar{B})P(\bar{B}) \]

• Bayes theorem

- Using the conditional probability formula leads to the Bayes rule:

\[ P(A/B) = \frac{P(B/A)P(A)}{P(B)} \]
Example: consider the probability of Disease given Symptom

\[ P(\text{Disease}/\text{Symptom}) = \frac{P(\text{Symptom}/\text{Disease})P(\text{Disease})}{P(\text{Symptom})} \]

\[ P(\text{Symptom}) = P(\text{Symptom}/\text{Disease})P(\text{Disease}) + P(\text{Symptom}/\overline{\text{Disease}})P(\overline{\text{Disease}}) \]

- The general form of the Bayes rule is given by:

\[ P(A_i/B) = \frac{P(B/A_i)P(A)}{P(B)} \]

where \( A_1, A_2, ..., A_n \) is a partition of mutually exclusive events and \( B \) is any event

\[ P(B) = \sum_{j=1}^{n} P(B/A_j)P(A_j) \text{ (law of total probability)} \]

- **Independence**

  - Two events \( A \) and \( B \) are independent iff:

    \[ P(A, B) = P(A)P(B) \]

  - From the above formula, we can also show that:

    \[ P(A/B) = P(A) \text{ and } P(B/A) = P(B) \]

  - \( A \) and \( B \) are conditionally independent given \( C \) iff:

    \[ P(A/B, C) = P(A/C) \]

  - The following formula can be shown easily:

    \[ P(A, B, C) = P(A/B, C)P(B/C)P(C) \]
• Random variables

- In many experiments, it is easier to deal with a summary variable than with the original probability structure.

*Example:* in an opinion poll, we ask 50 people whether agree or disagree with a certain issue.

* Suppose we record a "1" for agree and "0" for disagree.
* The sample space for this experiment has $2^{50}$ elements.
* Suppose we are only interested in the number of people who agree.
* Define the variable $X=$number of "1"’s recorded out of 50.
* Easier to deal with this sample space (has only 50 elements).

- A random variable (r.v.) is the value we assign to the outcome of a random experiment (i.e., a function that assigns a real number to each event).
- How is the probability function of the random variable being defined from the probability function of the original sample space?

(1) Suppose the sample space is $S = < s_1, \ldots, s_n >$

(2) Suppose the range of the random variable $X$ is $< x_1, \ldots, x_m >$

(3) We will observe $X = x_j$ iff the outcome of the random experiment is an $s_j \in S$ such that $X(s_j) = x_j$, i.e.,

$$P(X = x_j) = P(s_j \in S: X(s_j) = x_j)$$

- A discrete r.v. can assume only a countable number of values (e.g., consider the experiment of throwing a pair of dice):

$X =$ "sum of dice"

e.g., $X = 5$ corresponds to $A_5 = \{ (1,4), (4,1), (2,3), (3,2) \}$

$$P(X = x) = P(A_x) = \sum_{s: X(s) = x} P(s) \text{ or }$$

$$P(X = 5) = P((1,4)) + P((4,1)) + P((2,3)) + P((2,3)) = 4/36 = 1/9$$

- A continuous random variable can assume a range of values (e.g., most sensor readings).

**Why should we care about r.v.?**

- Every sensor reading is a random variable (e.g., thermal noise, etc.)

- Many things in the real world can be appropriately viewed as random events (e.g., start time of lecture).

- There is some degree of uncertainty in almost everything we do.

- Some synonymous terms for "random" are *stochastic* and *non-deterministic*
- **Probability distribution function (PDF)**

  - With every r.v., we associate a function called *probability distribution function* (PDF) which is defined as follows:

    \[ F(x) = P(X \leq x) \]

  - Some properties of the PDF are:
    
    1. \( 0 \leq F(x) \leq 1 \)
    
    2. \( F(x) \) is a non-decreasing function of \( x \)

  - If \( X \) is discrete, its PDF can be computed as follows:

    \[
    F(x) = P(X \leq x) = \sum_{k=0}^{x} P(X = k) = \sum_{k=0}^{x} p(k)
    \]

    \[
    F(0) = P(X \leq 0) = P(X = 0) = 1/8
    \]

    \[
    F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = 1/2
    \]

    \[
    F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 7/8
    \]

    \[
    F(3) = P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1
    \]
• **Probability mass (pmf) or density function (pdf)**

- The *pmf* of a discrete r.v. $X$ assigns a probability for each possible value of $X$:

$$ p(x) = P(X = x) \text{ for all } x $$

**Important note:** given two r.v.’s, $X$ and $Y$, their *pmf* or *pdf* are denoted as $p_X(x)$ and $p_Y(y)$; for convenience, we will drop the subscripts and denote them as $p(x)$ and $p(y)$, however, keep in mind that these functions are different!

- The *pdf* of a continuous r.v. $X$ satisfies

$$ F(x) = \int_{-\infty}^{x} p(t)dt \text{ for all } x $$

- Using the above formula it can be shown that:

$$ p(x) = \frac{dF}{dx} (x) $$

- Some properties of the pmf and pdf:

$$ \sum_{x} p(x) = 1 \text{ (pmf)} $$

$$ P(a < X < b) = \sum_{k=a}^{b} p(k) \text{ (pmf)} $$

$$ \int_{-\infty}^{\infty} p(x)dx = 1 \text{ (pdf)} $$

$$ P(a < X < b) = \int_{a}^{b} p(t)dt \text{ (pdf)} $$
Example: the Gaussian pdf and PDF

- The joint pmf and pdf

Discrete r.v.

- For \( n \) random variables, the joint pmf assigns a probability for each possible combination of values:

\[
p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)
\]

Important note: the joint pmf’s or pdf’s of the r.v.’s \( X_1, X_2, \ldots, X_n \) and \( Y_1, Y_2, \ldots, Y_n \) are denoted as \( p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) \) and \( p_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) \); for convenience, we will drop the subscripts and denote them as \( p(x_1, x_2, \ldots, x_n) \) and \( p(y_1, y_2, \ldots, y_n) \), keep in mind, however, that these are two different functions.

- Specifying the joint pmf requires an enormous number of values (e.g., \( k^n \) assuming \( n \) random variables where each one can assume one of \( k \) discrete values).

\( P(Cavity, Toothache) \) is a 2 x 2 matrix

<table>
<thead>
<tr>
<th></th>
<th>Toothache</th>
<th>not Toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity</td>
<td>0.04%</td>
<td>0.06%</td>
</tr>
<tr>
<td>not Cavity</td>
<td>0.01%</td>
<td>0.89%</td>
</tr>
</tbody>
</table>
- The univariate pmf is related to the joint pmf by:

\[ p(x) = \sum_y p(x, y) \text{ (marginalization)} \]

Continuous r.v.

- For \( n \) random variables \( X_1, \ldots, X_n \), the joint pdf is given by:

\[ p(x_1, x_2, \ldots, x_n) \geq 0 \]

- The univariate pmf is related to the joint pmf by:

\[ p(x) = \int_{-\infty}^{\infty} p(x, y)dy \text{ (marginalization)} \]

Some interesting results using the joint pmf/pdf

- The conditional pdf can be derived from the joint pdf:

\[ p(y|x) = \frac{p(x, y)}{p(x)} \text{ or } p(x, y) = p(y|x)p(x) \]

- The law of total probability:

\[ p(y) = \sum_x p(y|x)p(x) \]

- Knowledge about independence between r.v.'s is very powerful since it simplifies things a lot, e.g., if \( X \) and \( Y \) are independent, then:

\[ p(x, y) = p(x) \ p(y) \]

- The chain rule of probabilities:

\[ p(x_1, x_2, \ldots, x_n) = p(x_1/x_2, \ldots, x_n)p(x_2/x_3, \ldots, x_n) \ldots p(x_{n-1}/x_n)p(x_n) \]
• Why is the joint pmf (or pdf) useful?

- Any other probability relating to the random variables can be calculated.

\[ P(B) = P(B, A) + P(B, \bar{A}) \]  
*(marginalization)*

(we can compute the probability of any r.v. from its joint probability)

- Here is how to compute \( P(A/B) \) (conditional probability):

\[ P(A/B) = \frac{P(A, B)}{P(B)} = \frac{P(A, B)}{P(A, B) + P(\bar{A}, B)} \]

• Normal (Gaussian) distribution

- The Gaussian pdf is defined as follows:

\[ p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] \]

where \( \mu \) is the mean and \( \sigma \) the standard deviation.

- The multivariate Gaussian (\( x \) is a vector) is defined as follows:

\[ p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[ -\frac{1}{2} (x - \mu)^t \Sigma^{-1} (x - \mu) \right] \]

where \( \mu \) is the mean and \( \Sigma \) the covariance matrix.

- Linear combinations of jointly Gaussian distributed variables follow a Gaussian distribution:

if \( y = A^t x \), then \( p(y) \sim N(A^t \mu, A^t \Sigma A) \)

- Whitening transformation:

\[ A_w = \Phi \Lambda^{-1/2} \]

if \( y = A_w^t x \), then \( p(y) \sim N(A_w^t \mu, I) \), that is, \( \Sigma_w = I \)

where the columns of \( \Phi \) are the (orthonormal) eigenvectors of \( \Sigma \), and \( \Lambda \) is a diagonal matrix corresponding to the eigenvalues of \( \Sigma \).
- Shape and parameters of Gaussian distribution:

\[ d + d(d + 1)/2 \] parameters, shape determined by \( \Sigma \)
- Mahalanobis distance:

\[ r^2 = (x - \mu)^\text{T} \Sigma^{-1} (x - \mu) \]

- The multivariate normal distribution for \textit{independent} variables becomes:

\[
p(x) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left[ -\frac{(x - \mu_i)^2}{2\sigma_i^2} \right]
\]
• Expected value

- The expected value for a discrete r.v. $X$ is given by

$$E(X) = \sum_x x p(x)$$

*Example:* Let $X$ denote the outcome of a die roll

$$E(X) = 1 \frac{1}{6} + 2 \frac{1}{6} + 3 \frac{1}{6} + 4 \frac{1}{6} + 5 \frac{1}{6} + 6 \frac{1}{6} = 3.5$$

- The "sample" mean $\bar{x}$ for a r.v. $X$ is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where $x_i$ denotes the $i$-th measurement of $X$.

- The mean and the expected value are related by

$$E(X) = \lim_{n \to \infty} \bar{x}$$

- The expected value for a continuous r.v. is given by

$$E(X) = \int_{-\infty}^{\infty} x p(x) dx$$

*Example:* $E(X)$ for the Gaussian is $\mu$. 
• **Properties of the expected value operator**

  - The expected value of a function \( g(X) \) is given by:
    \[
    E(g(X)) = \sum_x g(x) p(x) \text{ (discrete case)}
    \]
    \[
    E(g(X)) = \int_{-\infty}^{\infty} g(x) p(x) dx \text{ (continuous case)}
    \]

  - Linearity property
    \[
    E(af(X) + bg(Y)) = aE(f(X)) + bE(g(Y))
    \]

• **Variance and standard deviation**

  - The variance \( Var(X) \) of a r.v. \( X \) is defined by
    \[
    Var(X) = E((X - \mu)^2), \text{ where } \mu = E(X)
    \]

  - The "sample" variance \( \overline{Var} \) for a r.v. \( X \) is given by
    \[
    \overline{Var}(X) = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2
    \]

  - The standard deviation \( \sigma \) of a r.v. \( X \) is defined by
    \[
    \sigma = \sqrt{Var(X)}
    \]

*Example:* The variance of the Gaussian is \( \sigma^2 \)
- Covariance

- The covariance of two r.v. $X$ and $Y$ is defined by:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$

- The correlation coefficient $\rho_{XY}$ between $X$ and $Y$ is given by:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- The "sample" covariance matrix is given by:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n-1} (x_i - \bar{x})(y_i - \bar{y})$$

- Covariance matrix

- The covariance matrix of 2 random variables is given by:

$$C_{XY} = \begin{bmatrix}
\text{Cov}(X, X) & \text{Cov}(X, Y) \\
\text{Cov}(Y, X) & \text{Cov}(Y, Y)
\end{bmatrix}$$

where $\text{Cov}(X, X) = \text{Var}(X)$, $\text{Cov}(Y, Y) = \text{Var}(Y)$

- The covariance matrix of $n$ random variables is given as:

$$C_X = \begin{bmatrix}
\text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \ldots & \text{Cov}(X_1, X_n) \\
\text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \ldots & \text{Cov}(X_2, X_n) \\
\ldots & \ldots & \ldots & \ldots \\
\text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \ldots & \text{Cov}(X_n, X_n)
\end{bmatrix}$$

where $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$ and $\text{Cov}(X_i, X_i) \geq 0$

Example: $\Sigma$ is the covariance matrix of the multivariate Gaussian.
• **Uncorrelated random variables**
  - $X$ and $Y$ are called *uncorrelated*, if:
    $$\text{Cov}(X, Y) = 0$$
  - $X_1, X_2, \ldots, X_n$ are called *uncorrelated*, if:
    $$C_X = \Lambda,$$
    where $\Lambda$ is a diagonal matrix.

• **Properties of the covariance matrix**
  - Since $C_X$ is symmetric, it has real eigenvalues $\geq 0$
  - Any two eigenvectors, with different eigenvalues, are *orthogonal*.
  - The eigenvectors corresponding to different eigenvalues define a *basis*.

• **Decomposition of the covariance matrix**
  - The covariance matrix $C_X$ can be decomposed as follows:
    $$C_X = \Phi \Lambda \Phi^{-1}$$
    (1) the columns of $\Phi$ are the eigenvectors of $C_X$
    (2) the diagonal elements of $\Lambda$ are the eigenvalues of $C_X$
• **Transformations between random variables**

- Suppose \( X \) and \( Y \) are vectors of random variables:

\[
X = \begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_n
\end{bmatrix}, \quad Y = \begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix}
\]

which are related through the following transformation:

\[
Y = \Phi^T X
\]

- The coordinates of \( Y \) are *uncorrelated*:

\[
C_Y = \Lambda \quad \text{(i.e.,} \quad \text{Cov}(Y_i, Y_j) = 0)\]

- The eigenvalues of \( C_X \) become the variances of \( Y_i \)'s:

\[
\text{Var}(Y_i) = \text{Cov}(Y_i, Y_i) = \lambda_i
\]

• **Moments of a r.v.**

- Definition of moments:

\[
m_n = E(x^n)
\]

- Definition of central moments:

\[
cm_n = E((x - \mu)^n)
\]

- Useful moments

\( m_1 \): mean  
\( cm_2 \): variance  
\( cm_3 \): skewness (measure of asymmetry of a distribution)  
\( cm_4 \): kurtosis (detects heavy and light tails and deformations of a distribution)