

Statistical Learning and Machine Learning

Lecture 2 - Probability Theory

August, 2025

Why Probability Theory? I

Uncertainty is a key concept in pattern recognition:

- noisy measurements
- uncertainty in process knowledge
- small (finite) size of data sets describing a phenomenon

Probability theory provides a consistent framework for:

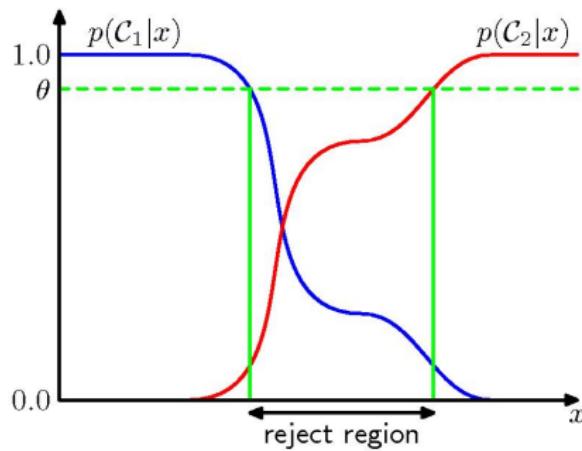
- the quantification of uncertainty
- the optimal use of all available information (even it may be incomplete and ambiguous)

Why Probability Theory? II

Input: x (X-ray image of patient)

Output: One of two classes, C_1 and C_2 (Cancer or No Cancer)

Goal: Finding a rule/function that assigns x to a particular class; the rule must be chosen based on probabilistic model



Key Concepts

- A **random experiment** is one in which outcome varies in unpredictable fashion when the experiment is repeated under similar conditions. It is specified by stating an experiment and a set of one or more measurements.
- A **Sample space** S of a random experiment is a set of all possible outcomes of the experiment. A single outcome or sample point of S , denoted by ς , is a result that cannot be decomposed further.
- Discrete vs Continuous Sample Space
- **Events** are set of points from a sample space that satisfy certain conditions

Key Concepts: The Axioms of Probability

- Probabilities are numbers assigned to events that indicate how likely it is that events will occur when a random experiment is performed.
- **Probability law** is a rule that assigns probabilities to outcomes/events of an experiment.
- The axioms of probability state conditions that a probability law must satisfy. Let E be an experiment with sample space S . The probability law assigns each event A a number $P[A]$, called the probability of A , which fulfills the following axioms:

① $P[A] \geq 0$

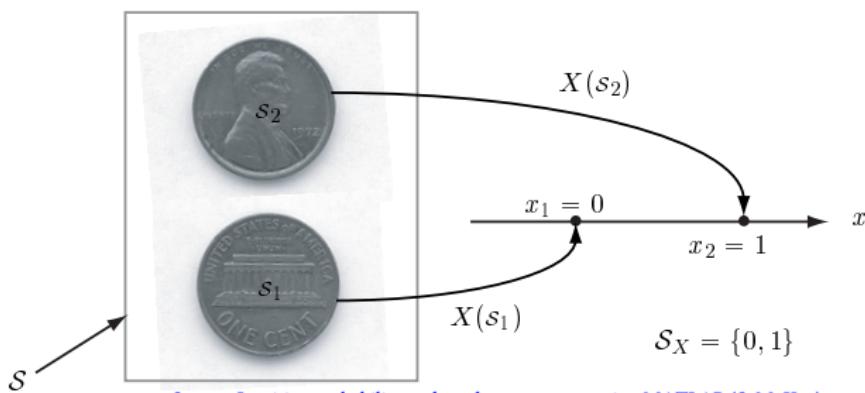
② $P[S] = 1$

③ If $A \cap B = \varphi$ then $P[A \cup B] = P[A] + P[B]$

④ If $A_i \cap A_j = \varphi$ for all $i \neq j$ then $P[\bigcup_{k=1}^{\infty} A_k] = \sum_{k=1}^{\infty} P[A_k]$

Random Variable

- A **Random variable** X is a function (or mapping) from sample space S of a random experiment to a set of real numbers i.e., X assigns numerical values to the outcomes of a random experiment.
- The sample space S is the domain while the new set S_X of all the values taken by X is the range of RV.
- The function X is fixed (deterministic) and the randomness in the observed values is due to the input argument.



Source: Intuitive probability and random processes using MATLAB (S. M. Kay)

Why do we need Random Variables?

- The events of interest in a random experiment involves measurement(s) or numerical attribute(s) of the outcomes
 - ① number of heads in n coin tosses
 - ② in a randomly selected computer job, we may be interested in the execution time of the job
 - ③ life time of a randomly selected computer chip
- **Mathematical convenience:** easier to deal with and manipulate numbers rather than non-numeric outcomes.

Cumulative Distribution Function

- **Cumulative distribution function (CDF)** of a RV X evaluated at x is defined by the probability of the event $\{X \leq x\}$:

$$F_X(x) = P[X \leq x] \text{ for } -\infty \leq x \leq +\infty \quad (1)$$

that is, the probability that X takes on values in the range $[(-\infty, x)]$.

- The probabilities corresponding to all intervals on real line can be computed from $F_X(x)$.

Properties of CDF

1

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

2

$$\lim_{x \rightarrow \infty} F_x(x) = 1$$

3

$$\lim_{x \rightarrow -\infty} F_x(x) = 0$$

4 $F_X(x)$ is a non-decreasing function of x i.e., if $a < b$, then $F_X(a) \leq F_X(b)$.

5

$$P[a < X \leq b] = F_X(b) - F_X(a)$$

6

$$P[X = b] = F_X(b) - F_X(b^-)$$

7

$$P[X > x] = 1 - F_X(x)$$

Example of CDF (Discrete RV)

A random experiment consists in counting the number of heads in 3 consecutive coin tosses. Let X denote the number of heads in 3 consecutive coin tosses. Find the CDF of X .

Probability Density Function (PDF)

- PDF of X is defined as the derivative of the CDF of X i.e.,

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

- Represents the ‘density’ of the probability of x in the following sense:

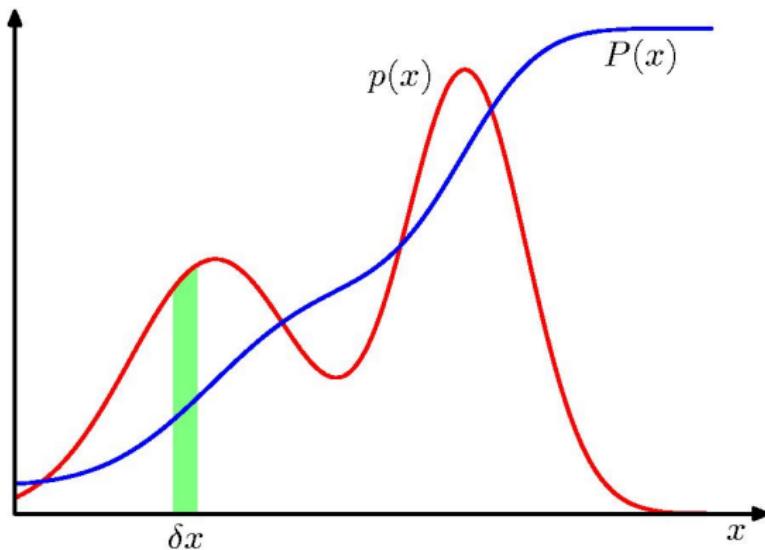
$$P[x < X \leq x + h] = F_X(x + h) - F_X(x) = \frac{F_X(x + h) - F_X(x)}{h} h.$$

If CDF has the derivative at x , then as h becomes very small,

$$P[x < X \leq x + h] = f_X(x)h.$$

Probability Density Function

Illustration of PDF as a measure of the 'probability density' of X at point x



Properties of PDF I

1

$$f_x(x) \geq 0$$

2

$$P[a \leq X \leq b] = \int_a^b f_X(x) dx$$

3

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

4

$$1 = \int_{-\infty}^{\infty} f_X(t) dt$$

Probability densities: multiple continuous variables

When we have several continuous variables forming a vector

$\mathbf{x} = [x_1, \dots, x_D]^T$ we define the joint probability $p(\mathbf{x}) = p(x_1, \dots, x_D)$:

- the probability of \mathbf{x} falling inside the space with volume $\delta\mathbf{x}$ containing \mathbf{x} is $p(\mathbf{x})\delta\mathbf{x}$.
- Properties:

$$p(\mathbf{x}) \geq 0 \quad \text{and} \quad \int p(\mathbf{x})d\mathbf{x} = 1$$

The integral is taken over the whole of \mathbf{x} space.

If \mathbf{x} is a discrete variable, then $p(\mathbf{x})$ is sometimes called a **probability mass function** because it can be regarded as a set of 'probability masses' concentrated at the allowed values of \mathbf{x} .

Expectations

The average value of a RV x under the probability distribution $p(x)$ is called *expectation* of x :

$$\mathbb{E}[x] = \sum_x xp(x)$$

For a continuous variable x :

$$\mathbb{E}[x] = \int xp(x)dx$$

When we are given N data points drawn from the probability distribution (or *probability density*) $p(x)$:

$$\mathbb{E}[x] \simeq \frac{1}{N} \sum_{n=1}^N x_n$$

Different forms of Expectations

- The average value of a function $f(x)$ under the probability distribution $p(x)$ is called *expectation* of $f(x)$:

$$\mathbb{E}[f] = \sum_x p(x)f(x)$$

- The expectation of a function $f(x, y)$ w.r.t. x is:

$$\mathbb{E}_x[f(x, y)]$$

- The *conditional expectation* w.r.t. a conditional distribution is given by:

$$\mathbb{E}[f|y] = \sum_x p(x|y)f(x)$$

Variance and Covariance

- The *variance* of x provides a measure of how much variability there is in x around its mean value $\mathbb{E}(x)$:

$$\begin{aligned}\text{var}[x] &= \mathbb{E} \left[(x - \mathbb{E}[x])^2 \right] \\ &= \mathbb{E}[x^2] - \mathbb{E}[x]^2\end{aligned}$$

- For two random variables x and y the *covariance* expresses the extent to which x and y vary together OR the linear relation between x and y

$$\begin{aligned}\text{cov}[x, y] &= \mathbb{E}_{x,y} \left[(x - \mathbb{E}[x])(y - \mathbb{E}[y]) \right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

- How will you model the covariance of a random vector with n components?

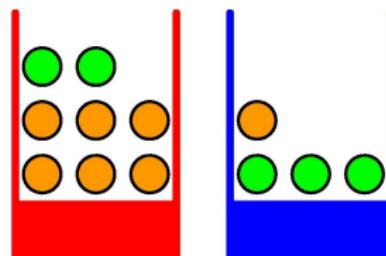
Key Concepts: Illustration with Example

We have two boxes:

- red box having 2 apples and 6 oranges
- blue box having 3 apples and 1 orange

We repeat the following process multiple times:

- ① we randomly pick one of the boxes
- ② we pick a piece of fruit
- ③ we observe the what type of fruit it is
- ④ we replace it in the box



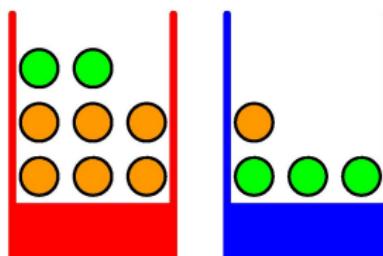
Key Concepts: Illustration with Example

Let us assume that:

- we pick the red(blue) box 40%(60%) of the times
- we are equally likely to select any of the pieces of fruit in any box

In the above experiment:

- the identity of the chosen box is a random variable B having two values:
 - r when the red box is chosen
 - b when the blue box is chosen
- the identity of the fruit is a random variable F having two values:
 - a when an apple is chosen
 - o when an orange is chosen



Probability Theory: Example

Probability of an event is the fraction of times that event occurs out of the total number of trials (in the limit that the total number of trials goes to infinity).

The probabilities of selecting the red and the blue boxes are:

$$p(B = r) = 4/10 = 0.4 \quad \text{and} \quad p(B = b) = 6/10 = 0.6$$

Properties:

- Probability values must lie in the interval $[0, 1]$
- The summation of all possible outcomes of *mutually exclusive* events is equal to 1.
- Probabilities follow two rules: the *sum rule* and the *product rule*

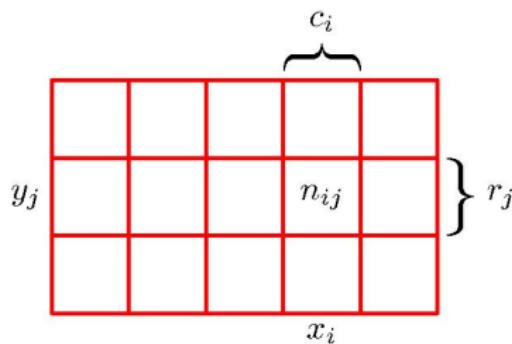
Probability Theory: Sum and Product rules

Let us consider two random variables:

- X which can take values $x_i, i = 1, \dots, M$
- Y which can take values $y_j, j = 1, \dots, L$

We make N trials and we count:

- the number n_{ij} of times we obtain $X = x_i$ and $Y = y_j$
- the number c_i of times we obtain $X = x_i$
- the number r_j of times we obtain $Y = y_j$



Probability Theory: Sum and Product rules

We define the following probabilities:

- The probability that $X = x_i$ and $Y = y_j$:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

- The probability that $X = x_i$ (irrespectively to the value of Y):

$$p(X = x_i) = \frac{c_i}{N}$$

Because $c_i = \sum_{j=1}^L n_{ij}$ we get the *marginal probability* of $X = x_i$:

$$p(X = x_i) = \sum_{j=1}^L p(X = x_i, Y = y_j)$$

The above is called the *sum rule* of probability.

Probability Theory: Sum and Product rules

We define the following probabilities:

- The fraction of instances for which $X = x_i$ corresponding to $Y = y_j$ is the *conditional probability*:

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

- Combining the marginal and the conditional probabilities we obtain the *joint probability* of $X = x_i$ and $Y = y_j$:

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i)p(X = x_i) \end{aligned}$$

The above is called the *product rule* of probability.

Probability Theory: Sum and Product rules

The Rules of Probability

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(Y|X)p(X).$

Continuous RVs:

$$\begin{aligned} p(x) &= \int p(x, y) dy \\ p(x, y) &= p(y|x)p(x) \end{aligned}$$

Probability Theory: The Bayes's Theorem

Using the product rule of probability and the symmetry property $p(X, Y) = p(Y, X)$ we get:

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

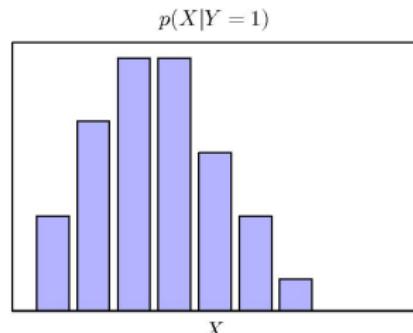
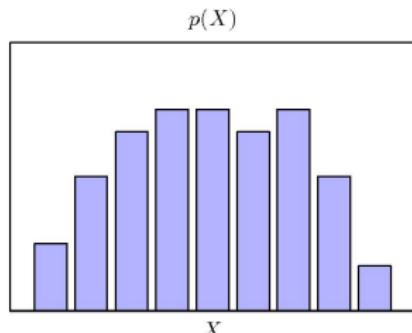
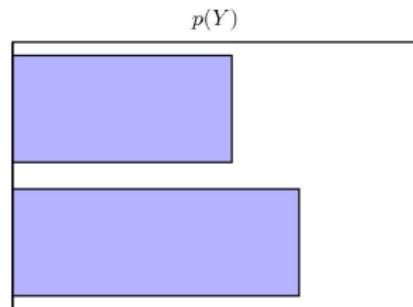
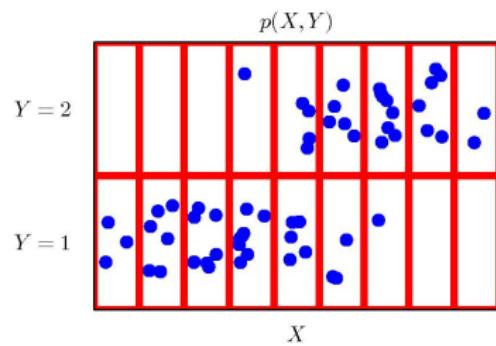
which is called *Bayes's theorem*.

Using the sum rule the Bayes's theorem becomes:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$$

We see that the denominator acts as a normalization term restricting the probability to the interval $[0, 1]$.

Probability Theory: The Bayes's Theorem



Probability Theory: Example

In the previous example:

$$p(B = r) = 4/10$$

$$p(B = b) = 6/10$$

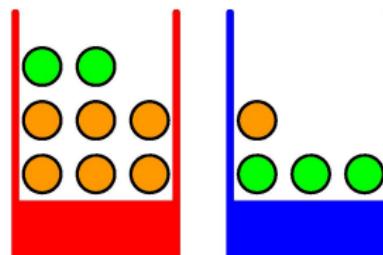
The conditional probabilities of fruit types given a box selection:

$$p(F = a|B = r) = 1/4$$

$$p(F = o|B = r) = 3/4$$

$$p(F = a|B = b) = 3/4$$

$$p(F = o|B = b) = 1/4$$



Probability Theory: Example

The conditional probabilities are normalized:

$$p(F = a|B = r) + p(F = o|B = r) = 1$$

$$p(F = a|B = b) + p(F = o|B = b) = 1$$

What is the probability of selecting an apple?

$$\begin{aligned} p(F = a) &= p(F = a|B = r)p(B = r) + p(F = a|B = b)p(B = b) \\ &= \frac{1}{4} \frac{4}{10} + \frac{3}{4} \frac{6}{10} = \frac{11}{20} \end{aligned}$$

What is the probability of selecting an orange?

Probability Theory: Example

Given that we selected an orange:

- what is the probability that we selected the red box?

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)} = \frac{3}{4} \frac{4}{10} \frac{20}{9} = \frac{2}{3}$$

- what is the probability that we selected the blue box?

$$p(B = b|F = o) = ?$$

Joint probability of independent variables

If for two variables X and Y we have:

$$p(X, Y) = p(X)p(Y)$$

then from the product rule:

$$p(Y|X) = \frac{p(X, Y)}{p(X)} = \frac{p(X)p(Y)}{p(X)} = p(Y)$$

Thus, X and Y are said to be *independent*.

Are RVs B and F independent in the previous example?