CS 452 Machine Learning
Ashesi University

Probability theory

“

The 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 approaches infinity

”

BISHOP, 2006, p.12

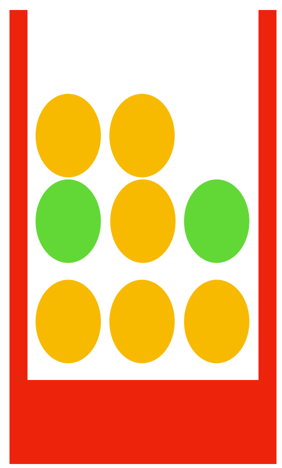
Key Terms/Concepts

- Sum rule

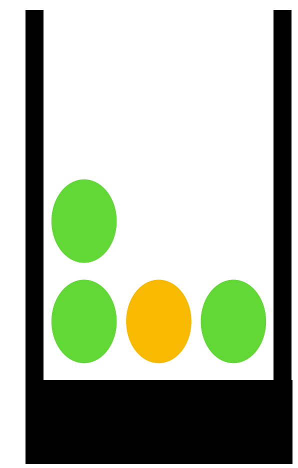
- Product rule

- Bayes' theorem

As a simple example, consider the two boxes below containing apples and oranges.



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

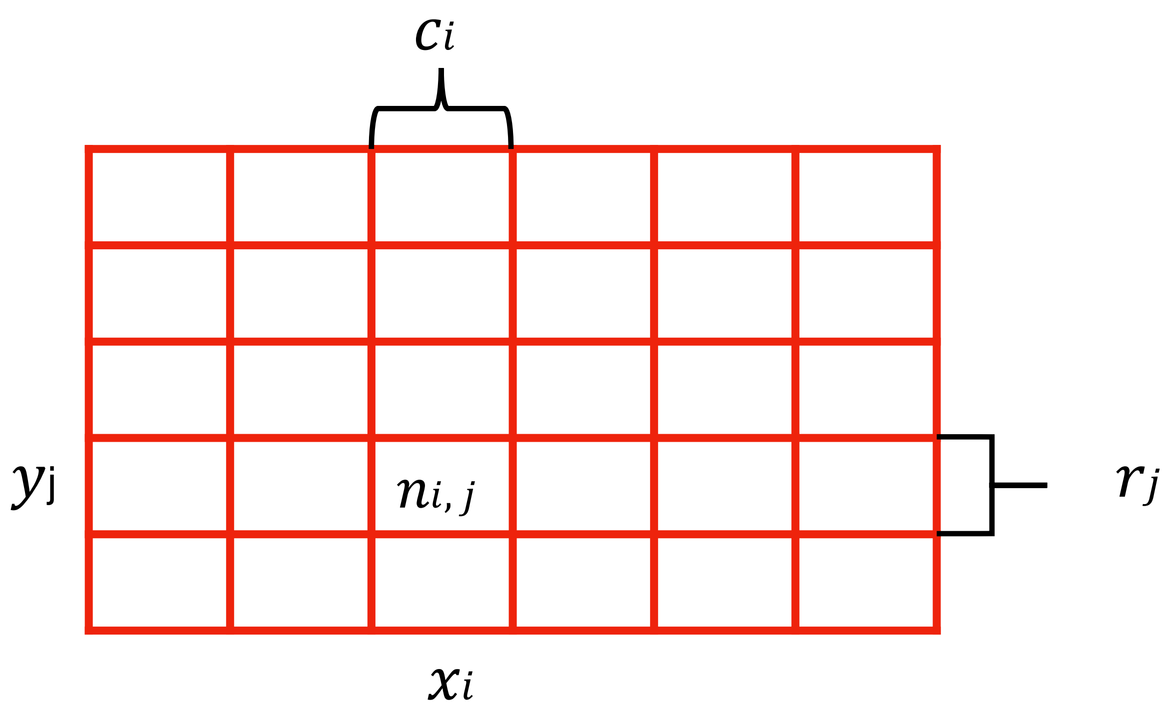


$p(B = r) = 6/10$

Supposing selecting any of the fruits from a box is equally likely, we can then ask questions such as what is the probability

- of selecting an orange from the red box?
- that the red box was chosen if an orange is selected?
- of selecting an apple?

To answer the above question and drive home some important rule, consider the diagram below



The grid above depicts two random variables X, which takes the values $\{x_1, \ldots, x_M\}$ and Y, which takes the values $\{y_1, \ldots, y_L\}$. If we consider a total number N of instances of these variables, then we denote the number of instances where $X = xi\ and\ Y = y_j$ to be $n_{i,j}$ where $i = 1, \ldots, M$ and $j = 1, \ldots, L$. $n_{i,j}$ is put in the corresponding cell of the array. The number of points in column i, where $X = xi$ is c_i and the number of points in row j where $Y = y_j$ is r_j.

Joint Probability

The probability that $X = xi$ and $Y = y_j$ called the joint probability is denoted as $p(X = x_i, Y = y_i)$.

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

Sum Rule

The probability that $X = xi$ irrespective of the values of Y is:

$$P(X = xi) = ci/N$$

Also,

$$c_i = \sum_{j=1}^{L} n_{i,j}$$

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

The sum Rule of Probability

Sometime called the marginal probability

Product Rule

If we consider only the instances for which $X = x_i$, then the fraction of such instances for which $Y = y_j$ written as $p(Y = y_j | X = x_i)$. This is called the conditional probability of $Y = y_j$.

$$\boxed{p(Y = y_j | X = x_i) = \frac{n_{i,j}}{c_i}}$$

conditional probability of $p(Y = y_j)$given $X = xi$

$$p(X = x_i, Y = y_j) = \frac{n_{i,j}}{N} = \frac{n_{i,j}}{c_i} \cdot \frac{c_i}{N}$$

$$\boxed{= p(Y = y_j | X = x_i) p(X = x_i)}$$

The product rule of probability

Bayes' Theorem

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

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

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

$$\Rightarrow p(Y|X)p(X) = p(X|Y)p(Y)$$

Sometimes called the prior

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

Sometimes called the posterior probability

Bayes' Theorem

Independent events

Two events are said to be independent if their joint distribution factorizes into the product of the marginals

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

From the product rule, it can be shown that

$$p(Y|X) = p(Y)$$

Thus, knowledge of one event does not affect the distribution of the other

Summary of Rules

Sum Rule: $$p(X) = \sum_{Y} p(X, Y)$$

Product Rule: $$p(X, Y) = p(Y|X)p(X)$$

Bayes' Theorem: $$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

Probability Densities

- Usually for continuous events.
- If the probability of a continuous variable x falling in the interval $(x, x+\ \delta x)$ is given by $p(x)\delta x \text{ for } \delta x \to 0$, then $p(x)$ is called the probability density over x.
- When x is discrete, $p(x)$ is called a probability mass function
- The probability that x is in the interval (a, b) is

$$p(\in (a, b) = \int_a^b p(x)dx$$

Probability density over x

Probability Densities

Any probability density MUST meet the following conditions:

$$\int_{-\infty}^{\infty} p(x)\delta x = 1$$

$$p(x) \geq 0$$

Note that $p(x)$ isn't required to be less than or equal to 1

Probability Densities

The sum and the product rule described in Lecture 2 also apply to probability densities, with the summations replaced with integrals

Sum rule: $p(x) = \int p(x, y)\mathrm{d}y$

Product rule: $\mathrm{p}(x, y) = p(y|x)p(x)$

Expectations

- A measure of averages of functions

- The average of a function $f(x)$ under a probability distribution $p(x)$ is the expectation of $f(x)$.

discrete variables:

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

continuous variables :

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

Variances

- A measure of variability of functions around their mean (average or expectation)

$$\mathbf{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

$$= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

- For a single variable x, the variance can be expressed as:

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

Covariances

- A measure of the extent to which two random variables vary together.
- The covariance of independent events is zero
- For two random variables x and y, their covariance is given as:

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

For two vectors of random variables x and y, the covariance is

$$\mathrm{cov}[\mathrm{x}, \mathrm{y}] = \mathbb{E}_{x,y}[\{\mathrm{x} - \mathbb{E}[\mathrm{x}]\}\{y^{\mathrm{T}} - \mathbb{E}[y^{\mathrm{T}}]\}]$$

$$= \mathbb{E}_{x,y}[\mathrm{x}y^{\mathrm{T}}] - \mathbb{E}[\mathrm{x}]\mathbb{E}[y^{\mathrm{T}}]$$