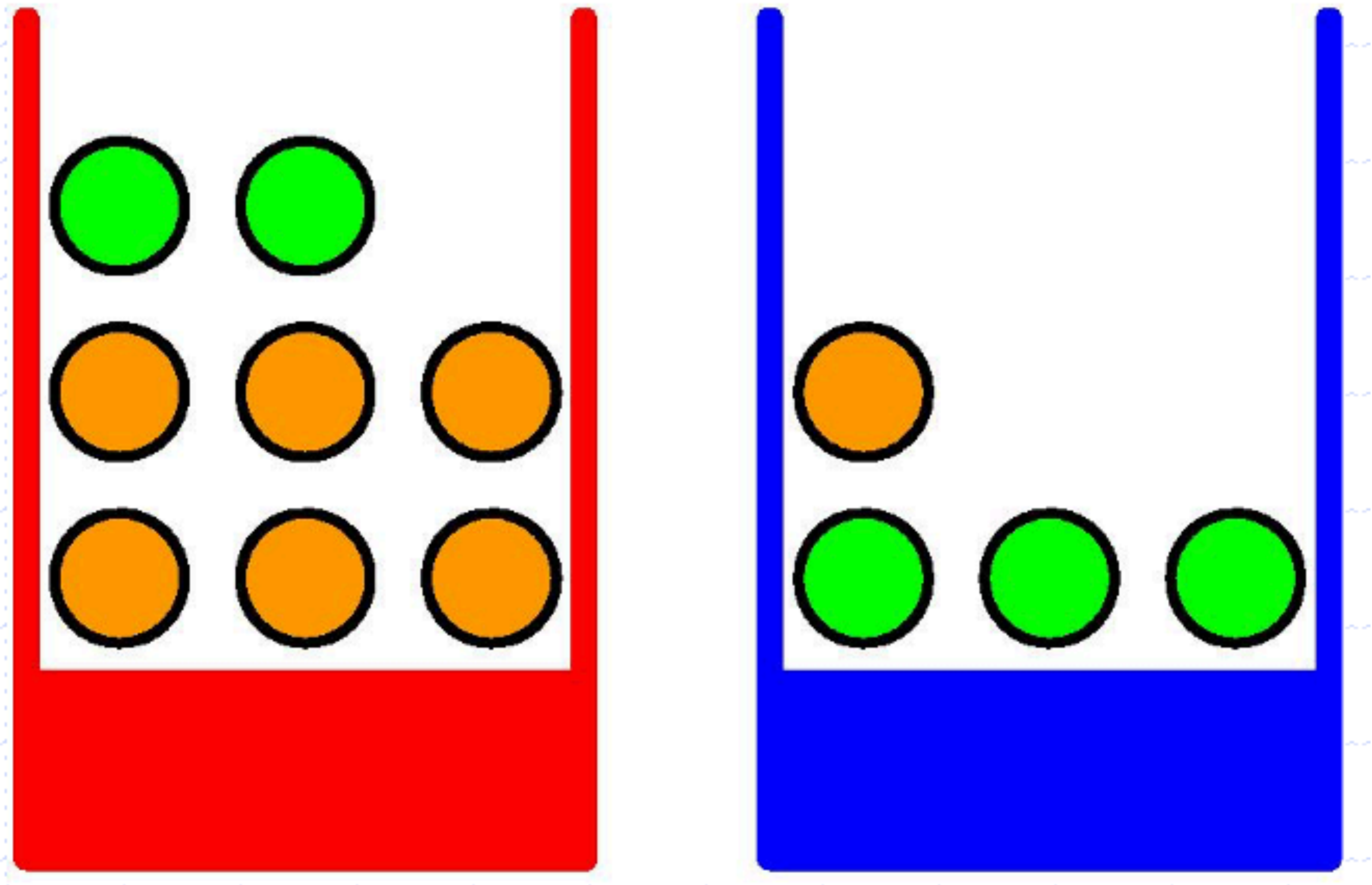
INM431: Machine Learning

Probability theory and Bayes Theorem

Pranava Madhyastha (pranava.madhyastha@city.ac.uk)

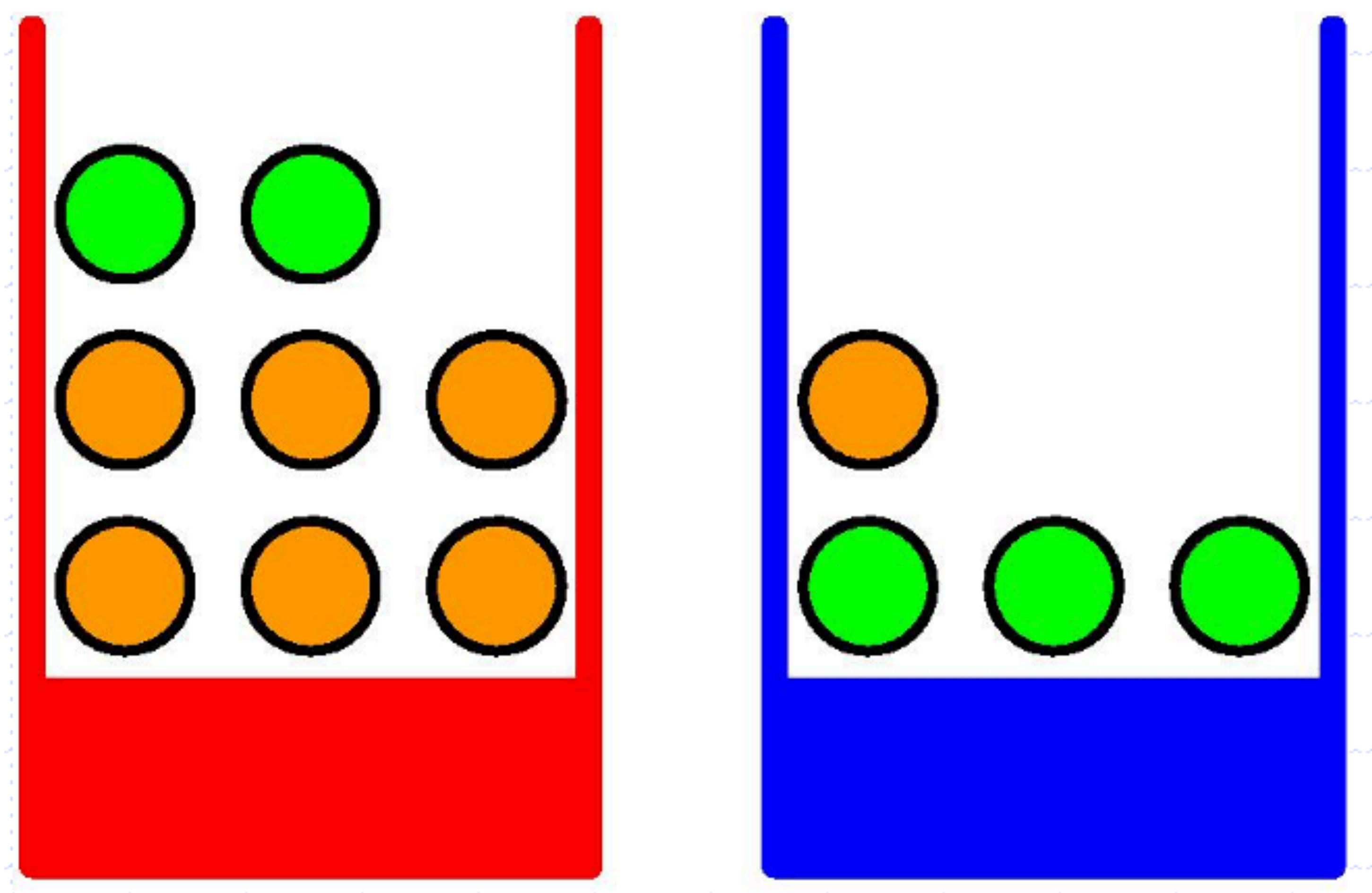
Based on materials by Artur S. d'Avila Garcez



Apples and oranges: picking an item from the box

Random variables: BOX = {red, blue}; FRUIT = {apple, orange}

What is the probability of picking an apple?



What is the probability of picking an apple?

$$P(\text{Fruit=Apple}) = P(\text{Fruit=apple,Box=red}) + P(\text{Fruit=apple,Box=blue})$$

$$P(\text{Fruit=apple,Box=red}) = P(\text{Fruit=apple}|\text{Box=red}) \times P(\text{Box=red})$$

$$P(\text{Fruit=apple,Box=blue}) = P(\text{Fruit=apple}|\text{Box=blue}) \times P(\text{Box=blue})$$

Basic rules of probability

Probability $\in [0,1]$

$$P(X) = 1 - \neg P(X)$$

Basic rules of probability

Sum rule: $$P(X) = \sum_{Y} P(X, Y)$$

Product rule: $$P(X \cap Y) = P(Y \mid X) P(X)$$

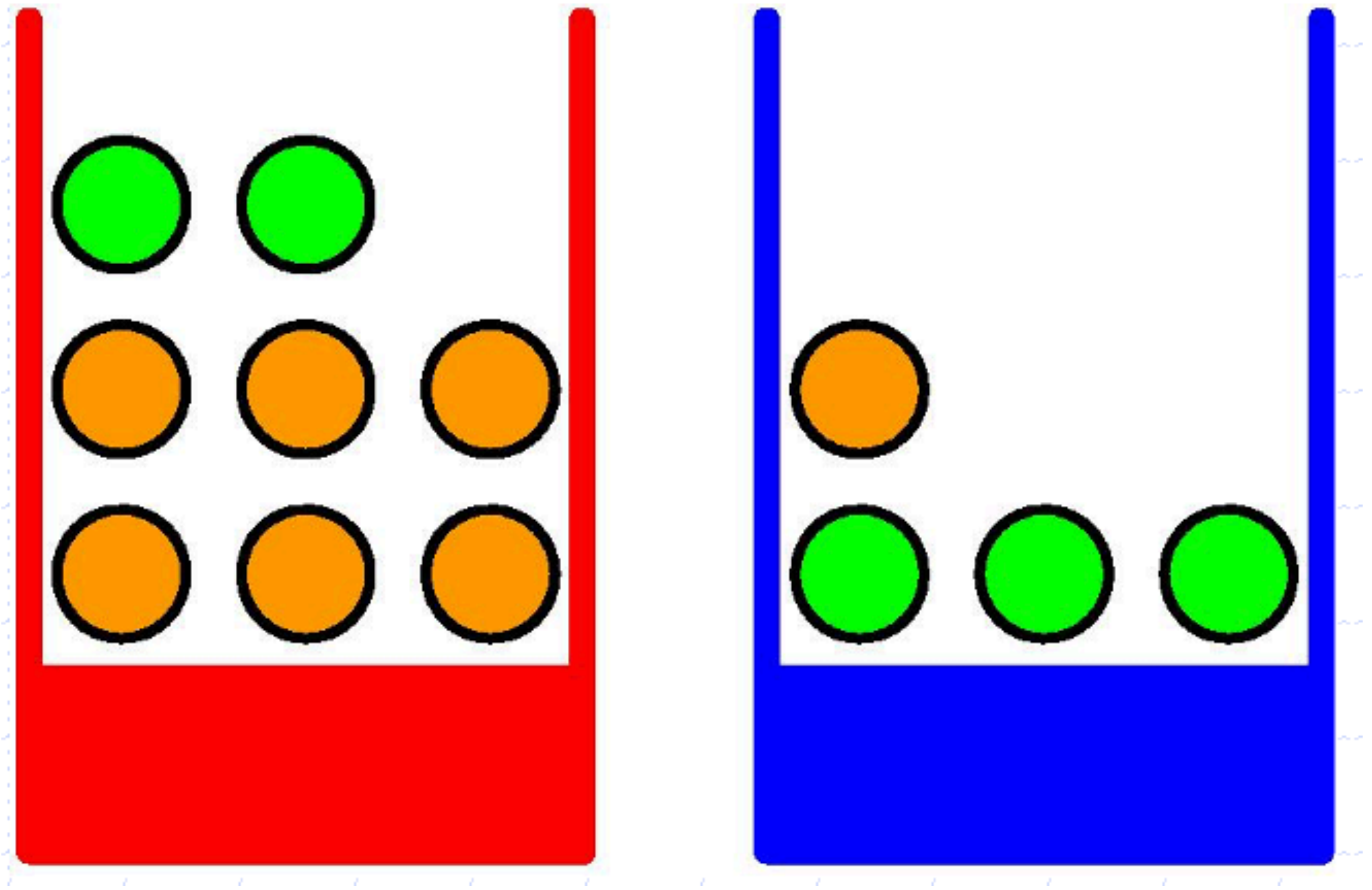
$$P(X, Y)$$

Joint probability

Conditional probability

Marginal probability

Quick quiz



What is the probability that the fruit was from BLUE box if an orange was picked?

Try resorting to sum & product rules (and nothing else).

Most general form of Bayes' theorem

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

where, $P(X) = \sum_{Y} P(X \mid Y)P(Y)$ (normalisation constant)

posterior $\propto$ likelihood $\times$ prior

Correlation: Normalised co-variance

The magnitude of the covariance is a bit tricky to interpret.

Correlation: $\rho[x, y] = \dfrac{\mathbb{E}_{x,y}[\{x - \mathbb{E}[x]]\}\{y - \mathbb{E}[y]\}]}{(\sigma_x \sigma_y)}$

where, σ_x is the standard deviation in x and σ_y is the standard deviation in y

also, $\rho[x, x] = 1$

Bayesian learning

Every training example can incrementally decrease or increase the estimated probability about a certain hypothesis

Prior knowledge can be combined with observations to obtain the probability of a certain event. This can be done through:

- a prior probability for every candidate hypothesis
- a probability distribution over the observations for each hypothesis

Bayes' theorem

We want to obtain the best hypothesis (h) from a large event space $\mathscr{H}$, given some observed data $\mathscr{D}$

In bayesian learning, best hypothesis = most probable hypothesis, given the observations $\mathscr{D}$ + prior information about the various hypotheses in $\mathscr{H}$

Bayes' theorem helps us compute the probability of a hypothesis based on its prior probability, the probabilities of observing data under a given hypothesis and the observed data.

Bayes' theorem

$P(h)$ = prior probability of hypothesis h

- initial probability of the hypothesis, before we have any data

$P(\mathscr{D})$ = prior probability of training data $\mathscr{D}$

- without any knowledge about any hypothesis

$P(h \mid \mathscr{D})$ = posterior probability of h given $\mathscr{D}$

- contains the confidence that h is true after having access to the dat.

$P(\mathscr{D} \mid h)$ = the likelihood of the data $\mathscr{D}$ given that h is true

The Bayes' theorem for Bayesian learning

$$P(h \mid \mathscr{D}) = \frac{P(\mathscr{D} \mid h)P(h)}{P(\mathscr{D})}$$

posterior ∝ likelihood × prior

In general, in frequentist ML, we are interested in $P(h \mid \mathscr{D})$, i.e., probability that h is true given the data $\mathscr{D}$

In bayesian learning methods, we have the opportunity to calculate $P(h \mid \mathscr{D})$ using both $P(h)$ and $P(\mathscr{D})$, but computing this is not always tractable

Exercise

A test for salmonella is made available to chicken farmers. The test will correctly show a positive result for salmonella 95% of the time. However the test also shows positive results 15% of the time in salmonella free chickens. 10% of chickens have salmonella.

If a chicken tests positive, what is the probability that it has salmonella?

Expectation

The average value (or the mean) of any function $f(x)$ under a probability distribution $P(x)$ is called the "**expectation**" (or the expected value) of $f(x)$, denoted by:

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

In the case of continuous variables:

$$\mathbb{E}[f] = \int_x P(x)f(x)dx$$

Further, with a finite number of points N drawn from a probability distribution, the expected value is computed by using:

$$\mathbb{E}[F] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

Covariance

$\mathrm{Cov}[x, y]$ denotes the extent to which the random variables x and y vary together (co-vary).

It is a measure of linear dependence.

If x and y are **independent** then $\mathrm{Cov}[x, y] = 0$, however note that the converse may not necessarily be true, e.g., $y = x^2$.

In case of vectors $\mathbf{x}$ and $\mathbf{y}$, covariance is typically represented with a matrix with a symbol $\sum$ or $K_{\mathbf{xy}}$.

Computing covariance

For random variables:

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

For vectors:

$$\begin{aligned}\mathrm{Cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]]\}\{\mathbf{y} - \mathbb{E}[\mathbf{y}]\}] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^\top]\end{aligned}$$

Example (covariance with random variables)

x	y	xy
3	2	6
-1	4	-4
1	1	1
$\mathbb{E}[x] = 1$	$\mathbb{E}[y] = 7/3$	$\mathbb{E}[xy] = 1$

$$\begin{array}{c c} & \begin{array}{cc} x & y \end{array} \\ \begin{array}{c} x \\ y \end{array} & \begin{bmatrix} \mathrm{Var}[x] & \mathrm{Cov}[x, y] \\ \mathrm{Cov}[y, x] & \mathrm{Var}[y] \end{bmatrix} \end{array}$$

$$\mathrm{Cov}[x, y] = 1 - (2.33) \times 1$$

$$\mathrm{Cov}[x, x] = \mathrm{Var}[x]$$

The covariance of a variable with itself is called its variance

Homework: vectors

$\mathbf{x}$	$\mathbf{y}$	$\mathbf{x}\mathbf{y}^\top$
(3,2)	(2,3)	?
(-1,4)	(-4,-5)	?
(1,1)	(2,3)	?
$\mathbb{E}[x] = ?$	$\mathbb{E}[y] = ?$	$\mathbb{E}[xy] = ?$

$$\begin{array}{c c} & \begin{array}{cc} x & y \end{array} \\ \begin{array}{c} x \\ y \end{array} & \left[\begin{array}{c} ? \end{array} \right] \end{array}$$

$$\mathrm{Cov}[x, y] = ?$$

$$\mathrm{Cov}[x, x] = ?$$