Probability Review

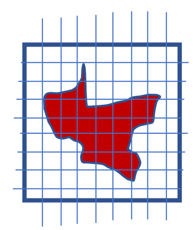
Outline of the Lecture

This lecture is intended at revising probabilistic concepts that play an important role in the design of machine learning and data mining algorithms. It is expected that you will learn and master the following three topics:

- **Frequentist definition of probability**
- **Conditioning**
- **Marginalization**

Probability as measure

Imagine we are throwing darts at a wall of size 1x1 and that all darts are guaranteed to fall within this 1x1 wall. What is the probability that a dart will hit the shaded area?



Probability is a measure of certainty of an event taking place, i.e. in the example above we were measuring the changes of hitting the shaded area

Probability as frequency

Consider the following questions:

What is the probability that when I flip a coin it is “heads”?

Why?

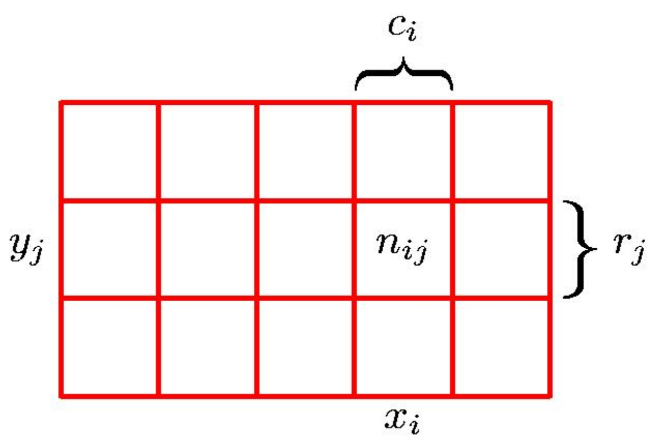
What is the probability that the Golden Bridge in San Francisco collapse before the term is over?

The frequentist view is very useful, but it seems that we also use domain knowledge to come up with probabilities. Moreover, it seems that probability can be subjective (different people have different probabilities for the same event)

Probability

Probability is the formal study of the laws of chance. Probability allows us to manage uncertainty.

Probability Theory



Marginal Probability

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

Joint Probability

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

Conditional Probability

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

The axioms

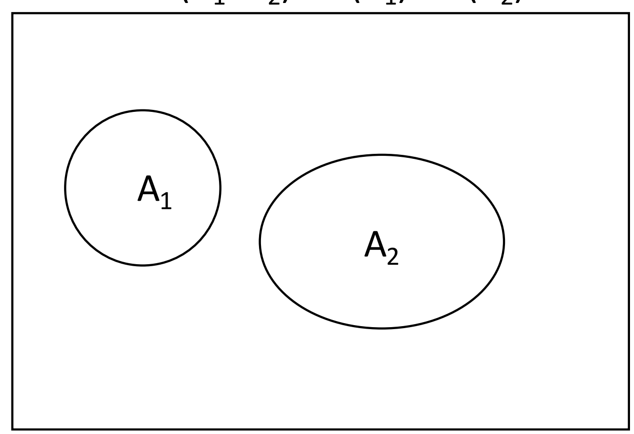
The following two laws are the key axioms of probability:

1. $$P(\emptyset) = 0 \leq p(A) \leq 1 = P(\Omega)$$

2. For **disjoint sets** A_n, $n \geq 1$, we have

$$P\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

$P(A_1+A_2) = P(A_1) + P(A_2)$

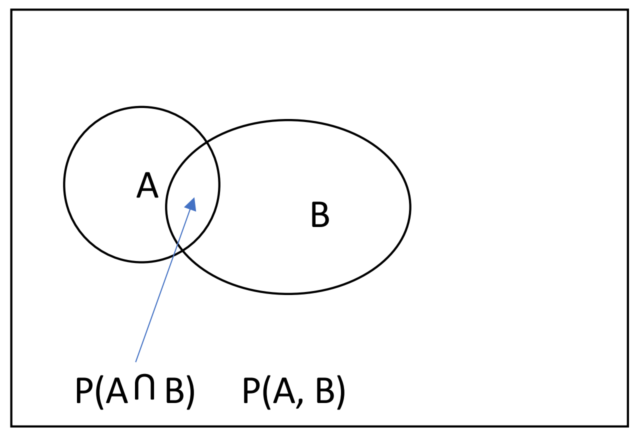


Or and And operations

Given two events, A and B, that are not disjoint, we have:

P(A or B) = P(A)+P(B) – P(A And B)

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



Joint Probability

Conditional Probability

Assuming (given that) B has been observed (i.e. there is no uncertainty about B), the following tell us the probability that A will take place:

P(A given B) = P(A and B) / P(B)

$P(A| B) = P(A, B) /P(B)$

That is, in the frequentist interpretation, we calculate the ratio of the number of times both A and B occurred and divide it by the number og times B occurred.

The Rules of Probability

Marginalization

- Sum Rule $$p(X) = \sum_Y p(X, Y)$$
- Product Rule $$p(X, Y) = p(Y|X)p(X)$$

Proof

All the possible value of Y

$$\sum_Y p(X, Y)$$

$P(XY_1) + P(XY_2) + P(XY_3)$ ...
$= P(Y_1|X).P(X) + P(Y_2|X).P(X) + P(Y_3|X).P(X) + \ldots$
$= P(Y_1 + Y_2 + Y_3 \ldots | X) . P(X)$
$= P(X)$

Independence

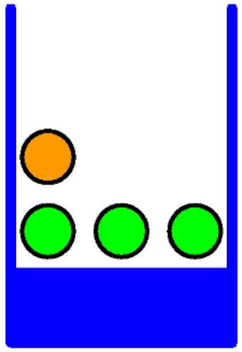
$$
\begin{aligned}
P(X, Y) &= P(X|Y)P(Y) \\
&= P(X)P(Y)
\end{aligned}
$$

Conditional probability example

Assume we have a dark box with 3 green balls and 1 orange ball. That is, we have the set {g,g,g,o}. What is the probability of drawing 2 green balls in the first 2 tries?

$P(B_1 = g, B_2 = g) = ?$

$P(B_2 = g, B_1 = g) = P(B_2 = g | B_1 = g).P(B_1 = g) = \frac{3}{4} \times 2/3 = \frac{1}{2}$



$P(B_2 = g) = ?$

$P(B_2 = g) = P(B_2 = g, B_1 = g) + P(B_2 = g, B_1 = o) = \frac{1}{2} + 1/4 = 3/4$

The Rules of Probability

Marginalization

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

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

Matrix notation

Assume that X can be 0 = g or 1 = o. we use the notation X$\in$ {0,1}

$$\begin{array}{c} x1 \\ \begin{bmatrix} 3/4 & 1/4 \end{bmatrix} \end{array} \quad \begin{array}{c} x2 \\ \begin{bmatrix} 2/3 & 1/3 \\ 1 & 0 \end{bmatrix} \\ x1 \end{array} = \begin{array}{c} x2 \\ \begin{bmatrix} 3/4 & 1/4 \end{bmatrix} \end{array}$$

$$\sum_{x_1 \in \{0,1\}} P(x_1) P(x_2 | x_1) = P(X_2)$$

For short, we write this using vectors and a stochastic matrix

Matrix notation

$$\sum_{x_1 \in \{0,1\}} P(x_1)P(x_2|x_1) = P(X_2)$$

For short, we write this using vectors and a stochastic matrix

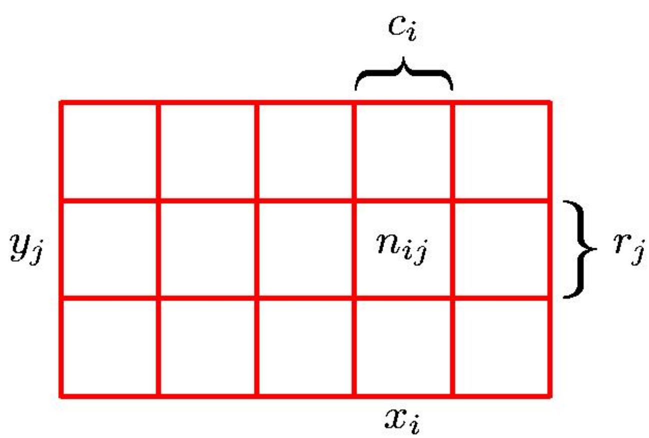
[1x2] [2x2] [2x1]

$$X_1 G = X_2$$

$$X_1 G = X_2 \qquad X_2 G = X_3 \qquad X_3 G = X_4$$

For a very large number k, after k iterations, the value of X stabilizes.
If $X_k = X$ Then $X G = X$
That is, X_k is an eigenvector of G with eigenvalue 1.
$Ax = x$

Probability Theory



Marginal Probability

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

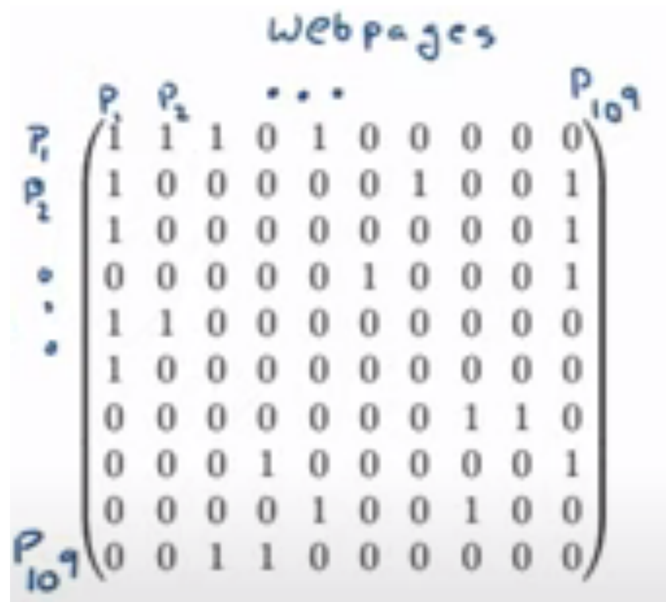
Joint Probability

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

Conditional Probability

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

Google’s website search algorithm (pageRank)



The largest Eigenvector is the probability of each page or as google says, its page rank.