Foundations of Machine Learning

Revisiting Prerequisite of Maths: Probability

Sayanta, Rishabh

MTech (RA)
Department of Artificial Intelligence
Indian Institute of Technology, Hyderabad

Table of Contents

Table of Contents

Experiments, Outcomes and Events

- Experiment - A repeatable process
- Outcome (ω) - The result of the experiment
- Outcome space or sample space ($\mathbf{\Omega}$) - The set of all possible outcomes of an experiment
- Event (A) - A subset of $\mathbf{\Omega}$

Experiments, Outcomes and Events

- Experiment - Two consecutive (independent) coin tosses
- Outcome (ω) - HT
- Outcome space or sample space ($\boldsymbol{\Omega}$) - {HH, TT, HT, TH}
- Event (A) - Exactly one H and one T: {HT, TH}

Random Variables

A random variable (RV) is a real-valued function of the outcome of an experiment.

$$X : \mathbf{\Omega} \to \mathbb{R}$$

Random Variables

Consecutive coin tosses:

- Number of heads

$$X(HH) = 2$$
$$\omega = HH \Rightarrow X = 2$$

- There is at least one head

$$X(HH) = 1$$
$$X(HT) = 1$$
$$X(TH) = 1$$
$$X(TT) = 0$$

Random Vectors

A random vector is a function of the outcome of an experiment that gives a vector of reals.

$$X : \mathbf{\Omega} \to \mathbb{R}^n$$

It can also be thought of as a collection of n real valued functions of the outcome (n random variables).

Random Vectors

If X_1 is a random variable "number of heads" and X_2 is a random variable "number of heads" as described earlier, then $[X_1 X_2]^T$ is a 2-dimensional random vector.

Table of Contents

Probability Measure

If Ω is the sample space and $\mathcal{F}$ is the set of events that can arise during the experiment, then a probability measure $\mathbb{P}$ can be defined as:

$$\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$$

and must satisfy the following properties:

- $\mathbb{P}(\Phi) = 0$, where Φ is the impossible event (an empty set)
- $\mathbb{P}(\Omega) = 1$
- If $A_1, A_2, ...$ is a collection of disjoint events, then

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Table of Contents

Discrete Random Variables

- A random variable whose range is either finite or countably infinite.
- Example - coin toss experiment

$$X(H) = 1$$
$$X(T) = 0$$

Probability Mass Function

- A function that assigns probability to each value in the image of X.
- Example - coin toss experiment

X	0	1
p(X)	0.2	0.8

- Example - number of coin tosses until a Head is seen. Range is $\mathbb{N}$

Bernoulli Random Variable

- Discrete random variable whose PDF can be parameterized by its "bias" θ.

$$p(X = 1) = \theta$$
$$p(X = 0) = 1 - \theta$$

- Sometimes denoted $\mathrm{Ber}(\theta)$

Categorical Random Variable

- Generalizes Bernoulli RV from 2 values to k values.
- Has k parameters

$$p(X = k) = \theta_k$$

X	1	2	...	k
p(X)	0.1	0.2	...	0.1

- Sometimes denoted $\text{Cat}(\theta_1, ..., \theta_k)$

Table of Contents

Cumulative Distribution Function

- A function that is defined as follows:

$$F(x) = \mathbb{P}\{X \leq x\}$$

- A CDF is non-decreasing.

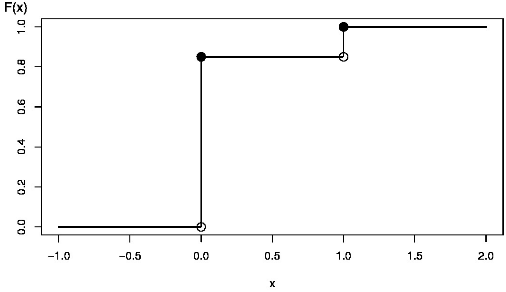


Table of Contents

Continuous Random Variables

- A random variable whose image is uncountably infinite (typically an interval).
- Example - wait time for packets in a router buffer. Range is $[0, \infty)$
- Example - error in measurements of a thermometer

Probability Density Function

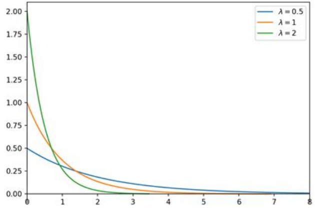
- Wish to assign probability to intervals of values taken by a continuous random variable.
- Exists if the CDF of a continuous RV is differentiable at all points on the real line.

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

Continuous Random Variables

- Wait time in a queue - Exponential RV

$$p(X) = \lambda e^{-\lambda X}$$



- Measurement error in a screw gauge - Gaussian RV

$$p(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

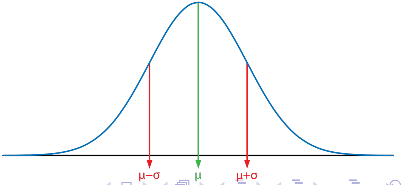
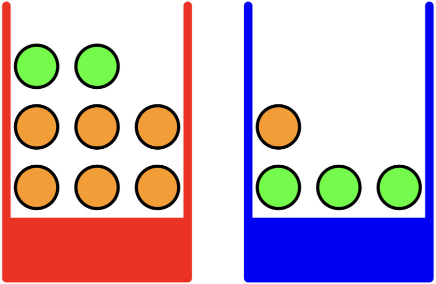


Table of Contents

Conditional Probability

- Two bowls of fruit. The red bowl contains 6 oranges and 2 pears and the blue bowl contains 1 orange and 3 pears.
- Experiment - Pick one bowl. Red is picked 60% of the time. Then pick a fruit from that bowl.



Conditional Probability

- A conditional distribution assigns probability to the values of one random variable given a particular value of another random variable.
- Example - What is the probability that a pear was picked given that we picked the red bowl?
- $p(F = pear|B = red)$

F | B=red	Pear	Orange
p(F | B=red)	2/8	6/8

Joint Probability

- Joint distribution over a pair of random variables assigns probability to every possible pair of values taken by the RVs.
- $p(B = red, F = pear)$

Product Rule

- Can be used to compute joint probability if conditional distributions are known.
- Useful for experiments that are sequential processes, such as fruit bowl experiment.

$$p(A, B) = p(A)p(B|A)$$

Joint Probability

- Joint distribution over a pair of random variables assigns probability to every possible pair of values taken by the RVs.
- $p(B = red, F = pear)$

P(B, F)	Pear	Orange
Red	$1/10$	$3/10$
Blue	$9/20$	$3/20$

Marginal Probability and Sum Rule

- Sum Rule: $p(B) = \Sigma_A P(A, B)$
- Marginal distribution assigns probability to a subset of variables. Can be computed using the sum rule.
- $p(F = pear)$

F	Pear	Orange
p(F)	11/20	9/20

Bayes' Rule

- Product Rule: joint distribution of 2 RVs has two factorizations

$$p(A, B) = p(A)p(B|A) = p(B)p(A|B)$$

- Together with sum rule, this gives us a rule to compute $p(A|B)$ if $p(A)$ and $p(B|A)$

$$\begin{aligned} p(A|B) &= \frac{p(A)p(B|A)}{p(B)} \\ &= \frac{p(A)p(B|A)}{\Sigma_A p(A, B)} \\ &= \frac{p(A)p(B|A)}{\Sigma_A p(A)p(B|A)} \end{aligned}$$

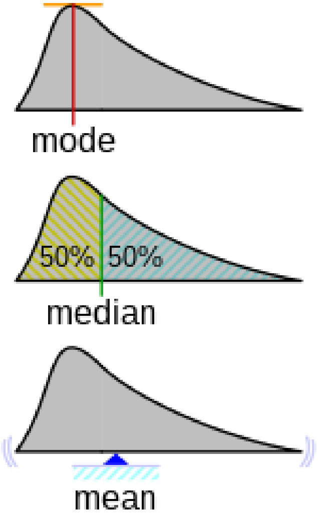
Table of Contents

Mode & Mean

Mode: The most probable value in the support of the distribution.
Mean: Given any RV, X with its PMF or PDF, we find the mean or average as :

$$\mathbb{E}[X] = \sum_{x} x p_X(x)$$

Mean can be thought as a *"centre of mass"* of the distribution.

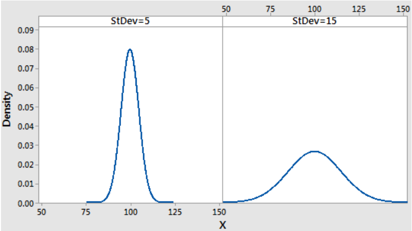


Variance

Variance of a distribution provides us the **deviation or spread** of that RV. It also give us a way to quantify how likely the RV can take values far from Mean.
Variance is calculated as:

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_x (x - \mu_x)^2 p_X(x)$$

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + \mathrm{E}[X]^2\right] \\ &= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + \mathrm{E}[X]^2 \\ &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \end{aligned}$$



Readings

- Introduction to Probability Models (Ross) - Chapters 1-3
- Pattern Recognition and Machine Learning (Bishop) - Chapter 1 (Section 1.2)