

# Machine Learning - Notes

Dom Hutchinson

October 8, 2019

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Motivation . . . . .	2
1.2	Probability Theory . . . . .	2
<b>2</b>	<b>Distributions</b>	<b>5</b>

## General

Lecturer - Carl Henrik Ek

Course Website - <http://carlhenrik.com/COMS30007/>

Course Repo - <https://github.com/carlhenrikek/COMS30007>

Course Subreddit - <https://www.reddit.com/r/coms30007/>

# 1 Introduction

## 1.1 Motivation

### Definition 1.1 - *Deductive Reasoning*

A method of reasoning in which the premises are viewed as supplying all the evidence for the truth of the conclusion.

### Definition 1.2 - *Inductive Reasoning*

A method of reasoning in which the premises are viewed as supplying some evidence for the truth of the conclusion, rather than all the evidence. This allows for the conclusion of the *Inductive Reasoning* to be false.

### Remark 1.1 - *Free-Lunch Theorem*

There are infinite number of hypotheses that perfectly explain the data. Adding a data point removes an infinite number of possibilities, but still leaves infinite possibilities.

### Remark 1.2 - *The Task of Machine Learning*

When proposing to use machine learning on a task, one should consider the following questions:

- i) How can we formulate beliefs and assumptions mathematically?
- ii) How can we connect our assumptions with data?
- iii) How can we update our beliefs?

### Remark 1.3 - *Useful Models are not always True*

Our goal is to understand realisations of a system. If we can then we can equate our model to the system. It is important to note that our model does not need to be perfectly true to be useful.

## 1.2 Probability Theory

### Definition 1.3 - *Stochastic/Random Variable*

A variable whose value depends on outcomes of random phenomena.  
e.g.  $x \sim \mathcal{N}(0, 1)$ .

### Definition 1.4 - *Probability Measure, $\mathbb{P}$*

A function with signature  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ , where  $\mathcal{F}$  is a sample space of rv  $X$ , and fulfils  $\int_{-\infty}^{\infty} \mathbb{P}(x) dx = 1$ .

### Definition 1.5 - *Joint Probability Distribution*

A *Probability Measure* for multiple variables,  $\mathbb{P} : X \times Y \rightarrow [0, 1]$ .

Let  $n_{ij}$  be the number of outcomes where  $X = x_i$  and  $Y = y_j$  then

$$\mathbb{P}(X = x_i, Y = y_j) = \frac{n_{ij}}{\sum_{i,j} n_{ij}}$$

### Definition 1.6 - *Marginal Probability Distribution*

A *Probability Measure* for one variable when the sample space is over multiple variables.

Let  $n_{ij}$  be the number of outcomes where  $X = x_i$  and  $Y = y_j$  then

$$\mathbb{P}(X = x_i) = \frac{\sum_j n_{ij}}{\sum_{i,j} n_{ij}}$$

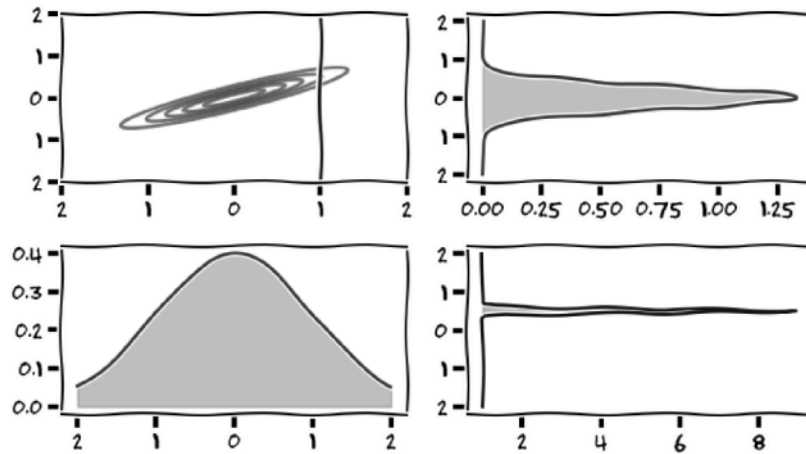
**Definition 1.7 - Conditional Probability Distribution**

A *Probability Measure* for a variable, given another variable has a defined value. Let  $n_{ij}$  be the number of outcomes where  $X = x_i$  and  $Y = y_j$  then

$$\mathbb{P}(Y = y_j | X = x_i) = \frac{n_{ij}}{\sum_j n_{ij}}$$

**Example 1.1 - Joint, Marginal & Conditional Probability**

The below image shows two marginals distributions in the bottom-left,  $X$ , & top-right,  $Y$ , their joint distribution in the top-left and a conditional in the bottom right  $\mathbb{P}(Y|X = 1)$ .

**Theorem 1.1 - Product Rule**

For random variables  $X$  &  $Y$

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(Y = y | X = x) \mathbb{P}(X = x)$$

**Theorem 1.2 - Sum Rule**

For random variables  $X$  &  $Y$

$$\mathbb{P}(X = x) = \sum_j \mathbb{P}(X = x, Y = y_j)$$

**Theorem 1.3 - Bayes' Theorem**

For random variables  $X$  &  $Y$

$$\mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(Y = y | X = x) \mathbb{P}(X = x)}{\mathbb{P}(Y = y)}$$

**Definition 1.8 - Elements of Bayes' Theorem**

The elements of *Bayes' Theory* can be broken down to explain parts of the model.

$$\underbrace{\mathbb{P}(\theta | Y)}_{\text{Posterior}} = \frac{\overbrace{\mathbb{P}(Y | \theta)}^{\text{Likelihood}} \overbrace{\mathbb{P}(\theta)}^{\text{Prior}}}{\underbrace{\mathbb{P}(Y)}_{\text{Evidence}}}$$

Posterior	Which parameters of the model do I believe produce distributions have generated the data $Y$
Likelihood	How likely is the data to come from the model specifically indexed by $\theta$
Prior	What distribution do I think parameter $\theta$ has
Evidence	How likely do I think data $Y$ is for all models.

*N.B.* The *Evidence* normalises this function.

**Definition 1.9 - Conjugate Prior**

If the *Prior* is the same probability distribution family as the *Posterior* then the *Prior* is called a *Conjugate Prior*.

**Definition 1.10 - Expectation Value,  $\mathbb{E}$** 

The mean value a random variable will produce from a large number of samples.

Continuous	Discrete
$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\mathbb{P}(X)dx$	$\mathbb{E}(X) = \sum_{-\infty}^{\infty} x\mathbb{P}(X)dx$
$\mathbb{E}(f(X)) = \int_{-\infty}^{\infty} f(x)\mathbb{P}(X)dx$	$\mathbb{E}(f(X)) = \sum_{-\infty}^{\infty} f(x)\mathbb{P}(X)dx$

**Definition 1.11 - Variance**

Describes the amount of spread in the values a single random variable will produce.

$$\text{var}(X) = \mathbb{E}(x - \mathbb{E}(x))^2 = \mathbb{E}(X^2) - \left(\mathbb{E}(X)\right)^2$$

**Definition 1.12 - Covariance**

Describes the joint variability between two random variables.

$$\text{cov}(X, Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right)$$

**Definition 1.13 - Marginalisation**

The process of summing out the probability of one random variable using its joint probability with another random variable.

$$\begin{array}{ll} \text{Continuous} & \mathbb{P}(X = x) = \int (X = x, Y = y)dy \\ \text{Discrete} & \mathbb{P}(X = x) = \sum_i \mathbb{P}(X = x, Y = y_i) \end{array}$$

**Definition 1.14 - Likelihood Function**

Define  $\mathbf{X} \sim f_n(\cdot; \theta^*)$  for some unknown  $\theta^* \in \Theta$  and let  $\mathbf{x}$  be an observation of  $\mathbf{X}$ .

A *Likelihood Function* is any function,  $L(\cdot; \mathbf{x}) : \Theta \rightarrow [0, \infty)$ , which is proportional to the PMF/PDF of the observed realisation  $\mathbf{x}$ .

$$L(\theta; \mathbf{x}) := C f_b(\mathbf{x}; \theta) \quad \forall C > 0$$

*N.B.* Sometimes this is called the *Observed Likelihood Function* since it is dependent on observed data.

**Definition 1.15 - Log-Likelihood Function**

Let  $\mathbf{X} \sim f_n(\cdot; \theta^*)$  for some unknown  $\theta^* \in \Theta$  and  $\mathbf{x}$  be an observation of  $\mathbf{X}$ .

The *Log-Likelihood Function* is the natural log of a *Likelihood Function*

$$\ell(\theta; \mathbf{x}) := \ln f_n(\mathbf{x}; \theta) + C, \quad C \in \mathbb{R}$$

**Definition 1.16 - Maximum Likelihood Estimation**

The *Maximum Likelihood Estimate* is an estimate for a parameter of a probability distribution which is the value which maximises the *Likelihood Function* (or the *Likelihood Function*).

$$\hat{\theta} := \text{argmax}_{\theta} L(\theta; \mathbf{x})$$

**Definition 1.17 - Central Limit Theorem**

The distribution of the sum (or mean) of a large number of independent, identically distributed random variables can be approximated to a normal distribution, regardless of the distributions of the random variables.

## 2 Distributions

### Definition 2.1 - Bernoulli Distribution

Models an event with a binary outcome (0 or 1) with parameter  $p$  st  $\mathbb{P}(X = 1) = p$ . Let  $X \sim \text{Bernoulli}(p)$ . Then

$$\begin{aligned} f_X(x) &= \begin{cases} p & , x = 1 \\ 1 - p & , x = 0 \\ 0 & \text{otherwise} \end{cases} \\ F_X(x) &= \begin{cases} 0 & , x < 0 \\ 1 - p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases} \\ \mathbb{E}(X) &= p \\ \text{Var}(X) &= p(1 - p) \end{aligned}$$

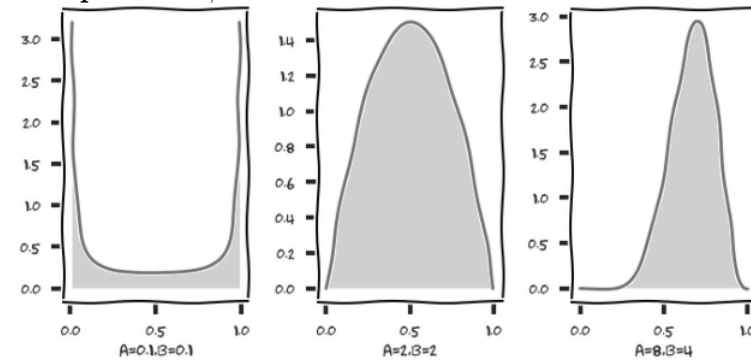
### Definition 2.2 - $\beta$ -Distribution

A  $\beta$ -Distribution is a continuous distribution over interval  $[0, 1]$  which is parameterised by two positive *shape parameters*,  $\alpha$  &  $\beta$ . A  $\beta$ -Distribution can be used to encode assumptions as a *Prior*.

Let  $X \sim \beta(\alpha, \beta)$ . Then

$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

### Example 2.1 - $\beta$ -Distribution



### Definition 2.3 - Dirichlet Distribution

Let  $X \sim \text{Dir}(\alpha)$ . Then

$$f_X(x) := \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \times \dots \times \Gamma(\alpha_N)} \prod_{i=1}^N x_i^{\alpha_i-1}$$

### Definition 2.4 - Exponential Distribution Family

The *Exponential Distribution Family* is a set of probability distributions which fit the form.

$$\mathbb{P}(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta})e^{\boldsymbol{\theta}^T \mathbf{u}(\mathbf{x})}$$

With conjugate prior

$$\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)g(\boldsymbol{\chi})^\nu e^{\nu \boldsymbol{\theta}^T \boldsymbol{\chi}}$$

### Definition 2.5 - Multivariate Normal Distribution

Let  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{N/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

*N.B.* Also known as *Gaussian Distribution*.

**Remark 2.1** - *Conjugate Prior for Normal Distribution*

For a *Normal Distribution* the conjugate prior of  $\boldsymbol{\mu}$  is *Gaussian* and the conjugate prior of  $\boldsymbol{\Sigma}$  is *Inverse-Wishard*.