

Chapter 3

Probability and Information Theory

Table of Contents

1. [Why Probability?](#)
2. [Random Variables](#)
3. [Probability Distributions](#)
 - a. [Discrete Variables and Probability Mass Functions](#)
 - b. [Continuous Variables and Probability Density Functions](#)
4. Marginal Probability
5. Conditional Probability
6. The Chain Rule of Conditional Probabilities
7. Independence and Conditional Independence
8. Expectation, Variance, and Covariance
9. Common Probability Distributions
 - a. Bernoulli Distribution
 - b. Multinomial Distribution
 - c. Gaussian Distribution
 - d. Exponential and Laplace Distributions
 - e. The Dirac Distribution and Empirical Distribution
 - f. Mixture of Distributions
10. Useful Properties of Common Functions
11. Bayes' Rule
12. Technical Details of Continuous Variables
13. Information Theory
14. Structured Probabilistic Models

Important terms:

- Probability
- Information theory
- Degree of Belief
- Frequentist prob
- Bayesian prob
- Random Variable
- [Probability Mass Function](#)
- [Probability Density Function](#)

Introduction

Probability -> means to represent uncertainty

Information Theory -> means to quantify amount of uncertainty

Why Probability?

Unlike other branches of Comp. Sc., Machine Learning normally deals with uncertain and stochastic quantities.

Commented [Fs1]: aka non-deterministic

Three possibilities of uncertainty:

1. **Model stochasticity**, e.g. dynamics of a sub-atomic particle
2. **Incomplete observability**, e.g. [Monty-Hall Problem](#)
3. **Incomplete modelling**, e.g. When a continuous quantity is binned, we lose some information

Why prob? Its more practical to be somewhat uncertain rather than much complex

Commented [Fs2]: Example: "Many birds fly"

Random Variables

Random Variable: a variable that can have different possible values. Random means not able to be predicted.

Types of random variables:

1. Continuous random variable
2. Discrete random variable

Probability Distributions

A random variable can take any possible state, but to quantify which state is it more likely to be in; we must use probability distributions. It can be for

- Discrete variables (described as probability mass function P)
- Continuous variables (described as probability density function p)

Probability Mass Function

$P(x = x)$ where P is PMF over x

Properties of PMF:

- Domain of $P \in \{\text{possible values of } x\}$
- $\forall x \in x, 0 \leq P(x) \leq 1$
- $\sum_{x \in x} P(x) = 1$

Commented [Fs3]: This property is called "normalized"

A special kind of PMF is **joint probability distribution**, which models many variables at the same time. For eg. $P(x = x, y = y)$ denotes the probability that $x = x$ and $y = y$ simultaneously.

Probability Density Function