# Machine learning handbook

Andrija Djurisic

February 2017

## 1 Probability and Information Theory

**Probability theory** is a mathematical framework for representing uncertain statements. It provides a means of quantifying uncertainty and axioms for deriving new uncertain statements. In artificial intelligence applications, we use probability theory in two major ways. First, the laws of probability tell us how AI systems should reason, so we design our algorithms to compute or approximate various expressions derived using probability theory. Second, we can use probability and statistics to theoretically analyze the behavior of proposed AI systems.

While probability theory allows us to make uncertain statements and reason in the presence of uncertainty, **information theory** allows us to quantify the amount of uncertainty in a probability distribution.

### 1.1 Random Variables

A **random variable** is a variable that can take on different values randomly. A random variable is just a description of the states that are possible; it must be coupled with a probability distribution that specifies how likely each of these states are. Random variables may be discrete or continuous.

### 1.2 Probability Distributions

A **probability distribution** is a description of how likely a random variable or set of random variables is to take on each of its possible states. The way we describe probability distributions depends on whether the variables are discrete or continuous.

#### 1.2.1 Discrete Variables and Probability Mass Functions

A probability distribution over discrete variables may be described using a **probability mass function** (PMF). The probability mass function maps from a state of a random variable to the probability of that random variable taking on that state. Probability mass functions can act on many variables at the same time. Such a probability distribution over many variables is known as a joint

probability distribution. $P(X = x, Y = y)$ denotes the probability that $X = x$ and $Y = y$ simultaneously. We may also write $P(x, y)$ for brevity.

To be a probability mass function on a random variable $x$, a function $P$ must satisfy the following properties:

1. The domain of $P$ must be the set of all possible states of $X$.
2. $\forall x \in X, 0 \leq P(x) \leq 1$
3. $\sum_{x \in X} P(x) = 1$

# 2 Machine Learning Basics

## 2.1 Support Vector Machines

We are going to start with logistic regression, and show how we can modify it, and get what is essentially the support vector machine.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)})) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

where $h_\theta(x) = \frac{1}{1-e^{-\theta^T x}}$.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} y^{(i)}(-log(\frac{1}{1 - e^{-\theta^T x^{(i)}}})) + (1 - y^{(i)}))(-log(1 - \frac{1}{1 - e^{-\theta^T x^{(i)}}})) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

Notation standard for SVMs is that we define control parameter for general loss instead of parameter for regularization. We can think of parameter $C = \frac{1}{\lambda}$.

$$J(\theta) = \frac{C}{m}\sum_{i=1}^{m} y^{(i)}(-log(\frac{1}{1 - e^{-\theta^T x^{(i)}}})) + (1 - y^{(i)}))(-log(1 - \frac{1}{1 - e^{-\theta^T x^{(i)}}})) + \frac{1}{2m}\sum_{j=1}^{n}\theta_j^2$$

# 3 Definitions

- A **deterministic system** is a system in which no randomness is involved in the development of future states of the system. A deterministic model will thus always produce the same output from a given starting condition or initial state.
- **Probability** is the measure of the likelihood that an event will occur.
- **Stochastic** or random.