

GENERAL NOTES

ANDREW S. HARD¹

November 15, 2015

CONTENTS

1	Introduction	2
2	Basic Statistics	2
2.1	Deduction and Induction	2
2.2	Conditional Probability and Bayes' Theorem	2
3	Data Structures and Associated Algorithms	4
3.1	Lists	4
3.2	Trees	4
3.3	Graphs	4
4	Numerical Methods	4
5	Machine Learning Introduction	4
6	Feature Selection	4
7	Clustering	5
8	Regression	5
9	Support Vector Machines	5
10	Decision Trees	5
11	Neural Networks	5
11.1	Use cases, overview	5
11.2	MLP: Multi-Layer Perceptrons	5
11.3	RBM: Restricted Boltzmann Machines	5
11.4	CNN: Convolutional Neural Networks	5
11.5	RNN: Recurrent Neural Networks	5
11.6	LSTM: Long Short Term Memory	5
11.7	Neural Turing Machines	5
12	Auto-Encoders and Deep Belief Networks	5
13	Lasso Method	5

LIST OF FIGURES

LIST OF TABLES

¹ Department of Physics, University of Wisconsin, Madison, United States of America

1 INTRODUCTION

As a fifth year graduate student, I have come to the realization that significant portions of my skillset and knowledge base are highly specialized. I intend to pursue a career outside of my field of study (experimental high-energy particle physics). In order to enhance my future job prospects, I decided that it would be useful to review basic concepts in computer science, statistics, mathematics, and machine learning.

2 BASIC STATISTICS

2.1 Deduction and Induction

Deduction: if $A \rightarrow B$ and A is true, then B is true.

Induction: if $A \rightarrow B$ and B is true, then A is more plausible.

2.2 Conditional Probability and Bayes' Theorem

The **conditional probability** of an event A assuming that B has occurred, denoted $P(A|B)$, is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (1)$$

Via multiplication:

$$P(A \cap B) = P(A|B)P(B). \quad (2)$$

This makes sense. The probability of A and B is the probability of A *given* B times the probability of B . Or, just as reasonably, the probability of A and B is the probability of B *given* A times the probability of A . As an aside, the equation above can be generalized:

$$P(A \cap B \cap C) = P(A|B \cap C)P(B \cap C) = P(A|B \cap C)P(B|C)P(C). \quad (3)$$

Back to the derivation, since $P(A \cap B) = P(B \cap A)$, we can use equation 2:

$$P(A|B)P(B) = P(B|A)P(A). \quad (4)$$

And re-arranging gives **Bayes' Theorem**:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)}{P(B)}P(B|A). \quad (5)$$

In this equation, $P(A)$ is the **prior probability**, the initial degree of belief in A , or the probability of A before B is observed. It is essentially the probability that hypothesis A is true before evidence is collected. The **posterior probability** $P(A|B)$ is the degree of belief after taking B into account.

Bayes' theorem describes the probability of an event, based on conditions that might be relevant to the event. With the Bayesian probability interpretation the theorem expresses how a subjective degree of belief should rationally change to account for evidence (**Bayesian inference**). Note: in the **Frequentist Interpretation**, probability measures a proportion of outcomes, not probability of belief.

Bayes' theorem can also be interpreted in the following manner: A is some hypothesis, and B is the evidence for that hypothesis.

The equation makes intuitive sense, particularly in the last form in which it is presented. If $P(B) \gg P(A)$, the probability $P(A|B)$ will be small because B *usually* occurs without A occurring. Even if $P(B|A)$ were 100%, the posterior probability would still be small. On the other hand, if $P(A) \gg P(B)$, $P(A|B)$ is high, even if the conditional probability $P(B|A)$ is smaller.

Nate Silver has a discussion of Bayesian statistics in "The Signal And The Noise", p245. Let x be the initial estimate of the hypothesis likelihood (**prior probability**). In the book, x is the initial estimate that your spouse is cheating. y is a conditional probability - the probability of an observation being made given that the hypothesis is true. So y in his case was the probability of underwear appearing conditional on the spouse cheating. z is the probability of an observation being made given that the hypothesis is false. So z was the probability of underwear appearing if the spouse was not cheating. The **posterior probability** is the revised estimate of the hypothesis likelihood. In Silver's case, the likelihood that the spouse is cheating on you, given that underwear was found. The posterior probability is given by the expression:

$$x' = \frac{xy}{xy + z(1 - x)}. \quad (6)$$

In Silver's formulation, the denominator is equivalent to the total probability of an observation being made (total probability of underwear being found). So it is the probability of the hypothesis being true times the associated conditional probability of the observation of underwear plus the probability of the hypothesis being false times the associated conditional probability of the observation.

Note that this method can be applied recursively by substituting x' for x in the expression.

EVERYTHING BELOW NEEDS REWORKING

There is a recursive form of Bayes' theorem, to account for series of data. WHAT IS IT?

$$S \equiv \bigcup_{i=1}^N A_i, \quad (7)$$

so that A_i is an event in S , and $A_i \cap A_j = \emptyset \forall i \neq j$. Then:

$$A = A \cap S = A \cap \left(\bigcup_{i=1}^N A_i \right) = \bigcup_{i=1}^N (A \cap A_i) \quad (8)$$

Then we can compare the probabilities using the Law of Total Probability:

$$P(A) = P\left(\bigcup_{i=1}^N (A \cap A_i)\right) = \sum_{i=1}^N P(A \cap A_i) = \sum_{i=1}^N P(A_i)P(A|A_i) \quad (9)$$

Using substitution into Bayes' Theorem (equation 5), we have a recursive form of Bayes' Theorem:

$$P(A_j|A) = \frac{P(A_j)P(A|A_j)}{\sum_{i=1}^N P(A_i)P(A|A_i)}. \quad (10)$$

3 DATA STRUCTURES AND ASSOCIATED ALGORITHMS

3.1 Lists

3.2 Trees

3.3 Graphs

4 NUMERICAL METHODS

5 MACHINE LEARNING INTRODUCTION

This section should contain a general description of data science principles, as well as use cases for different algorithms. Many tools for many problems.

Basic problems in machine learning are classification (labeling) and regression (function estimation).

How to do importance ranking of features? Feature importance, global loss function.

6 FEATURE SELECTION

Garbage in gives garbage out. Interactions between features. Interactions are not the same as correlations. Filters versus wrappers.

7 CLUSTERING

8 REGRESSION

9 SUPPORT VECTOR MACHINES

10 DECISION TREES

11 NEURAL NETWORKS

11.1 Use cases, overview

11.2 MLP: Multi-Layer Perceptrons

11.3 RBM: Restricted Boltzmann Machines

11.4 CNN: Convolutional Neural Networks

11.5 RNN: Recurrent Neural Networks

11.6 LSTM: Long Short Term Memory

11.7 Neural Turing Machines

12 AUTO-ENCODERS AND DEEP BELIEF NETWORKS

13 LASSO METHOD