GENERAL NOTES

ANDREW S. HARD[1]

November 16, 2015

CONTENTS

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 occured, denoted $P(A|B)$, is given by:

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

Via multiplication:

$$P(A \cap B) = P(A|B)P(B). \tag{2}$$

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:

$$P(A \cap B \cap C) = P(A|B \cap C)P(B \cap C) = P(A|B \cap C)P(B|C)P(C). \tag{3}$$

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

$$P(A|B)P(B) = P(B|A)P(A). \tag{4}$$

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

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

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 appearaing 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:

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

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. We can also come up with a recurrent form of Bayes theorem as formulated in 5 by letting $P(A_{n+1}) = P(A_n|B)$.

The derivation below is from mathworld's article on Bayes' Theorem. In this example, let A and S be sets. Furthermore, let

$$S \equiv \bigcup_{i=1}^{N} A_i, \tag{7}$$

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

$$A = A \bigcap S = A \bigcap (\bigcup_{i=1}^{N} A_i) = \bigcup_{i=1}^{N} (A \cap A_i) \tag{8}$$

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

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

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

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

What is the meaning of this? The denominator is simply $P(A)$, expressed as the sum of its constituents. So this reduces exactly to Bayes theorem as in 5, with $A \to A_j$ and $B \to A$.

2.3 Correlation and Covariance

Correlation and covariance are similar. Both concepts describe the degree to which two random variables or sets of random variables tend to deviate from their expected values in similar ways.

Let X and Y be two random variables, with means μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively. The **covariance** is defined (using $\langle\rangle$ to denote expectation value):

$$\sigma_{XY} = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle. \tag{11}$$

Similarly, the **correlation** of the two variables is defined:

$$\rho_{XY} = \frac{\langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle}{\sigma_X \sigma_Y}. \tag{12}$$

So correlation is dimensionless, while covariance is the multiple of the units of the two variables. **Variance** is simply the covariance of a variable with itself (σ_{XX} is usually denoted σ_X^2, the square of the standard deviation). The correlation of a variable with itself is always 1, except in the degenerate case where the two variances are zero and the correlation does not exist.

3 DATA STRUCTURES

3.1 Lists

3.2 Trees

3.3 Graphs

4 NUMERICAL METHODS

4.1 Minimization

4.2 Regression

4.3 Matrix Algebra

5 MACHINE LEARNING

5.1 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.

5.2 Feature Selection

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

Explain the use cases of each type of algorithm.