# Machine Learning

Oliver Bonham-Carter

December 19, 2013

## 1 Introduction

Machine learning, originally from artificial intelligence, is technique for deriving information from sets of data. Notable applications of machine learning concern sorting information based on specific kinds of criteria. For instance, a bank may use credit card usage information to determine the risk-level of its customers. A high risk may mean that the customer is unlikely to pay his bills. Evidence of the customer's credit will likely be found by his previous transactions.

Deriving frequency information concerning a specific event is one way to determine likely outcomes of the event in the future. For instance, in weather prediction or stock market analysis, unknown events such as annual rainfall or the price of a stock in light of many market sellers, is often predicted using machine learning methods. These methods work to capture the basic trends behind event which are found in the data relating to past occurrences of the event. When the weather man says that this year could be rainy, he may be making this assessment using last year's elevated rainfall as reference data which has been processed by instruments of machine learning.

## 2 The Number Game

Study the python program given to you by the instructor. The script is a game which asks the user to enter odd and even numbers. The program stops asking for inputs when the user types a zero and then calculates the probability for entered odd or even numbers. This output is the machine learning the trends of entering odd or even numbers. The program uses them to predict the likeliness that the user will enter odd or even numbers.

Sample output

```
$./learning.py start
  Enter a number or zero to exit :1
  Polarity : Odd
  Enter a number or zero to exit :1
  Polarity : Odd
  Enter a number or zero to exit :2
  Polarity : Even
  Enter a number or zero to exit :0
  Polarity : Even

  Results:

   Total values entered:  4
   The general preference for numbers is: odd
   Frequency of picking an Odd number  : 0.666666666667
   Frequency of picking an Even number : 0.333333333333

   The user is likely to enter an odd number
```

## 2.1 Primer Design

Bioinformatics relies on techniques in machine learning to determine the likelihood that a sequence going to melt under certain kinds of temperatures. For instance, in polymerase chain reaction (PCR) tasks, small amounts of the genetic material can now be amplified to be able to a identify, manipulate DNA, detect infectious organisms, including the viruses that cause AIDS, hepatitis, tuberculosis, detect genetic variations, including mutations, in human genes and numerous other tasks. There are three steps of the process: Denaturation, Annealing and Extension. First, the genetic material is denatured, converting the double stranded DNA molecules to single strands. The primers are then annealed to the complementary regions of the single stranded molecules. In the third step, they are extended by the action of the DNA polymerase. All these steps are temperature sensitive and the common choice of temperatures is 94, 60 and 70 (Celsius) respectively. Good primer design is essential for successful reactions.

For the process, a primer is a small sequence which uniquely binds to a specific region of DNA to be amplified and must be determined to be able to withstand the heat that is applied to the DNA to break the bonds connecting one strand to its complement to separate them. Two primers are added (the forward and the reverse) which must remoain connected to their specific regions of DNA. To ensure that they remain bonded to their regions of DNA, the primer melting temperatures must be higher than that of the rest of the DNA. This is done by determining that there are more G-C bonds in the primers than A-T bonds since G-C bonds require more energy to break than A-T bonds.

## 2.2 Experiment!

Write a script program to determine the frequency of G-C and A-T bonds in an inputted sequence. Compare these bond frequencies to determine whether your sequence is likely to be a strong primer which requires more energy than a primer largely made up of A-T bonds. Try changing the sequence to have mainly A-T bonds. What happens to the G-C frequency?