

Tutorial 7:

Machine learning basics

Name	Student no.

For this tutorial, you can discuss the questions in small groups (up to 4 students). Individually submit the answers to each of the 3 Questions.

Question 1

Have you ever wondered if you could win money by recording all LOTTO draws? For simplicity assume that there are only 2 numbered balls and you have to guess which one will roll out first. Let's say that you watched all televised draws during 20 weeks and recorded the outcomes:

Ball B1: 3 wins

Ball B2: 17 wins

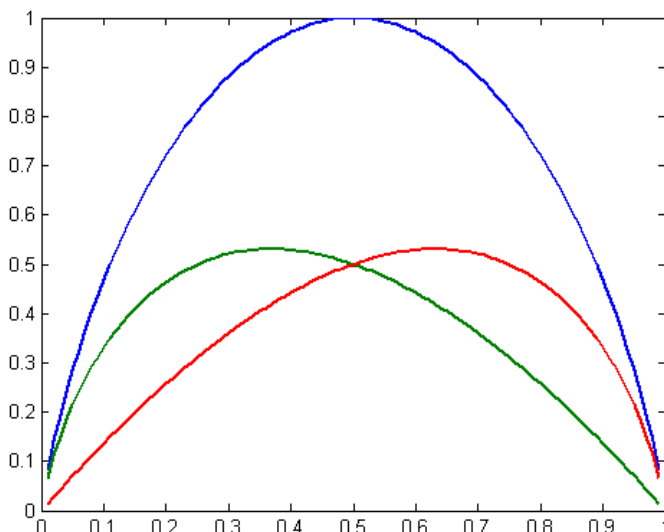
- a) Which ball would it be best to bet on in week 21?

Clearly, there is something suspicious here (and you are unlikely to find something like this in real lotto games). Having watched those 20 weeks of LOTTO draws we have actually learned some useful information. The outcomes do not seem to be uniformly random - there appears to be a pattern to the probabilities. Estimate the information content/entropy of the data.

$$I(P(v_1), \dots, P(v_m)) = \sum_{i=1}^m -P(v_i) \log_2 P(v_i)$$

$$I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

p is the number of positives (say, number of B1 wins) and n is the number of negatives (B2 wins), m is the number of possible outcomes, v_i is the i th type of outcome or result possible - here B1 or B2. If you don't have \log_2 on your calculator, note that $\log_2(x) = \log_A(x) / \log_A(2)$ for any base $A > 0$, $A \neq 1$. And $\log_{10}(2) = 0.30103$. If you don't have a calculator, use the following graph (applicable for 2-class data only).



Probability of event (horizontal axis) vs Information content per example (vertical axis)

The blue, upper line is:

$$f_1(P) = f_2(P) + f_3(P) = -P \log_2(P) - (1-P) \log_2(1-P).$$

The green and red lines correspond to

$$f_2(P) = -P \log_2(P) \text{ and}$$

$$f_3(P) = -(1-P) \log_2(1-P), \text{ respectively.}$$

- b) Calculate the information content / entropy of the data.

- c) Verify that the data for B1 and B2 can be swapped without changing the information / entropy content.

- d) For comparison, calculate the information content if we had instead seen: B1: 10 wins, B2: 10 wins

Question 2

Let's now say that you also recorded (1) the weather {Sunny, Rain and Cloudy} and (2) the colour of the tie {Black, Red, Green} of the person who pulls the handle.

Think about what a data set could look like if we were to support that the weather is more *important* than the colour of the tie for predicting the outcome. In other words, if you knew the weather, you would know the outcome. If you knew the colour of the tie, you would have to guess the outcome. To fill in the weather and tie values in the table is optional. You only need to submit an explanation of how you reason to come up with the values.

Draw	Weather	Tie	Outcome
1			B2
2			B2
3			B2
4			B2
5			B1
6			B2
7			B2
8			B2
9			B2
10			B2
11			B2
12			B2
13			B1
14			B2
15			B2
16			B1
17			B2
18			B2
19			B2
20			B2

Question 3

Assume you have 100 examples of handwritten digits (D001-D100). You will evaluate 10 different configurations of machine learning models (M01-M10) to find a preferred configuration (a model that successfully recognises digits). However, the machine learning model will be tested live *after* you have submitted your preferred configuration.

- a) Which of the following strategies would you use to train and select a model?
 1. Train all models/configurations M01-M10 on D001-D100 and select the one which performs best on average over D001-D100.
 2. Train all models/configurations M01-M10 on D001-D050 and select the one which performs best on average over D051-D100.
 3. Train all models/configurations M01-M10 on D001-D099 and select the one which performs best on D100 only.

- b) Discuss why you have suggested this strategy. Suggest an alternative strategy if necessary.