

## Tutorial 8:

### Machine learning basics

#### Question 1

Ball B2 looks like it is far more likely to roll out first, so if betting, it would be the one to pick. We should not forget that we have only a limited amount of data (20 observations) and even if B1 and B2 were 50/50 chances, it is possible we could see these sorts of proportions (but very unlikely).

Given the data, our estimates of the probabilities of B1 and B2 are  $3/20 = 0.15$  and  $17/20 = 0.85$ , respectively.

(in the following  $\log_2$  means log to base 2).

Estimated information content/entropy of the data, ie: average information per new lotto draw:

$$\begin{aligned} & I(P(B1),P(B2)) \\ &= I(0.15,0.85) \\ &= -0.15 \log_2(0.15) - 0.85 \log_2(0.85) \\ &\approx 0.61 \text{ bits/draw} \end{aligned}$$

Swapping B1 and B2 data: fairly clear that  $I(P(B1),P(B2)) = I(P(B2),P(B1))$  since both contain the same terms in the sum, so this won't change the estimated average information content.

If we had instead seen B1:10 wins, B2: 10 wins:

$$\begin{aligned} & I(P(B1),P(B2)) \\ &= I(0.5,0.5) \\ &= -0.5 \log_2(0.5) - 0.5 \log_2(0.5) \\ &= 1 \text{ bit/draw} \end{aligned}$$

[Note that  $\log_{10}(2)$  is actually 0.30103, not 0.693147 as stated on the tutorial sheet]

#### Question 2

If weather helps in predicting the data -> more often one type of weather is associated with one type of out come.

For example, B1 may be more common on days that are sunny while B2 may be more common on days that are rainy or cloudy.

If tie colour does not help in predicting the data -> approximately equal distribution of tie colour and results.

#### Question 3

There is no "right" answer for question 3, as too much depends on what you know about your machine learning models and how they operate, and the characteristics of

your data. Of the three options, strategy 2 provides the best balance between training and testing data sets.

Another option is leave-one-out cross-validation, in which we fit our model 100 times to 99 data points and test on the left out point, each time leaving out a different point. So this is an extension of option 3. The final model is then usually fit to all 100 data points, so the cross-validated estimate of e.g. error rate is then slightly conservative (all those cases were based on only 99 data points). One can instead use all 100 models and take the majority vote as the prediction, but this is typically no better than using the one model based on 100 data points.

Cross-validation is fairly accurate and efficient in its use of the data. A number of useful variations exist.