**Harry Langford**
hjel2@cam.ac.uk

# MLRWD Session notes

You want to find the **features** which makes the thing the most accurate.

$$\hat{c} = \text{argmax}_{c \in C} \log P(c) + \sum_i \log P(f_i|c)$$

Choose the most appropriate $f_i$! This may either be individual objects or the presence or absence of a type.

Often considering a feature to be the presence or absence of a type.

Mention testing things empirically.

Consider the presence or absence of types of trees. It's often more appropriate to measure absence when there is a small number of types of which most are almost always present.

## 1 Cross-Fold Validation

Cross-fold-validation is a way of enhancing the training data you do have by "re-using" it.

There are two types of cross-fold validation: stratified and sampled.

Stratified is when you make sure the proportion of each class in each fold is approximately right.

We should always use stratified cross validation. Random cross-fold validation ends up being pretty poor.

## 2 Testing

The testing process involves evaluation! Make sure to mention it in questions discussing testing.

## 3 Read the question!

2018P3Q7e:

Build a lexicon classifier and use a combination of this classifier and the naïve bayes classifier to predict the actual result.

Use semi-supervised learning with the sentiment classifier to label data for the naïve bayes classifier.

Smoothing parameter. If I don't see a tree species with a class in my data and I also don't see it in the dictionary then I assume it will never appear (this stops shifting probability mass and gives better parameters).

Something to do with "tuning some parameters" – doing something in terms of development.

The only parameter to smooth is the smoothing parameter.

Significance tests! Sign test! Relearn the formula!

Justify which you should use – one tailed or two tailed!

## 4 Parameters

Parameter estimation is **going to be simple** – absolutely do not overcomplicate things!