Introductory Applied Machine Learning, Tutorial Number 2

School of Informatics, University of Edinburgh, Instructors: Chris Williams, Oisin Mac Aodha, Hiroshi Shimodaira

September 2020

1. Suppose a hypothetical UK railservice from Edinburgh to Oldfort is often subject to delays. The train service is run by three different train operating companies (TOC). Over the course of a year, a random sample of the services was taken. The following data was obtained

	Weather	Season	TOC	Day	Lateness
Case 1	Windy	Summer	RotRail	Weekday	On time
Case 2	Windy	Winter	GNAF	Weekday	Delayed
Case 3	Windy	Autumn	GNAF	Weekday	Delayed
Case 4	Calm	Summer	Virgo	Weekend	Delayed
Case 5	Windy	Winter	RotRail	Weekend	Delayed
Case 6	Calm	Summer	Virgo	Weekday	Delayed
Case 7	Calm	Spring	RotRail	Weekday	On time
Case 8	Windy	Autumn	GNAF	Weekend	Delayed
Case 9	Calm	Winter	Virgo	Weekend	Delayed
Case 10	Calm	Spring	Virgo	Weekday	Delayed
Case 11	Windy	Autumn	GNAF	Weekday	Delayed
Case 12	Windy	Spring	GNAF	Weekday	On time
Case 13	Windy	Summer	RotRail	Weekday	On time
Case 14	Calm	Autumn	RotRail	Weekday	On time
Case 15	Windy	Winter	RotRail	Weekday	Delayed
Case 16	Calm	Autumn	Virgo	Weekday	Delayed
Case 17	Windy	Summer	Virgo	Weekday	Delayed
Case 18	Windy	Spring	Virgo	Weekend	Delayed
Case 19	Calm	Winter	GNAF	Weekday	On time
Case 20	Calm	Spring	GNAF	Weekend	On time

Find the root (top) node selected using the maximum information gain tree building procedure to classify whether a train will be delayed or on time. Show that it selects according to which TOC is providing the service.

You might find the following table a helpful starter

	Delayed	On time
Calm	5	4
Windy	8	3
Summer	3	2
Winter	4	1
Autumn	4	1
Spring	2	3

	Delayed	On time
RotRail	2	4
GNAF	4	3
Virgo	7	0
Weekday	8	6
Weekend	5	1

The maximum information gain tree building procedure creates the following first two layers of the tree. Suppose the whole tree were pruned to this level (2 layers). Find the final decision tree by filling in the missing classification values and missing classification ratios below

```
                             TOC
              |---------------^-------------|-----------------|
            RotRail                        GNAF              Virgo
              |                             |                 |
            Season                        Season           Delayed
   |------|---^--|------|        |------|---^--|------|
Summer Winter Autumn Spring    Summer Winter Autumn Spring
  |       |      |      |        |      |      |      |
On time Delay                    ?      ?    Delay
        (2:0)  (0:1)  (0:1)    (0:0)  (1:1)
```

2. Using your decision tree from question 1, how would you classify

	Weather	Season	TOC	Day	Lateness
Example 1	Windy	Autumn	RotRail	Weekday	on time
Example 2	Calm	Summer	Virgo	Weekday	Delayed
Example 3	Calm	Spring	GNAF	Weekend	On Time

3. A training set consists of one dimensional examples from two classes. The training examples from class 1 are $\{0.5, 0.1, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.35, 0.25\}$ and from class 2 are $\{0.9, 0.8, 0.75, 1.0\}$. Fit a (one dimensional) Gaussian using Maximum Likelihood to each of these two classes. You can assume that the variance for class 1 is 0.0149, and the variance for class 2 is 0.0092. Also estimate the class probabilities p_1 and p_2 using Maximum Likelihood. What is the probability that the test point $x = 0.6$ belongs to class 1?

Thinking:
find the mean and variance, find μ and σ, using the probability density function and then bring in with the Bayesian formula?

4. Two students are working on a machine-learning approach to spam detection. Each student has their own set of 100 labeled emails, 90% of which are used for training and 10% for validating the model. Student A runs the Naive Bayes algorithm and reports 80% accuracy on his validation set. Student B experiments with over 100 different learning algorithms, training each one on his training set, and recording the accuracy on the validation set. His best formulation achieves 90% accuracy. Whose algorithm would you pick for protecting a corporate network from spam? Why?

I think the A may better with no parameters required.
And B is not desirable, B may exist that over fitting, only picks models on the validation set, with randomness, but does not prove 90% on the test dataset, so I think B is not necessarily better than A.