Introductory Applied Machine Learning Tutorial 2 Answers

September 2017

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. 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

Let s denote each possible class (late, on time). Let M be the classification based on all the

data. Let M_i be the classification based on just looking at the data corresponding to value i of some attribute A (eg A=weather, i=calm). Then the information gain is given by

$$\begin{aligned} Gain(M,A) &= Ent(M) - \sum_{i\in A}\frac{|M_i|}{|M|}Ent(M_i) \\ &= Ent(M) + \sum_{i\in A}\frac{|M_i|}{|M|}\sum_s\frac{|M_i^s|}{|M_i|}\log\frac{|M_i^s|}{|M_i|} \\ &= Ent(M) + \sum_{i\in A}\frac{1}{|M|}\left[\left(\sum_s |M_i^s|\log|M_i^s|\right) - |M_i|\log|M_i|\right] \end{aligned}$$

Because $Ent(M)$ and $|M|$ are fixed, maximising the information gain is equivalent to maximising

$$\sum_{i\in A}\left(\sum_s |M_i^s|\log|M_i^s|\right) - |M_i|\log|M_i|$$

Calculating this for each attribute

Weather

$$5\log 5 + 4\log 4 - 9\log 9 + 8\log 8 + 3\log 3 - 11\log 11 = -12.63\ nats = -18.22\ bits$$

Season

$$\begin{gathered} 3\log 3 + 2\log 2 - 5\log 5 + 4\log 4 + 1\log 1 - 5\log 5 + 4\log 4 + \\ 1\log 1 - 5\log 5 + 2\log 2 + 3\log 3 - 5\log 5 = -11.73\ nats = -16.92\ bits \end{gathered}$$

TOC

$$2\log 2 + 4\log 4 - 6\log 6 + 4\log 4 + 3\log 3 - 7\log 7 + 7\log 7 + 0\log 0 - 7\log 7 = -8.60\ nats = -12.40\ bits$$

Day

$$8\log 8 + 6\log 6 - 14\log 14 + 5\log 5 + 1\log 1 - 6\log 6 = -12.26\ nats = -17.69\ bits$$

(To get information gain, divide these values by $|M| = 20$ and add the entropy of the split of the whole data set on the *Delayed* and *On time* categories ($13/20$ and $7/20$ resp) $= 0.65\ nats = 0.93\ bits$. This gives the list $(0.0159\ 0.0609\ 0.2174\ 0.0344)$ nats or $(0.0229\ 0.0879\ 0.3136\ 0.0496)$ bits).

The largest information gain comes from choosing to classify according to the train operating company (TOC).

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
   |------|---^--|------|       |------|---^--|------|   (7:0)
Summer Winter Autumn Spring    Summer Winter Autumn Spring
   |      |      |      |        |      |      |      |
On time Delay  Ontime Ontime     ?      ?    Delay  Ontime
(0:2)   (2:0)  (0:1)  (0:1)    (0:0)  (1:1)  (3:0)  (0:2)
```

The remaining values can be obtained simply from adding up the number of cases in each class with the relevant attributes.

2. Using your decision tree in 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

To get this we just check each attribute of the tree in the turn, following the relevant branches of the tree, and outputting the final classification given at the leaf of the tree.

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?*

The maximum likelihood estimator for the mean of each Gaussian is given by $\frac{\sum_i x_i}{n}$:

$$\begin{aligned}\hat{\mu}_1 &= 0.26\,\text{(add up the 10 numbers and divide by 10)},\\ \hat{\mu}_2 &= 0.8625\,\text{(add up the 4 numbers and divide by 4)},\end{aligned}$$

with variances as in the question:

$$\begin{aligned}\hat{\sigma}_1^2 &= 0.0149,\\ \hat{\sigma}_2^2 &= 0.0092\,.\end{aligned}$$

Class probabilities are:

$$\begin{aligned}\hat{p}_1 &= \frac{10}{14} = 0.7143,\\ \hat{p}_2 &= 1 - \hat{p}_1 = \frac{4}{14} = 0.2857\,.\end{aligned}$$

Now, the probability that a point x belongs to class 1 is given by:

$$p(c_1 \,|\, x) = \frac{p_1 p(x \,|\, c_1)}{p_1 p(x \,|\, c_1) + p_2 p(x \,|\, c_2)}\,,$$

where,

$$p(x \,|\, c_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{1}{2}\frac{(x-\mu_k)^2}{\sigma_k^2}\right]\,.$$

Crunching the numbers we obtain $p(c_1 \,|\, x = 0.6) = 0.6305$. Note that $\hat{\mu}_2$ is nearer to $x = 0.6$ than $\hat{\mu}_1 = 0.26$, but that $\hat{\sigma}_1^2 = 0.0149$ is broader than $\hat{\sigma}_2^2 = 0.0092$.

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?*

The question is ill-posed: each student has **their own** set of emails. Accuracy figures cannot be directly compared against two different testing sets. If we assume the two students are using the same dataset, and the same training / validation split, the following argument could be made.

There are only 100 labelled emails, so the validation set has only 10 emails in it. It is true that the *expected* performance on the validation set equals the generalisation error, but note that with such a small validation set one would expect quite large variance relative to the true generalisation error.

When student B selects from 100 learning algorithms, he will very likely select on the basis of idiosyncracies in the validation set rather than the true generalisation error.