March 23, 2018

Homework 3

Sanchit Singhal

CE 395R – Data Mining with Carlos Caldas

Spring 2018

University of Texas at Austin

1. Importance of Data Exploration in Data Mining

Data exploration can be described as a preliminary examination of the data to better understand its characteristics. It is an important tool for data mining and there are several motivations for it. It aides in the identification of the best approach by helping to select the right tool for both preprocessing – which can sometimes be the most time-consuming task – as well, as the analysis. Apart from gaining efficiency in time, data exploration can also increase the effectiveness of the findings by helping to choose analytical methods that best suit the data and how it relates to the problem at hand. Further, it can also make use of human's ability to inherently recognize patterns that would otherwise be very difficult to capture through data analysis tools.

2. Neural Networks

Advantages:

- Can provide computational efficiency through massive parallelism
- Model testing can be very fast
- Are universal approximators that can be used for a multitude of problems
- Can handle redundant attributes because weights are automatically learnt

Disadvantages:

- Model building is often very time consuming
- Can be sensitive to noise in training data
- Difficult to handle missing attributes
- Can suffer from overfitting if network is too large

3. Accuracy as a Performance Evaluator for a classifier

Accuracy is the most widely-used metric that tells us about the proportion of correct classifications to the total number of cases. However, a limitation of Accuracy as a metric for performance evaluation of a classifier can be that it is misleading sometimes. For example, when there are a large number of samples, n, and the model predicts everything to be a in a certain class then the accuracy score can still seem really high (99% and above) even though there are examples that it got wrong simply because the model did not detect any of the other class's examples.

④

Training Dataset

Example No	Color	Type	Origin	Stolen?	
1	Red	Sports	Domestic	Yes	C_1
2	Red	Sports	Domestic	No	C_2
3	Red	Sports	Domestic	Yes	C_1
4	Yellow	Sports	Domestic	No	C_2
5	Yellow	Sports	Imported	Yes	C_1
6	Yellow	SUV	Imported	No	C_2
7	Yellow	SUV	Imported	Yes	C_1
8	Yellow	SUV	Domestic	No	C_2
9	Red	SUV	Imported	No	C_2
10	Red	Sports	Imported	Yes	C_1

Test example

Example No	Color	Type	Origin	Stolen?
11	Red	SUV	Domestic	?

Naïve Bayes Approach: $P(C_i \mid A_n) = P(A_n \mid C_i) \cdot P(C_i)$

$\Rightarrow P(A_1 \mid C_i) \cdot P(A_2 \mid C_i) \cdot P(A_3 \mid C_i) \cdot P(C_i)$

If C_1 = Yes & C_2 = No,

- $P(C_1 \mid \text{Color} = \text{Red}, \text{Type} = \text{SUV}, \text{Origin} = \text{Domestic})$
  $= P(\text{Color} = \text{Red} \mid C_1) \cdot P(\text{Type} = \text{SUV} \mid C_1) \cdot P(\text{Origin} = \text{Domestic} \mid C_1) \cdot P(C_1)$

- $P(C_2 \mid \text{Color} = \text{Red}, \text{Type} = \text{SUV}, \text{Origin} = \text{Domestic})$
  $= P(\text{Color} = \text{Red} \mid C_2) \cdot P(\text{Type} = \text{SUV} \mid C_2) \cdot P(\text{Origin} = \text{Domestic} \mid C_2) \cdot P(C_2)$

Calculated Probabilities

- $P(C_1) = 5/10 = 0.5$
- $P(\text{Color=Red} \mid C_1) = 3/5 = 0.6$
- $P(\text{Type=SUV} \mid C_1) = 1/5 = 0.2$
- $P(\text{Origin=Domestic} \mid C_1) = 2/5 = 0.4$

- $P(C_2) = 5/10 = 0.5$
- $P(\text{Color=Red} \mid C_2) = 2/5 = 0.4$
- $P(\text{Type=SUV} \mid C_2) = 3/5 = 0.6$
- $P(\text{Origin=Domestic} \mid C_2) = 3/5 = 0.6$

$\Rightarrow P(C_1 \mid \text{Color=Red, Type=SUV, Origin=Domestic}) = P(\text{Color=Red} \mid C_1) \cdot P(\text{Type=SUV} \mid C_1) \cdot P(\text{Origin=Domestic} \mid C_1) \cdot P(C_1)$

$= 0.6 \times 0.2 \times 0.4 \times 0.5$

$= 0.024$

$\Rightarrow P(C_2 \mid \text{Color=Red, Type=SUV, Origin=Domestic}) = P(\text{Color=Red} \mid C_2) \cdot P(\text{Type=SUV} \mid C_2) \cdot P(\text{Origin=Domestic} \mid C_2) \cdot P(C_2)$

$= 0.4 \times 0.6 \times 0.6 \times 0.5$

$= 0.072$

Since $0.072 > 0.024$, the probability of the test example being in C_2 is greater than C_1 & therefore should be classified as Stolen = No.

$\therefore$

Example No	Color	Type	Origin	Stolen?
11	Red	SUV	Domestic	NO

(5) Decision Tree:

measure of node impurity = Entropy $= -\sum_{i=1}^{N} P_i \log(P_i)$

Total Entropy

C_1 : Stolen = Yes = 5/10 = 0.5

C_2 : Stolen = No = 5/10 = 0.5

$$\text{entropy} = -\left[\frac{5}{10}\log_2(5/10) + 5/10\log_2(5/10)\right] = \underline{1}$$

Split by Color

- Color = Red : $C_1 = 3$, $C_2 = 2$ [5]

$$\text{entropy} = -\left[3/5\log_2(3/5) + 2/5\log_2(2/5)\right] = \underline{0.971}$$

- Color = Yellow : $C_1 = 2$, $C_2 = 3$ [5]

$$\text{entropy} = -\left[2/5\log_2(2/5) + 3/5\log_2(3/5)\right] = \underline{0.971}$$

$\Rightarrow$ weighted average $= 5/10\,(0.971) + 5/10\,(0.971) = \underline{0.971}$

$\Rightarrow$ Information Gain $= 1 - 0.971 = \boxed{0.029}$

Split by Type

- Type = SpoAS : $C_1 = 4$, $C_2 = 2$ [6]

$$\text{entropy} = -\left[4/6\log_2(4/6) + 2/6\log_2(2/6)\right] = \underline{0.918}$$

- Type = SUV : $C_1 = 1$, $C_2 = 3$ [4]

$$\text{entropy} = -\left[1/4\log_2(1/4) + 3/4\log_2(3/4)\right] = \underline{0.811}$$

$\Rightarrow$ weighted average $= 6/10\,(0.918) + 4/10\,(0.811) = \underline{0.875}$

$\Rightarrow$ Information Gain $= 1 - 0.875 = \boxed{0.125}$

Split by Origin

- Origin = Domestic: $C_1 = 2, C_2 = 3$ [5]

  entropy = $-[2/5 \log_2(2/5) + 3/5 \log_2(3/5)] = \underline{0.971}$

- Origin = Imported: $C_1 = 3, C_2 = 2$ [5]

  entropy = $-[3/5 \log_2(3/5) + 2/5 \log_2(2/5)] = \underline{0.971}$

⇒ weighted average = $5/10\ (0.971) + 5/10\ (0.971) = \underline{0.971}$

⇒ Information Gain = $1 - 0.971 = \boxed{0.029}$

First Splitting Attribute

Information Gains:

- Color = 0.029
- Type = 0.125
- Origin = 0.029

Since splitting by feature Type gives the highest information gain, it should be the root node.