Computer Science 477

More on Entropy

Lecture 11

Twenty Questions

- Game in which possible values are determined by asking a series of yes/no questions.
- Answers are mutually exclusive (such as mutually exclusive classes)
- Assume that all M values are equally likely and that M is an exact power of 2, say 2^N , where $N \geq 1$.
- Example: an unknown capital city from the eight possibilities: London, Paris, Berlin, Warsaw, Sofia, Rome, Athens and Moscow (here $M = 8 = 2^3$). There are many possible ways of asking quest
- Random guessing:
  - Is it Warsaw? No
  - Is it Berlin? No
- With luck, might guess correctly

Strategies for Guessing

- Make our guesses in a fixed order: London, Paris, Berlin etc. until we guess the correct answer
- Never guess further than Athens, as a ‘no’ answer will tell us the city must be Moscow
  - If the city is London, we need 1 question to find it.
  - If the city is Paris, we need 2 questions to find it.
  - $\vdots$
  - If the city is Rome, we need 6 questions to find it.
  - the city is Athens, we need 7 questions to find it.
  - the city is Moscow, we need 7 questions to find it.
- Average number of guesses needed: $(1+2+3+4+5+6+7+7)/8 = 35/8 = 4.375$

Best Strategy

- Best strategy is to keep dividing the possibilities into equal halves.
- Thus:
  - Is it London, Paris, Athens or Moscow? No
  - Is it Berlin or Warsaw? Yes
  - Is it Berlin?
- To determine which 8 cities, never need more than 3 questions.
- If we start with 8 possibilities and halve the number by the first question, that leaves 4 possibilities.
- $8 = 2^3$
- The smallest number of yes/no questions needed to determine an unknown value from $M = 2^N$ equally likely possibilities is N.

Generalizing

- There are M^k possible sequences of k values.
- If M is not a power of 2, the number of questions needed, V_{kM} is the next integer above $\log_2 M^k$.
- Lower and upper bounds on the value of V_{kM} by the relation

$$\log_2 M^k \leq V_{kM} \leq \log_2 M^k + 1$$

- which leads to

$$k \cdot \log_2 M \leq V_{kM} \leq k \cdot \log_2 M + 1$$

- and so

$$\log_2 M/k \leq V_{kM}/k \leq \log_2 M/k + 1/k$$

Encoding with Bits

- Units of information can also be looked at as the amount of information that can be coded using only a zero or a one.
- If we have two possible values, say male and female
  - 0=male
  - 1=female
- Four values: man, woman, dog, cat
  - 00 = man
  - 01 = woman
  - 10 = dog
  - 11 = cat

Encoding Eight Values

- Eight values, say the eight capital cities, we need to use three bits:
  - 000 = London
  - 001 = Paris
  - 010 = Berlin
  - 011 = Warsaw
  - 100 = Sofia
  - 101 = Rome
  - 110 = Athens
  - 111 = Moscow
- Three questions required to discern which city

M Not a Power of 2

- Consider not just one value out of M possibilities
- Sequence (permutation & combination) of k such values (each one chosen independently of the others).
- Denote the smallest number of yes/no questions needed to determine a sequence (permutation) of k unknown values drawn independently from M possibilities
  - The entropy, by V_{kM}
- Identical to the number of questions needed to discriminate amongst M_k distinct possibilities

Example

- Identify a sequence of six days of the week, for example {Tuesday, Thursday, Tuesday, Monday, Sunday, Tuesday}.
- M is 7 and k is 6.
- Possible question might be

  Is the first day Monday, Tuesday or Wednesday and the second day Thursday and the third day Monday, Saturday, Tuesday or Thursday and the fourth day Tuesday, Wednesday or Friday and the fifth day Saturday or Monday and the sixth day Monday, Sunday or Thursday?

Example, Continued

- There are $7^6 = 117649$ possible sequences of six days.
- The value of $\log_2 117649$ is 16.84413.
- Between 16 and 17
- To determine which possible sequence of 6 days of the week would take 17 questions.
- Average number of questions for each of the six days of the week is $17/6 \ = \ 2.8333$.
- Close to $\log_2 7$ (approximately 2.8074)

Improved Approximation

- Choose a larger value of k, say $21.$
- Now $\log_2 M^k$ is $\log_2 7^{21} = 58.95445$, so 59 questions are needed for the set of 21 values
- Average number of questions per value of $59/21 = 2.809524$ - very close to $\log_2 7$
- Let k = 1000
- $\log_2 M^k = \log_2 7^{1000} =$ 2807.3549
- So 2808 questions are needed determine 1000 values, making an average per value of 2.808, which is very close to $\log_2 7$
- For M^k possible sequences of k values and M is not a power of 2, the number of questions needed, V_{km} is the ceiling of $\log_2 M^k$

Encoding Values That Are Not Equally Likely

- M possible values are equally likely the entropy has previously been shown to be $\log_2 M$
- Frequency with which the ith of the M values occurs as p_i where i varies from 1 to M.
- Then we have $0 \leq p_i \leq 1$ for all p_i and $\sum_{i=1}^{i=m} p_i = 1$

Example - Values That Are Not Equally Likely

- Suppose that p_i the reciprocal of powers of 2
  - $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$
- Four values, A, B, C and D with frequencies $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$ (so $M = 4$)
- Standard 2-bit encoding:
  - A 10
  - B 11
  - C 00
  - D 01

A	1
B	01
C	001
D	000

- Improve by choosing variable length encoding.

Variable Length Encoding

- To determine a value of A, need to examine only one bit.
- For B, examine two bits
- For C or D, examine three bits
- Any other representation, requires more bits to be examined
- Average: $1/2 \times 2 + 1/4 \times 1 + 1/8 \times 3 + 1/8 \times 3 = 2$
  - Same as 2-bit representation
- Average: s $1/2 \times 3 + 1/4 \times 4 + 1/8 \times 5 + 1/8 \times 6 = 3.875$

A 01
B 1
C 001
D 000

A 101
B 0011
C 10011
D 100001

When Frequencies are not $\frac{1}{2^k}$

- M values with frequencies $p_1, p_2, \ldots, p_M$
  - Average number of bit that need to be examined
  - I.e., the entropy
- "frequency of occurrence of the ith value multiplied by the number of bit that need to be examined if that value is the one to be determined, summed over all values of i from 1 to M"

$$E = \sum_{i=1}^{M} p_i \log_2(1/p_i)$$

- Or

$$E = -\sum_{i=1}^{M} p_i \log_2(p_i)$$

Entropy of a Training Set

- Knowing that the entropy of a training set is E, doesn’t mean that we can find an unknown classification by E well-chosen yes/no questions
- Instead, ask a series of questions about the value of a set of attributes
- Asking about the value of an attribute, splits the training set
- Connection: determine which attribute gets us to a classification
  - Diminishes the uncertainty about what the classification is the most likely

Information Gain can be Zero

- Consider:

X	Y	Class
1	1	A
1	2	B
2	1	A
2	2	B
3	2	A
3	1	B
4	2	A
4	1	B

- Entropy is $E_{start} = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = -\log_2(\frac{1}{2}) = \log_2(2) = 1$

Information Gain can be Zero

- Splitting on attribute X gives this frequency table:

Class	Attribute value			
	1	2	3	4
A	1	1	1	1
B	1	1	1	1
Total	2	2	2	2

- $E_{new} = 1$
- Splitting on attribute Y results in this frequency table:

Class	Attribute value	
	1	2
A	2	2
B	2	2
Total	4	4

- Both result in an information gain of $E_{new} - E_{new} = 1 - 1 = 0$

Information Gain for Feature Reduction

- Data sets often have attributes that contribute little to classification.
- Informally, ‘how much information is gained about the classification by knowing the value of the attribute a?’
- Only the attributes for which the information gain is high are retained.
- Three stages:
  - 1. Calculate the value of information gain for each attribute in the original dataset.
  - 2. Discard all attributes that do not meet a specified criterion.
  - 3. Pass the revised dataset to the preferred classification algorithm.

Information Gain for Feature Reduction

- We know a method of calculating information gain for categorical attributes using frequency tables
- Also know modification that enables the method to be used for continuous attributes by examining alternative ways of splitting the attribute values into two parts
- (The method also returns a ‘split value’, i.e. the value of the attribute that gives the largest information gain.
- This value is not needed when information gain is used for feature reduction.
- It is sufficient to know the largest information gain achievable for the attribute with any split value.

Policies

- Criteria for attribute retention:
- – Only retain the best 20 attributes
- – Only retain the best 25% of the attributes
- – Only retain attributes with an information gain that is at least 25% of the highest information gain of any attribute
- – Only retain attributes that reduce the initial entropy of the dataset by at least 10%.
- There is no one choice that is best in all situations

Genetics Dataset

- Three classifications
  - Distributed 767, 768 and 1655
  - Amongst the three classes for the 3190 instances.
- The relative proportions are 0.240, 0.241 and 0.519,
- Entropy is: $-0.240 \times \log_2(0.240) - 0.241 \times \log_2(0.241) - 0.519 \times \log_2(0.519) = 1.480.$
- The values of information gain for some of the attributes A0 to A59

Attribute	Information Gain
A0	0.0062
A1	0.0066
A2	0.0024
A3	0.0092
A4	0.0161
A5	0.0177
A6	0.0077
A7	0.0071
A8	0.0283
A9	0.0279
....	......
A27	0.2108
A28	0.3426
A29	0.3896
A30	0.3296
A31	0.3322
....	......
A57	0.0080
A58	0.0041
A59	0.0123

Information Gain Re-expressed

- Adjust table dividing information gain by largest value
  - Now a proportion from 0 to 1
- Multiply each value by 100
- A29 largest, much larger than any others
- Only small number are even 50% as large

Attribute	Info. Gain (adjusted)
A0	1.60
A1	1.70
A2	0.61
A3	2.36
A4	4.14
A5	4.55
A6	1.99
A7	1.81
A8	7.27
A9	7.17
....	....
A27	54.09
A28	87.92
A29	100.00
A30	84.60
A31	85.26
....	....
A57	2.07
A58	1.05
A59	3.16

Alternative View - Frequencies

- Divide number of adjusted values into bins labeled 10, 20, 30, … , 100.
- First bin corresponds to 0 to 10
- Second bind corresponds to 10 to 20
- Rightmost column is the cumulative frequency
- 41 of 60 attributes have information gain no more than the max at A29
- Only 6 attributes are more than 50% of A29

Bin	Frequency	Cumulative frequency	Cumulative frequency (%)
10	41	41	68.33
20	9	50	83.33
30	2	52	86.67
40	2	54	90.00
50	0	54	90.00
60	2	56	93.33
70	0	56	93.33
80	0	56	93.33
90	3	59	98.33
100	1	60	100.00
Total	60		

Experimental Result

- Using TDIDT with 10-fold cross-validation and all 60 attributes
  - 89.5% accuracy
- Using only the best six attributes
  - 91.8%
- Reduces the chance of overfitting.

Bcst96 Dataset

- The bcst96 dataset comprises 1186 instances (training set)
  - and a further 509 instances (test set).
- Each instance corresponds to a web page, which is classified into one of two possible categories, B or C,
- Using the values of 13,430 attributes, all continuous.
- There are 1,749 attributes that each have only a single value for the instances in the training set and so can be deleted, leaving 11,681 continuous attributes.
- Number of attributes is 11 times number of instances.
- A large number do not impact classification.

Bcst96 Dataset Information Gain

- Initial entropy is .996, indicating that the classes are equally distributed.
- Eliminate all attributes that have the same value for all training instances.
- 11,681 (95.33%) have information in the 5% bin
- Almost 99% are in the 5% and 10% bin.

Bin	Frequency	Cumulative frequency	Cumulative frequency (%)
5	11,135	11,135	95.33
10	403	11,538	98.78
15	76	11,614	99.43
20	34	11,648	99.72
25	10	11,658	99.80
30	7	11,665	99.86
35	4	11,669	99.90
40	1	11,670	99.91
45	2	11,672	99.92
50	1	11,673	99.93
55	1	11,674	99.94
60	2	11,676	99.96
65	2	11,678	99.97
70	0	11,678	99.97
75	1	11,679	99.98
80	0	11,679	99.98
85	1	11,680	99.99
90	0	11,680	99.99
95	0	11,680	99.99
100	1	11,681	100.00
Total	11,681		

Results

- Using TDIDT with the entropy attribute selection criterion for classification, the algorithm generates 38 rules from the original training set and uses these to predict the classification of the 509 instances in the test set.
- 94.9% accuracy (483 correct and 26 incorrect predictions).
- Discarding all but the best 50 attributes, the same algorithm generates a set of 62 rules,
- Also 94.9% predictive accuracy on the test set (483 correct and 26 incorrect predictions).
- All the attributes the TDIDT will examine approximately 1, 186 × 11, 681 = 13, 853, 666 attribute values at each node.
- If only the best 50 attributes are used the number drops to just 1, 186 × 50 = 59, 300.