Homework 1

Jarod Klion

CAP5771

Chapter 1:

1. 1. No, this is a database query
   2. No, this is a database query after some calculations
   3. No, this is a simple calculation
   4. No, this is a database query
   5. No, because the dice is fair, this is a probability calculation
   6. Yes, this is using predictive methods/regression
   7. Yes, this is a classification problem
   8. Yes, this is a classification problem
   9. No, this is signal processing
2. Data mining techniques like classification or anomaly detection can help an Internet search engine company greatly. Classification could assign search results to pre-defined categories which could speed up search speeds. Additionally, anomaly detection techniques can be used to detect when something is being searched at unusually high rates.

Chapter 2:

* 1. Binary, qualitative, ordinal
  2. Continuous, quantitative, ratio
  3. Discrete, qualitative, ordinal
  4. Continuous, quantitative, ratio
  5. Discrete, qualitative, ordinal
  6. Continuous, quantitative, ratio
  7. Discrete, quantitative, ratio
  8. Discrete, qualitative, nominal
  9. Discrete, qualitative, ordinal
  10. Discrete, qualitative, ordinal
  11. Continuous, quantitative, ratio
  12. Discrete, quantitative, ratio
  13. Discrete, qualitative, nominal
  14. We need to transform the original data into asymmetric binary variables by introducing a new item for each distinct attribute-value pair.
  15. There would be 400 of these asymmetric binary attributes because there are 100 questions with 4 possible answers each.

1. Temporal autocorrelation means that the values of two measurements are often very similar when the measurements are close in time. Thus, daily temperature is likely to show more temporal autocorrelation because rainfall varies based on many factors while temperatures usually relate to the time of day and the preceding temperature.
2. Advantages:
   1. Text files are more easily viewed on almost all systems
   2. Text files are also more easily edited with text editing software
   3. If there are duplicates, the order of the duplicates in the nearest neighbor list depends entirely on the order of the objects in the original data set. Another problem if there are duplicates is that having too many duplicates may lead to a case in which the nearest neighbor list is entirely populated by duplicates.
   4. To fix the problem of duplicates, we could keep only one value for each duplicate
   5. Hamming distance: # of bits different between 2 binary vectors = 3

Jaccard similarity: # of 1-1 matches / # of not-both-zero values = 2/5 = 0.4

* 1. Hamming is most like SMC because they both look at all the data and find where the data is different or similar, but SMC looks for similarities while Hamming looks for differences. On the other hand, Jaccard is more similar to the cosine measure because they both ignore 0-0 matches.
  2. Jaccard would be more appropriate as we’re looking for the shared genes only.
  3. Because humans share 99.9% of the same genes, we should look for mainly differences when comparing them, so we should use the Hamming distance.
  4. cosine:

correlation:

Euclidean:

1. cosine:

correlation:

Euclidean:

Jaccard:

1. cosine:

correlation:

Euclidean:

1. cosine:

correlation:

Jaccard:

1. cosine:

correlation:

* 1. Range of cosine is [-1, 1].
  2. Not necessarily since their magnitudes could be different, so the attributes can differ by a constant factor.
  3. cos(x,y) = corr(x,y) if the means of x and y are both 0.
  4. There is an inverse relationship between the Euclidean distance and cosine similarity when vectors have an L2 norm of 1.
  5. There is an inverse relationship between Euclidean distance and correlation when the vectors have been standardized to have a mean of 0 and a standard deviation of 1.
  6. Euclidean distance:
  7. For standardized vectors: mean = 0 and standard deviation = 1, we get

Euclidean distance:

* 1. clustering time series:
     1. For clustering, we want to maximize similarity in time series clusters, so we want high correlation values together and minimize similarity across clusters. We can achieve this by applying a transformation such that,
  2. one time series to another:
     1. For behavior from one to another, we want to consider all correlations, positive and negative, so we can map correlation values using the following possible transformation:

1. To go from similarity in [0,1] to dissimilarity in [0, ∞], we can apply the transformations shown in the book: