CS 484 FINAL EXAM REVIEW

NATHANIEL BECHHOFER

1. Types of Attributes

- **Categorical (Qualitative) attributes**
  - **Nominal attributes**: Only naming. Examples include zip codes, gender
  - **Ordinal attributes**: Also includes order. Examples include year of birth, street numbers
- **Numeric (Quantitative) attributes**
  - **Interval attributes**: Meaningful units which allow comparison of differences between values. Examples include dates, temperature
  - **Ratio attributes**: Also includes meaningfulness of ratios. Examples include income, height, age
- **Discrete attributes**: Can be categorical or numeric (e.g. counts), but has a finite (or countably infinite, which means corresponding to any integer) number of possible values. Examples include number of children, marital status, gender
  - Special case is **binary attrubutes**, which can only take on two values such as yes/no or 0/1
- **Continuous attributes**: Real numbers, which imply a potentially infinite number of possible values. Examples include height, weight, temperature

2. Noise vs Outliers

Noise is anything that is not the true data, also defined as the random component of a measurement error. It may have values close to the true data. An outlier is something that is much different than the other values with regards to either its characteristics or attribute values.

3. Measures of Similarity or Dissimilarity

3.1. **Euclidean Distance.** The **Euclidean distance** between two points x and y is given as:

$$\sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

where n is the number of dimensions and x_k and y_k are the k^{th} attributes of x and y.

3.2. **Hamming Distance.** The **Hamming distance** is the number of bits that are different between two objects that have only binary attributes. (Manhattan distance is Euclidean distance without the squared and square root.)

3.3. **Binary Similarity Measures.** If we have two objects x and y that consist of n binary attributes, we can denote counts f_{ij} which measure how many attributes measure 0 or 1 for the x and y objects.

This gives us a way to write the formula for the **simple matching coefficient**:

$$SMC = \frac{\text{number of matching attribute values}}{\text{number of attributes}} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

If binary attributes are asymmetric, one can use the **Jaccard coefficient**:

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

3.4. **Cosine Similarity.** For non-binary vectors, we can do something similar to the Jaccard with the **cosine similarity**:

$$\cos(x, y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

3.5. **Correlation.** We can use **Pearson's correlation coefficient** (often denoted using ρ) to measure the linear relationship between attributes of many types of objects.

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{s_X s_Y}$$

4. Impurity Measures

4.1. **Gini.** The Gini index is defined as $1 - \sum_{i=0}^{c-1}[p(i|t)]^2$ where c is the number of classes and $p(i|t)$ denotes the fraction of records belonging to class i at a given node t. For multiway split, we calculate using the weighted average for each category.

4.2. **Entropy.** Entropy is defined as $-\sum_{i=0}^{c-1} p(i|t)\log_2 p(i|t)$ where c is the number of classes and $p(i|t)$ denotes the fraction of records belonging to class i at a given node t. (We consider $0\log_2 0 = 0$ for entropy calculations.)

4.3. **Gain.** The gain can be found using a formula:

$$I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

where $I(.)$ is the impurity measure of a given node, N is the number of records at the parent node, k is the number of attribute values, and $N(v_j)$ is the number of records associated with the child node v_j. Gain is called **information gain** when using entropy as the impurity measure.

5. Bayes, the theorem

One standard form:

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B \mid A)P(A) + P(B \mid \neg A)P(\neg A)}$$

6. Naive Bayes

6.1. **Conditional independence.** Naive Bayes only works with conditional independence. X is conditionally independent of Y given Z if $P(X|Y,Z) = P(X|Z)$.

6.2. **Classification.** To choose how to classify a record, one merely finds the class that maximizes $P(Y)\prod_{i=1}^{d} P(X_i|Y)$.

6.3. **M-estimate.** If the classifier gets zeros, there may be an issue with categorization. An alternative uses the following formula:

$$P(x_i|y_j) = \frac{n_c + mp}{n + m}$$

where n is the total number of instances from class y_j, n_c is the number of training examples from y_j that take on the value x_i, m is a parameter known as the equivalent sample size, and p is a user-specified parameter.

7. Association rules

In order to select interesting rules from the set of all possible association rules, constraints on various measures of significance and interest are used. The best-known constraints are minimum thresholds on support and confidence. Let X be an item-set, $X \Rightarrow Y$ be an association rule and T a set in a given database.

Support is an indication of how frequently the item-set appears in the database.

The support value of $X$with respect to T is defined as the proportion of records in the database which contains the item-set X.

Confidence is an indication of how often the rule has been found to be true.

The confidence value of a rule, $X \Rightarrow Y$, with respect to a set T, is the proportion of the records that contain X and also Y.

Confidence is defined as

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}.$$