CSCI 4022 Spring 2021

Load nb0708

nb0708 day. Announcements:

"kmeans" & "EM"

HW 3 Due Monday! Some **hints:**

- Check out: companion notebooks with consolidated "solutions" to k-means and 1-Dim EM. This should make your "by-hand" implementations easier.

2a Sample covariance is almost the same calculation as sample variance! But you should do this by hand, since np.var/np.cov won't do *probability-weighted* calculations. mean: $\frac{1}{n}\sum x_i \cdot p_i$

2b To compute distance-from-point-to-component, you can either choose **the most likely component** for a single distance, or do **probability weighted distance** to *all* components. The latter will perform better as k increases, since there will be more "uncertain" points. Or come up with your own distance measure!

3 It should not surprise you if plotting mpg versus $disp$ makes unnormalized clusters **look** better than normalized clusters. You should already know why based on your answer in 3B! But consider other plots to demonstrate this, at least to yourself.

Market Basket Analysis

Definition: The *support* for itemset I is the number of baskets that contain all items in I. Often, support is expressed as a fraction of the total number of baskets. Given a *support threshold* s, the sets of items that appear in at least s baskets are called *frequent itemsets*.

Definition: The *confidence* of the association rule $I \rightarrow J$ is the ratio of the support for $I \cup \{j\}$ to the support for I.
$conf(I \rightarrow J) = \frac{support(I \cup \{j\})}{support(I)}$

Definition: The *interest* of the association rule $I \rightarrow J$ is the difference between its confidence and the fraction of baskets that contain j :
$interest(I \rightarrow J) = conf(I \rightarrow J) - P(j)$

Association Rules: Top down

Suppose we have an assoc. rule $I \to j$ with support s, and high confidence c. Then $I \cup j$ has support of at least cs because

$$conf(I \to j) = \frac{support(I \cup \{j\})}{support(I)} \Leftrightarrow c = \frac{support(I \cup \{j\})}{s}$$

This suggests a top-down mining algorithm to list off rules given set frequencies.

1. Find all itemsets with support at least cs (Set 1)
2. Find all itemsets with support at least s (Set 2, which will be a subset of Set 1 since $s \geq cs$)
3. Loop: For each itemset J of Set 1...
   3.1 Consider the $support(J) = s_2$ (we would have previously computed this)
   3.2 For each element $j \in J$, remove j and compute $support(J - \{j\}) = s_1$
   3.3 If $s_1/s_2 \geq c$ then $J - \{j\} \to j$ is an acceptable association rule.

Market Basket: Storing Counts

May be a preliminary step of data processing to encode names of items as numbers (e.g., through a bar code, or hash table). Then we store counts!

- The function:

$$a[k] = (i)\left(n - \frac{i+1}{2}\right) + j - i - 1$$

  will (0-indexed) store item counts for the pair i, j, where $1 \leq i < j \leq n$. This is a **triangular array**, because it saves exactly the information of the upper triangle (column$j$ >row i) of a matrix where i, j is the support of $\{i, j\}$.

- Alternatively, store counts as a list of triples $[i, j, c]$ where c is the count of $\{i, j\}$, $j > i$. Upside here: no saving "0" when i and j don't ever overlap.

Usually we'll have to preface this with a hash table to translate items as they appear in a file to integers.

We deal with these concepts today in notebooks 7 and 8.