

DMC@ISU: Iowa State University's 2015 Data Mining Cup Team

Categorical Similarity Measures

Spring 2015,

Name: Ian Mouzon

email: `imouzon@iastate.edu`

Due Date: May 8, 2015

0.1 Categorical Similarity

Oh My Gosh - did you know that you can compare categorical variables to each other? You can create kind of like a distance on them.

One simple example based on the Jaccard Measure:

Let \mathbf{u} and \mathbf{v} be two multidimensional categorical variables taking values in $A = \{a_1, a_2, \dots, a_n\}$. Then we can think of these as subsets of the the set A in which case we can describe a similarity between the coupons using

$$J(\mathbf{u}, \mathbf{v}) = \frac{|\mathbf{u} \cap \mathbf{v}|}{|\mathbf{u} \cup \mathbf{v}|}$$

.