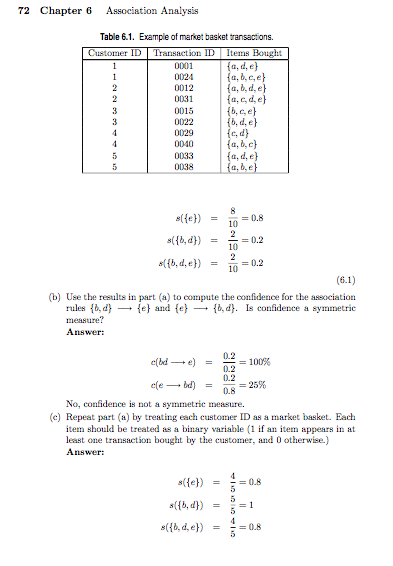
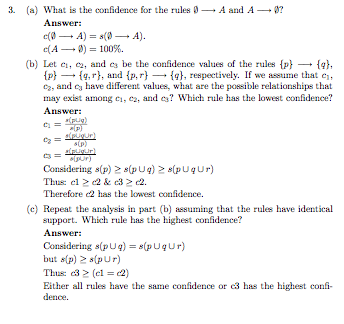
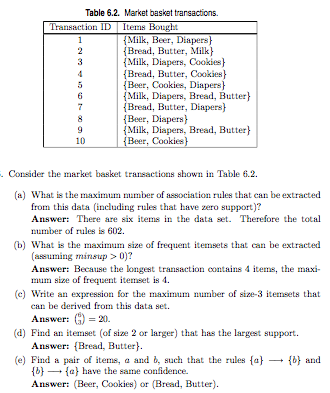
Normalization: Goal is to make an entire set of values have a particular property, A traditional example is that of “standardizing a variable” in statistics. If different variables are to be combined in some way, then they need to be transformed (by standard deviation for example) so to avoid having a variable with large values dominate the results of the calculation. Age vs Income example.. income dominates the calculation. Mean and standard deviation are strongly affected by outliers.

K-NN: instance based and save every training observation then make prediction by searching for most similar training observations. They are memory intensive, bad for high dimensional data, and the correct distance function. CLASSIFICATION: SUPERVISED learning task for modeling and predicting categorical variables (churn, email spam, financial fraud) Ensemble: decision trees. Perform well in practice, able to model non-linear decision boundaries thanks to hierarchical structure. Weaknesses: unconstrained, prone to overfitting. Naïve Bayes: model is a probability table that gets updated through training data. ASSUMES CONDITIONAL INDEPENDENCE (rarely holds true in real world). Strengths: perform well in practice, easy to implement and scale. Weaks: NB too simple. CLUSTERING: **UNSUPERVISED** learning task for finding natural groupings of observations (clusters) based on inherent structure within dataset. (customer segmentation, grouping similar items in ecommerce). Clustering is unsupervised meaning there is no “right answer” data visualization usually used to implement results. K-means: clusters made based on geometric distances between points. Grouped around centroids causing globular similar sized clusters. Strengths: fast simple flexible. Weaks: user must specify number of clusters, if clusters in data are not globular than not good clusters. Hierarchical: start with each point in its own cluster, for each cluster merge it with another based on some critera, repeat until one cluster left and hierarchy of clusters. Streng: non globular clusters. WEaks: user mus choose # of clusters. DBSCAN: makes clusters for dense regions of points. Does not assume globular clusters and does not require every point be assigned to a cluster. Weaks: user must define epsilon and Min\_Pts to define density of clusters.

SUPERVISED: All data us labeled, and algorithms learn to predict the output from the input data.

UNSUPERVISED: All data is unlabeled, and algorithms learnt to inherent structure from input data.

