Bill Chambers - 25912237
Andy (Yu) Wang - 24106507

CS289 Final Project Initial Report
Applied Distance Metrics and Categorical Data Manipulations

After exploring a variety of datasets, we have settled on working with a Kaggle competition set up by Walmart. This competition is primarily focused on classifying a given trip type given a set of products purchased on that trip. There are a lot of data points to learn off of (~650,000 and 95000 distinct visits) and an output metric that will be easy to measure via kaggle. However we have identified one of the core challenges, there are only 5 input variables and they are all categorical. For example, the UPC code column contains 97715 unique values. This presents a unique challenge because all variables are categorical variables. We have done some basic exploration of the data and run a random forest classifier without any data manipulation in order to compute a minimum baseline metric (another of which is supplied by kaggle). We aim to continue to improve our random forest method.

We also recognize that a tree-based method is the obvious approach. Therefore, we want to experiment to see if we can beat this baseline through clever manipulation of the categorical variables. After speaking with Professor Guyon, she recommended that we explore the chi-square kernel to try and find a way to manipulate categorical variables in a way that would make them amenable for use in other methods like neural networks. We have found some preliminary research below that might lead us towards a way to explore this challenge.

Our hope is that through clever usage of some of the research methods included below (primarily concerned with addressing high dimension and sparse spaces with distance metrics, the chi-square kernel, etc) we can match or beat a random forest/tree based method. As Professor Guyon mentioned in her theoretical challenges, kernel methods and deep learning are not easy to use with missing data and categorical data. We hope that by reviewing and trying out some of the below mentioned methods, we might be able to shed more light on approaches in preprocessing that help push us towards better usage of categorical variables with these other methods. While this is certainly a lofty goal, we feel that the implications would be extremely powerful. In conclusion we hope to explore usage of the random forest classifier and attempt to beat it with other methods that take advantage of clever handling of categorical variables.

References:

http://www-users.cs.umn.edu/~sboriah/PDFs/BoriahBCK2008.pdf
http://www.stat.tamu.edu/~hart/compare.pdf
http://www.umass.edu/landeco/teaching/multivariate/readings/McCune.and.Grace.2002.chapter6.pdf
http://www.cc.gatech.edu/~fli/chebyshev_cvpr12.pdf
http://papers.nips.cc/paper/5075-sign-cauchy-projections-and-chi-square-kernel.pdf
http://www.mtome.com/Publications/CiML/CiML-v3-book.pdf
http://www.marc-boulle.fr/publications/BoulleMLDM05.pdf (Marc Boullé)