**Updates 8/12/2016**

* Quick Background: I am aiming to formulate a Metric Learning cost function that will learn the appropriate distance metric for data with probabilistic or distribution labels. The expectation is that using this metric (instead of using just Euclidean distance, for example) will result in a better performance of algorithms such as KNN. In addition, the metric will also indicate what features are significant and what are not.
* Important restrictions when formulating a cost function:
  + It needs to be convex – so that there is a global minima that can be found
  + It needs to be specifically suited for KNN – priority must be given to those data points that are closer than those that are farther away and have similar labels.
* Current Formulation:

Subject to: A ≥ 0 (psd)

**Sij­** is the similarity between the labels of points i and j

dij is the distance between the features of points i and j

A is the Matrix of weights we are trying to solve for

xi and xj are the two vectors of feature values for points i and j

* + The formulation is affine (linear). However it collapses to A = 0.
  + We need to add a constraint to this so that it doesn’t collapse
  + Another important issue – What measure do we use for s and d? They need to be between 0 and 1. And­need to have an even distribution so that we can distinguish between points or labels that are similar and not similar.
* Proposed Formulation:

(This should solve the collapsing problem)

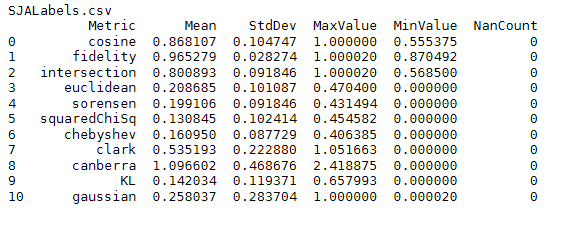
Where, the j’s are in the neighborhood of i. The neighborhood is determined by the Sij/ dij ratio and value of k selected.

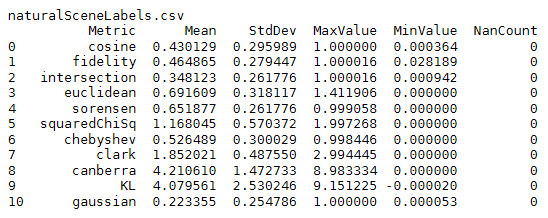
Subject to:

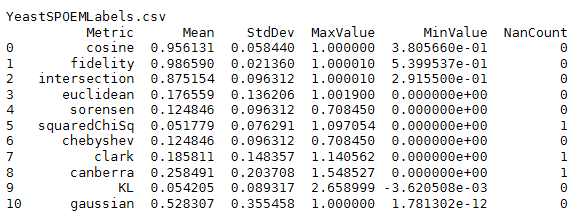
* + It is important to note here that a just because a point A is very close to point B it does not mean that they are in the same neighborhood. They can have very different labels. We want to push such points away.
  + However, even for this formulation we need to use the right measures to calculate Sij, dij

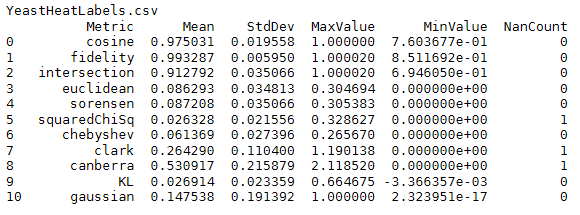
Here are some measure and their performance on datasets with distribution labels:

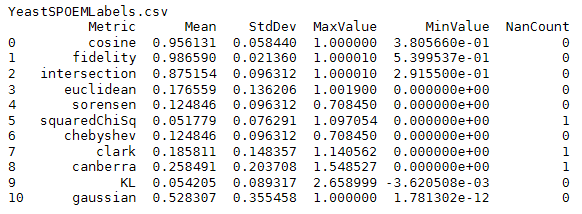
* Table summaries of datasets:











* Histograms by Measure:

