**Distance Metrics Notes:**

* **Statistical distance**: the distance between two statistical objects, which can be two random variables, or two probability distributions or samples, or the distance can be between an individual sample point and a population or a wider sample of points
  + A distance between populations can be interpreted as measuring the distance between two probability distributions and hence they are essentially measures of distances between probability measures
  + Where statistical distance measures relate to the differences between random variables, these may have statistical dependence,[1] and hence these distances are not directly related to measures of distances between probability measures
* **Geodesic**: a curve representing in some sense the shortest path between two points in a surface, or more generally in a Riemannian manifold
  + For a spherical Earth, it is a segment of a great circle
  + In a Riemannian manifold or submanifold geodesics are characterised by the property of having vanishing geodesic curvature
* **Minkowski distance**: a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance.
* **Manhattan**: measures distance following only axis-aligned directions.
* **Euclidean**: straight-line distance between two points in Euclidean space.
* **Cosine similarity**: a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them
  + a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors oriented at 90° relative to each other have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude
  + particularly used in positive space, where the outcome is neatly bounded in [0,1].
  + measure cohesion within clusters in the field of data mining
  + One advantage of cosine similarity is its low-complexity, especially for sparse vectors: only the *non-zero dimensions* need to be considered
* **Jaccard similarity**: a statistic used for gauging the similarity and diversity of sample sets
  + The Jaccard distance, which measures dissimilarity between sample sets, is complementary to the Jaccard coefficient and is obtained by subtracting the Jaccard coefficient from 1
  + measure of similarity for the two sets of data, with a range from 0% to 100%
  + J(X,Y) = |X∩Y| / |X∪Y|
  + Tanimoto similarity and Tanimoto distance is equivalent to Jaccard similarity, but the distance function is not the same as Jaccard distance.Correlation based
* **Pearson correlation coefficient**: measure of the *linear* correlation between two variables
  + the covariance of the two variables divided by the product of their standard deviations
* **Spearman's Rho**: a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.
  + While Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not)
  + If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other
* **Distance correlation or distance covariance** is a measure of dependence between two paired random vectors of arbitrary, not necessarily equal, dimension.
  + The population distance correlation coefficient is zero if and only if the random vectors are independent
  + measures both linear and nonlinear association between two random variables or random vectors, in contrast to Pearson's correlation
* **Kendall's Tao rank distance** is a metric that counts the number of pairwise disagreements between two ranking lists. The larger the distance, the more dissimilar the two lists are
  + called **bubble-sort distance** since it is equivalent to the number of swaps that the bubble sort algorithm would take to place one list in the same order as the other list
  + In order to calculate the Kendall tau distance, pair each person with every other person and count the number of times the values in list 1 are in the opposite order of the values in list 2
* **Mahalanobis distance**: measures distance between a point , P and a distribution D
  + Multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D; distance is zero if P is at the mean of D, and grows as P moves away from the mean along each principal component axis
  + If each of these axes is re-scaled to have unit variance, then the Mahalanobis distance corresponds to standard Euclidean distance in the transformed space
  + unitless and scale-invariant and takes into account the correlations of the data set.
* **Cook's distance**: Data points with large residuals (outliers) and/or high leverage (measure of how far away the independent variable values of an observation are from those of the other observations) may distort the outcome and accuracy of a regression. Cook's distance measures the effect of deleting a given observation.
  + Estimate of the influence of a data point when performing a least-squares regression analysis
  + indicate influential data points that are particularly worth checking for validity
  + to indicate regions of the design space where it would be good to be able to obtain more data points
* **Chi-Square test**: statistically significant difference (i.e., a magnitude of difference that is unlikely to be due to chance alone) between the expected frequencies and the observed frequencies in one or more categories of a so-called contingency table.
  + ***Central limit theorem (CLT)*** *establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a bell curve) even if the original variables themselves are not normally distributed*
* Distribution based
* **KS test:** a nonparametric test of the equality of continuous (or discontinuous), one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test)
  + quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples
  + *Note: The empirical distribution function is an estimate of the cumulative distribution function that generated the points in the sample*.
  + can be modified to serve as a goodness of fit test
* **Kullback–Leibler divergence (relative entropy)**: how one probability distribution is different from a second, reference probability distribution
  + it is a measure of surprise
  + does not satisfy the triangle inequality
  + distribution-wise asymmetric measure and thus does not qualify as a statistical metric of spread
  + ***Information entropy, or Shanon entropy****, is a basic quantity in information theory associated to any random variable, which can be interpreted as the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes*
  + ***Variation of information or shared information distance*** *is a measure of the distance between two clusterings (partitions of elements)*
  + ***Conditional entropy (or equivocation)*** *quantifies the amount of information needed to describe the outcome of a random variable Y given that the value of another random variable X is known*