STA9890 - Test 3 - Formula Sheet

Clustering Scores:

- Silhouette Score (point):

$$s_i = \frac{b_i - a_i}{\max\{b_i, a_i\}}$$

  where a_i is average distance to points in the same cluster as i and b_i is average distance to points in the nearest cluster to i

- Silhouette Score (Entire Data Set):

$$S = \max_k \frac{1}{|C_k|} \sum_{k' \in C_k} s_{k'}$$

  That is, the maximum average silhouette score across all clusters.

- Within-Cluster Sum of Squares:

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{i:\boldsymbol{x}_i \in C_k} \|\boldsymbol{x}_i - \mu_k\|^2 = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \mu_{C_i}\|^2$$

Clustering Linkages:

- Single Linkage: $d(A,B) = \min_{a\in A, b\in B} d(a,b)$
- Complete Linkage: $d(A,B) = \max_{a\in A, b\in B} d(a,b)$
- Average Linkage: $d(A,B) = \frac{1}{|A||B|} \sum_{a\in A, b\in B} d(a,b)$
- Centroid Linkage: $d(A,B) = d(\overline{a}, \overline{b}) = d\left(|A|^{-1} \sum_{a\in A} a, |B|^{-1} \sum_{b\in B} b\right)$

where the pointwise distances $d(a,b)$ are typically based on ℓ_p norms: $d(a,b) = \|a-b\|_p$

Convex Clustering:

$$\arg\min_{\boldsymbol{u}_i \in \mathbb{R}^p (i\in 1,\ldots,n)} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{u}_i\| + \lambda \sum_{i,j} \|\boldsymbol{u}_i - \boldsymbol{u}_j\|$$

PCA:

- Variance explained by PC-k: d_k^2
- Percent variance explained (marginally) by PC-k: $d_k^2 / \sum_i d_i^2$
- Cumulative percent variance explained up to (and including) PC-k: $(\sum_{i=1}^{k} d_i^2)/(\sum_{j=1}^{\text{rank}(\boldsymbol{X})} d_j^2)$
- Built upon SVD (see above):

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \sum_{j=1}^{\text{rank}(\boldsymbol{X})} d_j \boldsymbol{u}_j \boldsymbol{v}_j^\top$$