CCA and Clustering Review

Stephen Bates

July 25, 2019

1 Canonical Correlation Analysis

Goal

Suppose we have data points $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}^q$ for $i = 1, \ldots, n$. Canonical Correlation Analysis (CCA) is a technique for finding linear combinations of X_i and Y_i that are highly correlated. If we are primarily interested in the relationship between X_i and Y_i, then we can reduce dimension by restricting our attention to directions that exhibit dependence in the sample.

Formalizing the problem

Formally, let

$$\mathrm{Cov}(X, Y) = \Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$$

be the empirical covariance matrix. We then consider the following optimization problem:

$$\text{maximize}_{(a_1,b_1):a_1^T\Sigma_{XX}a_1\leq 1, b_1^T\Sigma_{YY}b_1\leq 1} a_1^T\Sigma_{XY}b_1.$$

We see that the solution will be directions a_1 and b_1 such that $a_1^T X_i$ and $b_1^T Y_i$ are maximally coordinated. We can then continue with this strategy by solving

$$\begin{aligned} \text{maximize}_{\ (a_k,b_k)} \quad & a_k^T\Sigma_{XY}b_k \\ \text{subject to} \quad & a_k^T\Sigma_{XX}a_k \leq 1 \\ & b_k^T\Sigma_{YY}b_k \leq 1 \\ & a_i^T\Sigma_{XX}a_j = 0 \qquad i < j \leq k \\ & b_i^T\Sigma_{YY}b_j = 0 \qquad i < j \leq k, \end{aligned}$$

for steps $k = 1, \ldots, K$. Notice that the last two constraints are enforcing a form of orthogonality between the old directions and the new directions. The resulting pairs $a_k^T X_i, b_k^T Y_i$ are known as the canonical variables, and they are encoding "orthogonal" directions in X_i and Y_i that are dependent with each other.

Solving via the SVD

We can reformulate the above optimization problem into an eignevalue problem to solve for several CCA directions simultaneously. Let $u_i = \Sigma_{XX}^{1/2} a_i$ and $v_i = \Sigma_{YY}^{1/2} b_i$ for $i = 1, \ldots, K$. Then we see that step k of the procedure above is simply:

$$\begin{aligned} \text{maximize}_{\ u_k, v_k} \quad & u_k^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} v_k \\ \text{subject to} \quad & u_k^T u_k \leq 1 \\ & v_k^T v_k \leq 1 \\ & u_i^T u_j = 0 \qquad i < j \leq k \\ & v_i^T v_j = 0 \qquad i < j \leq k \end{aligned}$$

which is equivalent to finding the matrices U and V from the SVD of $\Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}$.

2 Clustering

Suppose we have a set of data points $X_i \in \mathbb{R}^p, i = 1, \ldots, n$. Often, we expect that the data points exhibit some form of grouped structure, and so it is of interest to find procedures that would uncover this. One way to go about this is to partitioning the data points into K clusters such that points within a cluster are more similar than points in different clusters.

Top-down approaches

Some clustering algorithms can be thought of as "top-down" approaches: we start with a set of clusters and then try to iteratively update the clusters based on their membership. The two most common examples are

- K-means clustering
- K-mediods clustering

Both of these algorithms are fast to compute, which makes them popular. Note that both methods require some form of initialization, and typically several different starting values are tried.

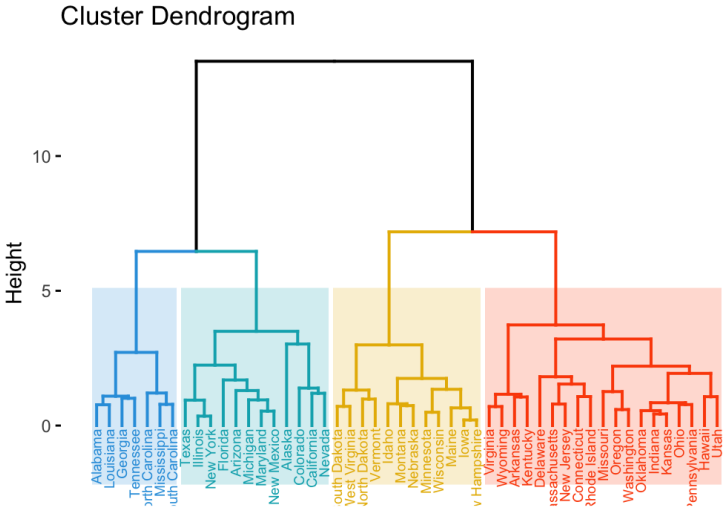
Bottom-up approaches

An alternative way of producing clusters is to go bottom up: we start with each point in its own cluster and then successively merge clusters based on the distance between each cluster. Different choices of distances lead to different procedures, and three common choices are summarized below.

- single linkage hierarchical clustering: cluster distance is the distance of the closest pair of points, one from each cluster.
- group average hierarchical clustering: cluster distance is the average distance of points in each cluster.

- complete linkage hierarchical clustering: cluster distance is the distance of the farthest pair of points, one from each cluster

These method all have the benefit that they produce a dendrogram, such as the one below. In addition to giving a nice representation of the clusters across several scales, the dendogram can give visual guidance about the choice of the number of clusters to consider.



Model-based approaches

A last approach to clustering is to specify a generative model for the data, such as a mixture of Gaussian distributions. After fitting the model, points can be assigned to the mixture component of highest probability to form a partition of the data points. This approach is similar to K-means, but has the benefit that goodness-of-fit statistics (like BIC) can help guide the choice of K.

Remarks

Note that the first two types clustering methods **rely heavily on a distance metric**. Euclidean distance is the default choice, but in that case scaling the variables to have equal variances is important. If specialized distance metrics are available, they may give better clusters.