View on GitHub

stats-learning-notes

Notes from Introduction to Statistical Learning

Previous: Chapter 09 - Support Vector Machines

Chapter 10 - Unsupervised Learning

Unlike most of the other learning methods discussed so far, in the unsupervised learning scenario, though there are still p predictors and n observations, there is no response variable Y. Instead, the goal is to discover interesting properties of the observations $X_1, X_2, \ldots, X_n$. Two popular unsupervised learning techniques are principal component analysis and clustering.

Unsupervised learning is often performed as part of exploratory data analysis. Results tend to be more subjective without clear goals and without accepted mechanisms for validating results. In the unsupervised learning scenario there is no correct answer to check predictions against.

Principal Component Analysis

Principal component analysis (PCA) allows for summarizing a large set of correlated variables with a smaller number of representative variables that collectively explain most of the variability in the original set of variables.

As discussed earlier, the principal component directions are the directions of the feature space along which the data are highly variable. These directions also define lines and subspaces that are as close as possible to the data cloud.

Principal component regression is the result of using principal components as the predictors in a regression model instead of the original set of variables.

Principal component analysis refers to the process used to compute the principal components and subsequent use of the components to understand the data.

Principal component analysis finds a low-dimensional representation of the data set that contains as much of the variation as is possible. Though each observation exists in a p-dimensional space, only a subset of those p-dimensions is interesting, where interesting is measured by the amount that the observations vary along each dimension. Each dimension found by principal component analysis is a linear combination of the p features.

The first principal component of a set of features is the normalized linear combination of the features:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ \ldots \ + \phi_{p1}X_p$$

that has the largest variance. Normalization is achieved by adhering to the constraint

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

The coefficients $\phi_{11}, \ldots, \phi_{p1}$ are referred to as the loadings of the first principal component. Together the loadings make up the principal component loading vector, $\phi_1 = (\phi_{11}, \ldots, \phi_{p1})$.

The loadings are constrained such that their sum of squares is equal to one to avoid setting the elements to be arbitrarily large in absolute value which could result in skewing the variance to be arbitrarily large.

Given an $n \times p$ data set X, the first principal component can be computed as follows.

First, each of the variables in X should be centered to have a mean of zero. Next, an optimization problem is solved that yields the optimal loading vector for the linear combination

$$z_{ij} = \phi_{11} X_{i1} + \phi_{21} X_{i2} + \ \ldots \ + \phi_{p1} X_{ip}$$

subject to the constraint

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

that has the largest sample variance. The optimization problem is defined as

$$\text{Maximize}_{\phi_{11}, \ldots, \phi_{pi}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\}$$

subject to

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

This objective can be restated as

$$\frac{1}{n} \sum_{i=1}^{n} z_{i1}^2.$$

Also, since variables in x have been centered to have a mean of zero, the average of $z_{11}, \ \ldots, \ z_{n1}$ will be zero also. Hence, the objective is to maximize the sample variance of the n values of z_{i1}.
$z_{11}, \ z_{21}, \ \ldots, \ z_{n1}$ are referred to as the scores of the first principal component.

The given optimization problem can be solved via an eigendecomposition, a standard technique in linear algebra not covered here.

It is worth noting that units are important for principal component analysis, so standardization is often recommended.

When interpreted in a geometric setting, the first principal component's loading vector ϕ_1 with elements $\phi_{11}, \ \ldots, \ \phi_{p1}$ defines a direction in the feature space along which the data vary the most. When the n data points $x_1, \ \ldots, \ x_n$ are projected onto this direction, the projected values are the principal component scores $z_{11}, \ \ldots, \ z_{n1}$ themselves.

As many as $\text{Min}(n-1, p)$ principal components can be computed.

The second principal component, Z_2 is the linear combination of $x_1, \ \ldots, \ x_p$ that has the maximal variance out of all linear combinations that are uncorrelated with Z_1. The scores for the second principal component, $z_{12}, z_{22}, \ \ldots, \ z_{n2}$, take the form

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \cdots + \phi_{p2} x_{ip}$$

where ϕ_2 is the second principal component loading vector with elements $\phi_{12}, \ \phi_{22}, \ \ldots, \ \phi_{p2}$. It works out that constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction of ϕ_2 to be orthogonal or perpendicular to the direction of ϕ_1.

Another interpretation of principal components is that they provide low-dimensional linear surfaces that are closest to the observations. Under such an interpretation, the first principal component has a very special property: it is the line in p-dimensional space that is closest to the n observations using average squared Euclidean distance as the metric for closeness. This is appealing because a single dimension of the data that lies as close as possible to all of the data points will likely provide a good summary of the data.

This interpretation extends beyond the first principal component. For example, the first two principal components define the plane that is closest to the n observations in terms of average squared Euclidean distance. Similarly, the first three principal components span the three dimensional hyperplane that is closest to the n observations.

Under this interpretation, the first M principal component score vectors, combined with the first M principal component loading vectors provide the best M-dimensional approximation (in terms of Euclidean distance) to the ith observation, x_{ij}. This representation takes the form

$$x_{ij} \approx \sum_{m=1}^{M} Z_{im}\phi_{jm}$$

assuming the original data matrix, X, is column centered. This means that when M is sufficiently large, the M principal component score vectors and loading vectors can provide a good approximation of the data. When $M = \text{Min}(n-1,\ p)$, the representation is exact:

$$x_{ij} = \sum_{m=1}^{M} Z_{im}\phi_{jm}.$$

The results obtained by performing principal component analysis depend on whether the variables have been individually scaled (each multiplied by a different constant). As such, variables are typically scaled to have a standard deviation of one before performing principal component analysis.

Principal components are unique and consistent, though signs may vary depending on the calculation method.

The portion of variance explained provides a means of determining how much of the variance in the data is not captured by the first M principal components.

The total variance in the data set assuming the variables have been centered to have a mean of zero is defined by

$$\sum_{j=1}^{p} \text{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2.$$

The variance explained by the mth principal component is defined as

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2.$$

From these equations it can be seen that the portion of the variance explained for the mth principal component is given by

$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

To compute the cumulative portion of variance explained by the first m principal components, the individual portions should be summed. In total there are $\text{Min}(n-1,\ p)$ principal components and their portion of

variance explained sums to one.

A scree plot can be useful for determining how many principal components is enough, but there is no well accepted, objective way to decide the appropriate number of principal components. Another means is to keep taking principal components while each new principal component explains a sizable portion of the variance. This doesn't always work well.

In supervised learning scenarios, cross validation can be used to tune the appropriate number of principal components.

Applying other statistical learning methods to the principal components instead of the original variables can yield less noisy results.

Clustering methods

Clustering refers to a broad set of techniques for finding subgroups or clusters in a data set. What constitutes similar or different tends to depend on the domain in question. Clustering can be an unsupervised problem in scenarios where the goal is to discover structure and that structure is not known in advance. Clustering looks for homogeneous subgroups among the observations.

There are many kinds of clustering. Two of the most popular clustering approaches are k-means clustering and hierarchical clustering. In general, observations are clustered by features in order to identify subgroups among the observations or features can be clustered by observations to try to find subgroups among the features.

K-Means Clustering

K-means clustering aims to partition a data set into K distinct, non-overlapping clusters, where K is stipulated in advance.

The K-means clustering procedure is built on a few constraints. Given sets containing the indices of the observations in each cluster, $C_1, \ \ldots, C_K$, these sets must satisfy two properties:

1. Each observation belongs to at least one of the K clusters: $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ \ldots, \ n\}$
2. No observation belongs to more than one cluster. Clusters are non-overlapping.
   $C_k \cap C_{k'} = \{\}$ for all k, $k \neq k'$

In the context of K-means clustering, a good cluster is one for which the within-cluster variation is as small as possible. For a cluster C_k, the within-cluster variation, $W(C_k)$, is a measure of the amount by which the observations in a cluster differ from each other. As such, an ideal cluster would minimize

$$\sum_{k=1}^{K} W(C_k).$$

Informally, this means that the observations should be partitioned into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

In order to solve this optimization problem, it is first necessary to define the means by which within-cluster variation will be evaluated. There are many ways to evaluate within-cluster variation, but the most common choice tends to be squared Euclidean distance, defined as

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

where $|C_k|$ denotes the number of observations in the kth cluster.

Combined with the abstract optimization problem outlined earlier yields

$$\text{Minimize}_{C_1, \ldots, C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Finding the optimal solution to this problem is computationally infeasible unless K and n are very small, since there are almost K^n ways to partition n observations into K clusters. However, a simple algorithm exists to find a local optimum:

1. Assign each observation to a cluster from 1 to K.
2. Iterate until cluster assignments stop changing:
    1. Compute the cluster centroid for each of the K clusters. The kth cluster centroid is the vector of p feature means for the observations in the kth cluster.
    2. Assign each observation to the cluster whose centroid is the closest, where closest is defined using Euclidean distance.

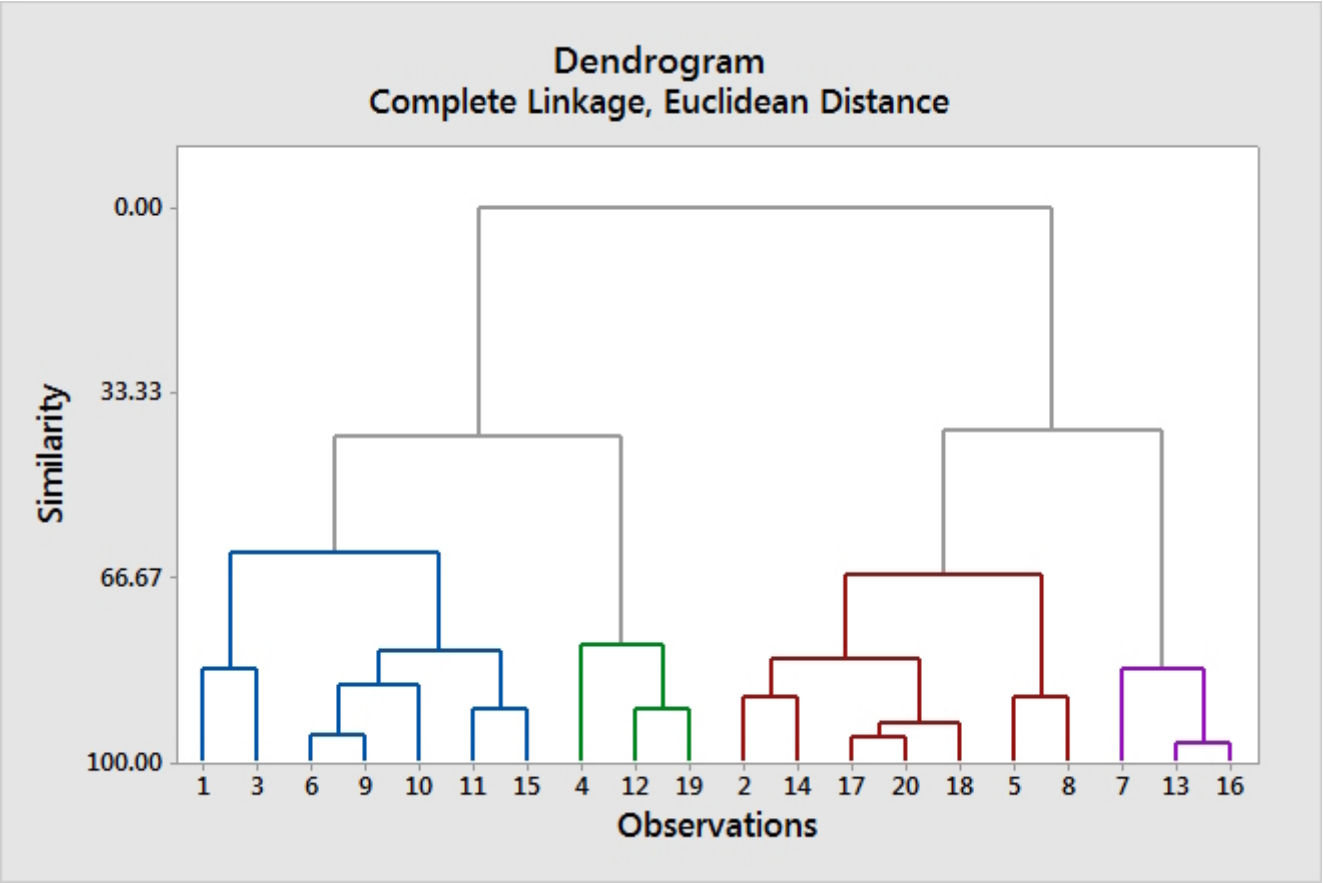
Though it may not seem like it, this algorithm is guaranteed to decrease the value of the objective function at each step. This is because the algorithm repeatedly relocates observations based on minimizing the sum of squared deviations, resulting in the sum of squared deviations getting smaller with each iteration. When the result of the algorithm no longer changes, this means a local optimum has been found.

K-means clustering gets its name from the fact that the cluster centroids are computed as means of the observations assigned to each cluster.

Because the K-means algorithm finds a local optimum instead of a global optimum, the results obtained will depend on the initial randomized cluster assignments of each observation. For this reason, it is important to run the algorithm multiple times from different initial configurations and select the solution for which the objective is smallest.

Hierarchical clustering

One disadvantage of K-means clustering is that it requires K to be stated in advance. Hierarchical clustering has the advantage over K-means clustering that it does not require committing to a particular choice of K and in addition, it results in a tree-based representation of the observations called a dendrogram.



The most common type of hierarchical clustering is bottom-up or agglomerative clustering in which the dendrogram is built starting from the leaves and combining clusters up to the trunk. Based on this, it is not hard to see that the earlier observations or branches fuse together, the more similar they are. Observations that fuse later, near the top of the tree, can be quite different. The height at which two observations fuse together is indicative of how different the observations are.

No conclusions about similarity should be made from horizontal proximity as this can vary.

Clusters can be extracted from the dendrogram by making a horizontal cut across the dendrogram and taking the distinct sets of observations below the cut. The height of the cut to the dendrogram serves a similar role to K in K-means clustering: it controls the number of clusters yielded.

Dendrograms are attractive because a single dendrogram can be used to obtain any number of clusters.

Often people will look at the dendrogram and select a sensible number of clusters by eye, depending on the heights of the fusions and the number of clusters desired. Unfortunately, the choice of where to cut the dendrogram is not always evident.

The term hierarchical refers to the fact that the clusters obtained by cutting the dendrogram at the given height are necessarily nested within the clusters obtained by cutting the dendrogram at any greater height.

The hierarchical structure assumption is not always valid. For example, splitting a group of people in to sexes and splitting a group of people by race yield clusters that aren't necessarily hierarchical in structure. Because of this, hierarchical clustering can sometimes yield worse results than K-means clustering.

The hierarchical clustering dendrogram is obtained by first selecting some sort of measure of dissimilarity between the each pair of observations; often Euclidean distance is used. Starting at the bottom of the dendrogram, each of the n observations is treated as its own cluster. With each iteration, the two clusters that are most similar are merged together so there are $n - 1$ clusters. This process is repeated until all the observations belong to a single cluster and the dendrogram is complete.

The dissimilarity between the two clusters that are merged indicates the height in the dendrogram at which the fusion should be placed.

One issue not addressed is how clusters with multiple observations are compared. This requires extending the notion of dissimilarity to a pair of groups of observations.

Linkage defines the dissimilarity between two groups of observations. There are four common types of linkage: complete, average, single, and centroid. Average, complete and single linkage are most popular among statisticians. Centroid linkage is often used in genomics. Average and complete linkage tend to be preferred because they tend to yield more balanced dendrograms. Centroid linkage suffers from a major drawback in that an inversion can occur where two clusters fuse at a height below either of the individual clusters in the dendrogram.

Complete linkage uses the maximal inter-cluster dissimilarity, calculated by computing all of the pairwise dissimilarities between observations in cluster A and observations in cluster B and taking the largest of those dissimilarities.

Single linkage uses the minimal inter-cluster dissimilarity given by computing all the pairwise dissimilarities between observations in clusters A and B and taking the smallest of those dissimilarities. Single linkage can result in extended trailing clusters where each observation fuses one-at-a-time.

Average linkage uses the mean inter-cluster dissimilarity given by computing all pairwise dissimilarities between the observations in cluster A and the observations in cluster B and taking the average of those dissimilarities.

Centroid linkage computes the dissimilarity between the centroid for cluster and A and the centroid for cluster B.

Choice of Dissimilarity Measure

Correlation based distance considers two observations to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance. Correlation based distance focuses on the shapes of observation profiles rather than their magnitudes.

The type of data and the scientific question at hand should determine what dissimilarity measure is used for hierarchical clustering.

The choice of whether or not to scale the variables before computing the dissimilarity measure depends on the application at hand.

With K-means clustering and hierarchical clustering, there's rarely one right answer and a few small decisions can have big consequences:

- Should observations be standardized in some way? For example, observations could be centered to have a mean of zero and/or scaled to have a standard deviation of one.
- In the case of K-means clustering:
    - How many clusters are desired?
- In the case of hierarchical clustering:
    - What dissimilarity measure should be used?
    - What type of linkage should be used?
    - At what height should the dendrogram be cut?

There are methods for assigning a p-value to a cluster in order to assess whether there is more evidence for a cluster than would be expected due to chance, however, there's been no consensus on a single best approach.

Because clustering can be non-robust to changes in the data set, it's recommended to cluster subsets of the data to get a sense of the robustness of the yielded clusters.

Other forms of clustering exist that do not force every observation into a cluster, which can be useful in data sets that contain outliers that do not belong to any cluster. Mixture models can help address such outliers. A soft version of K-means clustering can be effective in these scenarios.

stats-learning-notes maintained by tdg5

Published with GitHub Pages