ch10_notes

August 10, 2019

Table of Contents

---

1 Unsupervised Learning

1.1 The Challenge of Unsupervised Learning

- ***Unsupervised learning*** is learning in the absence of a response. It is often part of ***exploratory data analysis*** (EDA).
- Without a response, we aren't intested in prediction or classification, rather we are interested in discovering interesting things about the data. This can be difficult because such a goal is somewhat subjective.
- Objective performance critera for unsupervised learning can also be challenging.

1.2 Principal Components Analysis

- Principal components were discussed earlier as a dimensional reduction methof in the Section **??**. They provide a low-dimensional representation of the data that contains as much variation as possible.
- ***Principal Components Analysis*** is the process of computing principal components and using them in data analysis.

1.2.1 What Are Principal Components?

- The ***first principal component*** of features $X_1, \dots, X_p$ is the normalized linear combination

$$Z_1 = \hat{\phi}_1^\top X$$

where $X = (X_1, \dots, X_p), \hat{\phi}_1 \in \mathbb{R}^p$ and $||\hat{\phi}|| = 1$. The vector $\hat{\phi}_1$ is called the ***loading*** vector (its entries are called the ***loadings***) and

$$\hat{\phi}_1 = \underset{\substack{\phi \in \mathbb{R}^p \\ ||\phi|| = 1}}{\operatorname{argmax}} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\phi^\top x_i \right)^2 \right)$$

- Assume we have data X_i with features $X_1, \dots, X_p$ which is centered in the features (each feature has mean zero). The objective function in the above optimization problem can be rewritten

$$\hat{\phi}_1 = \underset{\phi \in \mathbb{R}^p}{\operatorname{argmax}} \left(\frac{1}{n} \sum_{i=1}^{n} ||z_i||^2 \right)$$

which is just the sample variance. The z_{i1} are called the ***scores*** of the first principal component Z_1.

- The first principal component has a nice geometric interpretation 85. The loading vector ϕ_1 defines a direction in $\mathbb{R}^p$ along which the variation is maximized. The principal component scores z_{i1} are the projections of the data x_i onto ϕ_1 – that is, the components of the x_i along this direction.

- For $j = 2, ..., p$ we can compute the j-th principal component ϕ_j recursively

$$\hat{\phi}_j = \underset{\phi \in \mathbb{R}^p}{\operatorname{argmax}} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\phi^\top x_i \right)^2 \right)$$

subject to 86

$$\phi_j^\top \phi_{j-1} = 0$$

. - We can plot the principal components against each other for a low-dimensional visualization of the data. For example a ***biplot*** plots both the scores and the loading vectors 87.

1.2.2 Another Interpretation of Principal Components

- Principal components can also be seen as providing low-dimensional surfaces that are "closest" to the observations.
- The span of the first M loading vectors $\phi_1, \dots, \phi_M$ can be seen as the M-dimensional linear subspaces of $\mathbb{R}^p$ which is closest to the observations x_i 88
- Together the principal components $Z_1, \dots, Z_M$ and loading vectors $\phi_1, \dots, \phi_M$ can be seen as an M-dimensional approximation89 of each observation

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

1.2.3 More on PCA

- PCA requires that the variables are centered to have mean zero
- PCA is sensitive to scaling, so we usually scale each variable to have standard deviation 1.
- Scaling to standard deviation 1 is particularly important when variables are measured in different units, however if they are measured in the same units we may not wish to do this.

Uniqueness of the Principal Components The loading vectors and score vectors are unique up to sign flips.

The Proportion of Variance Explained

- How much of the information in a given data set is lost by projecting onto the principal components? More precisely, what is the ***proportion of variance explained*** (PVE) by each principal component?

- Assuming centered data, the ***total variance*** 90 is

$$\text{var}_{total} := \sum_{j=1}^{p} \mathbb{V}(X_j) = \sum_{i=1}^{p} \left(\frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 \right)$$

  while the ***variance explained*** by the m-th principal component is

$$\text{var}_m := \frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{i=1}^{p} \phi_{jm} x_{ij} \right)^2$$

  .

- The PVE of the m-th component is then

$$\text{PVE}_m := \frac{\text{var}_m}{\text{var}_{total}}$$

  and the cumulative PVE of the first M components 91 is

$$\sum_{m=1}^{M} \text{PVE}_m$$

Deciding How Many Principal Components to Use

- In general choose we may not be interested in using all principal components, but just enough to get a "good" understanding of the data 92.
- A ***scree plot***, which plots PVM_m vs. m, can help identify a good number of principal components to use, is one visual method for identifying a good number of principal components. We look for an ***elbow*** - a value of m such that PVM_m drops off thereafter.
- In general, the question of how many principal components are "enough" is ill-defined, and depends on the application and the dataset. We maybe look at the first few principal components in order to find interesting patterns. If none are evident, then we conclude further components are unlikely to be of use. If some are evident, we continue looking at components until no more interesting patterns are found.
- In an unpervised setting, these methods are all ad hoc, and reflect the fact that PCA is generally used in EDA 93.

1.2.4 Other Uses for Principal Components

- Many statistical techniques (regression, classification, clustering) can be adapted to the $n \times M$ PCA matrix with columns the first $M << p$ principal component score vectors.
- The PCA matrix can be seen as a "de-noising" 94 of the original data, since the signal (as opposed to the noise) is weighted towards the earlier principal components

1.3 Clustering Methods

- This is a broad set of techniques for finding ***clusters*** (or subgroups) of the data set.
- Observations should be "similar" within clusters and dissimilar across clusters. The definition of "similar" is context dependent.
- Clustering is popular in many fields, so there exist a great number of methods.

1.3.1 K-Means Clustering

- ***K-means clustering*** seeks to partition the data into a pre-specified number K of distinct, non-overlapping clusters.
- More precisely, we seek a partition $\hat{C}_1, \dots \hat{C}_K$ of the set of indices $\{1, \dots n\}$

$$\hat{C}_1, \dots \hat{C}_K = \underset{C_1, \dots, C_k}{\operatorname{argmin}} \left(\sum_{k=1}^{K} W(C_k) \right)$$

where $W(C_k)$ is some measure of the variation within cluster C_k. - A typical choice of $W(C_k)$ is the average 95 squared Euclidean distance between points in C_k:

$$W(C)_k = \frac{1}{|C_k|} \sum_{i,i' \in C_k} ||x_i - x_i'||^2$$

- A brute force algorithm for finding the global minimum is $O(K^n)$ but there is a much faster algorithm which is guaranteed to find a local minimum. It uses a random initialization so it should be performed several times.

Algorithm: K-Means Clustering

1. Initialize by randomly assigning a cluster number $1, \dots K$ to each observation.
2. While the cluster assignments change:

    1. For each $k = 1, \dots K$, compute the ***centroid*** of the k-th cluster (the vector of feature means for the observations in the cluster).
    2. Assign to each observation the number of the cluster whose centroid is closest.

Advantages

Disadvantages

1.3.2 Hierarchical Clustering

- ***Hierarchical clustering*** is an alternative clustering method which doesn't require a specified number of clusters and results in an attractive tree-based representation of the data called a ***dendrogram***.
- ***Bottom-up*** or ***agglomerative*** hierarchical clustering builds a dendrogram from the leaves up to the trunk.

Interpreting a Dendrogram

- A dendrogram is a tree (visualized as upside down) with leaves corresponding to observations.
- As we move up the tree, similar observations fuse into branches, and similar branches again fuse.
- The earlier fusions occur, the more similar the corresponding groups of observations. The height at which two observations are joined by this fusing is a measure of this similarity.
- At each height in the dendrogram, a horizontal cut splits the observations into k clusters (corresponding to each of the branches cut) where $1 \leqslant k \leqslant n$.
- The best choice of cut (hence number k of clusters) is often obtained by inspecting the diagram.

The Hierarchical Clustering Algorithm

- This algorithm uses a notion of dissimilarity defined for clusters, called a ***linkage***.
- Let A, B be clusters, and let $d(a,b)$ be a dissimilarity measure 95 for observations a, b. A linkage defines a dissimilarity measure $d(A, B)$ between the clusters A, B. The four most common types of linkage are
  - ***complete***:
    $$d_{comp}(A,B) = \max_{(a,b)\in A\times B} d(a,b)$$
  - ***single***:
    $$d_{sing}(A,B) = \min_{(a,b)\in A\times B} d(a,b)$$
  - ***average***
    $$d_{avg}(A,B) = \frac{1}{|A||B|}\sum_{(a,b)\in A\times B} d(a,b)$$
  - ***centroid***
    $$d_{cent}(A,B) = d(x_a, x_b)$$
    ,
    where x_a (resp. x_b) is the centroid of A (resp. B).
- Average, complete, and single linkages are preferred by statisticians. Average and complete linkages are generally preferred as they result in more balanced dendrograms.
- Centroid linkage is often used in genomics, but suffers from the possibility of an ***inversion***, in which two clusters are fused at a height *below* the individual clusters, which makes interpretation difficult.

Choice of Dissimilarity Measure

- The squared Euclidean distance is often used as a dissimilarity measure.
- An alternative is the ***correlation-based distance***
- The choice of dissimilarity measure is very important and has a strong effect on the resulting dendrogram. The choice of measure should be determined by context.
- One should consider scaling the data before choosing the dissimilarity measure.

Algorithm: Hierarchical Clustering

```
1. Initialize with $n$ clusters, one for each observation, and compute the dissimilarities $d(x_i, x_j)$ for each pair.
2. For $i = n, \dots, 2$:
    1. Compute all dissimilarites among the $i$ clusters, and fuse the two clusters which are the least dissimilar. This dissimilarity is the height in the dendrogram where the fusion is placed.
    2. Compute the dissimilarities among the new $i - 1$ clusters.
```

Advantages
Disadvantages

1.3.3 Practical Issues in Clustering

Small Decisions with Big Consequences

- Should observations or features be standardized in some way?
- For hierarchical clustering:
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
  - Where should we cut the dendrogram to determine the number of clusters?
- For K-means clustering, what is the choice of K?

Validating the Clusters Obtained

- It is important to decide whether the clusters obtained reflect true subgroups in the data or are a result of "clustering the noise".
- There exist techniques for making this decision, such as obtaining p-values for each cluster.

Other Considerations in Clustering

- Both K-means and hierarchical clustering assign all observations to some cluster. This can be problematic, for example in the presence of outliers that don't clearly belong to any cluster.
- "Mixture models" are an attractive approach to accommodating outliers (they amount to a "soft" clustering approach).
- Clustering methods are not robust to perturbations.

A Tempered Approach to Interpreting the Results of Clustering

- Clustering can be a very useful and valid statistical tool if used properly.
- To overcome the sensitivity to hyperparameters, is recommended to try hyperparameter optimization.
- To overcome the sensitivity to perturbations, it is recommended to cluster on subsets of the data.
- Finally, results of cluster analysis should be considered a part of EDA and not taken too seriously

---

1.4 Footnotes

85. The linear algebra interpretation is also nice

86. This constraint is equivalent to

$$\mathrm{corr}(Z_j, Z_{j-1}) = 0$$

87. See book figure 10.1 and corresponding discussion.

88. That is, the linear subspace in $\mathbb{R}^p$ which minimizes the sum of the squared euclidean distances to the points x_i.

89. When $M = \min\{n-1, p\}$, the approximation is exact.

90. More accurately, the sum on the right is an estimate of the sum on the left.

91. In general there are $\min\{n-1, p\}$ principal components and

$$\sum_{m=1}^{\min\{n-1,p\}} \mathrm{PVE}_m = 1$$

92. Indeed, if we take $M < p$ principal components, then we are truly doing dimensional reduction.

93. In a supervised setting, however, we can treat the number of components as a tuning parameter.

94. There is a nice information-theoretic interpretation of this statement.

95. That we are taking an average is probably the reason for the "means" in "K-means".

96. For example, the commonly used squared Euclidean distance. See Choice of Dissimilarity Measure