Unsupervised Learning

Sravani Vadlamani

Schedule

Date	Topic
April 12	Principal Component Analysis, Examples and Applications
April 14	Clustering Methods, K-means algorithm , Hierarchical Clustering, Examples and Applications
April 19	Exam Review
April 21	Introduction to Deep Learning
April 26	Final Exam in Class
May 5 (8 – 10 AM IST 1017)	Final Project Presentations

Final Project - Deadlines

Date	Deliverable
March 22	Project Proposal Due
April 3	Exploratory Data Analysis
April 10	Analysis/Results
April 17	Draft Final Report
April 27	Final Report Due
April 27	Final Presentation Due

Unsupervised Learning

- Agenda
  - What is Unsupervised Learning?
  - Principal Component Analysis
  - Clustering

Supervised vs Unsupervised Methods

- Supervised Learning
  - Goal is to predict the response variable Y using a set of features X_1, X_2, ….X_p measured on n observations
  - Examples include regression and classification methods
- Unsupervised Learning
  - There is no response variable, and we are not interested in prediction
  - The goal is to discover interesting subgroups or patterns among the variables or among the observations. Another way of visualizing the data

Unsupervised Methods

- Two Methods
  - Principal Component Analysis
    - A useful tool for data visualization and data pre-processing before applying supervised techniques
  - Clustering
    - A range of methods to discover unknown subgroups in the data

Challenges of Unsupervised Methods

- Supervised methods include well developed set of tools to estimate trained models and assess their quality using test sets and cross validation techniques
- Unsupervised methods are more subjective as there is no specific goal like prediction of a response variable. They are often performed as part of exploratory data analysis and the results are hard to assess.
- Unsupervised methods are used in various fields
  - Subgroup breast cancer patients by their gene expression measurements
  - Group shoppers by their browsing and purchase histories
  - Group movies by ratings assigned by viewers

Principal Components Analysis (PCA)

- Allows you to summarize a large set of correlated variables with a smaller number of representative variables that collectively explain most of the variability in the original set
- PCA refers to the process of computing the principal components and subsequently using these to understand the data.
- PCA produces derived variables to use in supervised learning methods. It also serves as a tool for data visualization and data imputation (filling missing values)

Principal Components

- Are low dimensional representation of the data set that contains as much variation as possible. Although n observations lie in p-dimensions, all these dimensions are not interesting. PCA aims to find the subset of the p-dimensions that are interesting. The concept of interesting is measured by the amount that the observations vary in each dimension.
- Each of the dimensions found by PCA is a linear combination of the p features.

Principal Components

- The first principal component of the set of features X_1, X_2, ….X_p is the normalized linear combination of the features that has the largest variance given by

$$Z_1 = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \emptyset_{31}X_3 \ldots\ldots\ldots\ldots + \emptyset_{p1}X_p$$

- Normalization is achieved by constraining the loadings such that their sum of squares is equal to one. This constraint avoids setting the elements to be arbitrarily large in absolute value that could result in an arbitrarily large variance

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

- The coefficients $\emptyset_{11}$, $\emptyset_{21}$, …. $\emptyset_{p1}$ are called the loadings of the first principal component. Together, the loadings make up the principal component loading vector $\emptyset_1 = (\emptyset_{11}, \emptyset_{21}\ldots\ldots\ldots \emptyset_{p1})$

Principal Components

- Given a nxp dataset X, first principal components are computed as follows
- Each of the variable in X should be centered to have mean zero. We then look for the linear combination of the sample feature values of the form below that has the largest variance

$$Z_{i1} = \emptyset_{11}X_{i1} + \emptyset_{21}Xi_2 + \emptyset_{31}X_{i3} \ldots\ldots\ldots\ldots + \emptyset_{p1}X_{ip}$$

- Subject to the constraint

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

Principal Components

- The optimization problem is defined as

$$maximize_{\emptyset_{11}, \emptyset_{12}, \ldots \emptyset_{11},} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \emptyset_{j1} \, x_{ij} \right)^2 \right\} subject\ to \sum_{j=1}^{p} \emptyset_{j1}^2 = 1$$

- This objective function can be rewritten as $\frac{1}{n} \sum_{i=1}^{n} z_{i1}^2$
- Since the variables in X have been centered to have a mean zero, the average of z_{11} , ……. z_{n1} will also be zero. The objective is maximizing the sample variance of the n values of z_{i1} .
- z_{11} , ……. z_{n1} are the scores of the first principal component.
- The above optimization problem is solved using eigen decomposition

Principal Components

- Geometric interpretation of the first principal component
  - The loading vector $\emptyset_1$ with elements $\emptyset_{11}$, …….. $\emptyset_{p1}$ defines a direction in feature space along which the data vary the most. If we project the n data points x_1, ….. x_n onto this direction, the projected values are the principal component scores $z_{11,\ldots\ldots\ldots}$ z_{n1}themselves
- Number of principal components that can be computed
  - Minimum (n-1, p)

Second Principal Component

- Second principal component Z_2 is the linear combination of X_1, X_2, ….X_p that has maximal variance out of all linear combinations that are uncorrelated with Z_1 and is given by

$$Z_{i2} = \emptyset_{12}X_{i1} + \emptyset_{22}X_{i2} + \emptyset_{32}X_{i3} \ldots\ldots\ldots\ldots + \emptyset_{p2}X_{ip}$$

- Where Z_{i2} is the second principal component loading vector with elements $\emptyset_{12}, \emptyset_{22}\ldots\ldots\ldots \emptyset_{p2}$
- Constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction of $\emptyset_2$ to be orthogonal or perpendicular to $\emptyset_1$

Another Interpretation of Principal Component

- 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 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 the data points will likely provide a good summary of the data.

Proportion of Variance Explained

- The proportion of variance explained is a good way of capturing the amount of variance in the data not captured by the first M principal components.
- The total variance in the data set assuming that the variables have been centered to mean zero is defined as

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

- The variance explained by the m^{th} principal component is

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

Proportion of Variance Explained

- The PVE of the m^{th} principal component is given by

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

- The cumulative portion of variance explained by the first m principal components is calculated by summing the individual portions. There are a total of min(n-1, p) principal components and the portion of variance explained by them sums to one

Determine the number of principal components

- No objective way to determine the appropriate number of principal components.
- A scree plot is used to typically determine the number of principal components
- Another approach is to keep including principal components while each new one explains a sizeable portion of the variance although this approach does not always work well

Scaling the variables

- Variables should be scaled so that they are all in the same units.
- Variables are scaled to have mean zero and standard deviation one
- If all the variables are measured in the same units, no scaling is required
- Each principal component loading vector is unique up to a sign flip. Two different packages will yield the same principal components although the signs of the loading vectors may differ.