Unsupervised Learning Notes

Sameer Kesava PhD

Contents

Introduction

The purpose of unsupervised learning is to

1. Finding hidden structures in the data
2. Data visualization
3. Data compression
4. Data denoising
5. Finding correlations

Clustering

The goal of clustering is to find natural groupings in the data.

Since there are no ground truth labels in unsupervised learning, the use of intrinsic performance metrics and domain knowledge is required to evaluate the quality of clustering.

In general, observations can be clustered on the basis of features and vice-versa, i.e. *features can be clustered*.

2.1 ML Algorithms

The different clustering algorithms and their properties.

2.1.1 Principal Component Analysis

Popular applications of PCA:

1. EDA
2. Denoising of signals, e.g. stock market data, genome data and gene expression levels.

The features should be standardized to a σ=1 and, preferably, μ=0 before using PCA.

Kernels:

1. Linear
2. Radial (RBF): for non-linear datasets.
    - Gaussian kernel function: γ is the hyperparameter.

PCA tends to perform well in cases where most of the information is captured by the first few principal components. This can be obtained as variance explained ratios implemented in sklearn as pca.explained_variance_ratio_, and plotted as a **Scree plot** (elbow points in this plot gives an estimate for the choice of no. of PCs to choose for further processing).

Biplots: for visualizing PC scores and loading vectors.

2.1.2 Latent Dirichlet Allocation

- Unsupervised Learning
- For topic modeling, e.g. classifying a newspaper as a topic.
- scikit-learn lda module for batch and online learning.

2.1.3 KMeans

- This is a prototype-based clustering method where the clusters are represented by a centroid or mediod (from medians).
- Non-overlapping, i.e. an observation belongs to only 1 cluster.
- Elbow method and Silhouette plots (see Silhouette Analysis subsection 2.3.1) for determining the optimal number of clusters and the quality of clustering respectively.
- Once an optimal cluster number is chosen, the clustering should be repeated with different random initial configurations to check for consistent results since KMeans typically finds local optimum rather than global optimum.
- Furthermore, it is important to run the algorithm with a large **n_init** value (in sklearn), else an undesirable local optimum could be obtained.
- Important to also set random seed for reproducibility.
- "Within-cluster variation": measure of the amount by which the observations within a cluster differ from each other.

2.1.4 MeanShift

- Distance-based clustering
- Has the ability to detect outliers.

2.1.5 DBSCAN

Density-based spatial clustering of applications with Noise

- Hyperparameters:
  1. min_samples: controls how tolerant the algorithm is towards noise. Starting value: 2*dimension.
  2. ϵ: crucial parameter; controls the local neighborhood of points. A starting value can be chosen using elbow/knee point in nearest neighbor plot.
  3. ϵ-metric: euclidean, minkowski, etc.
- Has the ability to detect outliers: A sample that is not a core sample and is at least ϵ in distance from any sample is considered an outlier.
- Works well for non-linear data.
- Affected by the curse of dimensionality.
- OPTICS is a variant of DBSCAN, does not require ϵ to be set.

2.1.6 BIRCH

- A hierarchical clustering method.
- Hyperparameters
  1. Threshold
  2. Branching Factor
  3. Memory efficient, online learning algorithm
  4. If the number of data instances needs to be reduced or if one wants a large no. of subclusters either as a preprocessing step or otherwise, BIRCH is more useful than MiniBatchKMeans.

2.1.7 Agglomerative

- A hierarchical clustering method.
- Uses dendograms for visualization from which the number of clusters is chosen typically by eye.

- Scaling the features to have a $\sigma = 1$ is very important, else, some features might have no influence/weights.
- Hierarchical clustering does not work in cases where are clusters within clusters, e.g. a group of people with 50-50 gender split, which in turn, is evenly split among different nationalities.
- Hyperparameters
    1. Distance-metric, i.e. similarity between observations.
        (a) Euclidean
        (b) Minkowski
        (c) Correlation-based. E.g. Online retailer interested in clustering shoppers based on their past shopping histories would prefer to use correlation-based distance metric rather than Euclidean. Used when p>2.
    2. Linkage:
        (a) Average and complete linkage yield more balance dendograms, and hence, preferred by statisticians (may be also single linkage).
        (b) Centroid linkage often used in genomics but suffers with inversion drawback, where, sometimes, the clusters are fused at a height below either of the individual clusters.

2.1.8 Manifold Learning

2.1.9 Spectral Clustering

Advanced clustering method.

2.1.10 UMAP

- Feature extraction method
- Non-linear dimensionality reduction

2.1.11 t-SNE

t-SNE is only for visualization and not for data preprocessing.

2.1.12 triMAP

2.2 Small Dataset

n: number of samples
p: number of features/predictors
When p > n, then there is no unique solution - variance is ∞.

2.2.1 When p is comparable to or greater than n

1. Use feature extraction methods such as PCA or UMAP.

2.3 Diagnostic Tools

2.3.1 Silhouette Analysis

- Measure of how tightly the clusters are grouped.

2.4 Practical Issues with Clustering

1. There is no single right answer. Hence, domain knowledge is important.
2. Different hyperparameters should be explored and the results examined for patterns that consistently emerge.
3. Clustering is not robust to perturbations to the data. Hence, clustering on subsets of the data should be attempted and the results examined.
4. The clustering results should not be taken a s the absolute truth.
5. ?Mixture models are an attractive approach for acommodating the presence of outliers, e.g. soft version of KMeans and DBSCAN.

Autoencoders

This comes under self-supervised learning.

- Image-generation with DL is done by learning latent spaces/hidden unit values that capture statistical information about a dataset of images.

3.1 Variational Autoencoders

- For generating new data
- GANs have been shown to provide better quality results compared to VA, although, VA is easier to train.

3.2 Convolutional Autoencoders

- Beware of checkerboard artifacts arising due to overlapping regions upon upsampling.
- Upsampling: unpooling layer using nearest-neighbor interpolation.
- Use formula: s(n-1) + k -2p for figuring out the padding during upsampling.
- Variant: conv-ae with randomforests
    1. train the conv-ae
    2. use the encoder as inputs to the classifier

Applications

1. Cybersecurity
2. Medical sciences such as gene expression
3. Marketing