Machine Learning

Assignment 3 (Descriptive Questions)

Mohammad Rouintan
Student No: 400222042
Shahid Beheshti University

(Spring 1402)

Exercise 1:

What is the curse of dimensionality and how does it affect clustering?

Solution:

A large number of features greatly slows down the training process and can also prevent finding a good solution. This problem is known as the ***curse of dimensionality***. The way to solve this problem is called ***dimensionality reduction***. This technique can cause data loss. Therefore, although this technique increases the speed of training, it may decrease the efficiency of the model. In addition, dimensionality reduction is very suitable for data visualization because the human mind usually cannot easily understand more than three dimensions. Three of the most common dimensionality reduction techniques are: ***PCA***, ***Kernel PCA*** and ***LLE***

Exercise 2:

In what cases would you use regular PCA, incremental PCA, randomized PCA, or random projection?

Solution:

Regular PCA: This technique, which uses full ***SVD***, is the default technique. To run this algorithm, the entire training dataset must be loaded on the system memory.

Increamental PCA: All PCA algorithms have one big problem: to run these algorithms, the entire training dataset must be loaded into the system memory. This problem has been solved by developing the incremental PCA algorithm. This algorithm allows the training dataset to be divided into small pieces (mini-batches) and the algorithm is executed on one piece at a time. This feature is very suitable for processing very large datasets or online learning.

Randomized PCA: This algorithm reduces the computational complexity from the order of $O(m \times n^2) + O(n^3)$ for full SVD to $O(m \times d^2) + O(d^3)$, which if d is much smaller than n, the increase in speed is very significant. Randomized PCA is a variation of Principal Component Analysis (PCA) that is designed to approximate the first k principal components of a large dataset efficiently. Instead of computing the eigenvectors of the covariance matrix of the data, as is done in traditional PCA, randomized PCA uses a random projection matrix to map the

data to a lower-dimensional subspace. The first k principal components of the data can then be approximated by computing the eigenvectors of the covariance matrix of the projected data.
Advantages:

- ***Speed:*** Randomized PCA is much faster than traditional PCA for large datasets, making it more suitable for real-time applications.
- ***Low-rank approximation:*** Randomized PCA can be used to obtain a low-rank approximation of a large dataset, which can then be used for further analysis or visualization.
- ***Sparsity:*** Randomized PCA is able to handle sparse datasets, which traditional PCA is not able to handle well.
- ***Scalability:*** Randomized PCA can handle large datasets that are not possible to fit into memory using traditional PCA.

Random Projection: In mathematics and statistics, random projection is a technique used to reduce the dimensionality of a set of points which lie in Euclidean space. Random projection methods are known for their power, simplicity, and low error rates when compared to other methods.
This algorithm maps samples using random linear mapping in a low-dimensional space. Random imaging seems very strange, but mathematically it has been proven that it is actually very possible for random images to preserve distances well. Thus random projection is a suitable approximation technique for distance based method. The quality of dimension reduction depends on the number of samples and the number of target dimensions, but it has nothing to do with the initial dimensions of the dataset.

Exercise 3:

Does it make sense to chain two different dimensionality reduction algorithms?

Solution:

Yes. It can absolutely make sense to chain two different dimensionality reduction algorithms. A common example is using ***PCA*** to quickly get rid of a large number of useless dimensions, then applying another much slower dimensionality reduction algorithm, such as ***LLE***.

Exercise 4:

What are the main assumptions and limitations of PCA?

Solution:

PCA assumes that the data points are centered around the coordinate origin and there is a linear relationship between the features. The algorithm is not well suited to capturing non-linear relationships. When the features have a non-linear relationship, other algorithms such as Kernel PCA can be used.

Exercise 5:

How can clustering be used to improve the accuracy of the linear regression model?

Solution:

We can also add a new feature in the dataset We can first apply clustering algorithms on the dataset, then perform linear regression on different cluster groups. We can also add a new feature in the dataset, for example:

- Creating an input feature for cluster ids as an ordinal variable.
- Creating an input feature for cluster centroids as a continuous variable.
- Creating an input feature for cluster size as a continuous variable.

Exercise 6:

How is entropy used as a clustering validation measure?

Solution:

Entropy can also be used to verify clustering quality. It makes use of the probability of a record in the cluster i of being classified as class i. Smaller values of entropy indicate less disorder in a clustering, which means a better clustering.

Entropy of cluster k:

$$H(k) = -\sum_{c \in C} P(k_c) \log_2 P(k_c) \tag{1}$$

where:

-

Total Entropy of clustering:

$$H(\Omega) = -\sum_{k \in \Omega} H(k) \frac{N_k}{N} \tag{2}$$

where:

- $\Omega = \{k_1, k_2, \ldots, k_n\}$ is the set of clusters.
- N_k is the number of points in cluster k.
- N is the total number of points.

Exercise 7: (Extra Point)

What is label propagation? Why would you implement it, and how?

Solution:

Label propagation is a semi-supervised machine learning algorithm that assigns labels to previously unlabeled data points.
The Label Propagation algorithm is a fast algorithm for finding communities in a graph. It detects these communities using network structure alone as its guide, and doesn't require a pre-defined objective function or prior information about the communities. Label Propagation works by propagating labels throughout the network and forming communities based on this process of label propagation.
Label Propagation works as follows:

- Every node is initialized with a unique community label (an identifier).
- These labels propagate through the network.
- At every iteration of propagation, each node updates its label to the one that the maximum numbers of its neighbours belongs to. Ties are broken arbitrarily but deterministically.
- Label Propagation reaches convergence when each node has the majority label of its neighbours.
- Label Propagation stops if either convergence, or the user-defined maximum number of iterations is achieved.