Report (Group 9): Home Work 4

Course Code: Data Science Lab 3 (MTH312A)

Submitted by, Group 9

Anirban Ghosh(221271), Khyati Singh(221332)

Dasari Charithambika(210302)

Rajdeep Adhya(221385), Rohit Dutta(221396)

Instructor

Dr. Subhra Sankar Dhar

Associate Professor

Department of Mathematics and Statistics,

Indian Institute of Technology, Kanpur

Submission Date: 7 April, 2024

Question

1. Download Heart Disease and Breast Cancer Wisconsin (Diagnostic) data from https://archive.ics.uci.edu/datasets. Apply depth-based classifier, SVM-based classifier, K-NN based classifier (choose K with proper justification), Kernel density function based classifier on the aforesaid two data sets and compute empirical miss-classification probability for each classifier. Note that the final result will depend on the choice of training and test data set.

Classification Methods

- **Depth Based Classifier**

  To classify points, the Depth Based Classifier has the advantage of the notion of data depth. The centrality of a point inside a dataset is measured by data depth. Assume that $\mathcal{X} = \mathbf{x_1}, \cdots, \mathbf{x_n}$ and $\mathcal{Y} = \mathbf{y_1}, \cdots, \mathbf{y_n}$ represent the two classes in our dataset. If $D_{\mathcal{X}}(\mathbf{x}) > D_{\mathcal{Y}}(\mathbf{x})$, then assign a new observation, $\mathbf{x}$, to class $\mathcal{X}$. This technique performs well with high-dimensional data and is resistant to outliers. Furthermore, this approach has little chance of misclassification. A depth-based classifier was trained using R's `ddalpha` package.

- **Support Vector Machine (SVM) Based Classifier**

  For classification problems, the supervised machine learning method Support Vector Machine (SVM) is used. In order to maximize the margin—the distance between the hyperplane and the nearest data points from each class—it finds the hyperplane that best divides the various classes in the feature space (support vectors). Using a training dataset consisting of n samples $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ in which X_i is the feature vector and y_i is the class label ($y_i \in \{-1, 1\}$ for binary classification), By addressing the convex optimization problem, support vector machines (SVM) attempt to identify a hyperplane, represented by the equation $wx + b = 0$, that divides the data into two groups.

$$\begin{gathered}\min\left(\frac{1}{2}\|w\|^2\right) \\ \text{subject to } y_i(wx_i + b) \geq 1 \;\forall\; i = 1, \cdots, n\end{gathered} \tag{1}$$

Here, the bias component is denoted by b, and the weight vector w is perpendicular to the hyperplane.

The decision function of the SVM is given by $f(x) = sign(wx + b)$, where $sign$ is the sign function.

$$sign(x) = \begin{cases} -1 & ;\ x < 0 \\ 1 & ;\ x > 0 \end{cases}$$

The reason why SVM is so popular is because it works well in high-dimensional spaces, can handle non-linear decision boundaries by using the kernel method, and is resistant to overfitting, especially in high-dimensional areas. An SVM model with a linear kernel was trained using the R's `e1071` package.

- **K-Nearest Neighbourhood (KNN) Based Classifier**

  An easy-to-use but powerful classification system is the K-Nearest Neighbors (KNN) algorithm. By using the K nearest neighbors in the training set to cast majority votes, it classifies a new data point.

  KNN Algorithm

  1. **Input:** Labeled samples from the training dataset, a test sample for classification, and the number of neighbors, k.
  2. **Calculate Distance:** Determine the difference between the test and training samples for each sample in the training dataset. The Manhattan distance, Minkowski distance, and Euclidean distance are examples of common distance metrics.
     - **Euclidean Distance:** $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
     - **Manhattan Distance:** $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$

3. **Find Neighbors:** Choose the k training dataset samples that are nearest to the test sample (i.e., have the least distance) from the dataset.

4. **Majority Vote:** Count the number of samples in each class among the k closest neighbors for classification. Put the test sample in the class that has the highest level of common with its k closest neighbors. Calculate the average of the k nearest neighbors' target values for regression.

5. **Output:** The predicted class or value for the test sample.

Using cross-validation, we developed a custom function utilizing R's built-in `knn` function to determine the ideal value of k (number of neighbors).

- **Kernel Density Estimator (KDE) Based Classifier**

  The combined probability density function of the features for each class—benign and malignant—is estimated by the Kernel Density Estimator(KDE) Based Classifier. Based on the estimated densities, it determines the possibility of a fresh sample falling into each class and designates the class with the highest likelihood as the projected class. Suppose, we have binary data coming from two different distributions F and G (class 1 & 2 respectively) which are unknown to us. Here, the forms of F and G are difficult to guess. We assign a new observation $\mathbf{x}$ to class 1 if

  $$\frac{\hat{f}(\mathbf{x})}{\hat{g}(\mathbf{x})} > 1$$

  where $\hat{f}(.)$ and $\hat{g}(.)$ are the kernel density estimates of $F(.)$ and $G(.)$ respectively.

  This approach is appropriate for complex datasets since it is non-parametric and does not presume any particular distribution of the data. It can, however, be computationally costly, particularly when dealing with high-dimensional data. We trained the KDE classification model in R using the `ks` package.

Comparison among the four Methods

To evaluate the effectiveness of each classifier, we calculate the accuracy and empirical miss-classification probability.

A model's accuracy, which is defined as the percentage of properly identified occurrences, gives a clear indicator of how successful it is. A greater accuracy means that the model is more effective at differentiating between instances that are benign and those that are malignant, which is important for precise diagnosis and treatment planning.

Misclassification probability, on the other hand, indicates the chance that a model would categorize an instance wrongly. A lower misclassification probability is preferable since it denotes fewer cases of false positives, which lowers the possibility that patients would receive ineffective care or neglect essential actions.

1 Breast Cancer Data Classification

Globally, breast cancer is a major health problem, and improving outcomes depends heavily on early identification. In this effort, we use machine learning approaches to classify cases of breast cancer as either malignant (M) or benign (B). To assess how well four distinct classifiers work in categorizing breast cancer data, we investigate them: Depth Based, SVM Based, KNN Based, and KDE Based.

Data Description

The dataset used in this project contains information about various features extracted from breast cancer biopsies. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image. A few of the images can be found at http://www.cs.wisc.edu/~street/images/. The data includes 30 numeric features, such as radius, texture, area, smoothness, compactness, concavity, symmetry, and fractal dimension. The target variable is the diagnosis, which indicates whether the biopsy is benign (B) or malignant (M).

Data Source: https://archive.ics.uci.edu/datasets

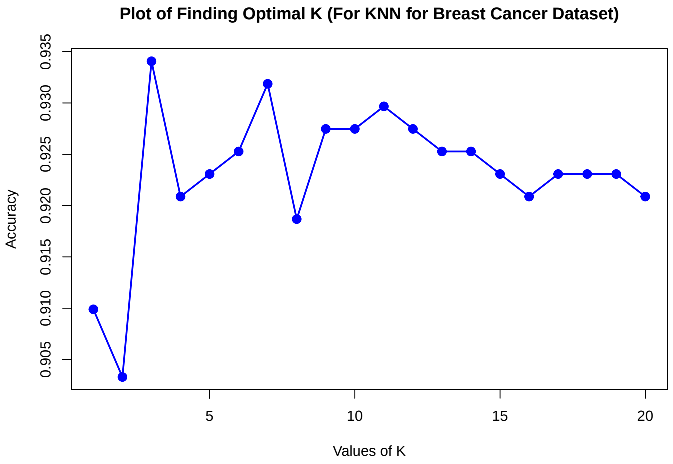

Figure 1: Line Diagram for finding the optimal value of K (for KNN classifier for Breast Cancer data)

Train-test split of the Dataset

The dataset was split using a random sampling approach. The training set comprises 80% of the total data, while the remaining 20% is allocated to the testing set. We have fitted the proposed classification models on the train dataset and checked the accuracy of the model on the test dataset.

Note 1. *To address the instability caused by the high dimensionality of the data, we implemented Principal Component Analysis (PCA) for dimensionality reduction. We selected the first 6 principal components, which collectively explain approximately 93% of the variability in the data. By reducing the dimensionality in this way, we aimed to stabilize the kernel density estimate (KDE) calculation. Subsequently, we performed KDE-based classification using these 6 transformed predictor variables.*

We have chosen the optimal value of K for KNN classifier which give highest accuracy by cross validation as shown in figure 1. Here, optimal value for K is found to be 3.

The final results are shown in the following table 1.

Classifier	Accuracy	Misclassification Probability
Depth Based	0.9649123	0.03508772
SVM Based	0.9912281	0.00877193
KNN Based	0.9561404	0.04385965
KDE Based	0.9035088	0.09649123

Table 1: Accuracy and Misclassification Probability for the four classifiers

Conclusion for Breast Cancer data

Clearly, the performance of the **SVM classifier** is best among all the classifiers discussed here.

2 Heart Disease Data Classification

Data Description

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to date. Here, we have considered only Cleveland data for our study. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

Data Source: https://archive.ics.uci.edu/datasets

Train-test split of the Dataset

The dataset was split using a random sampling approach. The training set comprises 80% of the total data, while the remaining 20% is allocated to the testing set. We have fitted the proposed classification models on the train dataset and checked the accuracy of the model on the test dataset.

The final results are shown in the following table 2 for Cleveland dataset.

Conclusion for Cleveland Data

Classifier	Accuracy	Misclassification Probability
Depth Based	0.7666667	0.2333333
SVM Based	0.8166667	0.1833333
KNN Based	0.65	0.35
KDE Based	0.689521	0.310479

Table 2: Accuracy and Misclassification Probability for the four classifiers

Clearly, the performance of the **SVM classifier** is best among all the classifiers discussed here.