EE 524 Machine Learning Lab

Assignment 6

18 November 2020

Support Vector Machine: Since, not all machine learning problems are binary classification problems, there are datasets where we have multiple classes. Multi class classification can be seen in classifying handwritten MNIST digits, multi class image classification and even the iris dataset.

1. **Import the Dataset** Download the iris flower dataset and import the csv file.
2. **Split the Dataset** Split the dataset into 4 parts. Training data, Training target, Test data and Test target. Split the data in such a way that the set (Training Data, Training Target) has 75% of the total dataset and (Test Data, Test Target) comprises of the rest of the dataset. Use scikit-learn for this.
3. **Training the SVM** Train a SVM Classifier using scikit-learn. Train the model using the Training Data and Training Target.
4. **Testing the model** Use the trained SVM Classifier to predict the labels on the Test Data.
5. **Performance of model** Use the predicted labels from the previous part and the Training Target to create a Confusion Matrix. A confusion matrix is a way to check whether the model has performed well or not on different cases.
6. **Accuracy** Report the accuracy of the model from the confusion matrix. Also try the same procedure for the splitting ratio as 80% Train and 20% Test Data.

Principal Component Analysis and Linear Discriminant Analysis This is one of the most important topics that one can encounter in a ML course. This is a feature reduction technique. Many times we have data that has a large number of features (1000's of them). It is not possible always to consider all possible features and carry out the training. So we use a feature reduction technique to make the same problem with less number of features. If we reduce the features, will the model be efficient enough? The answer lies in the fact, that we will choose features that are having the maximum variances.

1. **Import the dataset** Download the iris flower dataset and import it as a csv file.

2. **Visualize the Dataset** There are 3 classes of iris flowers. Take all the 4 features and create plots pairwise. For example: Take feature1 and feature 2 and plot all the 150 samples in 3 different colors to visualize. You will get a total of 6 different plots like this.

3. **Normalize the dataset** Normalize the iris flower dataset for training the model. Use min max normalization.

4. **Mean and Covariance Matrix** Calculate the Mean vector. The Mean vector is the mean of all the features so it will be a 4X1 vector. Once you calculate the Mean vector calculate the Covariance Matrix among the 4 features using the formula:

$$\sigma_{jk} = \frac{1}{n-1}\sum_{i=1}^{i=n}(x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k})$$

and

$$\Sigma = \frac{1}{n-1}((X - \bar{\mu})^T(X - \bar{\mu}))$$

where n is the number of samples, $\vec{\mu}$ is the mean vector and Σ is the covariance matrix. Remember that the covariance matrix is always dXd where d are the number of features.

5. **Eigen Decomposition** Calculate the Eigen values and Eigen vectors of the covariance matrix Σ. The Eigen Vectors denote the direction of spread and the Eigen Values denote the variance among that direction.

6. **Principal Components** Sort the Eigen values in descending order. Take the first 2 Eigen Values and the corresponding Eigen Vectors.Create a Projection Matrix P of size 4X2 using these vectors.

7. **Projecting to new feature space** Project the original normalized dataset using the matrix P to a new feature space such that the data has only 2 features. Use the relation:

$$Y = XxP$$

where Y is the new projected dataset with 2 features, X is the original dataset with 4 features and P is the 4X2 projection matrix.

8. **Final Visualization** Plot the final reduced dataset Y with the 2 principal components and visualize the result. Use 3 different colours for the 3 classes.