In the previous lecture, we developed an understanding of classifiers and the classification process. In this lecture, we continue from the previous lecture where we understood and applied KNN and other classifiers for classification and we wish to understand how to use Naive Bayes Classifier when the features are continuous instead of being discrete. We also go deep into PCA and its applications and discuss PCA feature extraction and how to choose the top principal components to transform the data to a lower dimension space. We also evaluate performance of PCA features using KNN multiclass classifier as discussed in the previous lab session.

For Naive Bayes Classifier, we will be working with the Pima Indians Diabetes Dataset. This dataset contains data of 768 patients split into two classes: diabetic and not-diabetic. We wish to classify if they are diabetic or not, thus formulating a two class classification problem. The dataset is good for classification using Naive Bayes as all the features are uncorrelated.

For PCA feature extraction, we use the CIFAR-10 dataset which we got familiar with in Lab03, Experiment03. We will go through four things. First, we will see what preprocessing is required to make the data compatible for PCA feature extraction. Second, we will learn how to extract the eigenvalues and eigenvectors for the data matrix. Third, we will learn how to choose the best N principal components to reduce the dimensionality of features in the data without losing too much information. Lastly, we will evaluate the PCA features using KNN multiclass classifier and experiment for different values of N, the number of principal components.

Lab 3A

Design of the Lab

The goal of this lab is to undersand following concepts:

1. Naive Bayes algorithm
2. PCA feature extraction and dimensionality reduction

1 Experiment 1

Naive Bayes Classifier: The data set of interest is the Pima Indians Diabetes Dataset which contains records of 768 patients over 2 classes: diabetic or not-diabetic in a csv file. Each data sample has 8 features.

1. Visualize and load the Pima Indians Diabetes Dataset.
2. Apply the Naive Bayes Classifer concepts learnt in class to derive the relation between posterior probability, likelihood and prior probabilities for each feature and each class.
3. Calculate the prior Probability.
4. Calculate the likelihood.
5. Finally, build the Naive Bayes Classifier by deriving the Posterior Probability and calculate the accuracy on the test set.

2 Experiment 2

PCA feature extraction Let us consider the previously used CIFAR-10 dataset for this experiment. Here, we will develop a deeper understanding of how PCA works, its general principle of how to capture the variance in data by deriving eigenvalues and eigenvectors equations and concepts. Finally, we learn how to choose the optimal number of principal components or basis eigenvectors to project the data matrix into lower dimensinal space.

1. Load and preprocess the data to prepare it for its eigen decomposition. Reshape the dataset to fit the $N \times K^2$ dimension where $K \times K$ is the size of each image and N is the number of samples used. Use min-max scaling to normalize the dataset.
2. Understand and derive the theory and equations to find the eigenvectors and the eigenvalues of the covariance matrix of our data.
3. Write a function to extract the eigenvalues and eigenvectors for our data based on the theory derived.

4. Understand how to choose a good value of N (the principal coponents or basis eigenvectors chosen to reduce the dimensionality of data) and write the code for it.
5. Write code to project the data matrix into a lower dimensional space using the principal components derived in the previous section.
6. Visualize and compare the variance captured by the first (PCA1) and second (PCA2) eigenvectors over small part of datasets.
7. Evaluate the PCA features using KNN multiclass classifier and experiment for different values of N, the number of principal components.