You have seen the idea of characterizing an object using a set of attributes or features, which is often represented as a column vector (called feature vector). You have also seen how similarity between objects or samples are computed as distances (metric of your choice) in the resulting feature space. In this session, we will extend these ideas to create a better understanding of classifiers and the classification process.

We will first look at how representations of objects generalizes from 2 to 3 or more attributes and how this extends our understanding of a feature space. We will also look at representations of lines as classification boundaries in 2D and their extensions to higher dimensions, where they are planes (in 3D) and hyperplanes (4 or more dimensions). The representations of hyperplanes also happen to be column vectors. Linear classification involves the dot product between the hyperplanes and object vectors. We can also transform the points in a feature space to another feature space (of possibly different dimensions) by multiplying with an appropriate transformation matrix.

Once a classifier is applied to a set of points (test samples) in a feature space, the result is a class label that is assigned to each test sample. We would like to characterize how well the classifier performed. Depending on the problem there are several metrics or measures that computed to understand how well the classifier performed. These include accuracy, confusion matrix, precision, recall, etc.

Lab 2

Design of the Lab

The goal of this lab is to undersand following concepts:

1. 2D Data Visualization through a plot
2. Visualization of higher-dimensional data in lower dimensions
3. Visualizing linear classifiers
4. Metrics to measure the accuracy of a classifier

1 Experiment 1

1. Let the data set of interest be 10 samples of 3 dimesnion $\mathcal{D} = \{Samples\ of\ fruit\ dataset\}$. Each example is 3D vector and a class label (1 or 0). Plot this in 3D with class 1 as red and class 0 as green.
2. Consider a "Plane" $w = [1\ \ 0\ \ 0\ \ -312]^T$ that can classify these samples. See $W^T X_i$ for all the samples. You see that all the samples are correctly classified.
3. Consider another plane $W_2 = [0.5\ \ 15\ \ -5\ \ -200]$. What is the classification accuracy?
4. Now project the samples to 2D X-Y by multiplying each of the samples with a simple 2×3 matrix M_1, M_2 and M_3 as below.

   $$x_i = M_1 X_i$$

   where $M_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$

   Plot the new set of 10 2D points with the same colour convention as above. Plot the line $w =$ in 2D and see that the samples are separable.
5. Now project the samples to Y-Z plane with the M_2 as $M_2 =$ You will find that the samples are NOT linearly separable.
6. Write M_3 that will project the samples to $X - Z$ plane. (Pick the right matrix from the given below three choices).

   (a) Identity 3 by 3 matrix (b) True (c) wrong 2 by 3
7. Consider the 2D separable projection we had in Q. $\{x_i\}$. What if the axes were multiplied by a constant factor? (i.e., scaled). Let us keep x axis constant and multiply y axis by t. In fact this is equivalent to multiplying by

   $$\begin{bmatrix} 1 & 0 \\ 0 & t \end{bmatrix}$$

2 Experiment 2

Fighting Cancer with ML! Let us consider the Breast Cancer Wisconsin Dataset (lots of code here). Given information of the "tumor" characteristics of a certain case, we wish to classify it as benign or malignant.

1. Use the Linear Discriminator Classifier learnt in class (a trained model is given to you) for classifiying this dataset. Calculate the accuracy on the test set.
2. Now, let us calculate a different metric to analyze our results, called the confusion matrix. The confusion matrix is a simple matrix of size equal to the number of classes. An element $A_{i,j}$ corresponds to number of samples belonging to class i classified as j.
3. Let's us consider two terms, **false positive** (fp) and **false negative** (fn). A false positive error is a result that indicates a given condition exists, when it does not. A false negative error, or in short a false negative, is a test result that indicates that a condition does not hold, while in fact it does. Which error is more problematic in our case? Also, tell the number of samples that exhibit that kind of error from the confusion matrix plotted above.
4. Instead of looking at tp, tn, fp and fn directly from the confusion matrix, metrics called the precision, recall and miss rate have been defined, such that,

$$p = \frac{tp}{tp + fp}$$

$$r = \frac{tp}{tp + fn}$$

$$m = 1 - r$$

Intrestingly, accuracy can also calculated from our confusion matrix,

$$a = \frac{tp + tn}{tp + tn + fp + fn}$$

Write the code to calculate these values and check with the earlier accuracy that we computed.

3 Experiment 3

Searching and Ranking Consider the CIFAR 10 Tiny Images dataset containing 60000 images. Let us try to write a simple image search engine over this. Given a query image drawn from the test set, we attempt to find top "relevant" images from the training set using the K-Nearest Neighbour Method we learned earlier. The algorithm is simple, we rank the images present in the training set by their distance from the query images. A retrieved image is considered "relevant" if the class of retrieved image is same as the query image.

1. Do you think accuracy is a valid metric to evaluate our search engine performance? If Yes, Explain.
2. Information Retrieval experts usually use two very closely related metrics called Precision@k and Recall@k to evaluate their search engine models where k corresponds to the top-k retrievals. Let's say q is the query, U is number of images in the training set, R is the set of "relevant" images in the training set and $T(k)$ is the retrieved set of images from our algorithm.

$$p@k = \frac{|T(k) \cap R|}{|T(k)|}$$

$$r@k = \frac{|T(k) \cap R|}{|R|}$$

   - Compute the precision@k and recall@k for k = 1,3,5,10. (see this and difference from earlier precision here)
   - Plot the Precision-Recall Curve.
   - Does precision increase or decrease as we increase k, what do you expect?
   - Is there a way to make recall@k = 1 for every query for some k? What is that value of k?
   - For real search engines, is finding recall@k feasible? Why or Why not? Is finding precision@k feasible?

4 Case Study

Spam Filtering We consider the SMS Spam Dataset (sample code to preprocess into features here, please do that before providing the dataset). Assume that the dataset is already transformed into feature vectors for you and for each SMS we know if it's spam or not. The data is given in this form

- Training set of xxxx SMS's received in 2016.
- Test set is divided into three parts, xxxx messages received in January 2017, xxxx messages received in February 2017 and xxxx message received in March 2017.

Let's try to make a spam filter using this dataset. We will be working with Naive Bayes Classifier for this task and will be modifying it for our use case.

1. We wish to estimate $p(c|x)$ where x is the feature vector and c is the class label. So, we decompose this to $\frac{p(c)p(x|c)}{p(x)}$ and discard the denominator as it doesnt' depend on c. Both the terms, $p(c)$ and $p(x|c)$ can be estimated from our training set under certain assumptions.
2.