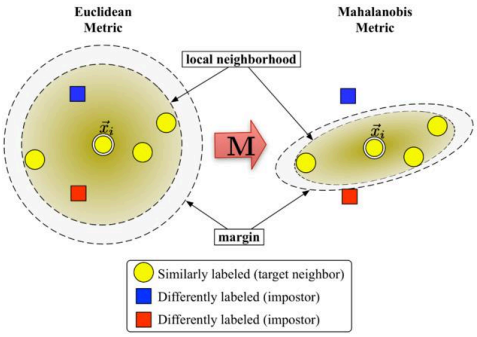
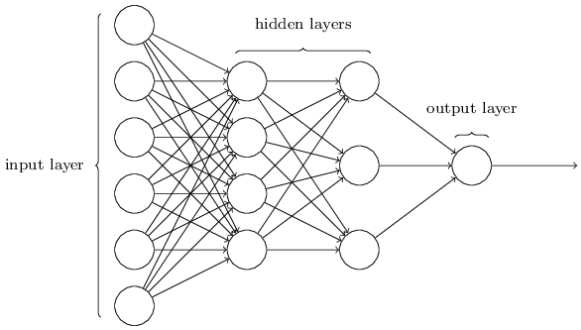
EE468/EE9SO29/EE9CS729

Pattern Recognition

Coursework on Distance Metrics and Neural Networks [50% mark]





Release on 5 Dec 2016, the report due on 23 Dec 2016 (midnight)

The course work requires Matlab programming. Use the provided face data (face.mat). In all questions, you can use any existing toolbox/code, unless specified. Some suggestions are below.

Submission instructions:

One joint report by each pair

Page limit: 4-6 A4 pages per report with 10 font size (use the IEEE standard double column paper format, either in MS word or latex).

http://www.pamitc.org/cvpr16/files/egpaper_for_review.pdf

http://www.pamitc.org/cvpr16/files/cvpr2016AuthorKit.zip

Give insights, discussions, and reasons behind your answers, on the scope of lectures. ***Quality and completeness of discussions within the page limit*** will be marked. Include formulas where appropriate, results presented in figures and their discussion.

Source code, such that the experiments and the results can be reproduced, is not mandatory, however it can help disambiguate in case the results are arguable: this can go to appendices, which do not count for the page limit.

Submit the report **in pdf** through the Blackboard system. No hard copy is needed. Write your full names and CID numbers on the first page.

If you have questions, please contact

Mr. Guillermo Garcia-Hernando (g.garcia-hernando@imperial.ac.uk)

Dr. Vassileios Balntas (v.balntas@imperial.ac.uk)

Download Wine recognition data from Blackboard. Its number of features are: class 1: 59, class 2: 71, class 3: 48. NOTE: 1st dimension out of 14 is class identifier (1-3).

Split the data into 3 sets for Training (118), Validation (20), and Testing (42). Move the class identifiers to separate vectors.

Q1. [15] Distance Metrics

A. Prepare 178 13-dimensional feature vectors by
   a) Using original unmodified feature vectors
   b) Normalizing feature vectors to unit norm L2
B. Estimate covariance matrices from training features in A.a and A.b
   a) All features from all classes
   b) Independently from features in class 1, class 2 and class 3.
C. Using means the covariance matrices analyse and compare the data distributions in different dimensions, within and across classes. Your report discussion should include numbers and figures e.g. with distributions of points in two dimensions, which support your observations.
D. For each test data point identify its nearest neighbour in training data and assign its class label to the test point. Repeat the experiment for features A.a and A.b with the metrics:
   a) L2, L1, Chi2, histogram intersection, correlation.
   b) Mahalanobis distance with covariance B.a and B.b from features in A.b Hint: convert covariance matrices into linear projections and project features from A.b, then use L2.

   Calculate the classification error as a fraction of incorrect labels assigned to test points. Identify the metric that gives the best classification results. Comment on the properties of various metrics which should be supported by examples from the experiment. Discuss effect of the dimensions' variances in distance calculations between data points.

Q2. [15] K-means clustering

Perform kmeans clustering of training features from Q1.A.b for k=3. Obtain the cluster centres and assign a different class label to each of the centres as the one that is the most representative for its members. For each test data point identify its nearest neighbour cluster centre and assign the cluster label to the test point. Calculate the classification error as in Q1.D. Repeat this experiment for features from Q1.A.b and the metrics available in Matlab's kmeans: 'sqeuclidean', 'cityblock', 'cosine', 'correlation' as well as features from Q1.D.b. Report the results in a figure. Comment on the classification performance of different metrics and compare to the results in Q1.D. Perform similar experiment with k-10 and comment on the results.

Q3. [20] Neural Network

Using Matlab Neural Network toolbox create a network, train and test with the wine data. Vary the parameters of the network to minimize the classification error. Report the observations and compare to the results from Q1 and Q2. Generate another random split of train/test/validation data and repeat the experiment.