Introduction to Machine Learning - Homework 3

Prof. Alan Yuille

Spring 2014

Due on Tuesday 27/May. 2014. Hand in papers in class.

Question 1. Decision Trees.

Describe the Decision Tree algorithm. Consider the task of deciding whether a customer is low-risk ($y = 1$) or high-risk ($y = -1$) depending on income x_1 and savings x_2. Suppose the set of tests are tests of form "*is* $x_1 > T_1$?" and "*is* $x_2 > T_2$?", where T_1 and T_2 are thresholds. The training set has low-risk ($y = 1$) points at (x_1, x_2) positions: $(2, 3), (3.5, 4), (2.5, 6), (6, 3.5), (7, 8)$ and high-risk ($y = -1$) points at $(7, 1.5), (1, 8), (1.5, 1.5), (2, 2), (3, 3)$. Derive the best decision tree where each node is pure, specifying the impurities and the test at the nodes. Use 2 as the logarithm base. Hint: At each node, you need to specify the form and the threshold of the test.

Question 2. Principal Component Analysis.

Suppose the data elements $\vec{x}$ is an M-dimensional vector. The vectors are of form $\vec{x} = a\delta_k = (0, ..., 0, a, 0, ...)^T$, where the a is in the k^{th} slot, and k, a are random variables. k is uniformly distributed over $1, .., M$ and $P(a)$ is arbitrary. Calculate the covariance matrix. Show that it has one eigenvector of form $(1, ..., 1)$ and that the other eigenvectors all have the same eigenvalue. Discuss whether PCA is a good way to select features for this problem. Hint: Use the expectation to compute covariance matrix $C = E[(\vec{x} - E[\vec{x}])(\vec{x} - E[\vec{x}])^T]$. The covariance matrix C of the signals $\vec{x}$ is of form $C_{i,j} = \lambda + \mu\delta_{i,j}$ for some λ, μ.

Question 3. Fisher's linear discriminant analysis.

Describes Fisher's linear discriminant. How is it used to discriminate between data from two classes.

Suppose we have $2M$-dimensional data from two classes. Each datapoint $\vec{x}$ in the first class is of form $\vec{x} = (x_1, ..., x_{2M})^T$ where components $x_i, i = 1, ..., 2M$ are i.i.d. from a Gaussian with zero mean and standard deviation σ. The datapoints in the second class are of form $\vec{x} = (x_1, ..., x_M, \rho + x_{M+1}, ..., \rho + x_{2M})^T$ where ρ is fixed and the x_i are also i.i.d. generated by a Gaussian with zero mean and standard deviation σ. What is the Fisher's linear discriminant between these two classes? Does this discriminant change if ρ changes?

Use PCA and Fisher's LDA for dimension reduction from 2-dimension to 1-dimension on "P3_A.txt" and "P3_B.txt". What is the projection direction of PCA and LDA on both datasets? Are they equal? Try to explain the result. Note: the last column of text file is class label (+1 or -1), and the rest are data. Ignore the class label when doing PCA.

Question 4. ICA: Signal whitening and non-Gaussianity

In MATLAB (or Octave), let's generate 1000 realizations of two i.i.d. random variables drawn from a uniform distribution U[0,1]. And let's store them in a matrix of size 2x1000 ($s = rand(2, 1000)$). Then let's mix them with the linear transformation given by $A = [2, 3; 2, 1], x = As$. Use eigendecomposition (eig function) to obtain the whitened version of x (name it xw) and s (name it sw). How much is the kurtosis of the first component $sw(1,:)$ and $xw(1,:)$? Which of them is farther from the Gaussian kurtosis? Use this definition of kurtosis, which is based on the 2nd and 4th moments. Hint: You can plot the joint distributions of x (or other variables) with the command: plot(x(1,:),x(2,:),'.').

What is the problem of Kurtosis as a non-Gaussianity criterion? Which Information Theory concept justifies the maximizing non-Gaussianity?

Question 5. Non-linear dimension reduction.

Describe the MDS algorithm. What is the relation and difference to PCA?

Briefly describe and compare the ISOMAP algorithm and LLE algorithm.