ISyE 8803 – Topics on High-Dimensional Data Analytics
Exam I – Spring 2020

Question 1 (25 points)

Predict whether a subject is running, or jogging based on time series of accelerometer data. Dataset is sampled from : https://archive.ics.uci.edu/ml/datasets/WISDM+Smartphone+and+Smartwatch+Activity+and+Biometrics+Dataset+

a) Find the optimal lambda using GCV for the smoothing splines fitted to the mean signal of the training data. Report the optimal lambda and the number of spline coefficients corresponding to the optimal lambda. (7 points)

b) Use functional PCA for dimension reduction and feature extraction from training data. How many FPC-scores are required to explain more 95% of variations? (7 points)

c) Develop prediction models to predict the activity type based on the extracted features from the training data. This can be done using random forest. (6 points)

d) Evaluate and compare the performance of the estimated prediction models using the test data. Which model do you recommend? Use the test data to evaluate the prediction performance by calculating accuracy as well as showing confusion matrix. (5 points)

Question 2 (25 points)

In this question, we are going to use optimization methods to find MLE estimates of the parameters of gamma distribution:

$$P(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

Where $\Gamma(\alpha)$ is the gamma function, α is a shape parameter, and β is a rate parameter.

a) Write down the log-likelihood function for estimating α, β. (5 points)
b) Drive the gradient and the Hessian of the log-likelihood function. (10 points)
c) A data set includes 1000 samples drawn from a Gamma distribution is given in data.csv. Report your estimated values of α, β. Plot the value of the parameters α, β versus the number of iterations. Plot the value of the log-likelihood function versus the number of iterations. For computing digamma $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ and trigamma $\psi'(\alpha)$, you can use built-in functions. (10 points)

- Accelerated Gradient descent
- Newton's method

Question 3 (35 points)

Two sets of images are given in the files 'SetA.mat' and 'SetB.mat'. Set A includes a sample of 50 noisy images with similar backgrounds. They represent the stress map of good quality silicon wafers produced from a semiconductor manufacturing process. Set B, however, is a mixed sample of both defective and non-defective silicon wafers. The main goal of this question is to see how Tucker can capture different variation modes in the data.

a) Apply Tucker decomposition with $R_1 = 3$, $R_2 = 3$, $R_3 = 3$ on the images in Set A. Use the same rank to decompose the images in Set B. (5 points)
b) Use the Tucker decomposition results from set A and the "imagesc" function in MATLAB to plot the Kronecker product of V1(:,1) and V2(:,1) as well as the Kronecker product of V1(:,2) and V2(:,2); where Vr is the rth factorizing matrix corresponding to the rth mode. Compare the resulting images with the mean image of set A and comment on the type of variations captured by each eigen-matrix (the results of the Kronecker product). (7 points)
c) For set A, plot all columns of matrix V3. What is the main source of the variability in these curves (what types of variability are captured by these vectors)? (2 points)
d) Use the Tucker decomposition results from set B and the "imagesc" function in MATLAB to plot the Kronecker product of V1(:,1) and V2(:,1) as well as the Kronecker product of V1(:,2) and V2(:,2); where Vr is the rth factorizing matrix. Compare the resulting images with the image you obtained in part b and comment on the difference on the types of variations captured. (7 points)
e) For set B, plot all columns of the matrix V3. What is the main source of the variability in these curves? (2 points)
f) Denoise both sets of images using a 2D Bspline basis and a median filtering with a 3x3 window. (5 points)
The median filtering method sorts the intensities in the $M \times N$ neighbourhood window of the reference pixel and calculates the median value of the sorted data. The original value at the reference pixel is then replaced by the median value. Figure 1 illustrates an example calculation. (5 points)

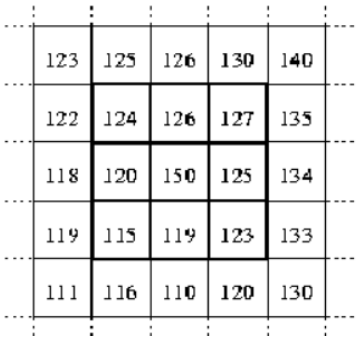

Figure 1: Median value of a local pixel neighborhood in 3×3 window mask.

- Neighbourhood values are 115, 119, 120, 123, 124, 125, 126, 127, 150
- Median is 124

g) If you were told that the noise standard deviation is 0.1, what denoising method would you pick. (Hint: you can answer this question by analyzing and comparing the image residuals.) (2 points)

Question 4 (15 points)

"Face_tensors.mat" contains tensors of 3 subjects containing multiple grayscale images per subject. Decide whether subject 3 is most similar to subject 1 or 2 using tensor decomposition. Dataset is sampled from :
https://web.archive.org/web/20190408145600/https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

a) Implement the alternative least-squares (ALS) algorithm for Tucker composition of third-order tensors (i.e. write your own code for the algorithm on the slide "Tucker Decomposition: Computation" on page 24 in the lecture notes for Module 3). (7 points)
Your codes need to have the following inputs:
A: the input tensor
rank: a 3x1 or 1x3 vector indicating the desired rank of the Tucker decomposition
tol: the tolerance for the stopping criterion
maxiter: maximum number of iterations of the ALS algorithm
And the following outputs:
B: the output tensor
err: the relative error $\|A-B\|_F/\|A\|_F$
iter: the number of iterations performed until the ALS algorithm stops
The stopping criterion is that ALS stops whenever either the change in relative error is smaller than the tolerance, i.e. delta $\|A=B\|_F/\|A\|_F$<**tol** or the maximum number of iterations "maxiter" is reached.

b) In order to reach the goal of determining the subject in tensor 3, use your own Tucker decomposition and the one provided by the tensor toolbox in MATLAB or rTensor package in R. Use AIC to choose the best rank for decomposition. Compare results obtained by the two Tucker decompositions. Present your conclusion and the steps that lead to it. (8 points)