ECE5258 Pattern Recognition (Fall 2020)
Mini-Project #1 (ver. 1.0)

Dr. Georgios C. Anagnostopoulos

Sunday 4th October, 2020

1 Objectives

The objective of Mini-Project (MP) I is to assess students' understanding in a variety of aspects pertaining to the Multi-Variate Gaussian (MVG) distribution as will as to the Quadratic Discriminant Analysis (QDA) and Linear Disciminant Analysis (LDA) classifiers.

Before putting together your work, carefully review the preparation guidelines (Section 3) and submission instructions (Section 4) that are provided.

2 Assignments

● **Task 1. [20 total points]**
Task 1 of this MP pertains to the LDA classifier. For each part/question provide a succinct, but complete derivation.

(a) **[10 points]** Assume a training set $\{(\mathbf{x}_n, \ell_n)\}_{n=1}^{N}$ of a C-class classification problem that contains N_k samples from class k, where $k = 1, 2, \ldots, C$. Derive the expression for LDA's Maximim Likelihood Estimator (MLEtor) for the covariance matrix $\mathbf{C}$, which is shared by all class-conditional MVGs, when these MVGs are of general form.

(b) **[5 points]** Repeat the previous part for the case, where the MVG distributions have independent/uncorrelated co-variates.

(c) **[5 points]** Repeat the previous part for the case, where the MVG distributions are isotropic.

● **Task 2. [60 total points]**
This task showcases training and predictions obtained via QDA and LDA classification.

Obtain the *Iris Flower Dataset* from the UCI Machine Learning Repository via https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data, which contains the sepal length, sepal width, petal length and petal width measurements (all measured in centimeters; this is the order in which they appear in the aforementioned file) for 3 subspecies of Iris flowers: *Setosa* (class 1), *Versicolor* (class 2) and *Viriginica* (class 3). Each subspecies is represented by 50 samples for a total of 150 dataset samples. For the purpose of this problem we are going to consider only the 2 petal measurements (length and width; columns #3 and #4) of each sample as our features.

For the parts shown below, let the 10 first samples of each species/class (as they appear in the file `iris.data`) constitute the *training* set, the next 20 samples of each class the *validation* set and the remaining 20 samples of each class the *test* set.

(a) **[20 points]** Train QDA classifiers using the *Iris Flower* training set for MVGs with general, independent and isotropic co-variates. Next, for each such case, produce a single plot for the feature region $[0.5, 7.5] \times [-0.5, 3.0]$ that (i) depicts the training samples that are color-coded per class, (ii) the sample mean vectors with a different marker and color-coded per class, (iii) a (50%-content, 99%-coverage) tolerance region for each class and (iv) the decision region for each class using similar shades/colors per class. State any overall remarks you may have concerning these 3 plots; are the observed results as expected and why?

(b) **[20 points]** Repeat the previous part, but, now, using LDA classifiers.

(c) **[20 points]** Present in a table the estimated misclassification rates of each of the 6 models as computed (i) on the training set and (ii) on the validation set. Among these models, identify the best performing model as reflected by its validation set performance. For this champion model, obtain an honest estimate of its performance by computing its estimated misclassification rate on the test set. State any observations and conclusions based on all these results.

● **Task 3. [20 total points]**
This task pertains to a special regularized version of QDA classification that considers MVGs of general type. In specific, let us assume that this version of QDA (let's call it *RQDA*) estimates the class prior probabilities and mean vectors of each class the same way as traditional QDA does. However, its estimators for the covariance of each class k are computed as

$$\breve{\mathbf{C}}_k \triangleq \nu \hat{\mathbf{C}}_k + (1 - \nu)\hat{\mathbf{C}}_{\mathbf{pooled}} \qquad \qquad \nu \in [0, 1]$$

where $\hat{\mathbf{C}}_k$ is QDA's MLEtor of $\mathbf{C}_k$ and $\hat{\mathbf{C}}_{\mathbf{pooled}}$ is LDA's MLEtor of the common covariance matrix shared by all class-conditional MVGs.

(a) **[10 points]** Train this RQDA classifier for $\nu \in \{0.1, 0.2, \ldots, 0.9\}$ and, for each such case, produce a decision region plot as you did in previous parts for QDA and LDA and comment on the obtained results.

(b) **[10 points]** For each value of ν considered in the previous part, select the one than yields a RQDA classifier minimizes the estimated misclassification rate on the validation set. This will be your champion RQDA classifier. Evaluate its misclassification performance on the test set and compare it to the champion QDA/LDA model of task 2(c).

3 Preparation Guidelines

Below are some general guidelines that you are asked to follow, when compiling a project report. These guidelines greatly facilitate your work's assessment by a grader and, at the same time, aim at helping you sidestep some major pitfalls that would prevent you from receiving the maximum credit for your work.

- **Task Statements:** Before attempting to address a particular task, ensure that you completely understand what is asked from you to perform and/or to produce. When in doubt, ask your instructional staff for clarifications! Also, make sure you did not omit your response to any of the parts that you have attempted. Finally, make sure that it is crystal clear, which response corresponds to which task/part.

- **Derivations & Proofs:** If you provide handwritten derivations and/or proofs, make sure you use your best handwriting. Each derivation should have a logical and organized flow, so that it is easy to follow and verify.

- **Code & Data:** The code that you author should be as well-organized as possible and amply commented. This is very useful for assessing your work, as well as for you, while you are debugging/or modifying it, or when you have to go back to it in the near future. **Caution:** You are not allowed to use any code that you have not produced without having/obtaining explicit prior permission, in which case the source(s) you have obtained this code from must be clearly indicated via comments inside your code as well in your report. You are deemed to be plagiarizing, if you fail to do so, which may have dire consequences. Finally, if a task asks you to generate data, keep them organized in a separate folder and document, *e.g.*, in a text file, the specifics of how they were generated.

- **Figures, Plots & Tables:** Plots should have their axes labeled and, if featuring several visual elements such as curves or types of points on the same graph, an appropriate legend should be used. Whether figures or tables, each one of these elements should feature a caption with sufficient information on what is being displayed and how were these results obtained (*e.g.*, under what experimental conditions or settings, etc.). You should ask yourself the question: if someone comes across it, will they understand what is being depicted? Apart from a concise description, major, relevant conclusions stemming from the display should also be included in the caption text.

- **Observations, Comments & Conclusions:** When stating observations about a particular result, do not stop at the obvious that anyone can notice (*e.g.*, "*... we see that the curve is increasing.*"). Instead, assess whether the result is expected, either by theory or intuition (*e.g.*, "*... This is as expected, because X is the integral of ...*"), or, if it is unexpected, offer a convincing reasoning behind it (*e.g.*, "*... We expected a decreasing curve ... All points to that I must have not been calculating X correctly ...*"). The latter is more preferable (*i.e.*, expect partial credit) than stopping at the obvious, which happens to be wrong (*i.e.*, do not expect partial credit). Next, descriptions and comments on results should be sufficient. Be concise, but complete. Finally, conclusions that you draw must be well-justified; vacuous conclusions will be swiftly discounted.

4 Submission Instructions

Kindly adhere to the conventions and submission instructions outlined below. Deviations from what is described here may cause unnecessary delays, costly oversights and immense frustrations related to the assessment of your hard work.

First, store all your project deliverables in a folder named `lastname_mpX`, where `lastname` should be your last name and `X` should be the number of the MP, like 1, 2, etc. The folder name should be all lower case. For example, Anagnostopoulos' folder for MP 1 would be named: `anagnostopoulos_mp1`.

Secondly, your `lastname_mpX` folder should have the following contents:

- A signed & dated copy of the Work Origination Certification page in Adobe PDF format. You can either scan such a page after you complete, date and sign it, or do so electronically, as long as your signature is not typed. If this page is missing from your report, your MP work will not be considered for assessment (grading) and will be assigned a default total score of 0/100.
- An Adobe PDF document named **lastname_report.pdf**, where, again, "lastname" should be replaced by your last name in all lower case, *e.g.*, *anagnostopoulos_report.pdf*. This document should contain your entire Mini-Project report as a single document. This will be the document that will be graded. Also, here are some important things to adhere by:
  - Your responses to the project's tasks and parts should be given in their expected sequential order, *i.e.*, task 1 part (a), task 1 part(b), etc., followed by task 2 part (a), task part(b), and so on. If you did not attempt a particular part, list it in its expected order and state that you have not attempted it as your response.
  - For tasks and parts that require you to show analytical work (*e.g.*, a derivation/proof), you are not obliged to typeset it. While it would be nice to do so, such effort may turn out to be quite time-consuming. Instead, you can scan your work into an image, as long as it is legible and well organized with a clear logical flow. When scanning your hand-written work, use a relatively low-resolution (DPI) setting, so your resulting PDF document does not become too big in size, which may prevent you from uploading your work to Canvas. Use a scanner, not a photo taken by a mobile device.
- A folder named `src`, which should contain all your code (*e.g.*, MATLAB or Python scripts, Jupyter Notebooks, etc.) that you authored and used for producing your results and the data sets that you created for this Mini-Project, if applicable. It is best, if you named your scripts according to the task and or task/part pair they produce results for.
- An optional folder named `docs`, in which you can include a MS Word version of your report and other ancillary material connected in one way or another to your Mini-Project report.

Finally, when you are done putting together all required project materials, compress your folder called `lastname_mpX` into a single `ZIP` archive named `lastname_mpX.zip` and upload it to Canvas by the specified deadline.

WORK ORIGINATION CERTIFICATION

By submitting this document, I, ____, the author of this deliverable, certify that

1. I have reviewed and understood Regulation UCF 5.015 of the current version of UCF's Golden Rule Student Handbook available at http://goldenrule.sdes.ucf.edu/docs/goldenrule.pdf, which discusses academic dishonesty (plagiarism, cheating, miscellaneous misconduct, etc.)
2. The content of this Major Project report reflects my personal work and, in cases it is not, the source(s) of the relevant material has/have been appropriately acknowledged after it has been first approved by the course's instructional staff.
3. In preparing and compiling all this report material, I have not collaborated with anyone and I have not received any type of help from anyone but the course's instructional staff.

Signature ____ Date _______________