ECE 5258 - Pattern Recognition (Fall 2016)
Mini-Project #1

Dr. Georgios C. Anagnostopoulos*

September 2016

1 Objectives

The objective of Mini-Project I is to expose the students to (i) to the k-Nearest Neighbor Classification Rule and (ii) to the Parzen Windows Classifier.

As usual, standard preparation guidelines (Section 4) and submission instructions (Section 5) are provided, which the students are expected to strictly adhere to. Finally, at the end of this document, a few, possibly helpful, references can be found.

2 Problem Setting

The classification problem to be considered is the "circle in the square" (CIS) toy problem. In words, the feature space is the unit square containing a circle centered at $(0.5, 0.5)$ with surface equal to 0.5. Uniformly sampled patterns that happen to fall inside the circle are labeled as/belong to class 1, otherwise to class 2. A variation of the problem called "noisy circle in the square" (NCIS) flips the labels of CIS samples with a probability of $p_{flip} < 0.5$ (label 1 becomes label 2 and vice versa).

In order to perform this assignment, you need to implement k-NN and the Parzen Windows Classifier (PWC) as follows:

- Author a function `knn_classify` with signature `[Ypred, PCP] = knn_classify(X, Xref, k, p, unknown_label)`, where `X` is an $N_x \times D$ matrix containing N_x D-dimensional test samples in rows, `Xref` is an $N \times D$ matrix containing N D-dimensional reference (training) samples in rows, `k` is the number of nearest-neighbors to be used, $p \geq 1$ specifies the L_p norm to be used for measuring distances and `unknown_label` is an integer that is going to be used to label test samples that cannot be labeled due to nearest-neighbor voting ties. Finally, `Ypred` is a N_x-dimensional vector containing the predicted labels of the test samples (in the same order, in which they appear as rows of `X`) and `PCP` is a $N_x \times C$ matrix containing the estimated posterior class probabilities for each class, when the total number of classes is C. Your function should assume that your reference samples are labeled as $1, 2, \ldots, C$, but it needs to establish what C is.
- Author another function, `pwc_classify`, with signature `[Ypred, PCP] = pwc_classify(X, Xref, kernel_type, spread, unknown_label)`, where the arguments of this function are similarly defined as in the case of `knn_classify`. When `kernel_type` is 1, a Gaussian interpolation kernel should be sued, while when it is equal to 2, a squared sinc should be used. Obviously, `spread` specifies the spread of the kernel to be used and needs to be strictly positive. Note that when the spread is very small relative to the pair-wise distances between training samples, computing posterior class probabilities may lead to 0/0 computations resulting in Not-A-Number (NaN) results. In such cases, it might

*georgio@fit.edu

be warranted to label the sample as `unknown_label`. Again, your function should assume that your reference samples are labeled as $1, 2, \ldots, C$, but it needs to establish what C is.

Finally, some useful functions in MATLAB that you may want to consider are:

- `rand`: to generate (N)CIS samples (generates uniform samples)
- `contourf`: to plot decision regions in the feature space (generates a filled contour plot of a bivariate function)
- `boxplot`: to create Box plots of your measurements
- `surfc`: to plot a 3D surface plot with contour lines

Note(s): Scalars are depicted in normal font, vectors in lower-case bold face and matrices in upper-case bold face. All vectors are considered column vectors. If $\mathbf{A}$ is a matrix, then $\mathbf{A}^T$ denotes its transpose. MATLAB keywords and/or code are depicted in `orange` font.

3 Assignments

Task 1. [25 total points]
This task deals with the characteristics of the NCIS classification problem.

(a) **[5 points]** Identify the prior probabilities, class conditional densities of the two NCIS classes 1 and 2.

(b) **[5 points]** Identify the posterior class probabilities of the optimal classifier for the NCIS classification problem and the decision boundaries it generates. Throughout this project assume that we are using a 0-1 loss matrix.

(c) **[5 points]** Why do we put the constraint $p_{flip} < 0.5$? More specifically, what happens when $p_{flip} = 0.5$? Finally, what happens when $p_{flip} > 0.5$?

(d) **[5 points]** Compute the Bayes error rate as a function of p_{flip}.

(e) **[5 points]** Draw 100 NCIS samples for $p_{flip} = 0.0, 0.1, 0.2, 0.3$ and 0.4 and plot them in the feature space (5 plots) along with the optimal decision boundary to verify your answers in parts Part (a) and Part (b).

Task 2. [15 total points]
Author a function that generates labeled samples (arranged in rows of a data matrix) from the NCIS problem given p_{flip} and the number of samples N. Then, draw a 100-point labeled sample from the NCIS population with $p_{flip} = 0.1$.

(a) **[5 points]** Provide plots of the feature space depicting the optimal decision boundary, the training patterns used and the decision regions generated by a 1-NN classifier for $N = 10, 30, 50, 75, 100$ training patterns and state your observations.

(b) **[5 points]** Repeat Part (a) with a 5-NN classifier and compare it to the results youve obtained in the previous part.

(c) **[5 points]** Repeat Part (a) but now use the L_∞ norm to measure distances instead and state your observations. Also explain the particular shape of the decision boundaries you obtain for $N = 10$.

Task 3. [15 total points]
Draw 100 labeled samples from NCIS with $p_{flip} = 0.1$ to form a test set.

(a) **[5 points]** For 30 times do the following: randomly draw $N = 10$ training samples from the same population. For each time use these patterns as a training set for a 1-NN and calculate its estimated error rate. Create a Box plot at $N = 10$ of the 30 estimated error rates youve obtained.

(b) **[5 points]** Repeat Part (a) for training set sizes $N = 10^a$ for $a = 1.2, 1.4, \ldots, 2.8, 3.0$ (obviously, round N to the nearest integer). Create a plot containing all 11 Box plot (including the one in Part (a)) and use a log_{10} scale for N (depicted on the x-axis). State your observations on the effect of N on the estimated misclassification error.

(c) **[5 points]** Repeat Part (a) and Part (b) using a 5-NN classifier and compare the obtained results with the ones in the previous to parts.

Task 4. [25 total points]
Draw a 100-point labeled sample from the NCIS population with $p_{flip} = 0.1$.

(a) **[10 points]** Provide plots of the feature space depicting the optimal decision boundary, the training patterns used and the decision regions generated by a PWC classifier with Gaussian kernel for $N = 10, 30, 50, 75, 100$ training patterns and state your observations. For each case of N showcase the effect of very small kernel spread value s, an appropriate value and a very large value. Comment on all your results.

(b) **[10 points]** Repeat Part (a) using the squared sinc kernel and compare with the results in Part (a).

(c) **[5 points]** Plot the posterior class probability functions in $3D$ as a surface for class 1 along with its contour lines underneath the $3D$ plot for their values equal to $0.25, 0.5$ and 0.75 in the case of the PWC classifiers of Part (a) and comment on them.

Task 5. [10 total points]

Draw 3 sample sets of sizes 30 (training), 100 (validation), 100 (testing) from the NCIS population with $p_{flip} = 0.1$.

(a) **[5 points]** Use the validation set to find the optimal value of nearest neighbors k for the k-NN classifier (using Euclidean distances, of course); use $k = 1, 2, \ldots, 15$. Plot the estimated error versus k. Refer to the winning k-NN classifier as the *champion* k-NN model and provide both small- and large-sample confidence intervals for its misclassification error. Furthermore, for the aforementioned 15 models plot the misclassification error as calculated using the validation set versus the misclassification error as calculated using the training set. Comment on your results. How does it compare to Bayes error?

(b) **[5 points]** Use the validation set to locate the optimal value of the kernels spread s for a PWC model using Gaussian interpolation kernel. Plot the estimated error versus the spread s. Refer to the winning PWC classifier as the *champion* PWC model and comment on your results. How does it compare to Bayes error?

Task 6. [10 total points]

Assume a four-class classification problem, whose populations are isotropic bivatiate Gaussians with common variance σ^2 and means $(d, d), (-d, d), (-d, -d)$ and $(d, -d)$, where $d \geq 0$. Derive a closed-form expression for the Bayes error of this problem in terms of the univariate Gaussian CDF $\Phi(x; \mu, \sigma)$. Show all the steps of your derivation.

Hint(s): As an intermediate step, it is easier to compute the accuracy, instead of the error, of the optimal classifier. Also, "isotropic" implies that the Gaussian variates of the same Gaussian distribution are independent.

4 Preparation Guidelines

Below are some general guidelines that should be followed, when compiling a Mini-Project report. I strongly encourage you to stick to them, so that you receive full credit for your correct responses.

- **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, come to ask me 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.
- **Material Presentation:** The material you generate for each task should be presented in your report in proper sequence by task and part number. If you have not attempted or completed a part, you need to indicate so at the appropriate location of your report.
- **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 if you have to go back to it in the near future. Driver scripts (scripts that may call other scripts or functions to accomplish a main task) should be named according to the part, for which they generate their numerical and/or graphical results; for example, the driver program for Task 1, Part (b) should be named **task1b.m**. Regarding the data you generate, keep them organize and document somehow (*e.g.*, in a text file) the specifics of how they were generated. **Caution:** You are not allowed to use any code and/or data that you have not produced without my explicit prior permission, in which case the sources you have obtained these from must be clearly indicated in your code or data description as well in your report. You are deemed to be plagiarizing, if you fail to do so, which may have dire consequences to your academic tenure here at Florida Tech!
- **Figures, Plots & Tables:** Plots should have their axes labeled and, if featuring several graphs, an appropriate legend should be used. Whether figures, plots 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 about 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.

5 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 Mini-Project deliverables in a folder named **lastname_mpX**, where "lastname" should be your last name and X should be the number of the Mini-Project, like 1, 2, etc. The folder name should be all lower case. For example, my folder for Mini-Project 1 would be named *anagnostopoulos_mp1*.

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

- 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 keep in mind:
  - The report must include a signed & dated copy of the Work Origination Certification page. You can either scan such a page and include it in your document, or sign and date it electronically, as long as your signature is not typed. If this page is missing from your report, or it does not comply with the aforementioned conditions, I reserve the right not to accept the report and assign a score of 0/100 for the relevant Mini-Project.
  - The Mini-Project may ask you to produce a variety of derivations, proofs, etc. You are not obliged to type such parts; it would be nice, but I realize that such effort would be quite time-consuming. Instead, you can import scanned images (or whole pages) of your handwritten work, as long as they are legible and well organized, so that the report has a clear logical flow. For example, it has to be clear where this hand-written work corresponds to (*e.g.*, which assignment it addresses).
  - Having said all this, you may want to consider to print out your typed work, appropriately merge it with any handwritten pages (don't forget the signed Work Origination Certification page!) and then scan the whole compilation into a single PDF report, say, in the Library. **Caution:** when scanning, 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.
- A folder named **src**, which should contain all your MATLAB scripts that you authored and used for producing your results and the data sets that you created for this Mini-Project, if applicable.
- 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.

Next, compress your **lastname_mpX** folder into a single ZIP archive named **lastname_mpX.zip**; *e.g.*, mine would be called *anagnostopoulos_mp1.zip*.

Finally, upload your ZIP archive to Canvas by the specified deadline using the appropriate drop box. You are done!

Work Origination Certification

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

1. I have reviewed and understood the Academic Honesty section of the current version of FITs Student Handbook available at http://www.fit.edu/studenthandbook/, which discusses academic dishonesty (plagiarism, cheating, miscellaneous misconduct, etc.)
2. The content of this 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 courses instructor.
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 from the courses instructor.

_____

Full Name (please PRINT)

_____

Signature Date