Bonus Assignment: Parzen Window and Bayesian classification

Due Sunday, April 19, 2020 at 11:59 pm EST

Description

In this problem, we will explore the use of Parzen window to non-parametrically estimate $p(\boldsymbol{x}|\omega_k)$, where $\boldsymbol{x}$ is the feature vector. This classifier will be used classify data entries in the Pima Indians Diabetes dataset from homework assignment 6 (note: for this dataset $\boldsymbol{x} \in \mathcal{R}^8$).

What to submit

Create a folder: `psB_LastName_FirstName` and add in your solutions:

`psB_LastName_FirstName/`

- input/ - input images, videos or other data supplied with the problem set
- output/ - directory containing output images and other files your code generates
- psB.m - code for completing each part, esp. function calls; all functions themselves must be defined in individual function files with filename same as function name, as indicated.
- *.m Matlab/Octave function files (one function per file), or any utility code. It's good practice to end your functions with the clause "end"
- psB_LastName_FirstName_debugging.m – one m-file that has all of your codes from all the files you wrote for this assignment. It should be a concatenation of your main script and all of your functions in one file (simply copy all the codes and paste them in this file). In fact, this file in itself can be executed and you can regenerate all of your outputs using it.
- psB_report.pdf - a PDF file with all output images and text responses

Zip it as `psB_LastName_FirstName.zip`, and submit on Canvas.

Guidelines

1. Include all the required images in the report to avoid penalty.
2. Include all the textual responses, outputs and data structure values (if asked) in the report.
3. Make sure you submit the correct (and working) version of the code.
4. Include your name and ID on the report.
5. **Comment your code appropriately**.
6. Please avoid late submission. Late submission is not acceptable.
7. Plagiarism is prohibited as outlined in the Pitt Guidelines on Academic Integrity.
    a. **Please don't share your codes with any of your colleagues.**

Questions

1. **Non-parametric estimation of $p(x|\omega_k)$ and Bayesian classification**

- Please read section 2.5.6 in the provided supplemental materials and consider the use of Gaussian kernel to estimate the likelihood of an $l-\text{dimensional}$ testing vector $\boldsymbol{x}$**:**

$$\hat{p}(\boldsymbol{x}|\omega_k) \;= \frac{1}{N_k}\sum_{i=1}^{N_k}\frac{1}{(2\pi)^{l/2}h^l}\ \exp\left(-\frac{\left(\boldsymbol{x}-\boldsymbol{x}_{i|k}\right)^T\left(\boldsymbol{x}-\boldsymbol{x}_{i|k}\right)}{2\,h^2}\right)$$

where N_k is the number of training samples belonging to class ω_k and $\boldsymbol{x}_{i|k} \in \mathcal{R}^l$ is the i^{th} training vector that belongs to ω_k.

- Divide your dataset into training and testing sets.
  i. Consider different values of N, the size of the training set. ($N \;=\; 30\%;\; 50\%;\; 70\%;$ and 90% of the dataset size).
  ii. Randomly pick N vectors for the trainig set and the remaining $768 - N$ vectors for testing.

- Consider different values of h, e.g., $h \;=\; 0.1;\; 0.4;\; 0.7;\; 1.0;\;$ and 1.5.

- Assume that $p(\omega_0) = 0.65$ and $p(\omega_1) = 0.35$; consider the classification rule given in eq. (2.114), pg. 65 (Don't use the risk parameters) and make predictions (classify) for the testing vectors in your testing set.

- Prepare a table that summarizes your classification accuracies for different pairs of (N, h).

- Your report must document the following:
  i. The classification accuracy for each pair (N, h).
  ii. Your observations and comments on the results.
  iii. Compare this approach to the naive Bayes classification implemented in assignment 6 in terms of accuracy and complexity.