

Problem Set 5 (b)

MAS 622J/1.126J: Pattern Recognition and Analysis

Due: Friday, November 14th, 2008

Second submission: Monday, November 17th, 2008

[Note 1: This problem set consists of two problems that all use the same data set. The training and testing data can be downloaded from the class web site. There are two classes with equal prior probabilities and scalar-valued features. Each problem focuses on a different classification technique. Each classification technique has free parameters you must estimate using *leave-one-out validation*.

[Note 2: All instructions to plot data or write a program should be carried out using either Python accompanied by the `matplotlib` package or Matlab. Feel free to use either or both, but in order to maintain a reasonable level of consistency and simplicity we ask that you do not use other software tools.]

Problem 1: Generalized Linear Discriminant

- a. Use the generalized linear discriminant (DHS second edition section 5.3) in conjunction with the pseudoinverse technique (DHS second edition section 5.8.1) to classify the data. That is, for a polynomial of degree k , let

$$\mathbf{y} = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix}$$

be the augmented data vector. Let the margin vector be $\mathbf{b} = \mathbf{1}$, so the weight vector \mathbf{a} can be computed from the pseudoinverse of the data matrix. The degree k is undetermined and is to be estimated using leave-one-out validation.

To estimate k , compute the evidence curve $v(k)$ and then choose the k that maximizes $v(k)$. For a particular value of k , $v(k)$ should be computed as follows:

- Break the training data set into two parts, A and B , where B contains only a single sample.

- Compute the generalized linear discriminant from A only.
- Determine if B (the sample left out) is classified correctly.
- Repeat this process for every possible choice of B . Define $v(h)$ as the number of times the sample left out was correctly classified.

In this process, it is helpful to multiply the augmented data from the second class by -1 (DHS second edition section 5.4).

Plot the evidence curve for $k = \{1, 2, \dots, 20\}$. Pick three values of k from different parts of the curve and plot the sign of the discriminant function over the scalar feature space; 1 when choosing the first class, -1 otherwise.

- Design a minimum error rate classifier using the k that maximizes $v(k)$. What is this classifier's performance on the test data set?

Problem 2: k -Nearest-Neighbor

- Use the k -nearest-neighbor method to classify the data (DHS second edition section 4.5.4). As before, the parameter k is undetermined and must be estimated using an evidence curve $v(k)$. For a particular value of k , compute $v(k)$ as follows:
 - Break the training data set into two parts, A and B , where B contains only a single sample.
 - Determine if B is correctly classified by its k -nearest-neighbors estimate based on A .
 - Repeat this process for every possible choice of B . Define $v(h)$ as the number of times the sample left out was correctly classified.

Plot the evidence curve for odd values of k in the range 1 to 19. Pick three values of k from different parts of the curve and plot the resulting discriminant function over the scalar feature space; 1 when choosing the first class, -1 otherwise. Intuitively speaking, what is the best value of k to use?

- Design a minimum error rate classifier using the k that maximizes $v(k)$. What is its performance on the test data?