POLITECNICO DI MILANO

Dipartimento di Elettronica e Informazione

Pattern Analysis and Machine Intelligence

M. Matteucci, L. Malagò, D. Eynard

10/07/2012

Answer the following questions identifying the key aspects and try not to exceed the 1.5 page limit per question.

- Use only the 3 sheets provided by the teacher
- **Write your answers on different sheets according to the question**
- Write your name and Student ID on each sheet you turn in
- English is the official language, however Italian is allowed
- Either pen and pencil are allowed no other technical mean to support yourself is allowed

In case you have special needs (e.g., being graded within a given time) please ***tell it to the teacher!***.

Question 1: Linear Discriminant Analysis (Answer on sheet 1 – 10 points)

With reference to the Linear Discriminant Analysis method/model for classification, describe:

1. Assumptions and analytical form of LDA
2. Its (dis)advantages wrt plain Linear Regression (on the indicator matrix), Logistic Regression, and Optimal Separating Hyperplanes
3. How do we train this classifier from data?
4. What is the difference between Linear Discriminant Analysis and Quadratic Discriminant Analysis
5. Does it work for multi-class problems? If yes how? If no, why?

Question 2: Selection & Projection (Answer on sheet 1 – 6 points)

Sometimes the space data live in is bigger than what we actually need for effective classification; for instance, data might use only a subspace of the input space, e.g., an affine subset or manifold, or we could have some input dimensions, i.e., features, not be useful at all.

1. An example of the first situation is Fisher projection in LDA. Describe it in details and discuss about its use.
2. An example of the second is the use of the parameters of Logistic Regression to understand how much each single input affects the probability of the output. Describe this for the case of binary classification.
3. Why it is useful to reduce the dimensionality of the input space in LDA or LR? Would be the performance of the classifier affected somehow by this reduced dimensionality? Why?

Question 3: Clustering (Answer on sheet 2 – 8 points)

Suppose you want to evaluate the results of a clustering algorithm and you plan to do that by calculating the sum of squared errors.

1. If you were to write an SSE function that calculates the sum of squared errors, what would its inputs and output be?
2. How would your function calculate the SSE? (code is optional, you can just describe it with your own words)
3. After running your function, you obtain the result SSE=115.3. How can you evaluate whether this is a good or bad result? What would you compare this result with?
4. Now suppose you have ground truth for your dataset (i.e., the true association between points and clusters). You run two different clustering algorithms on the same data and obtain the following results:

	SSE	Accuracy
Algorithm1	115.3	87%
Algorithm2	1285	95%

What is the meaning of these results? Which algorithm is better? What can you tell about the characteristics of the dataset?

Question 4: Regression (Answer on sheet 3 – 8 points)

Suppose you have a dataset with 2201 observations, each of them corresponds to a text document. The input variables corresponds to the number of occurrences of each word in the document. Follows that the X matrix, with dimensions 2201 times 101.709 is rather sparse, since each document contains only a subset of words of the dictionary.

The output variable is a continuous variable and describes how relevant is each document, with respect to a specific topic. Large (positive) values identify, at different levels, interesting documents, while small (or negative) values documents wich are not (so) related to the current topic.

The guy who was working on this regression problem before you first tried the ordinary least squares estimator for the output variable, but the results where not so good, in term of prediction error on the test set. So they asked you for some help. Some more details: the method that will be used has to be somehow efficient, since some task have larger number of documents, and such predictions have to me evaluated frequently. Moreover, it would be very nice (but not necessary at first) to be able to identify a set of relevant keywords for the given task, whose presence could help in identifying concepts or words related to the topic.

What would you do in this situation? Present a solution, motivate your choices, list plus and possible drawbacks of your approach. Suppose your first solution does not perform very well either. Which could be the reason? Then, what would be your second choice, and why? Motivate your answer, and do not write more than a page and a half.