Pattern Analysis and Machine Intelligence

Matteo Matteucci, Davide Eynard

19/02/2016

Abstract

Please write Ex1 and Ex3 on one sheet and Ex2 and Ex4 on a different one. Indicate clearly which exercise and question you are answering in you manuscript.

1 Generative vs. Discriminative Models (8 points)

A classical distinctions between classification models is the generative vs discriminative one. Answer the following about this distinction.

(a) What are discriminative and generative models? How do they differ? Which one should be preferred and why?

(b) Is Logistic Regression a discriminative model or a generative one? Why?

(c) Is Linear Discriminant Analysis a discriminative model or a generative one? Why?

(d) Is Support Vector Machines a discriminative model or a generative one? Why?

Let us consider the Support Vector Machine model for classification

(e) What is a Support Vector Machine? How is it defined (i.e., the optimization problem it solves) and how is it trained (i.e., the optimization problem is solved to train it)? How does the solution looks like and what this has to do with the name of the model?

(f) What is the kernel trick and how can it be applied to Support Vector Machines (i.e., what do you need to change with respect to the original algorithm)?

2 Linear regression (8 points)

(a) You have a dataset with n=1000 observations and try to fit different models on the data:

- a linear regression model, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$
- the polynomial regression model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$
- a smoothing spline (i.e. an even more flexible model than the previous two)

For each of the three models, you calculate both training and test RSS. How would you expect the values of RSS to be (both in the training and in the test case), supposing that the true relationship between X and Y is (a) linear or (b) cubic?

(b) What is the *additive assumption* in a linear regression model? Show how you would detect a possible interaction between variables and how you would model it. Finally explain, with an example, how the new model would take this interaction into account.

3 Classification (8 points)

In an attempt to convince students to regularly attend his classes, a professor and his TA collected data from their own students and studied the relationship between Class Hours (CH=X_1, i.e. the total number of hours attended in class for a given subject), Study Hours (SH=X_2, i.e. the total number of hours spent at home studying for that subject), and the probability Y of passing the final exam. Roughly, the more hours a student spends on a subject the higher the probability, but class hours tend to be worth more than the ones spent at home. After fitting a logistic regression, the following coefficients were estimated: $\hat{\beta}_0 = -8.75$, $\hat{\beta}_1 = 0.25$, and $\hat{\beta}_2 = 0.1$.

(a) Estimate the probability for a student with CH=35 and SH=20 to pass the exam

(b) Estimate how many hours of SH a student who could attend only CH=25 hours of classes needs to study to have that same probability to pass the exam

According to Statistical Decision Theory, the lowest error for a classifier is the Bayer Error, i.e., the error obtained by the Bayes Classifier, i.e., the classifier which selects the class according to

$$\arg\max_j P(Y = j | X = [x_1, x_2, \ldots, x_p])$$

(c) Under which posterior distributions does the Logistic Regression classifier obtains the lowest average error rate in the case of (i) binary classes and in the case of (ii) multiple classes (e.g., K classes)?

(d) What is the expected average error for the Bayes Classifier?

4 Clustering (8 points)

Given the two figures below (where blue diamonds are points from the same dataset and red dots ones different centroids starting points), calculate and show the different steps of the K-Means algorithm for each of the examples in the following way:

- At each step, specify the initial positions of the centroids
- Without actually calculating it (unless it is needed to verify distances you cannot tell apart at a glance), for each step specify which centroid the various dataset points belong to
- After you have assigned points to the different centroids, calculate their new positions and proceed to next step

NOTE: in both figures, in (1,1), you have a diamond AND a dot!!!

Tell how many iterations the algorithm needs to converge, compare its behavior in the two cases and write a comment about it: is the final situation the one you might expect/desire? If not, explain why.