FEUP | PDEEC0049 Machine Learning 2013/14 (1st Semester)
18-782 PP: Machine Learning 2013/14

Exam - 28/01/2014 Duration: 2h30min

1. We are dealing with samples x where x is a single value. We would like to test two alternative regression models:

Model1: y=ax+e

Model2: $y=ax+bx^2 + e$

a) Assume we have n samples in the training set $x_1, x_2, \ldots, x_n$ and the corresponding y values $y_1, y_2, \ldots, y_n$. You can use *a* (but not *e*) in the equation for b.

b) Which of the two models is more likely to fit the *training* data better? Why?

i) model1
ii) model2
iii) both fit equally well
iv) impossible to tell

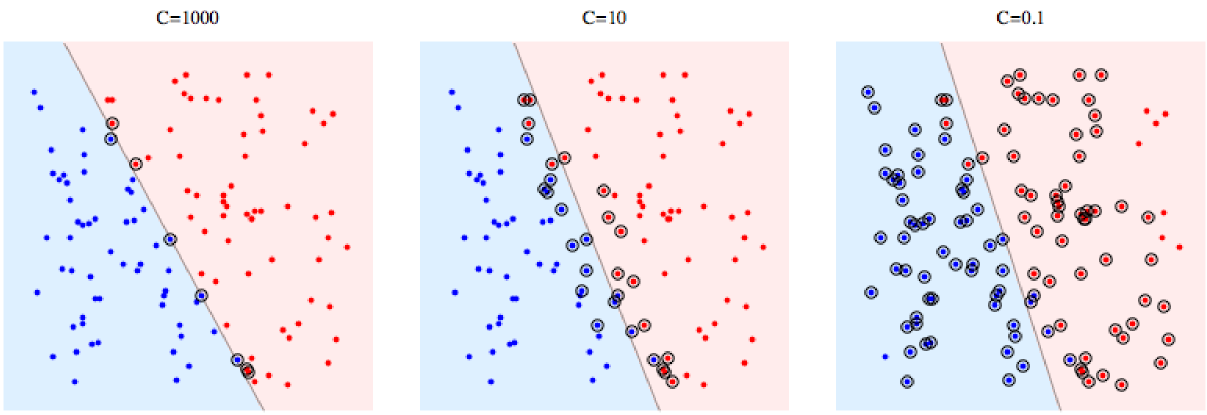
c) Which of the two models is more likely to fit the *test* data better? Why?

i) model1
ii) model2
iii) both fit equally well
iv) impossible to tell

2

a) Compare the <u>KNN and SVM classifiers</u> in terms of training and testing time. Justify your answer.

b) Consider the <u>SVM linear models</u> shown below, which were trained with different C values on the same dataset. Circled points show support vectors. Explain briefly the role of the C parameter in these SVM models. What happens when C=0 and when $C \rightarrow \infty$?



3. Bayes Decision Theory

Let the features $X=(x_1, x_2, \ldots, x_d)$ be binary valued (1 or 0). Let p_{ij} denote the probability that the feature x_i takes on the value 1 given class j. Assume that there are only two classes and that they are equally probable. Let the features be conditionally independent for both classes. Finally, assume that <u>d is odd</u> and the $p_{i1}=p>1/2$ and $p_{i2}=1-p$, for all i. Show that the optimal Bayes decision rule becomes: decide class one if $x_1+x_2+\ldots+x_d>d/2$, and class two otherwise.

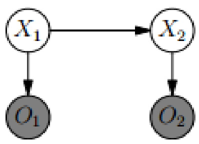
4. Trimmed Nearest Neighbor

The *relative neighborhood graph* of a set of *n* points *S* in space is defined as follows. The vertices in the graph are the *n* given points of *S*. Let L_{ij} denote the intersection of the two circles determined by two points x_i and x_j, such that they are centered at x_i and x_j, respectively, and each circle has radius equal to the distance between x_i and x_j. If no other points of *S* fall strictly inside L_{ij} then x_i and x_j are joined by an edge in the graph.

In the *relative neighborhood graph editing* algorithm all data points in *S* that have their relative neighbors in the same class are removed (in parallel) from *S*. The resulting condensed set (remaining points) is denoted by *C*. Prove or disprove that the relative neighborhood graph editing algorithm is *training-set consistent*. Recall that training-set-consistent means that using C as the classifier the *nearest-neighbor* decision rule classifies all points in S correctly.

5. HMM

Consider the following Hidden Markov Model.



X_1	$\Pr(X_1)$
0	0.3
1	0.7

X_t	X_{t+1}	$\Pr(X_{t+1}\|X_t)$
0	0	0.4
0	1	0.6
1	0	0.8
1	1	0.2

X_t	O_t	$\Pr(O_t\|X_t)$
0	A	0.9
0	B	0.1
1	A	0.5
1	B	0.5

Suppose that O_1 = A and O_2= B is observed.

a) Use the Forward algorithm to compute the probability distribution $\Pr(X_2, O_1 = A, O_2 = B)$. Show your work. You do not need to evaluate arithmetic expressions involving only numbers.

b) Compute the probability $\Pr(X_1= 1|O_1= A, O_2= B)$. Show your work.

6. Unconventional Classification Setting

Consider a learning setting where you do not know the individual label of each observation in the training set. Instead, training data is provided as a set of bags of observations, where the bag is labelled positive or negative. Each bag contains many instances. A bag is labeled negative only if all the instances in it are negative. A bag is labeled positive if there is at least one instance in it which is positive.

Under this setting, suggest how you would learn a binary classifier to label individual instances correctly.