TÉCNICO LISBOA

LEIC-T 2024/2025
Aprendizagem - Machine Learning
Homework 2
Deadline 7/10/2024 21:00
Submit on Fenix as pdf

I) Bayesian Classifier (8 pts)

Given a data set describing a sample

x_1	x_2	Class
0.6	0.4	A
1	1.1	A
1.6	1.5	A
1.8	1.8	A
2	0	B
2	1	B
3	0	B
4	1.2	B

And the query vector $x = (x_1, x_2)^T = (1,2)^T$

a) (3pts) Compute the most probable class for the query vector, under the Naive Bayes assumption, using 1-dimensional Gaussians to model the likelihoods. (Hint, the likelihood is described of each class is described by two Gaussians (Normal Distributions, each distribution is defined by a mean value and standard deviation.)
You can (should?) use your computer with Python, NumPy, MATLAB, Octave, Mathematica, etc. (whatever tool/language you like). Please indicate your results step by step.

b) (3 pts) Compute the most probable class for the query vector assuming that the likelihoods are 2-dimensional Gaussians. Are the results from 1 a) and 1 b) the same? Why are the same or not (one sentence, no mathematical proof required)

c) (1 pts) Given a data set

x_3	Class
0	A
1	A
1	A
0	A
1	B
1	B
0	B
1	B

And the query vector $x3 = True = 1$

TÉCNICO LISBOA

LEIC-T 2024/2025
Aprendizagem - Machine Learning
Homework 2
Deadline 7/10/2024 21:00
Submit on Fenix as pdf

Compute the most probable class, with x3 being a categorial class 1=True, 0=False.

d) (1pts) Given a data set describing a sample combining the data set before

x_1	x_2	x_3	Class
0.6	0.4	0	A
1	1.1	1	A
1.6	0.5	1	A
1.8	1.8	0	A
2	0	1	B
2	1	1	B
3	0	0	B
4	1.2	1	B

x1 and x2 are dependable and x3 is independent of x1 and x2. x3 is a categorial class.
And the query vector $x = (1,2,1)^T$ Compute the most probable class and indicate the estimated relative probability.

Hint,

$p(A, xquery) = p((1,2)|A) \cdot P(1|A) \cdot p(A)$

$p(B, xquery) = p((1,2)|B) \cdot P(1|B) \cdot p(B)$

you have already computed the values in b) and in c)

$P(1|A)$ =card(A.1)/card(A)=2/4
$P(1|B)$ =card(A.1)/card(B)=3/4

II Software Experiments (2pts)

Download the jupyter notebook HM2_kB.ipynb.

Split the data using the command (in the notebook)

```
digits = datasets.load_digits()
X, y = digits.data, digits.target
```

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=**0.4**,stratify=y,random_state=**your_group number**)

And do the experiments with kNN with k=3, k=30, and GaussNB as indicated in the file and *indicate the accuracy results*.
Load the wine data set wine = datasets.load_wine() and redo the experiments, *indicate the new accuracy values*.

Which method kNN, k=3,k=30, GaussNB gives better result for which data set? Do you know why? Please indicate in one/two sentence/s.