UNIVERSITY OF EASTERN FINLAND

Please carefully read and follow the general instructions regarding exercises. Failing to meet the requirements might lead to penalties. https://elearn.uef.fi/mod/page/view.php?id=293750

If you suspect that something is wrong with some exercise question, please contact the lecturer.

If you face persistent issues while working on an exercise, do ask for help, e.g. during a course meeting or by contacting the lecturer via email.

Consider the dataset consisting of twelve training instances (ids 1–12, on the left) and seven test instances (ids 13–19, on the right), with four variables v_1–v_4 and a class label y that can take one of three values, shown in Figure 1.

id	v_1	v_2	v_3	v_4	y
(1)	1	4	−1	0	●
(2)	3	6	−2	0	●
(3)	7	5	−6	0	●
(4)	2	5	1	0	●
(5)	0	4	6	0	●
(6)	4	6	2	0	●
(7)	6	2	−1	0	●
(8)	8	3	−6	0	●
(9)	7	1	1	1	●
(10)	3	2	6	1	●
(11)	5	1	2	1	●
(12)	1	3	2	1	●

id	v_1	v_2	v_3	v_4	y
(13)	7	4	−6	0	●
(14)	3	3	1	0	●
(15)	4	5	−2	0	●
(16)	6	5	6	1	●
(17)	0	3	−1	0	●
(18)	9	2	1	1	●
(19)	1	1	6	0	●

Figure 1: Dataset consisting of twelve training instances (left) and seven test instances (right)

Problem 1 (Naive Bayes classifier). We transform the dataset by mapping it to two dimensions as follows, where the first dimension is the sum of attributes $v_1 + v_3$, while attribute v_4 stands as the second dimension.

a) Write down the transformed training dataset.

We want to train a Bayes classifier on this transformed dataset. The aim is to tell red instances apart from blue and yellow ones.

We consider the first transformed dimension, denoted as v'_1 (i.e. $v'_1 = v_1 + v_3$), as a continuous variable and model it with a Gaussian distribution. The second dimension, denoted as v'_2 (i.e. $v'_2 = v_4$) is binary and should be modelled using a Bernoulli distribution, applying Laplacian smoothing with $\alpha = 1$.

On one hand, we can build a model with three components, that is, represent the blue and yellow classes with distinct distributions, then binarize the predicted class labels.

b) Compute the parameters of a naive Bayes classifier for the dataset while considering the three distinct classes.

c) Use this classifier to predict the class labels of the test instances, then binarize these predictions.

On the other hand, we can directly build a model with two components, representing the blue and yellow classes together with combined distributions.

d) Compute the parameters of a naive Bayes classifier for the dataset while considering two distinct classes (red vs. combined blue and yellow).

e) Use this classifier to directly predict binary class labels for the test instances.

f) Discuss the differences between the two approaches.

```
### Data matrix
## id, v1, v2, v3, v4, y
## training instances
1, 1, 4, -1, 0, 0
2, 3, 6, -2, 0, 0
3, 7, 5, -6, 0, 0
4, 2, 5, 1, 0, 1
5, 0, 4, 6, 0, 1
6, 4, 6, 2, 0, 1
7, 6, 2, -1, 0, 1
8, 8, 3, -6, 0, 1
9, 7, 1, 1, 1, 2
10, 3, 2, 6, 1, 2
11, 5, 1, 2, 1, 2
12, 1, 3, 2, 1, 2
## test instances
13, 7, 4, -6, 0, 0
14, 3, 3, 1, 0, 0
15, 4, 5, -2, 0, 1
16, 6, 5, 6, 1, 1
17, 0, 3, -1, 0, 1
18, 9, 2, 1, 1, 1
19, 1, 1, 6, 0, 2
```