KI Project C

Martijn Dekker 6013368
Egor Dmitriev 6100120
Cody Bloemhard 6231888
LaTeX

January 15, 2019

1 Exercise 1

1.a ex1-a

number of input features + what they represent Images are converted into 2 dimensional arrays with size 28x28 for digits and 60x74 for faces. In these arrays every element represents one pixel of the image.

the possible values and what these represent For every element in the array there are two options. For digits a 0 indicates a white pixel and a 1 indicates a gray/black pixel. For faces a 0 indicates no edge and a 1 indicates an edge.

the output labels and what they represent For digits there are 10 possible labels. These are the numbers 0 to 9, and they represent the numbers 0 to 9. For faces there are 2 possible labels: 1 and 0, because it's either a face or it's not. Here a 1 represents a face, and a 0 represents it's not a face

the frequency or probability distributions over the output labels

1.b ex1-b

Most frequent counts for every possible label how often it appears and then uses the most common label to classify an input. Naive bayes uses the log-joint distribution, so it also looks at how often a label is given, but unlike most frequent, it does so for every feature and normalizes the results it found.

1.c ex1-c

Most frequent does use supervised learning because it directly looks at the labels given to the training data, and uses the most common label to classify all future inputs. Naive bayes does also use supervised learning because it looks at the labels by counting how often a level A is given to a feature B.

2 Exercise 2

2.a ex2-a

What is important to know is that for digits there are ten possible labels and for faces there are only two possible labels. This means that if you randomly

Classifier	data	Validation	testing	k
Most frequent	digits	14%	14%	x
	faces	56%	53%	x
Naive Bayes	digits	69%	55%	2
	faces	77%	75%	2

Classifier	data	Validation	testing	k
Naive Bayes	digits	74%	65%	0.1
	faces	85%	82%	0.05

guess you will have a higher chance of guessing correctly with the faces data set. This is what happens for the most frequent classifier. The classifier found the most common label in the training set and apparently this label also appears with an above average frequency in the validation and testing set. Therefore the scores for digits is 14% instead of 10%, which would be the case if all labels were equally common in the test and validation set. And the same is true for faces, if both labels were equally common in the test and validation set, then the scores would be 50%, instead of 56% and 53%. The naive bayes classifier scores higher because it really looks at the image. It tries to find good features that really help distinguish between the different labels. The score for faces is higher for naive bayes because with the digits the different images will overlap, even though they are not necessarily the same label. This is because multiple digits have, for example, a curve in the top. And with faces it is easier to find a feature that distinguishes well, because there are only two possible options.

2.b ex2-b

Imagine we have 5 possible labels, these labels do not occur with the same frequency, but rather one label is quite rare. Then we can have a big training set where this label does not occur. And so the classifier will assign a probability of 0 to it. Then when we come across an input in our test data with this label, the classifier can not possibly assign the correct label. In order to prevent this, and make sure that every label has a chance to be assigne, we use laplace smoothing. This entails that we will always add a standard value to the frequency of a label. The probability used to be: frequency of a label divided by the total number of elements in the training data. But with laplace smoothing it becomes: (frequency of a label + 1) / (number of elements + number of labels).

2.c ex2-c

The autotune option finds the best value for k among these [0.001, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 20, 50]. And then the classifier uses the best k value that was found for the validation and test set. This option helps to improve the score for the naive bayes classifier and will in general always give better scores. It is better for both data sets, but there could be a dataset where a k of 2 would give the best scores and there the autotune would give worse scores. What does stay the same is the presence of a difference between accuracy on validation and testing set. Autotune will always choose a k that performs best on the validation set, but this is not necessarily the best k for the test set.

2.d ex2-d

For naive bayes the validation set can be used to get a good k value and make sure we do not get a classifier that does really well on the training set but poorly on other sets.

2.e ex2-e

3 Exercise 3

3.a ex3-a

3.b ex3-b

3.c ex3-c

3.d ex3-d

4 Exercise 4

4.a ex4-a

4.b ex4-b

4.c ex4-c

5 Exercise 5

5.a ex5-a

5.b ex5-b

6 Exercise 6

6..1 ex6-a

6..2 ex6-b

6..3 ex6-c

7 Exercise 7

7.a ex7-a

7.b ex7-b

7.c ex7-c

7.d ex7-d

8 Exercise 8

8.a ex8-a

8.b ex8-b

8.c ex8-c

9 Exercise 9