KI Project C

Martijn Dekker 6013368
Egor Dmitriev 6100120
Cody Bloemhard 6231888
LaTeX

January 18, 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.

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

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 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: first it gets the classifier by looking at the training set, then 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. And as last it is tested on the test set to get an indication of how the classifier will perform on real life data sets. For most frequent: it gets the classifier by looking at the training set. It looks which label is the most common and then uses that as classifier. It does nothing important with the validation set. It merely looks what score it gets on there. Then it uses the test set to get an indication of how the classifier will perform on real life data sets.

2.e ex2-e

Test set, because the validation set can be use to fine tune the classifier and rule out any form of overfitting. This happens with the naive bayes classifier when the autotune option enable. The validation test is used to get the best k value from the training data. The test set is always used to get an indication of how the classifier will perform on real life data sets.

3 Exercise 3

3.a ex3-a

Because there are a lot of features which vary between zero and one multiplying them together can yield very small numbers. These numbers in some cases cannot be represented using float and double types.
A workaround for this is representing probabilities a log probabilities. This way their values are between $-\infty$ and 0.
Also because of this we no longer use product, but use summation instead. This works since sum of logarithms is equivalent to the log of the product of all the probabilities.

3.b ex3-b

Conditional probability can also be written as $P(A\hat{B}) = P(A|B)P(B)$. With this we can see that joint probability yields the same result as posterior probability multiplied with probability of P(B).
Using log in joint probabilities achieves same result as described earlier because sum of logarithms of the probabilities is equivalent to the log of the product of all the probabilities.
TODO: I could be wrong here at my first point. Doublecheck

	0.55	0.60	0.65	0.70	0.75	0.80
accuracy test	56.0%	56.0%	56.0%	56.0%	56.0%	56.0%
accuracy validation	53.0%	53.0%	53.0%	53.0%	53.0%	53.0%

3.c ex3-c

Thresholding does not seem to to improve performance. It rather worsens the performance by introducing more false negatives. The posterior probabilities are all around 50is because features are of a low quality.//

TODO: this cant be right. Reproduce dataClassifier.py -c naiveBayes -d faces -k 0.1 –threshold=0.45,0.55

3.d ex3-d

The naive bayes classifier would be more biased towards classifying numbers as even numbers. Normally one would need to balance the dataset first, but this would require retraining. Since naive bayes works using frequency counting we can also correct the amount of even number occurences in the frequency table by reducing them by 50dataset out.

4 Exercise 4

4.a ex4-a

4.b ex4-b

4.c ex4-c

5 Exercise 5

5.a ex5-a

Sequence weights (a) are more representable of the perceptron because these represent features that are weighted the most in classifying the given label. Because these features are mostly used to determine whether and input corresponds to the label it is fair to conclude that perceptron sees weights 'a' as the ones that represent the labels the most.

5.b ex5-b

Perceptron is capable of correctly classifying separable data. In this case digit writings can not be considered separable data. This is because a lot of digits look very similar. They also depend on handwriting and because of this a lot of different digits can look alike when written by different people.

6 Exercise 6

6..1 ex6-a

By limiting weight change extreme outliers can not influence the weights as much. By not limiting the weight change large errors may cause weights to overshoot the minima which causes the weights converge slower. Because there is a large number of features, they may vary a lot in values which may cause such a slowdown. // TODO improve answer (many features)

6..2 ex6-b

Autotune trains the model with different values for constant C and picks the best value which achieves the best accuracy.
Autotune has used three different values for C 0.002, 0.004, 0.008. Between those value of 0.004 for C has given the highest accuracy on the validation set.
MIRA has given 68.0for validation and 48.0with k=0.1.
MIRA has better performance than perceptron but performs worse than best naive bayes.
MIRA performs better than perceptron because it limits the learning rate and tries to put a margin between the boundaries which gives it an advantage at classifying unseen data points.
MIRA itself is still perceptron based and does not handle inseparable data. Classify digits with pixels as features is an example of inseparable data and therefore causes MIRA perform worse than naive bayes.

6..3 ex6-c

7 Exercise 7

7.a ex7-a

No naive bayes classifier is not a suitable instrument for classifying digits because it assumes the values of the features are independent of the values of any other features given a class variable. This is not true because digits aren't made of independent pixels. They are made of lines / curve segements which together for a digit. Therefore it is possible that values of features are dependent on other features.