Assignment 2: Naive Bayes and Text Classi cation

Benedikt Riegel

May 10, 2021

1 Task 1: Simple Bayes (20 Points)

1.1 Task (10 Points)

De ne:

- P (Box x) := probability that Box x is chosen)
- P (Apple) := probability that an apple is chosen
- P (Orange) := probability that an orange is

chosen Given: P (Box 1) = P (Box 2)

	Apples	Oranges
Box 1	4	10
Box 2	6	8

We assume P (Box 1) + P (Box 2) = 1.
=) P (Apple j Box 1) = 4=14 = 2=7, P (Orange j Box 1) = 10=14 = 5=7,
P (Apple j Box 2) = 6=14 = 3=7 and P (Orange j Box 2) = 8=14 = 4=7.
P (Box 1) = P (Box 2) = 1=2.

What is the probability of choosing an apple?

$$\begin{aligned} P(\text{Apple}) &= P(\text{Apple, Box 1}) + P(\text{Apple, Box 2}) \\ &= P(\text{Box 1})\ P(\text{Apple} \mid \text{Box 1}) + P(\text{Box 2})\ P(\text{Apple} \mid \text{Box 2}) \\ &= \frac{1}{2}\frac{2}{7}+\frac{1}{2}\frac{3}{7} \\ &= \frac{5}{14} \end{aligned}$$

If an apple is chosen, what is the probability that it came from box 1?

$$
\begin{aligned}
P\,(\text{Box 1 j Apple}) &= \frac{P\,(\text{Box 1, Apple})}{P\,(\text{Apple})} \\
&= \frac{P\,(\text{Apple, Box 1})}{P\,(\text{Apple})} \\
&= \frac{\frac{1}{2}\quad\frac{2}{7}}{\frac{5}{14}} \\
&= \frac{1214}{2\ \ 7\ \ 5} \\
&= \frac{28}{70} \\
&= \frac{2}{5}
\end{aligned}
$$

1.2 Task (10 Points)

Given: Given are two M&M bags from 1994 and 1996 and the probabilities of nding a speci c colour in the two di erent bags. The probabilities are as follows:

	Yellow	Green	Other
1994	0.3	0.2	0.5
1996	0.16	0.24	0.6

Now scenario A :="one M&M of each bag is taken out, one is green and the other is yellow." occurs.
What is the probability of scenario B := "the yellow M&M came from the 1994 bag"? So it's asked for the probability

$$P\,(B \text{ j } A) = \frac{P\,(B;\,A)}{P(A)}$$

First take a look at P (B; A), notice that A and B implies that not only is the 1994 M&M yellow, but the 1996 M&M is green.

$$
\begin{aligned}
P\,(B;\,A) &= P\,(\text{1994 Bag was picked})\ \ P\,(\text{yellow M\&M}) \\
&\quad +P\,(\text{1996 Bag was picked})\qquad P\,(\text{green M\&M}) \\
&= 0{:}5\ \ 0{:}3 + 0{:}5\ \ 0{:}24 \\
&= 0{:}27
\end{aligned}
$$

Notice that because of statement A we have to take one out of each bag, so we have to treat the probabilities for each bag to be 0:5. Now P (A) can also be computed by summing up all the allowed combinations of M&M that ful l scenario A.

A.1 1994 M&M is yellow =) 1996 M&M is green.

A.2 1996 M&M is green =) 1996 M&M is yellow.

This means there are only scenario A:1 and A:2 that ful l A.

[A:1 , B and A] =) [P (A:1) = P (B; A) = 0:27]

Analogously P (A:2) can be computed to be

$$
\begin{aligned}
P (A:2) &= 0:5 \ \ 0:2 + 0:5 \ \ 0:16 \\
&= 0:18
\end{aligned}
$$

resulting in

$$
\begin{aligned}
P (A) &= P (A:1) + P (A:2) \\
&= 0:27 + 0:18 \\
&= 0:45
\end{aligned}
$$

Therefore the answer to our previous question is

$$
\begin{aligned}
P (B \ j \ A) &= \frac{P(B; A)}{P(A)} \\
&= \frac{0:27}{0:45} \\
&= 0:6
\end{aligned}
$$

2 Task 2: Fake News Classi cation with Naive Bayes (20 Points)

This task was answered in the jupyter-notebook.

3 Task 3: kNN for Text Classi cation (30 Points)

3.1 How do you represent the text?

These are just some thoughts and can be deleted.
We could add one dimension for each term, which represents the frequency of this speci c term. E.g.

- point [0; 0] represents a text with 0 'hi's and 0 'bye's (empty text []).
- point [2; 1] represents a text with 2 'hi's and 1 'bye' (['hi', 'hi', 'bye'], ['hi', 'bye', 'hi'], ['bye', 'hi', 'hi']).

We should use a min-max scaling on this data.

3.2 What distance function do you use?

Specifically, five different distance functions, which are Euclidean distance, cosine similarity measure, Minkowsky, correlation, and Chi square, are used in the k-NN classifier.

Euclidean dist. (p,q) :

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

It is used to calculate the distance between two data points in a plane.

Cosine Similarity:

$$sim\,(A,B) = \frac{\vec{A}\cdot\vec{B}}{\left|\vec{A}\right|\left|\vec{B}\right|}$$

This metric is used when the magnitude between vectors does not matter but the orientation.

Minkowsky:

$$dist\_Minkowsky\,(A,B) = \left(\sum_{i=1}^{m}|x_i - y_i|^r\right)^{1/r}$$

Correlation:

$$dist\_correlation\,(A,B) = \frac{\sum_{i=1}^{m}(x_i-\mu_i)(y_i-\mu_i)}{\sqrt{\sum_{i=1}^{m}(x_i-\mu_i)^2\sum_{i=1}^{m}(y_i-\mu_i)^2}}$$

Chi Square:

$$dist\_Chi\text{-}square\,(A,B) = \sum_{i=1}^{m}\frac{1}{sum_i}\left(\frac{x_i}{size_Q} - \frac{y_i}{size_I}\right)^2$$