In order to reduce my email load, I decide to implement a machine learning algorithm to decide whether or not I should read an email, or simply file it away instead. To train my model, I obtain the following data set of binary-valued features about each email, including whether I know the author or not, whether the email is long or short, and whether it has any of several key words, along with my final decision about whether to read it ($y = +1$ for "read", $y = -1$ for "discard").

x_1	x_2	x_3	x_4	x_5	y
know author?	is long?	has 'research'	has 'grade'	has 'lottery'	⇒ read?
0	0	1	1	0	-1
1	1	0	1	0	-1
0	1	1	1	1	-1
1	1	1	1	0	-1
0	1	0	0	0	-1
1	0	1	1	1	1
0	0	1	0	0	1
1	0	0	0	0	1
1	0	1	1	0	1
1	1	1	1	1	-1

In the case of any ties, we will prefer to predict class +1.
I decide to try a naïve Bayes classifier to make my decisions and compute my uncertainty.

1. Compute all the probabilities necessary for a naïve Bayes classifier, i.e., the class probability $p(y)$ and all the individual feature probabilities $p(x_i|y)$, for each class y and feature x_i `(10 points)`

1. $P(y=1) = \frac{4}{10} = \frac{2}{5}$

$P(y=-1) = \frac{6}{10} = \frac{3}{5}$

$P(x_1=1 \mid y=1) = \frac{P(x_1 \cap y=1)}{P(y=1)} = \frac{\frac{3}{10}}{\frac{2}{5}} = \frac{3}{4}$

$P(x_1=1 \mid y=-1) = \frac{3}{6} = \frac{1}{2}$

$P(x_1=0 \mid y=1) = \frac{1}{4}$

$P(x_1=0 \mid y=-1) = \frac{1}{2}$

$P(x_2=1 \mid y=1) = \frac{0}{4} = 0$

$P(x_2=1 \mid y=-1) = \frac{5}{6}$

$P(x_2=0|y=1) = 1$

$P(x_2=0|y=-1) = \frac{1}{6}$

$P(x_3=1|y=1) = \frac{3}{4}$

$P(x_3=1|y=-1) = \frac{4}{6} = \frac{2}{3}$

$P(x_3=0|y=1) = \frac{1}{4}$

$P(x_3=0|y=1) = \frac{1}{3}$

$P(x_4=1|y=1) = \frac{2}{4} = \frac{1}{2}$

$P(x_4=1|y=-1) = \frac{5}{6}$

$P(x_4=0|y=1) = \frac{1}{2}$

$P(x_4=0|y=-1) = \frac{1}{6}$

$P(x_5=1|y=1) = \frac{1}{4}$

$P(x_5=1|y=-1) = \frac{2}{6} = \frac{1}{3}$

$P(x_5=0|y=1) = \frac{3}{4}$

$P(x_5=0|y=-1) = \frac{2}{3}$

2. Which class would be predicted for $\underline{x} = (0\ 0\ 0\ 0\ 0)$? What about for $\underline{x} = (1\ 1\ 0\ 1\ 0)$? **(10 points)**

① $X = (00000)$

$$P(y=1|X) = P(y=1) \times P(x_1=0|y=1)$$
$$\times P(x_2=0|y=1) \times P(x_3=0|y=1)$$
$$\times P(x_4=0|y=1) \times P(x_5=0|y=1)$$
$$= \frac{2}{5} \times \frac{1}{4} \times 1 \times \frac{1}{4} \times \frac{1}{2} \times \frac{3}{4}$$
$$= \frac{3}{320} = 0.0094$$

$$P(y=-1|X) = P(y=-1) P(x_1=0|y=-1)$$
$$P(x_2=0|y=-1)\ P(x_3=0|y=-1)\ P(x_4=0|y=-1)$$
$$P(x_5=0|y=-1)$$
$$= \frac{3}{5} \times \frac{1}{2} \times \frac{1}{6} \times \frac{1}{3} \times \frac{1}{6} \times \frac{2}{3}$$
$$= \frac{1}{540} = 0.0019$$

$$P(y=1|X) > P(y=1|X)$$

Predicted for $x=(00000)$ is 1

② $X = (1\ 1\ 0\ 1\ 0)$

$$P(y=1|X) = P(y=1)P(x_1=1|y=1)$$
$$P(x_2=1|y=1)\ P(x_3=0|y=1)$$
$$P(x_4=1|y=1)\ P(x_5=0|y=1)$$
$$= \frac{2}{5} \times \frac{3}{4} \times 0 \times \frac{1}{4} \times \frac{1}{2} \times \frac{3}{4}$$
$$= 0$$

$$P(y=-1|X) = P(y=-1)P(x_1=1|y=-1)$$
$$P(x_2=1|y=-1)\ P(x_3=0|y=-1)$$
$$P(x_4=1|y=-1)\ P(x_5=0|y=-1)$$
$$= \frac{3}{5} \times \frac{1}{2} \times \frac{5}{6} \times \frac{1}{3} \times \frac{5}{6} \times \frac{2}{3}$$
$$= \frac{5}{108} = 0.047$$

$P(y=-1|x) > P(y=1|x)$

Predict for $x=(11010)$ is -1

3. Compute the posterior probability that $y = +1$ given the observation $\underline{x} = (1\ 1\ 0\ 1\ 0)$. **(5 points)**

$$P(y=1|X) = P(x_1=1|y=1)\,P(x_2=1|y=1)$$
$$P(x_3=0|y=1\;P(x_4=1|y=1)\;P(x_5=0|y=1)$$
$$= \frac{3}{4} \times 0 \times \frac{1}{4} \times \frac{1}{2} \times \frac{3}{4}$$
$$= \frac{9}{128}$$

4. Why should we probably not use a "joint" Bayes classifier (using the joint probability of the features x, as opposed to a naïve Bayes classifier) for these data? **(10 points)**

① High computational consumption

② As the number of features increases, the amount of data needed to reliably estimate the joint probability distribution grows exponentially.

5. Suppose that, before we make our predictions, we lose access to my address book, so that we cannot tell whether the email author is known. Should we re-train the model, and if so, how? (e.g.: how does the model, and its parameters, change in this new situation?) Hint: what will the naïve Bayes model over only features $x_2 \ldots x_5$ look like, and what will its parameters be? (10 points)

We need to retrain.

Because we lose X1's data. But Original model was trained with X1 would be available.

And Parameters will be the Probabilities associated with features X2 to X5.