Homework 2 Machine Learning

Oskar Hulthen 950801-1195 huoskar@student.chalmers.se
Alexander Branzell 931003-1977 alebra@student.chalmers.se

April 2018

1 Theoretical section

Naive Bayes

To compute a prediction using Bayes, we use the following formula:

$$P(t_{new} = k \mid x_{new}, \mathbf{X}, \mathbf{t}) = \frac{p(x_{new} \mid t_{new} = k, \mathbf{X}, \mathbf{t}) P(t_{new} = k)}{\sum_j p(x_{new} \mid t_{new} = j, \mathbf{X}, \mathbf{t}) P(t_{new} = j)}$$

According to the data we have the following probabilities for specific variables (where x_1 = rich, x_2 = married, x_3 = healthy, c = content and $\bar{k}$ indicates the negation of k):

$$P(x_1 \mid c) = \frac{3}{4},\ P(x_2 \mid c) = \frac{1}{2},\ P(x_3 \mid c) = \frac{3}{4}$$

$$P(x_1 \mid \bar{c}) = \frac{1}{4},\ P(x_2 \mid \bar{c}) = \frac{1}{4},\ P(x_3 \mid \bar{c}) = \frac{1}{4}$$

$$P(c) = \frac{1}{2},\ P(\bar{c}) = \frac{1}{2}$$

Due to us looking for combinations of variables using naive Bayes in the following tasks, we use the following formula to find the probability of the combination:

$$P((x_1,\ x_2,\ x_2) \mid c) = \prod_{i=1}^{3} P(x_i \mid c)$$

1.0.1 What is the probability that a person is "not rich", "married" and "healthy" is "content"?

In other words we want to find $P(c \mid (\bar{x_1},\ x_2,\ x_3))$ following the above formulas we get the following expression:

$$P(c \mid (\bar{x}_1,\ x_2,\ x_3)) = \frac{P((\bar{x}_1,\ x_2,\ x_3) \mid c) \cdot P(c)}{P((\bar{x}_1,\ x_2,\ x_3) \mid c) \cdot P(c) + P((\bar{x}_1,\ x_2,\ x_3) \mid \bar{c}) \cdot P(\bar{c})}$$

Where there are only two unknown probabilities, which we can both calculate:

$$P((\bar{x}_1,\ x_2,\ x_3) \mid c) = P(\bar{x}_1 \mid c) \cdot P(x_2 \mid c) \cdot P(x_3 \mid c) = \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{3}{4} = \frac{3}{32}$$

$$P((\bar{x}_1,\ x_2,\ x_3) \mid \bar{c}) = P(\bar{x}_1 \mid \bar{c}) \cdot P(x_2 \mid \bar{c}) \cdot P(x_3 \mid \bar{c}) = \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{3}{64}$$

Yielding the following result:

$$P(c \mid (\bar{x}_1,\ x_2,\ x_3)) = \frac{\frac{3}{32} \cdot \frac{1}{2}}{\frac{3}{32} \cdot \frac{1}{2} + \frac{3}{64} \cdot \frac{1}{2}} = \frac{\frac{3}{64}}{\frac{9}{128}} = \frac{6}{9} = \frac{2}{3}$$

In other words there is a 66.6% chance that a person who is "not rich", "married" and "healthy" is content.

1.0.2 What is the probability that a person who is "not rich" and "married" is content ?

In this case we want to find $P(c \mid (\bar{x_1},\ x_2))$, giving us the following expression:

$$P(c \mid (\bar{x_1},\ x_2)) = \frac{P((\bar{x_1},\ x_2) \mid c) \cdot P(c)}{P((\bar{x_1},\ x_2,) \mid c) \cdot P(c) + P((\bar{x_1},\ x_2) \mid \bar{c}) \cdot P(\bar{c})}$$

Where again there are two unknown probabilities that both can be calculated according to:

$$P((\bar{x_1},\ x_2) \mid c) = P(\bar{x_1} \mid c) \cdot P(x_2 \mid c) = \frac{1}{4} \cdot \frac{1}{2} \cdot = \frac{1}{8}$$

$$P((\bar{x_1},\ x_2) \mid \bar{c}) = P(\bar{x_1} \mid \bar{c}) \cdot P(x_2 \mid \bar{c}) = \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{16}$$

Yielding the following result to the above expression:

$$P(c \mid (\bar{x_1},\ x_2)) = \frac{\frac{1}{8} \cdot \frac{1}{2}}{\frac{1}{8} \cdot \frac{1}{2} + \frac{3}{16} \cdot \frac{1}{2}} = \frac{\frac{1}{16}}{\frac{5}{32}} = \frac{2}{5}$$

In other words there is a 40% chance that a person who is "not rich" and "married" is content.

Extending naive Bayes

The problem here is that only one of x_1, x_2 and x_3 are able to be set to true in a certain combination. A person can not be below 20 and between 20 and 30 at the same time. This means that only certain combinations are allowed. In the case of the given task, there are only six combinations that are allowed:

$$(1,0,0,0),\ (0,1,0,0),\ (0,0,1,0),\ (1,0,0,1),\ (0,1,0,1),\ (0,0,1,1)$$

In other words the people are only allowed to have one "age" variable. However if we were to use these given variables in our naive Bayes we would see that there would be probabilities for combinations that are not allowed to occur (for example: $P(R \mid (1,1,1,x) \neq 0)$. Therefore implying that the sum of the allowed probabilities for the allowed combinations would not sum to 1, meaning that the results would be wrong for the given task. The solution is to merge the "age" variables x_1, x_2 and x_3 into a single variable x_* which can take the values of $\{0,\ 1,\ 2\}$. This gives us a new set of allowed combinations.

$$(0,0),\ (1,0),\ (2,0),\ (0,1),\ (1,1),\ (2,1)$$

This set corresponds to each of the allowed combinations before the merge, but it has no space for combinations that are disallowed, thus all the probabilities will sum to 1 and the result will also be correct.