

COMP5318/COMP4318 Machine Learning and Data Mining

Week 4 Tutorial exercises Naïve Bayes

Exercise 1. Naïve Bayes for data with nominal features (to do in class)

Given is the following dataset where *loan default* is the class. Predict the class of the following new example using Naïve Bayes:

home owner = no, *marital status* = married, *annual income*=very high

	home owner	marital status	income	loan default
1	yes	single	very high	yes
2	no	married	high	yes
3	no	single	medium	no
4	yes	married	very high	no
5	yes	divorced	high	yes
6	no	married	low	no
7	yes	divorced	very high	no
8	no	single	high	yes
9	no	married	medium	no
10	no	single	low	yes

Dataset adapted from, Tan, Steinbach, Karpapne and Kumar, Introduction to Data Mining, Pearson, 2019

Solution:

E= home owner = no, marital status = married, annual income=very high

E1 is home owner = no, E2 is marital status = married, E3 is annual income=very high

We need to compute $P(\text{yes}|E)$ and $P(\text{no}|E)$ and compare them.

$$P(\text{yes}|E) = \frac{P(E_1|\text{yes})P(E_2|\text{yes})P(E_3|\text{yes})P(\text{yes})}{P(E)}$$

$$P(\text{no}|E) = \frac{P(E_1|\text{no})P(E_2|\text{no})P(E_3|\text{no})P(\text{no})}{P(E)}$$

$$P(\text{yes})=5/10$$

$$P(\text{no})=5/10$$

$$P(E_1|\text{yes})=P(\text{home owner=no}|\text{yes})=3/5$$

$$P(E_1|\text{no})= P(\text{home owner=no}|\text{no})=3/5$$

$$P(E_2|\text{yes})=P(\text{marital status=married}|\text{yes})=1/5$$

$$P(E_2|\text{no})=P(\text{marital status=married}|\text{no})=3/5$$

$$P(E_3|\text{yes})=P(\text{annual income=very high}|\text{yes})=1/5$$

$$P(E_3|\text{no})=P(\text{annual income=very high}|\text{no})=2/5$$

$$P(\text{yes}|E) = \frac{\frac{3}{5} \cdot \frac{1}{5} \cdot \frac{1}{5} \cdot \frac{5}{10}}{P(E)} = \frac{\frac{3}{250}}{P(E)} = \frac{0.012}{P(E)}$$

$$P(\text{no}|E) = \frac{\frac{3}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{5}{10}}{P(E)} = \frac{\frac{9}{125}}{P(E)} = \frac{0.072}{P(E)}$$

$P(\text{no}|E) > P(\text{yes}|E) \Rightarrow$ Naïve Bayes predicts **loan default = no** for the new example.

Exercise 2. Naïve Bayes for data with numeric features (to do in class)

The same task as in the previous exercise but now *annual income* is a numeric feature:

	home owner	marital status	income (in K)	loan default
1	yes	single	125	yes
2	no	married	100	yes
3	no	single	70	no
4	yes	married	120	no
5	yes	divorced	95	yes
6	no	married	60	no
7	yes	divorced	220	no
8	no	single	85	yes
9	no	married	75	no
10	no	single	90	yes

Gaussian Naïve Bayes assumes that the likelihood $P(x_i|y)$ follows the Gaussian distribution for each x_i given y . Therefore,

$$P(x_i|y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Use Naïve Bayes to predict the class of the following new example:

home owner = no, marital status = married, annual income=120

Solution:

1) Calculate the mean  and standard deviation  values for the numeric feature *income*:

$$\bar{x} = \frac{\sum_{i=1}^n X_i}{n} \quad s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{x})^2}{n-1}}$$

where $X_i, i=1..n$ – the i -th measurement, n -number of measurements

We need to calculate the mean and standard deviation separately for each class (yes and no) – separate the values of income:

class yes income		class no income
125		70
100		120
95		60
85		220
90		75
 _income_yes=99		 _income_no=109
 _income_yes=15.57		 _income_no=66.18

2) Calculate $P(\text{income}=120|\text{yes})$ and $P(\text{income}=120|\text{no})$ using the probability density function for normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$f(\text{income} = 120|\text{yes}) = \frac{1}{15.57\sqrt{2\pi}} e^{-\frac{(120-99)^2}{2*15.57^2}} = 0.01032$$

$$f(\text{income} = 120 | \text{no}) = \frac{1}{66.18 \sqrt{2\pi}} e^{-\frac{(120-109)^2}{2 \cdot 66.18^2}} = 0.00595$$

3) Calculating the probabilities $P(\text{yes}|E)$ and $P(\text{no}|E)$ using the Bayes Theorem; we already have the probabilities for the nominal attributes from the previous exercise:

$$P(\text{yes}|E) = \frac{\frac{3}{5} \cdot \frac{1}{5} \cdot 0.01032 \cdot \frac{5}{10}}{P(E)} = \frac{0.000619}{P(E)}$$

$$P(\text{no}|E) = \frac{\frac{3}{5} \cdot \frac{3}{5} \cdot 0.00595 \cdot \frac{5}{10}}{P(E)} = \frac{0.001071}{P(E)}$$

$P(\text{no}|E) > P(\text{yes}|E) \Rightarrow$ Naïve Bayes predicts **loan default = no** for the new example.