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, Karpatne 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(yes|E) and P(no|E) and compare them.

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

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

P(yes)=5/10

P(E1|yes)=P(home owner=no|yes)=3/5
P(E2|yes)=P(marital status=married|yes)=1/5
P(E3|yes)=P(annual income=very high|yes)=1/5

P(no)=5/10

P(E1|no)= P(home owner=no|no)=3/5
P(E2|no)=P(marital status=married|no)=3/5
P(E3|no)=P(annual income=very high|no)=2/5

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

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

P(no|E)>P(yes|E) => 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

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 μ and standard deviation σ values for the numeric feature *income*:

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n-1}$$

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
μ_income_yes=99		μ_income_no=109
σ_income_yes=15.57		σ_income_no=66.18

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

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

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

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

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

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

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

P(no|E)>P(yes|E) => Naïve Bayes predicts **loan default = no** for the new example.