COMP3308/COMP3608 Introduction to Artificial Intelligence

Weeks 6 Tutorial exercises
Naïve Bayes. Evaluating Classifiers.

Exercise 1. ***Naïve Bayes (Homework)***

Suppose you want to recognize good and bad items produced by your company. You are able to measure two properties of each item (P1 and P2) and express them with Boolean values. You randomly grab several items and test if they are good or bad, obtaining the following results:

P1	P2	result
Y	Y	good
Y	N	bad
N	N	good
N	Y	bad
Y	N	good
N	N	good

Use Naïve Bayes to predict the class, *good* or *bad,* of the following new item: P1=N, P2=Y. If there are ties, make a random choice.

Solution:

E is P1=N and P2=Y; E_1 is P1=N, E2 is P2=Y

We need to compute P(good|E) and P(bad|E) and compare them.

$$P(good \mid E) = \frac{P(E_1 \mid good)\,P(E_2 \mid good)\,P(good)}{P(E)}$$

P(good)=4/6=2/3 P(bad)=2/6=1/3

P(E1|good)=P(P1=N|good)=2/4=1/2 P(E1|bad)=P(P1=N|bad)=1/2

P(E2|good)=P(P2=Y|good)=1/4 P(E2|bad)=P(P2=Y|bad)=1/2

$$P(good \mid E) = \frac{\frac{1}{2}\frac{1}{4}\frac{2}{3}}{P(E)} = \frac{\frac{1}{12}}{P(E)} \qquad P(bad \mid E) = \frac{\frac{1}{2}\frac{1}{2}\frac{1}{3}}{P(E)} = \frac{\frac{1}{12}}{P(E)}$$

The two probabilities are the same. To resolve the tie we randomly choose between the 2 classes => e.g. class *good*.

Exercise 2. ***Naïve Bayes***

Why is the Naïve Bayesian classification called “naïve”?

Answer: Naïve Bayes assumes that the values of the attributes are independent of each other and that all attributes are equally important. These assumptions are unrealistic, that’s why it is called “Naïve”.

Exercise 3. ***Applying Naïve Bayes to data with both numerical and nominal attributes***

Given is the training data in the table below (the *weather* data with some numerical attributes, *play* is the class). Predict the class of the following new example using the Naïve Bayes classification: outlook=overcast, temperature=60, humidity=62, windy=false.

outlook	**temperature**	**humidity**	**windy**	**play**
sunny	85	85	false	no
overcast	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Solution:

First, we need to calculate the mean μ and standard deviation σ values for the numerical attributes, for each class (yes and no) separately.

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

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

You can use Excel, Python, Matlab etc. to calculate these values.

μ_temp_yes=73, σ_temp_yes=6.2; μ_temp_no=74.6, σ_temp_no=8.0
μ_hum_yes=79.1, σ_temp_yes=10.2; μ_hum_no=86.2, σ_temp_no=9.7

Second, to calculate f(temperature=60|yes), f(temperature=60|no), f(humidity=62|yes) and f(humidity=62|no) using the probability density function for normal distribution:

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

$$f(temperature = 60|yes) = \frac{1}{6.2\sqrt{2\pi}} e^{-\frac{(60-73)^2}{2*6.2^2}} = 0.0071$$

$$f(temperature = 60|no) = \frac{1}{8\sqrt{2\pi}} e^{-\frac{(60-74.6)^2}{2*8^2}} = 0.0094$$

$$f(humidity = 62|yes) = \frac{1}{10.2\sqrt{2\pi}} e^{-\frac{(62-79.1)^2}{2*10.2^2}} = 0.0096$$

$$f(humidity = 62|no) = \frac{1}{9.7\sqrt{2\pi}} e^{-\frac{(62-86.2)^2}{2*9.7^2}} = 0.0018$$

Third, we can calculate the probabilities for the nominal attributes:

P(yes)=9/14=0.643	P(no)=5/14=0.357
P(outlook=overcast|yes)=4/9=0.444	P(outlook=overcast|no)=1/5=0.2
P(windy=false|yes)=6/9=0.667	P(windy=false|no)=2/5=0.4

Fourth, we can calculate the final probabilities:

$$P(yes \mid E) = \frac{0.444 * 0.0071 * 0.0096 * 0.667 * 0.643}{P(E)} = \frac{12.97 * 10^{-6}}{P(E)}$$

$$P(no|E) = \frac{0.2 * 0.0094 * 0.0018 * 0.4 * 0.357}{P(E)} = \frac{4 * 10^{-7}}{P(E)}$$

Therefore, the Naïve Bayes classifier predicts *play=yes* for the new example.

Exercise 4. ***Bayes Theorem (Advanced only)***

Suppose that the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. If 1/5 of the University students are graduate students and the rest are undergraduates, what's the probability that a student who smokes is a graduate student?

Hint: Use the Bayes Theorem; you will need to calculate the denominator using the *law of total probability*, see its Wikipedia description.

Solution:

Suppose that:
X represents if the student smokes: {smoker, non-smoker} or abbreviated {S, NS} and
Y represents the type of student: {undergraduate, graduate} or abbreviated {UG, G}

Given: P(G)=1/5=0.2, P(UG)=4/5=0.8, P(S|UG)=0.15, P(S|G)=0.23

P(G|S)=?

$$P(G \mid S) = \frac{P(S \mid G)\, P(G)}{P(S)}$$

To calculate P(S) we will use the law of total probability. If {Y1, Y2, …, Yk} is the set of mutually exclusive and exhaustive outcomes of Y, then:

$$P(X) = \sum_{i=1}^{k} P(X, Y_i) = \sum_{i=1}^{k} P(X \mid Y_i) P(Y_i)$$

=> P(S)=P(S|G)P(G)+P(S|UG)P(UG)=0.23*(1/5)+0.15*(4/5)=0.166

=>P(G|S)=(0.23*0.2)/0.166=0.277

Exercise 5. ***Using Weka – Comparing Classifiers***

1. Load the iris dataset
2. Choose "Percentage split" mode for evaluation: 66% training set, 33% testing set
3. Run the Naïve Bayes and review Weka's output
4. For comparison, also run k-neatest neighbor with k=1 and 3 (IB1 and IBk), OneR and ZeroR. Which is the most accurate classifier?

5. Change the test mode to "Cross validation". Apply 10-fold cross validation instead of percentage split as evaluation mode and compare the classifiers.

- Which classifier produced the most accurate classification?
- Which evaluation strategy (percentage split or 10-fold cross validation) produced better results?
- Which evaluation strategy, percentage split or cross validation, is more statistically reliable and why?

6. Apply leave-one-out cross validation. Tip: You need to specify the number of folds in the WEKA's cross validation box.

7. Check the confusion matrix printed by WEKA for one of the classifiers, e.g. Naïve Bayes, and verify the accuracy, recall, precision and F1 measure. Note: Weka shows recall, precision and F1 for each class separately.

Answer: Leave-one-out is n-fold cross validation where n is the number of examples. Thus, the number of folds should be set to 150 for iris data.