UECS3213 / UECS3453 DATA MINING

SESSION: January 2019

TUTORIAL 3

Chapter 3 - Classification

k-NN

1. Consider the one-dimensional data set shown in Table 5.4.

Table 5.4. Data set

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	−	−	+	+	+	−	−	+	−	−

a) Classify the data point x = 5.0 according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote).

Answer:

- **1-nearest neighbor: +,** (4.9; 1 positive vs 0 negative)
- **3-nearest neighbor: –,** (4.9, 5.2, 5.3; 2 negative vs 1 positive)
- **5-nearest neighbor: +,** (4.9, 5.2, 5.3, 4.6, 5.5 or 4.5; 3 positive vs 2 negative)
- **9-nearest neighbor: –.** (4.9, 5.2, 5.3, 4.6, 5.5, 4.5, 3.0, 7.0, 0.5 or 9.5; 5 negative vs 4 positive)

b) Repeat the previous analysis using the *distance-weighted voting* approach.

Answer:

w_i= 1/dist(x, x_i)

- **1-nearest neighbor: +,** (positive: 1/0.1=10; negative: 1/0.2 = 5; positive > negative)
- **3-nearest neighbor: +,** (positive: 1/0.1 = 10; negative: (1/0.2 + 1/0.3)/2 = 4.17; positive > negative)
- **5-nearest neighbor: +,** (positive: (1/0.1 + 1/0.4 + 1/0.5)/3 = 4.83; negative: (1/0.2 + 1/0.3)/2 = 4.17; positive > negative)
- **9-nearest neighbor: +.** (positive: (1/0.1 + 1/0.4 + 1/0.5 + 1/0.5)/4 = 4.125; negative: (1/0.2 + 1/0.3 + 1/2 + 1/2 + 1/4.5)/5 = 1.91; positive > negative)

- Majority voting:
  $c^* = \arg\max_c g(c)$

- Weighted voting: weighting is on each neighbor
  $c^* = \arg\max_c \Sigma_i w_i \delta(c, f_i(x))$

  Where $\delta(c,f_i(x))$ is 1 if $f_i(x) = c$ and 0 otherwise

- Weighted voting allows us to use more training examples:

  e.g., $w_i = 1/\text{dist}(x, x_i)$

Decision Tree

2. Consider the training examples shown in Table 4.1 for a binary classification problem.

Table 4.1. Data set

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

a) Compute the Gini index for the overall collection of training examples.

- Gini = $1 - 2 \times 0.5^2 = 0.5$ (C0: $10/20 = 0.5$; C1: $10/20 = 0.5$)

b) Compute the Gini index for the Gender attribute.

- The Gini for Male is $1 - 2 \times 0.5^2 = 0.5$. The gini for Female is also $1 - 2 \times 0.5^2 = 0.5$.
- Therefore, the overall Gini for Gender is $0.5 \times 0.5 + 0.5 \times 0.5 = 0.5$.

c) Compute the Gini index for the Car Type attribute using multiway split.

Answer:
The gini for

- Family car = $1 - (1/4)^2 - (3/4)^2 = 0.375$
- Sports car = $1 - (8/8)^2 - (0/8)^2 = 0$
- Luxury car = $1 - (1/8)^2 - (7/8)^2 = 1 - 0.015625 - 0.765625 = 0.2188$.

- The overall gini = 0.375*4/20 + 0*8/20 + 0.2188*8/20 = **0.16252**

d) Compute the Gini index for the Shirt Size attribute using multiway split.

Answer:

- The gini for
  - Small shirt size = $1 - (2/5)^2 - (3/5)^2 = 1 - 0.16 - 0.36 = 1 - 0.52 = 0.48$
  - Medium shirt size = $1 - (3/7)^2 - (4/7)^2 = 1 - 0.1837 - 0.3265 = 0.4898$
  - Large shirt size = $1 - (2/4)^2 - (2/4)^2 = 0.5$
  - Extra Large shirt size = $1 - (2/4)^2 - (2/4)^2 = 0.5$.
- The overall gini for Shirt Size attribute = 5/20*0.48 + 7/20*0.4898 + 4/20*0.5 + 4/20*0.5 = **0.4914**.

e) Which attribute is **better**, Gender, Car Type, or Shirt Size?

Answer:

- Car Type because it has the lowest gini among the three attributes. Car Type is the best.

3. Consider the training examples shown in Table 4.2 for a binary classification problem.

Table 4.2. Data set

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

a) What is the *entropy* of this collection of training examples with respect to the positive class?

- There are four positive examples and five negative examples.
- Thus, P(+) = 4/9 and P(−) = 5/9.
- The entropy of the training examples is $-4/9 \log_2(4/9) - 5/9 \log_2(5/9)$
- -0.4444(-1.1701) – 0.5556(-0.8479) = 0.52 + 0.4711 **= 0.9911**

b) What are the *information gains* of a_1 and a_2 relative to these training examples?

Answer:

For attribute a_1, the corresponding counts and probabilities are:

a_1	+	-
T	3	1
F	1	4

The entropy for a_1 is

$$\frac{4}{9}\left[-(3/4)\log_2(3/4) - (1/4)\log_2(1/4)\right] + \frac{5}{9}\left[-(1/5)\log_2(1/5) - (4/5)\log_2(4/5)\right] = 0.7616.$$

Therefore, the information gain for a_1 is $0.9911 - 0.7616 = 0.2294$.

For attribute a_2, the corresponding counts and probabilities are:

a_2	+	-
T	2	3
F	2	2

The entropy for a_2 is

$$\frac{5}{9}\left[-(2/5)\log_2(2/5) - (3/5)\log_2(3/5)\right] + \frac{4}{9}\left[-(2/4)\log_2(2/4) - (2/4)\log_2(2/4)\right] = 0.9839.$$

Therefore, the information gain for a_2 is $0.9911 - 0.9839 = 0.0072$.

c) For a_3, which is a continuous attribute, compute the *information gain* for every possible split.

Answer:

	1		3		4		5		6		7		8			
Split	**0.5**		**2**		**3.5**		**4.5**		**5.5**		**6.5**		**7.5**		**8.5**	
	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
P(+)	0	4	1	3	1	3	2	2	2	2	3	1	4	0	4	0
P(-)	0	5	0	5	1	4	1	4	3	2	3	2	4	1	5	0

Entropy (1, target class) = P(≤0.5)*Entropy (0,0) + P(>0.5)*Entropy(4,5)

$$0 + \frac{9}{9}\left[-\frac{4}{9} * \log_2(\frac{4}{9}) - \frac{5}{9} * \log_2(\frac{5}{9})\right] = 0.9912$$

GAIN (1) = 0.9912 – 0.9912 = 0

Entropy (3, target class) = P(≤2)*Entropy (1,0) + P(>2)*Entropy(3,5)

$$\frac{1}{9}\left[-\frac{1}{1} * \log_2(\frac{1}{1}) - \frac{0}{1} * \log_2(\frac{0}{1})\right] + \frac{8}{9}\left[-\frac{3}{8} * \log_2(\frac{3}{8}) - \frac{5}{8} * \log_2(\frac{5}{8})\right] = 0.8484$$

GAIN (3) = 0.9912 – 0.8484 = 0.1428

Entropy (4, target class) = P(≤3.5)*Entropy (1,1) + P(>3.5)*Entropy(3,4)

$$\frac{2}{9}\left[-\frac{1}{2}*\log_2(\frac{1}{2})-\frac{1}{2}*\log_2(\frac{1}{2})\right]+\frac{7}{9}\left[-\frac{3}{7}*\log_2(\frac{3}{7})-\frac{4}{7}*\log_2(\frac{4}{7})\right]=0.9885$$

GAIN (4) = 0.9912 – 0.9885 = 0.0027

Entropy (5, target class) = P(≤4.5)*Entropy (2,1) + P(>4.5)*Entropy(2,4)

$$\frac{3}{9}\left[-\frac{2}{3}*\log_2(\frac{2}{3})-\frac{1}{3}*\log_2(\frac{1}{3})\right]+\frac{6}{9}\left[-\frac{2}{6}*\log_2(\frac{2}{6})-\frac{4}{6}*\log_2(\frac{4}{6})\right]=0.9183$$

GAIN (5) = 0.9912 – 0.9183 = 0.0729

Entropy (6, target class) = P(≤5.5)*Entropy (2,3) + P(>5.5)*Entropy(2,2)

$$\frac{5}{9}\left[-\frac{2}{5}*\log_2(\frac{2}{5})-\frac{3}{5}*\log_2(\frac{3}{5})\right]+\frac{4}{9}\left[-\frac{2}{4}*\log_2(\frac{2}{4})-\frac{2}{4}*\log_2(\frac{2}{4})\right]=0.9839$$

GAIN (6) = 0.9912 – 0.9839 = 0.0073

Entropy (7, target class) = P(≤6.5)*Entropy (3,3) + P(>6.5)*Entropy(1,2)

$$\frac{6}{9}\left[-\frac{3}{6}*\log_2(\frac{3}{6})-\frac{3}{6}*\log_2(\frac{3}{6})\right]+\frac{3}{9}\left[-\frac{1}{3}*\log_2(\frac{1}{3})-\frac{2}{3}*\log_2(\frac{2}{3})\right]=0.9728$$

GAIN (7) = 0.9912 – 0.9728 = 0.0184

Entropy (8, target class) = P(≤7.5)*Entropy (4,4) + P(>7.5)*Entropy(0,1)

$$\frac{8}{9}\left[-\frac{4}{8}*\log_2(\frac{4}{8})-\frac{4}{8}*\log_2(\frac{4}{8})\right]+\frac{1}{9}\left[-\frac{0}{1}*\log_2(\frac{0}{1})-\frac{1}{1}*\log_2(\frac{1}{1})\right]=0.8889$$

GAIN (8) = 0.9912 – 0.8889 = 0.1023

Answer:

a_3	Class label	Split point	Entropy	Info Gain
1.0	+	2.0	0.8484	0.1427
3.0	-	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0 5.0	- -	5.5	0.9839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0 7.0	+ -	7.5	0.8889	0.1022

The best split for a_3 occurs at split point equals to 2.

d) What is the best split (among a_1, a_2, and a_3) according to the information gain?

- Best of a_3 at GAIN (3) = 0.1427
- GAIN (a_3) = 0.1427
- GAIN (a_2) = 0.0072
- GAIN (a_1) = 0.2296
- We can see that, MAX (GAIN (a_3), GAIN (a_2), GAIN (a_1)) = GAIN (a1) = 0.2296.
- According to information gain, a_1 produces the best split.

e) What is the best split (between a_1 and a_2) according to the classification error rate?

- Classification Error Rate: = (FP + FN) / (TP + TN + FP + FN)
- Classification error rate a_1: (1+1)/ (3+1+1+4)=2/9= 0.2222
- Classification error rate a_2: (2+2)/ (2+3+2+2)=4/9= 0.4444
- Therefore, a_1 provides best split because of lower classification error rate

f) What is the best split (between a_1 and a_2) according to the Gini index?

Answer:
For attribute a_1, the gini index is

$$\frac{4}{9}\left[1-(3/4)^2-(1/4)^2\right]+\frac{5}{9}\left[1-(1/5)^2-(4/5)^2\right]=0.3444.$$

For attribute a_2, the gini index is

$$\frac{5}{9}\left[1-(2/5)^2-(3/5)^2\right]+\frac{4}{9}\left[1-(2/4)^2-(2/4)^2\right]=0.4889.$$

Since the gini index for a_1 is smaller, it produces the better split.

Naive Bayes

4. Consider the data set shown in Table 5.1

Table 5.1. Data set

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

a) Estimate the *conditional probabilities* for $P(A|+)$, $P(B|+)$, $P(C|+)$, $P(A|-)$, $P(B|-)$, and $P(C|-)$.

Answer:

- $P(A = 1|-) = 2/5 = 0.4$, $P(B = 1|-) = 2/5 = 0.4$,$P(C = 1|-) = 1$,
- $P(A = 0|-) = 3/5 = 0.6$, $P(B = 0|-) = 3/5 = 0.6$, $P(C = 0|-) = 0$;
- $P(A = 1|+) = 3/5 = 0.6$, $P(B = 1|+) = 1/5 = 0.2$, $P(C = 1|+) = 4/5 = 0.8$,
- $P(A = 0|+) = 2/5 = 0.4$, $P(B = 0|+) = 4/5 = 0.8$, $P(C = 0|+) = 1/5 = 0.2$.

b) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample $(A = 0, B = 1, C = 0)$ using the Naive Bayes approach.

Let $P(A = 0, B = 1, C = 0) = K$.

$$
\begin{aligned}
& P(+|A = 0, B = 1, C = 0) \\
= & \frac{P(A = 0, B = 1, C = 0|+) \times P(+)}{P(A = 0, B = 1, C = 0)} \\
= & \frac{P(A = 0|+)P(B = 1|+)P(C = 0|+) \times P(+)}{K} \\
= & 0.4 \times 0.2 \times 0.2 \times 0.5/K \\
= & 0.008/K
\end{aligned}
$$

$$\begin{aligned}
& P(-|A=0,B=1,C=0) \\
= & \frac{P(A=0,B=1,C=0|-) \times P(-)}{P(A=0,B=1,C=0)} \\
= & \frac{P(A=0|-) \times P(B=1|-) \times P(C=0|-) \times P(-)}{K} \\
= & 0/K
\end{aligned}$$

The class label should be '+'.

c) Estimate the *conditional probabilities* using the m-estimate approach, with p = 1/2 and m = 4.

$$\hat{P}(a_i \mid v_j) \leftarrow \frac{n_c + mp}{n+m}$$

where

- ***n*** is number of training examples for which ***v*** = ***v****j*
- ***n****c* number of examples for which ***v*** = ***v****j* and ***a*** = ***a****i*
- ***p*** is the prior estimate of the prob. We wish to determine P`(***a****i*|***v****j*)
- ***m*** is a constant called the ***equivalent sample size***, which determines how heavily to weight ***p*** relative to the observed data. (i.e. adding ***m*** "virtual" examples distributed according to ***p***)
- A typical method for choosing ***p*** in the absence of other information is to assume uniform priors.

Answer:

m = 4, *p* = 0.5

- P(A = 0|+) = (2 + 2)/(5 + 4) = 4/9, n_c = 2, *n* = 5
- P(A = 0|–) = (3+2)/(5 + 4) = 5/9, n_c = 3, *n* = 5
- P(B = 1|+) = (1 + 2)/(5 + 4) = 3/9, n_c = 1, *n* = 5
- P(B = 1|–) = (2+2)/(5 + 4) = 4/9, n_c = 2, *n* = 5
- P(C = 0|+) = (1 + 2)/(5 + 4) = 3/9, n_c = 1, *n* = 5
- P(C = 0|–) = (0 + 2)/(5 + 4) = 2/9, n_c = 0, *n* = 5

d) Repeat part (b) using the conditional probabilities given in part (c).

Let P(A = 0,B = 1, C = 0) = K

$$
\begin{aligned}
& P(+|A=0,B=1,C=0) \\
= \ & \frac{P(A=0,B=1,C=0|+) \times P(+)}{P(A=0,B=1,C=0)} \\
= \ & \frac{P(A=0|+)P(B=1|+)P(C=0|+) \times P(+)}{K} \\
= \ & \frac{(4/9) \times (3/9) \times (3/9) \times 0.5}{K} \\
= \ & 0.0412/K
\end{aligned}
$$

$$
\begin{aligned}
& P(-|A=0,B=1,C=0) \\
= \ & \frac{P(A=0,B=1,C=0|-) \times P(-)}{P(A=0,B=1,C=0)} \\
= \ & \frac{P(A=0|-) \times P(B=1|-) \times P(C=0|-) \times P(-)}{K} \\
= \ & \frac{(5/9) \times (4/9) \times (2/9) \times 0.5}{K} \\
= \ & 0.0274/K
\end{aligned}
$$

The class label should be '+'.

e) Compare the two methods for estimating probabilities. Which method is **better** and why?

Answer:

- When one of the conditional probability is zero, the estimate for conditional probabilities using the m-estimate probability approach is better, since we don't want the entire expression becomes zero.
- Example: $P(C = 0|-)$ in the example above.

5. Consider the data set shown in Table 5.2.

Table 5.2. Data set for Exercise 8.

Instance	A	B	C	Class
1	0	0	1	−
2	1	0	1	+
3	0	1	0	−
4	1	0	0	−
5	1	0	1	+
6	0	0	1	+
7	1	1	0	−
8	0	0	0	−
9	0	1	0	+
10	1	1	1	+

a) Estimate the conditional probabilities for $P(A = 1|+)$, $P(B = 1|+)$, $P(C = 1|+)$, $P(A = 1|-)$, $P(B = 1|-)$, and $P(C = 1|-)$ using the same approach as in the previous problem.

Answer:

- $P(A = 1|+) = 3/5 = 0.6$
- $P(B = 1|+) = 2/5 = 0.4$
- $P(C = 1|+) = 4/5 = 0.8$
- $P(A =1|-) = 2/5 = 0.4$
- $P(B = 1|-) = 2/5 = 0.4$
- $P(C = 1|-) = 1/5 = 0.2$

b) Use the conditional probabilities in part (a) to predict the class label for a test sample ($A = 1, B = 1, C = 1$) using the naive Bayes approach.

Answer:

- Let R : ($A = 1, B = 1, C = 1$) be the test record.
- To determine its class, we need to compute $P(+|R)$ and $P(-|R)$. Using Bayes theorem,
- $P(+|R) = P(R|+)P(+)/P(R)$ and $P(-|R) = P(R|-)P(-)/P(R)$.
- Since $P(+) = P(-) = 0.5$ and $P(R)$ is constant, R can be classified by comparing $P(+|R)$ and $P(-|R)$.
- For this question,
  - $P(R|+) = P(A = 1|+) \times P(B = 1|+) \times P(C = 1|+) = 0.6 \times 0.4 \times 0.8 = 0.192$
  - $P(R|-) = P(A = 1|-) \times P(B = 1|-) \times P(C = 1|-) = 0.4 \times 0.4 \times 0.2 = 0.032$
- Since $P(R|+)$ is larger, the record is assigned to (+) class.

c) Compare $P(A = 1)$, $P(B = 1)$, and $P(A = 1,B = 1)$. State the relationships between A and B.

Answer:

- $P(A = 1) = 0.5$, $P(B = 1) = 0.4$ and $P(A = 1, B = 1) = P(A) \times P(B) = 1/5 = 0.2$
- Therefore, A and B are independent.

d) Repeat the analysis in part (c) using $P(A = 1)$, $P(B = 0)$, and $P(A = 1, B = 0)$.

Answer:

- $P(A = 1) = 0.5$, $P(B = 0) = 0.6$, and $P(A = 1, B = 0) = P(A = 1) \times P(B = 0) = 0.3$.
- A and B are still independent.

e) Compare $P(A = 1, B = 1|Class = +)$ against $P(A = 1|Class = +)$ and $P(B = 1|Class = +)$. Are the variables *conditionally independent* given the class?

Answer:

- Compare $P(A = 1, B = 1|+) = 0.2$ against $P(A = 1|+) = 0.6$ and $P(B = 1|Class = +) = 0.4$.
- Since $P(A = 1|+) \times P(A = 1|-) \neq P(A = 1, B = 1|+)$, A and B are not conditionally independent given the class.

Support Vector Machine

6. Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.
   a) If you remove the following any one red points from the data. Does the decision boundary will change? Yes/No
   b) If you remove the non-red circled points from the data, the decision boundary will change? True/False

Answer:

a) Yes. These three examples are positioned such that removing any one of them introduces slack in the constraints. So the decision boundary would completely change.

b) False. On the other hand, rest of the points in the data won't affect the decision boundary much.

7. What do we mean by generalization error in terms of the SVM?

   a) How far the hyperplane is from the support vectors
   b) How accurately the SVM can predict outcomes for unseen data
   c) The threshold amount of error in an SVM

Answer: B. Generalisation error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

8. What do we mean by a hard margin?

a) The SVM allows very low error in classification
b) The SVM allows high amount of error in classification
c) None of the above

Answer: A. A hard margin means that an SVM is very rigid in classification and tries to work extremely well in the training set, causing overfitting.

9. The SVM's are less effective when:

a) The data is linearly separable
b) The data is clean and ready to use
c) The data is noisy and contains overlapping points

Answer: C. When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

10. Which of the following are real world applications of the SVM?

a) Text and Hypertext Categorization
b) Image Classification
c) Clustering of News Articles
d) All of the above

Answer: D. SVM's are highly versatile models that can be used for practically all real world problems ranging from regression to clustering and handwriting recognition.

The End