Knowledge Discovery and Data Mining

Unit # 5

Acknowledgement

- Most of the slides in this presentation are taken from course slides provided by
  - Han and Kimber (Data Mining Concepts and Techniques) and
  - Tan, Steinbach and Kumar (Introduction to Data Mining)

Supervised Feature Discretization Technique: Chimerge

- Chimerge is one automated discretization algorithm that analyzes the quality of multiple intervals for a given feature by using χ^2 statistics.
- The algorithm consists of three basic steps:
  - Sort the data for the given feature in ascending order.
  - Define initial intervals so that every value is in a separate interval.
  - Repeat until no χ^2 of any two adjacent intervals is less then threshold value.

Chimerge Formula

- $\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} (A_{ij} - E_{ij})^2 / E_{ij}$
  - K = number of classes
  - Aij = number of instances in the i-th interval, j-th class
  - Eij = expected frequency of Aij, computed as (Ri . Cj)/N
  - Ri = number of instances in the i-th interval
  - Cj = number of instances in the j-th class
  - N = total number of instances
- If either Ri or Cj is 0, Eij is set to a small value.

	Class 1	Class 2	
Interval I	A_{11}	A_{12}	R_1
Interval 2	A_{21}	A_{22}	R_2
Σ	C_1	C_2	Σ

Chimerge Example

- For this example, interval points for feature F are 0, 2, 5, 7.5, 8.5, 10, etc.

	Class 1	Class 2	
[7.5, 8.5]	1	0	1
[8.5, 10]	1	0	1
Σ	2	0	2

- $\chi^2 = (1-1)^2/1 + (0-0.1)^2/0.1 + (1-1)^2/1 + (0-0.1)^2/0.1 = 0.2$
- For the degree of freedom d=1, $\chi^2 = 0.2 < 2.706$ (for $\alpha = 0.1$). We can conclude that there are no significant differences in relative class frequencies and that the selected intervals can be merged.

F	K
1	1
3	2
7	1
8	1
9	1
11	2
23	2
37	1
39	2
45	1
46	1
59	1

Chimerge Example (Cont'd)

- After several iterations we won't be able to merge intervals further.

	Class 1	Class 2	
[0, 10]	4	1	5
[10, 42]	1	3	4
Σ	5	4	9

- $\chi^2 = (4-2.78)^2/2.78 + (1-2.22)^2/2.22 + (1-2.22)^2/2.22 + (3-1.78)^2/1.78 = 2.72$
- For the degree of freedom d=1, $\chi^2 = 2.72 > 2.706$ (for $\alpha = 0.1$). The conclusion is that significant differences exist between two intervals and merging is not recommended.

ChiMerge Exercise

- Apply the ChiMerge technique to reduce the number of values for numeric attributes (Slide # 30, Unit # 2)
  - Reduce the number of numeric values for feature I1 and find the final, reduced number of intervals.
  - Reduce the number of numeric values for feature I2 and find the final, reduced number of intervals.
  - Reduce the number of numeric values for feature I3 and find the final, reduced number of intervals.

Bayes Theorem

- $P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$

  $= \frac{P(B \mid A)\,P(A)}{P(B|A)P(A)+P(B|\neg A)P(\neg A)}$
- P(A) is the prior probability and P(A | B) is the posterior probability.
- Suppose events $A_1, A_2, \ldots\ldots, A_k$ are mutually exclusive and exhaustive; i.e., exactly one of the events must occur. Then for any event B:

  $P(A_i \mid B) = \frac{P(B \mid A_i)\,P(A_i)}{\sum P(B \mid A_i)\,P(A_i)}$

Example I

- According to American Lung Association, 7% of the population has lung cancer. Of these people having lung disease, 90% are smokers; and of those not having lung disease, 25.3% are smokers.
- Determine the probability that a randomly selected smoker has lung cancer.

Bayesian Classifiers

- Consider each attribute and class label as random variables
- Given a record with attributes (A_1, A_2,…,A_n)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes P(C| A_1, A_2,…,A_n)
- Can we estimate P(C| A_1, A_2,…,A_n) directly from data?

Bayesian Classifiers

- Approach:
  - compute the posterior probability P(C | A_1, A_2, …, A_n) for all values of C using the Bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - Choose value of C that maximizes
    P(C | A_1, A_2, …, A_n)
  - Equivalent to choosing value of C that maximizes
    P(A_1, A_2, …, A_n|C) P(C)
- How to estimate P(A_1, A_2, …, A_n | C)?

Naive Bayes

- Naïve Bayes classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes.
- This assumption is called class conditional independence.
- It is made to simplify the computations involved and, in this sense, is considered "naïve".

Naïve Bayes Classifier

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_j)\, P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$
  - Can estimate $P(A_i \mid C_j)$ for all A_i and C_j.
  - New point is classified to C_j if $P(C_j) \prod P(A_i \mid C_j)$ is maximal.

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Class: $P(C) = N_c/N$
  - e.g., P(No) = 7/10, P(Yes) = 3/10
- For discrete attributes:

$$P(A_i \mid C_k) = |A_{ik}| / N_c$$

  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
  - Examples:

P(Status=Married|No) = 4/7
P(Refund=Yes|Yes)=0

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Naïve Bayes

Classification: Mammals vs. Non-mammals

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

- Train the model (learn the parameters) using the given data set.
- Apply the learned model on new cases.

Naïve Bayes

Classification: Mammals vs. Non-mammals

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Example: Play Tennis

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(P) = 9/14
P(N) = 5/14

Outlook	Temperature	Humidity	Windy	Class
rain	hot	high	false	?

outlook	
P(sunny|p) = 2/9	**P(sunny|n) = 3/5**
P(overcast|p) = 4/9	**P(overcast|n) = 0**
P(rain|p) = 3/9	**P(rain|n) = 2/5**
temperature	
P(hot|p) = 2/9	**P(hot|n) = 2/5**
P(mild|p) = 4/9	**P(mild|n) = 2/5**
P(cool|p) = 3/9	**P(cool|n) = 1/5**
humidity	
P(high|p) = 3/9	**P(high|n) = 4/5**
P(normal|p) = 6/9	**P(normal|n) = 2/5**
windy	
P(true|p) = 3/9	**P(true|n) = 3/5**
P(false|p) = 6/9	**P(false|n) = 2/5**

Characteristics of Naïve Bayes Classifiers

- They are robust to isolated noise points because such points are averaged out when estimating conditional probabilities from data.
- Naïve Bayes classifiers can also handle missing values by ignoring the example during model building and classification.
- They are robust to irrelevant attributes. If Xi is an irrelevant attribute, then P(Xi | Y) becomes almost uniformly distributed.
- Correlated attributes can degrade the performance of naïve Bayes classifiers because the conditional independence assumption no longer holds for such attributes.

How Effective are Bayesian Classifiers?

- Various empirical studies of this classifier in comparison to decision tree and neural network classifiers have found it to be comparable in some domain.
- In theory, Bayesian classifiers have the minimum error rate in comparison to all other classifiers.
- However, in practice this is not always the case, owning to inaccuracies in the assumptions made of its use, such as class conditional independence, and the lack of available probability data.

Weka Demo

Accuracy or Error Rates

- Partition: Training-and-testing
  - use two independent data sets, e.g., training set (2/3), test set(1/3)
  - used for data set with large number of examples

Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

Metrics for Performance Evaluation…

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

- Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10
- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Cost Matrix

		PREDICTED CLASS	
	C(i|j)	**Class=Yes**	**Class=No**
ACTUAL CLASS	**Class=Yes**	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

C(i|j): Cost of misclassifying class j example as class i

Cost Matrix (Cont'd)

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	5
	False	1	14

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	3
	False	3	14

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	6
	False	0	14

All three confusion matrices have the same accuracy value, i.e., **24 / 30**

What if the cost of misclassification is not the same for both type of errors?

Cost Matrix (Cont'd)

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	5x5
	False	1	14

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	3x5
	False	3	14

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	6x5
	False	0	14

Suppose the cost of misclassifying True as False is 5 while the cost of misclassifying False as True is 1.

Accuracy values are:
24/50, 24/42, 24/54

Cost Matrix (Cont'd)

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	5x4
	False	1	14

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	3x4
	False	3	14

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	6x4
	False	0	14

Suppose the cost of misclassifying True as False is **4** while the cost of misclassifying False as True is 1.

Accuracy values are:
24/45, 24/39, 24/48

Cost-Sensitive Measures

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F-measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

Recall and Precision

Actual	Prediction
T	T
T	F
F	T
F	F
F	T
T	T
T	T
T	F
F	T
T	T

Recall and Precision

Actual	Prediction
T	T
T	F
F	T
F	F
F	T
T	T
T	T
T	F
F	T
T	T

- Recall = 4 / 6

Recall and Precision

Actual	Prediction
T	T
T	F
F	T
F	F
F	T
T	T
T	T
T	F
F	T
T	T

- Recall = 4 / 6
- Precision = 4 / 7
- F-Measure = 8 / 13