Knowledge Discovery and Data Mining

Unit # 3

Unsupervised Feature Discretization Techniques

- The task of feature discretization techniques is to discretize the values of continuous features into a small number of intervals, where each interval is mapped to a discrete symbol.
- Suppose the set of values for a given feature are {3, 2, 1, 5, 4, 3, 1, 7, 5, 3}. After sorting, these values can be placed into three bins
  - {1, 1, 2, 3, 3, 3, 4, 5, 5, 7}

Value Reduction

- One of the main problems of the previous method is to find the best cutoffs for bins.
- The value-reduction problem can be stated as an optimization problem in the selection of k bins: given the number of bins k, distribute the values in the bins to minimize the average distance of a value from its bin mean or median.
- The distance is usually measured as the absolute distance for a bin mean/median.

Value Reduction – A Heuristic Algorithm

- Sort all values for a given feature.
- Assign approximately equal number of sorted adjacent values (vi) to each bin, where the number of bins is given in advance.
- Move a border element vi from one bin to the next (or previous) when that reduces the global distance error (ER) (the sum of all distances from each vi to the mean or mode of its assigned bin).

Working of the Algorithm

- The set of values for a feature f is {5, 1, 8, 2, 2, 9, 2, 1, 8, 6}.
- Split them into three bins (k = 3), where the bins will be represented by their modes.
- Initial bins are {1, 1, 2 2, 2, 5 6, 8, 8, 9}
- Means for the three bins are {1.33, 3, 7.75}. The error, ER, is 0.33+0.33+0.67+1+1+2+1.75+0.25+0.25+1.25=
- After moving two elements from BIN2 into BIN1 and one element from BIN3 to BIN2 in the next three iterations, the final distribution of elements are {1, 1, 2, 2, 2, 5, 6, 8, 8, 9}
- The total minimized error, ER, is _____.

Value Reduction Exercise

- Perform Bin-based values reduction with the best cutoffs for the following:
  - The feature Attribute 2 (in slide # 18) using mean values as representatives for two bins.
  - Repeat the same exercise for three bins

Chi Square Test Example (Source: Stattrek.com)

- A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent).
- Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences?

	Voting Preferences			Row total
	Republican	Democrat	Independent	
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

Hypothesis

- H_0: Gender and voting preferences are independent.
  H_a: Gender and voting preferences are not independent.

Calculations

- DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2

$E_{r,c} = (n_r * n_c) / n$
$E_{1,1} = (400 * 450) / 1000 = 180000/1000 = 180$
$E_{1,2} = (400 * 450) / 1000 = 180000/1000 = 180$
$E_{1,3} = (400 * 100) / 1000 = 40000/1000 = 40$
$E_{2,1} = (600 * 450) / 1000 = 270000/1000 = 270$
$E_{2,2} = (600 * 450) / 1000 = 270000/1000 = 270$
$E_{2,3} = (600 * 100) / 1000 = 60000/1000 = 60$

Calculations (Cont'd)

- $X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$
  $X^2 = 16.2$
- From the Chi-Square Distribution we find $P(X^2 > 16.2) = 0.0003$.
- **Interpret results**. Since the P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

Another Example (Source: math.hws.edu)

- Relationship between location and type of malaria.

	Asia	Africa	South America	Totals
Malaria A	31	14	45	90
Malaria B	2	5	53	60
Malaria C	53	45	2	100
Totals	86	64	100	250

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

- $\chi^2 = (1\text{-}1)^2/1 + (0\text{-}0.1)^2/0.1 + (1\text{-}1)^2/1 + (0\text{-}0.1)^2/0.1 = 0.2$
- 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.

Attribute1	Class
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.

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:

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

Imbalance Class Problem

- An imbalance class problem occurs when one or more classes have very low proportions in the training data as compared to the other classes.
  - In online advertising, an advertisement is presented to a viewer which creates an impression. The click through rate is the number of times an ad was clicked on divided by the total number of impressions and tends to be very low.

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

KNIME Demo

Terminology

- True Positive: The number of positive examples correctly predicted by the classification model.
- False Negative: The number of positive examples wrongly predicted as negative by the classification model.
- False Positive: The number of negative examples wrongly predicted as positive by the classification model.
- True Negative: The number of negative examples correctly predicted by the classification model.

Terminology (Cont'd)

- The true positive rate (TPR) or sensitivity is defined as TPR = TP / (TP + FN).
- The true negative rate (TNR) or specificity is defined as TNR = TN / (TN + FP).
- The false positive rate (FPR) is defined as FPR = FP / (TN + FP).
- The false negative rate (FNR) is defined as FNR = FN / (TP + FN).

ROC (Receiver Operating Characteristic)

- Developed in 1950s for signal detection theory to analyze noisy signals
  - Characterize the trade-off between positive hits and false alarms
- ROC curve plots TPR (on the y-axis) against FPR (on the x-axis)
- Remember that TPR represents "sensitivity" while FPR represents "100 – specificity".

ROC Curves

- Suppose sensitivity in a given scenario is poor (40%) while specificity is fairly high (92.9%).
- The values are calculated from classes that are determined with the default 50% probability threshold.
- Lowering the threshold to 30% results in a model with 60% sensitivity and 79.3% specificity.

ROC Curve (Cont'd)

- The ROC curve is created by evaluating the class probabilities for the model across a continuum of thresholds.
- For each candidate threshold, the resulting true-positive rate (senstivity) and the false-positive rate (1-specificity) are plotted against each other.

ROC Curve (Cont'd)

- It is important to remember that altering the threshold only has the effect of making samples more positive (or negative as the case may be).
- In the confusion matrix, it cannot move samples out of both off-diagonal table cells. There is almost always a decrease in either sensitivity or specificity as 1 is increased.

ROC Curve (Cont'd)

- The optimal model should be shifted towards the upper left corner of the plot.
- Alternatively, the model with the largest area under the ROC curve would be the most effective.
- The ROC curve is only defined for two-class problems but has been extended to handle three or more classes.

How to Construct an ROC curve

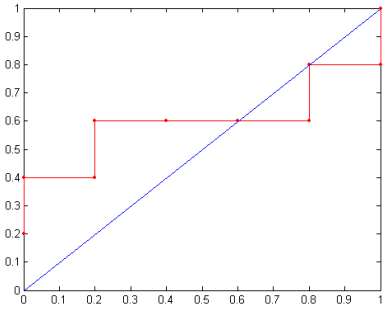
Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

- Use classifier that produces posterior probability for each test instance P(+|A)
- Sort the instances according to P(+|A) in decreasing order
- Apply threshold at each unique value of P(+|A)
- Count the number of TP, FP, TN, FN at each threshold
- TP rate, TPR = TP/(TP+FN)
- FP rate, FPR = FP/(FP + TN)

How to construct an ROC curve

Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
→ TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
→ FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

ROC Curve:



Lift and Gain Charts

- Very commonly used in the marketing research.
- **Lift** is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.
- A lift chart consists of a lift curve and a baseline
- The greater the area between the lift curve and the baseline, the better the model

Lift Charts

- Lift charts are a visualization tool for assessing the ability of a model to detect events in a data set with to classes.
- Suppose a group of samples with M events is scored using the event class probability. When ordered by the class probability, one would hope that the events are ranked higher than the nonevents.
- Lift charts do just this: rank the samples by their scores and determine the cumulative event rate as more samples are evaluated.

Lift Charts (Cont'd)

- In the optimal case, the M highest-ranked samples would contain all M events. When the model is non-informative, the highest-ranked X% of the data would contain, on average, X events.
- The lift is the number of samples detected by a model above a completely random selection of samples.

Steps

- Predict a set of samples that were not used in the model building process but have known outcomes
- Determine the baseline event rate, i.e., the percent of true events in the entire data set.
- Order the data by the classification probability of the event of interest.
- For each unique class probability value, calculate the percent of true events in all samples below the probability value.
- Divide the percent of true events for each probability threshold by the baseline event rate.

Example

http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html

- Using the response model P(*x*)=100-AGE(*x*) for customer *x* and the data table, construct the cumulative gains and lift charts. Ties in ranking should be arbitrarily broken by assigning a higher rank to who appears first in the table.

Customer Name	*Height*	*Age*	*Actual Response*
Alan	70	39	N
Bob	72	21	Y
Jessica	65	25	Y
Elizabeth	62	30	Y
Hilary	67	19	Y
Fred	69	48	N
Alex	65	12	Y
Margot	63	51	N
Sean	71	65	Y
Chris	73	42	N
Philip	75	20	Y
Catherine	70	23	N
Amy	69	13	N
Erin	68	35	Y
Trent	72	55	N
Preston	68	25	N
John	64	76	N
Nancy	64	24	Y
Kim	72	31	N
Laura	62	29	Y

Example: Steps 1 & 2

1. Calculate P(*x*) for each person *x*
2. Order the people according to rank P(*x*)

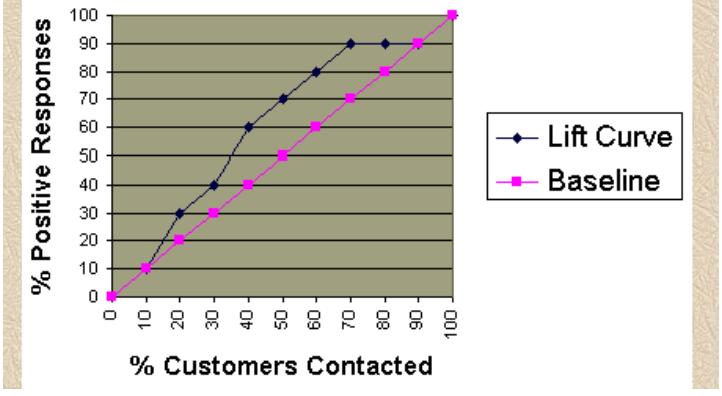
Customer Name	*P(x)*	*Actual Response*
Alex	88	Y
Amy	87	N
Hilary	81	Y
Philip	80	Y
Bob	79	Y
Catherine	77	N
Nancy	76	Y
Jessica	75	Y
Preston	75	N
Laura	71	Y
Elizabeth	70	Y
Kim	69	N
Erin	65	Y
Alan	61	N
Chris	58	N
Fred	52	N
Margot	49	N
Trent	45	N
Sean	35	Y
John	24	N

Example: Step 3

- Calculate the percentage of total responses for each cutoff point
  - Response Rate = Number of Responses / Total Number of Responses (10)

Total Customers Contacted	Number of Responses	Response Rate
2	1	10%
4	3	30%
6	4	40%
8	6	60%
10	7	70%
12	8	80%
14	9	90%
16	9	90%
18	9	90%
20	10	100%

Example: Lift Charts



Similarity with ROC Curves

- Like ROC curves, the lift curves for different models can be compared to find the most appropriate model and the area under the curve can be used as a quantitative measure of performance.