Knowledge Discovery and Data Mining

Unit # 4

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*)

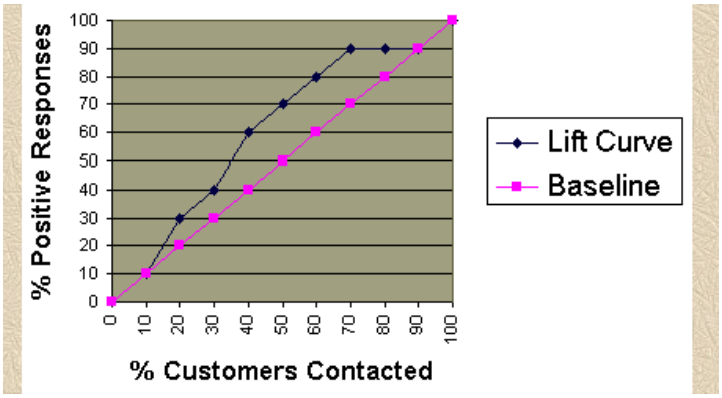
Customer Name	*P(x)*	*Actual Response*
Alex	88	Y
Amy	87	N
Hilary	81	Y
Philip	80	Y
Bob	79	Y
Catherine	77	N
Nancy	76	Y
Jessica	75	Y
Preston	75	N
Laura	71	Y
Elizabeth	70	Y
Kim	69	N
Erin	65	Y
Alan	61	N
Chris	58	N
Fred	52	N
Margot	49	N
Trent	45	N
Sean	35	Y
John	24	N

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.

KIME Discussion

Entropy-based Measure for Feature Ranking

- The distribution of all similarities for a given data set is a characteristic of the organization and order of data in an n-dimensional space. This may be measured by entropy.
- The proposed technique compares the entropy measure for a given data set before and after removal of a feature. If the two measures are close, then the reduced set of features will satisfactorily approximate the original set.
- $E = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (S_{ij} \times \log S_{ij} + (1 - S_{ij}) \times \log (1 - S_{ij}))$

Entropy-based Measure for Feature Ranking (Cont'd)

- $Sij = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}| \right) / n$
- Where $|x_{ik} - x_{jk}|$ is 1 if $x_{ik} = x_{jk}$, and 0 otherwise.
- For mixed data, we can discretize numeric values and transform numeric features into nominal features before we apply this similarity measure.

Example

- $E = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (S_{ij} \times \log S_{ij} + (1 - S_{ij}) \times \log (1 - S_{ij}))$

Sample	F1	F2	F3
R1	A	X	1
R2	B	Y	2
R3	C	Y	2
R4	B	X	1
R5	C	Z	3

	R1	R2	R3	R4	R5
R1		0/3	0/3	2/3	0/3
R2			2/3	1/3	0/3
R3				0/3	1/3
R4					0/3

Algorithm: Entropy based Ranking (Sequential Backward Ranking)

1. Start with the initial full set of features F.
2. For each feature $f \in F$, remove one feature F and obtain a subset F_f. Find the difference between entropy for F and entropy for all F_f.
3. Let f_k be a feature such that the difference between entropy for F and entropy for f_{fk} is minimum.
4. Update the set of features $F = F - \{f_k\}$.
5. Repeat steps 2-4 until there is only one feature.

Entropy-based Feature Ranking Exercise

- Given four-dimensional samples where the first two dimensions are numeric and last two are categorical

X1	X2	X3	X4
2.7	3.4	1	A
3.1	6.2	2	A
4.5	2.8	1	B
5.3	5.8	2	B
6.6	3.1	1	A
5.0	4.1	2	B

- Apply a method for unsupervised feature selection based on entropy measure to reduce one dimension from the given data set

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

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.