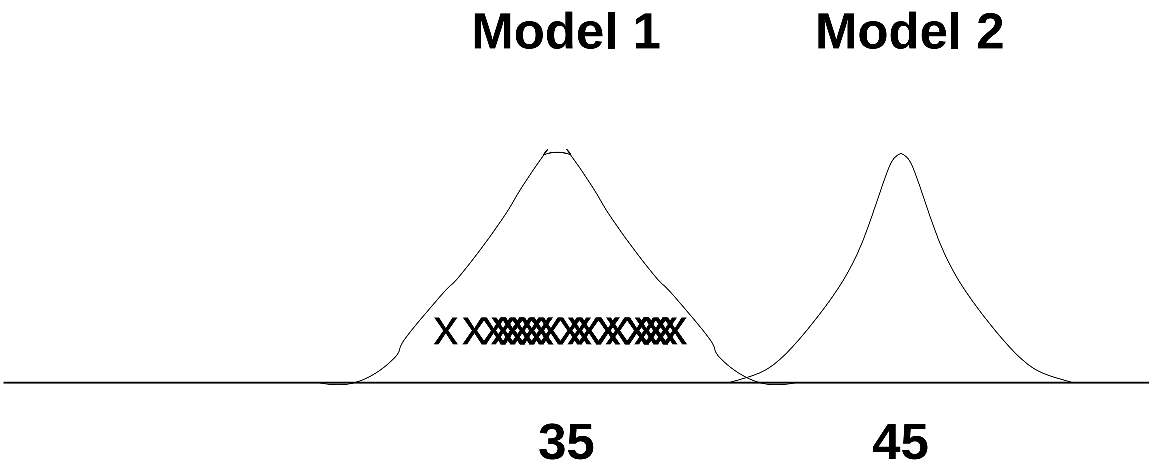
CLUSTERING WITH EM 1

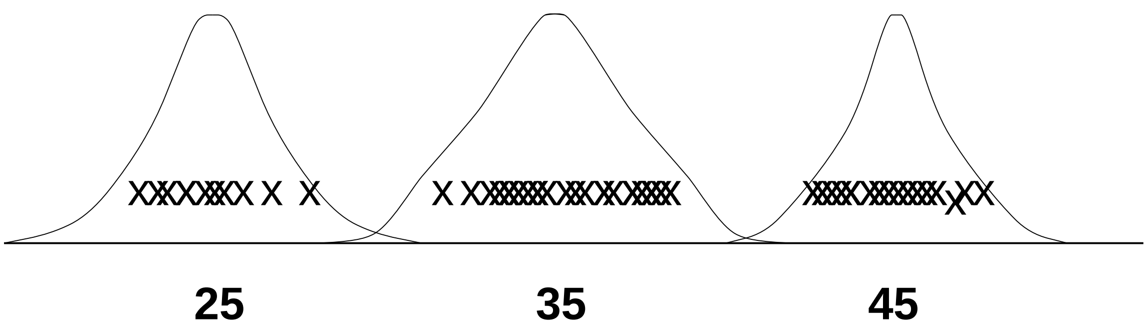


- Fitting a probability model to data
- Model 1, Normal, mean 35, SD 2, is a better fit
- than Model 2, Normal, mean 45, SD 1.5
- Goodness of fit can be calculated

CLUSTERING WITH EM 2

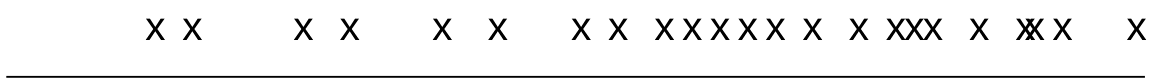
- Consider the following ages of people in a DB:

- It is reasonably clear that there are 3 clusters and that 3 gaussians would be a reasonable fit.

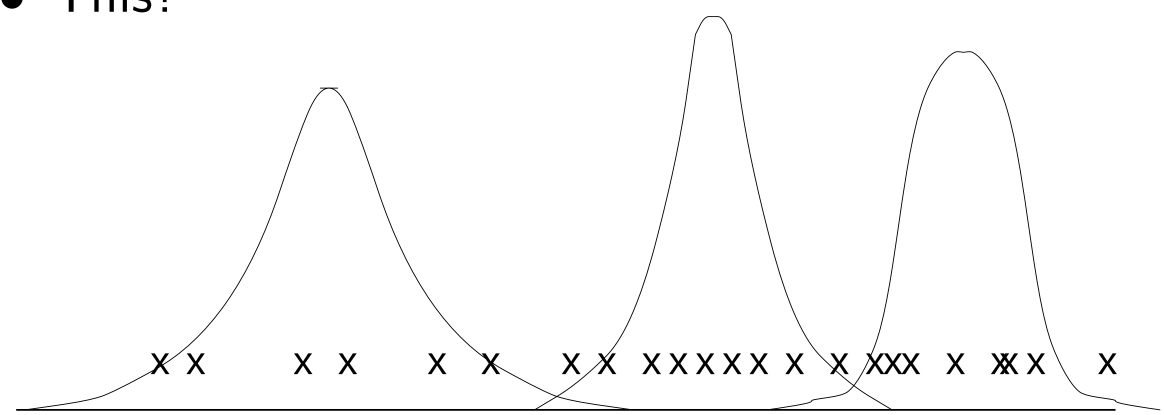


CLUSTERING WITH EM 3

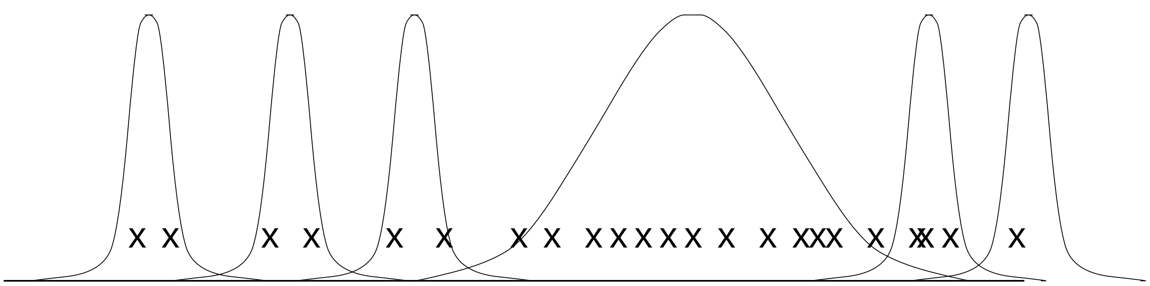
- What distributions best fit the following?



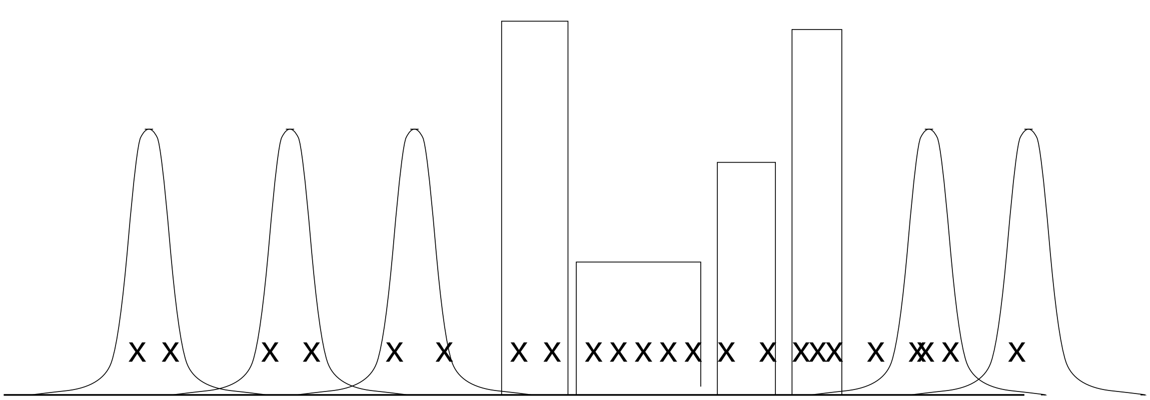
- This?



- Or this?

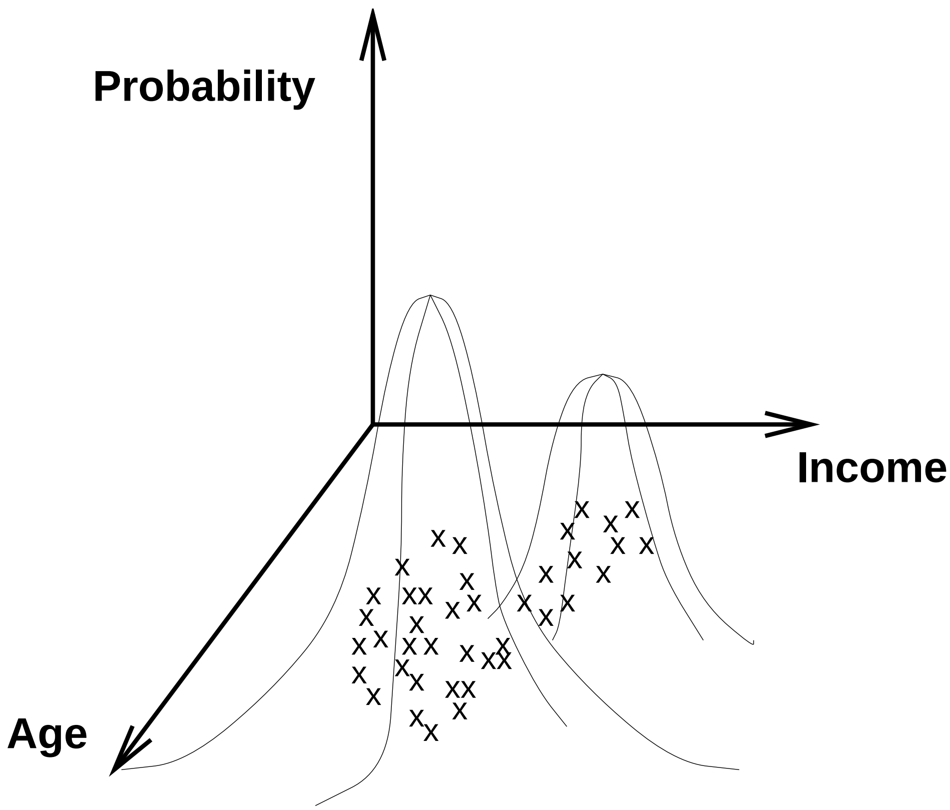


- Or this?

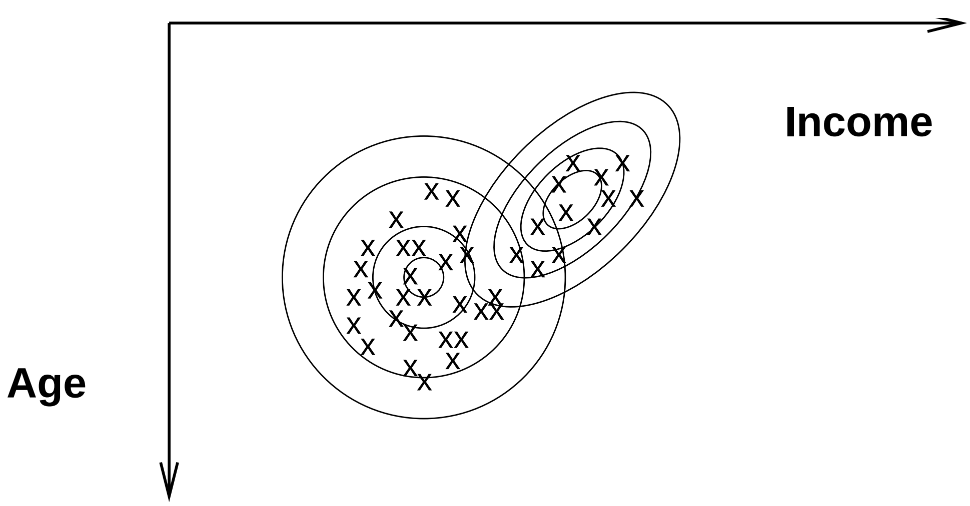


CLUSTERING WITH EM 3

In general we will have points in n dimensions.



Probability ‘isolines’ looking down from top



OUTLINE OF EM ALGORITHM

EM = Expectation Maximisation

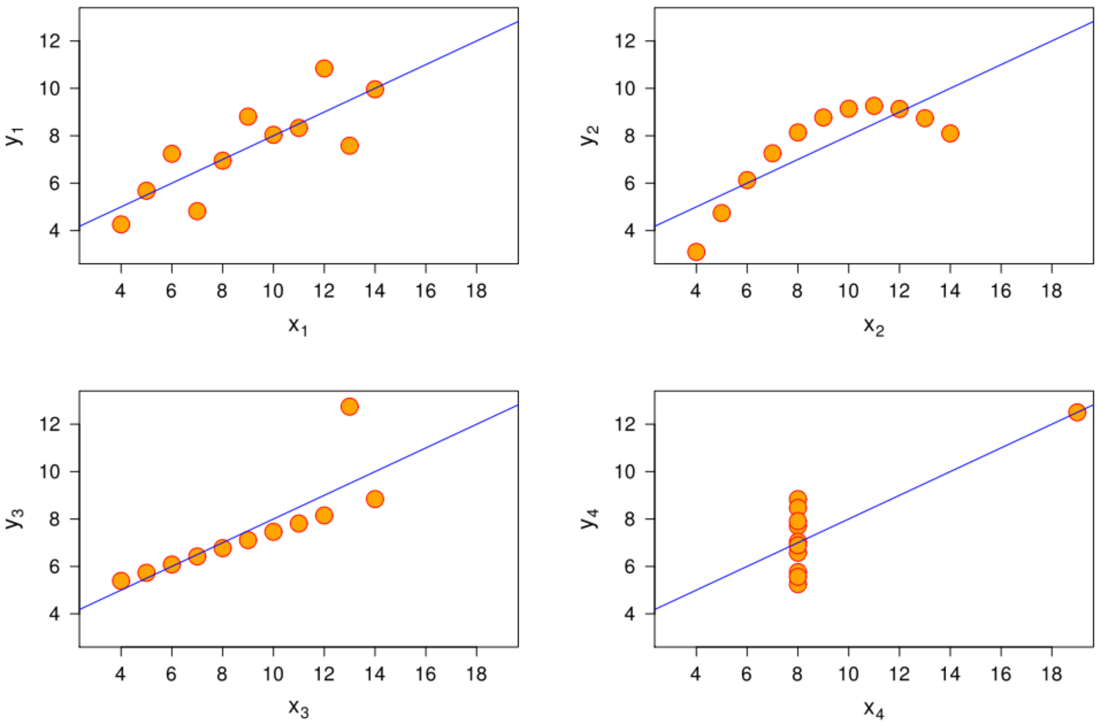
1. Based on fitting probability distributions to the data
2. Set K = 1
3. Randomly generate K means and standard deviations.
4. Measure how well the distributions fit the data
5. If the fit can be improved compute new means and SDs and go to 4
6. K = K + 1
7. If clustering can be improved go to 4

- Theory behind EM
  - Finite Mixture Models
  - Optimization, Local optimum vs global optimum
  - Search

CLUSTERING IN PRACTICE

- You don't necessarily need to get the exact number of clusters to get something useful.
- You can be happy if you get a small number of meaningful clusters.
- There is no single measure of the best clustering result.
- Clustering algorithms don't scale well with number of records. Might need to sample.
- Clustering algorithms don't scale well with number of attributes. Use domain knowledge to select attributes

NUMBERS CAN BE MISLEADING



Four datasets for which the statistical properties mean, variance, correlation and regression line are the same.

Property	Value
Mean of each x variables	9.0
Variance of each x variables	11.0
Mean of each y variables	7.5
Variance of each y variables	4.12
Correlation between each x and y variable	0.816
Regression line	$y = 3 + 0.5x$

Source: `http://upload.wikimedia.org/wikipedia/commons/thumb/b/b6/Anscombe.svg/\\1000px-Anscombe.svg.png`

ASSOCIATION FINDING

- Finding inherent regularities in data
- Frequent Patterns
- Market Basket Analysis
  - People who buy milk often buy bread
  - People who buy beer often by potato chips
  - People who buy beer often buy nappies
  - What products are often purchased together?
- What are the subsequent purchases after buying a PC?
- What kinds of DNA are sensitive to this new drug?
- Web browsing patterns
- Applications Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

MARKET BASKET ANALYSIS

- Uses
  - If two items are often purchased together locate them close together. [Shopper will come for one item, see and buy the other]
  - If two items are often purchased together locate them far apart. [Shopper will come for one item, buy other items while they look for the second.]
  - If shopper buys one item, suggest that the they might be interested in related items. [Amazon]

ITEMS and ITEM SETS

- Item: Presence of something (in a transaction)
  bread, milk, coffee, sugar, eggs

- Item: A combination of attribute and value (in an arff file)
  sex=m, sex=f, class=verginica

- Item set: A set of items
  {bread,milk}
  {bread,coffee,eggs}

  {sex=m}
  {sex=f,class=verginica}

- Frequent Item set: Occurs with a minimum support (coverage)

ASSOCIATION RULE

```
if milk then bread
       [Coverage=5%,Accuracy=60%]
milk ==> bread
       [Coverage=5%,Accuracy=60%]

sex=m and student=no ==>  movie=action
       [Coverage=3%,Accuracy=80%]

sex=m ==> student=no and  movie=action
       [Coverage=2%,Accuracy=70%]
```

Coverage/Support Percentage of transactions/records to which the rule applies.
In 5% of all transactions people bought milk.
In 3% all records (sex=m and student=no)

Accuracy/Confidence The percentage of times the consequent appears with antecedent.
60% of the time that a person bought bread, they also bought milk.

70% of the times that sex=m and student=no then movie=action

WEATHER DATA

- Will I play golf?

Outlook	Temp	Humidity	Windy	Play
Rainy	Mild	High	True	No
Overcast	Hot	Normal	False	Yes
Overcast	Mild	High	True	Yes
Sunny	Mild	Normal	True	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Mild	High	False	No
Overcast	Cool	Normal	True	Yes
Rainy	Cool	Normal	True	No
Rainy	Cool	Normal	False	Yes
Rainy	Mild	High	False	Yes
Overcast	Hot	High	False	Yes
Sunny	Hot	High	True	No
Sunny	Hot	High	False	No

Note: This is the file `weather.nominal.arff` in the Weka distribution

ITEM SETS FROM WEATHER D

One-item sets	Two-item sets	Three-item sets	F
Outlook=Sunny(5)	Outlook=Sunny Temperature=Hot(2)	Outlook=Sunny Temperature=Hot Humidity=high(2)	O T H P
Temp=Cool(4)	Outlook=Sunny Humidity=High(3)	Outlook=Sunny Humidity=High Windy=False(2)	O T W P
...	...	...	.

- In total: 12 one-item sets, 47 two-item sets, 39 three-iter
  and 0 five-item sets (with minimum support of two)

GENERATING RULES FROM ITEM SETS

- First get all of the item sets
- Example:

```
Humidity = Normal, Windy = False, Play = Yes (4)
```

- Seven $(2^N - 1)$ potential rules

```
If Humidity=Normal and Windy=False then Play=Yes  4/4
If Humidity=Normal and Play=Yes then Windy=False  4/6
If Windy=False and Play=Yes then Humidity=Normal  4/6
If Humidity=Normal then Windy=False and Play=Yes  4/7
If Windy=False then Humidity=Normal and Play=Yes  4/8
If Play=Yes then Humidity=Normal and Windy=False  4/9
If True then Humidity=Normal and Windy=False
   and Play=Yes                                    4/12
```

ASSOCIATION RULES FOR WEATHE

- Rules with support > 1 and confidence=100%

	Rule	
1	Humidity=Normal Windy=False	==> Play=Yes
2	Temperature=Cool	==> Humidity=Norma
3	Outlook=Overcast	==> Play=Yes
4	Temperature=Cold Play=Yes	==> Humidity=Norma
...	...	...
58	Outlook=Sunny Temperature=Hot	==> Humidity=High

- In Total:
  3 rules with support four
  5 with support three
  50 with support two

RULES FROM THE SAME ITEM

- Item set

  Temperature = Cool, Humidity = Normal, Windy = Fals

- Resulting rules (all with 100% confidence):

  Temperature = Cool, Windy = False **==>** Humidity =
  Temperature = Cool, Windy = False Humidity = Nor
  Temperature = Cool, Windy = False, Play = Yes **==>** H

- Due to the following 'frequent' item sets:

  Temperature = Cool, Windy = False (2) Temperature
  Normal, Windy = False (2) Temperature = Cool, Wind
  (2)

FREQUENT ITEM SETS

- A $frequent$ item set is an item set that meets a previously specified minimum support/coverage
- A $large$ item set is the same as a frequent item set
- Use of $large$ is historical

EFFICIENT GENERATION OF ITEM SETS

- Finding one-item sets is easy
- Basic idea: Use one-item sets to generate two-item sets, two-item sets to generate three-item-sets
- Theorems:
  - If {A,B} is a frequent item set, then {A} and {B} must be frequent.
  - If X is a frequent k-item set, then all $(k-1)$ item subsets of X must be frequent.
- Compute k-item set by merging $(k-1)$ item sets

EFFICIENT GENERATION OF ASSOCIATION RULES

- Many transactions contain may items
- There may be many possible items
- Data is sparse, many items are not purchased in supermarket trip
- There may be many transactions, too much for main memory
- Finding association rules requires a lot of search
- Good data structures and algorithms are needed.
  - Still a major research area

APRIORI in WEKA

1. Set minimum support to 100%
2. Set number of rules required
3. Set minimum confidence
4. Generate rules
5. If not time to stop
   Decrease confidence by 5%
   Go to 4
6. Stop if
   Enough rules have been generated
   Minimum confidence is reached
   Support reaches 10%

APRIORI in WEKA

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D
Relation:     cluster1.csv
Instances:    200
Attributes:   3
              Sex
              Student
              MovieType
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.1 (20 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 7
Size of set of large itemsets L(2): 10
Size of set of large itemsets L(3): 4
Best rules found:
 1. Student=y MovieType=action 41 ==> Sex=m 40
    <conf:(0.98)> lift:(1.82) lev:(0.09) [18] conv:(9.53)
 2. MovieType=action 86 ==> Sex=m 82
    <conf:(0.95)> lift:(1.78) lev:(0.18) [35] conv:(8)
 3. Student=n MovieType=action 45 ==> Sex=m 42
    <conf:(0.93)> lift:(1.74) lev:(0.09) [17] conv:(5.23)
 4. Sex=f Student=y 48 ==> MovieType=romance 44
   <conf:(0.92)> lift:(1.95) lev:(0.11) [21] conv:(5.09)
```

GENERATED ITEM SETS

```
Size of set of large itemsets L(1): 7
Large Itemsets L(1):
Sex=f 93
Sex=m 107
Student=n 97
Student=y 103
MovieType=action 86
MovieType=horror 20
MovieType=romance 94
Size of set of large itemsets L(2): 10
Large Itemsets L(2):
Sex=f Student=n 45
Sex=f Student=y 48
Sex=f MovieType=romance 82
Sex=m Student=n 52
Sex=m Student=y 55
Sex=m MovieType=action 82
Student=n MovieType=action 45
Student=n MovieType=romance 45
Student=y MovieType=action 41
Student=y MovieType=romance 49
Size of set of large itemsets L(3): 4
Large Itemsets L(3):
Sex=f Student=n MovieType=romance 38
Sex=f Student=y MovieType=romance 44
Sex=m Student=n MovieType=action 42
Sex=m Student=y MovieType=action 40
```

ASSOCIATIONS NOT ALWAYS USEFUL

```
Apriori
=======

Minimum support: 0.95 (4396 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 9

Size of set of large itemsets L(3): 6

Size of set of large itemsets L(4): 1

Best rules found:

 1. mutton=f 4604 ==> salads=f 4598    <conf:(1)
 2. cigarette cartons=f 4590 ==> salads=f 4584    <conf:(1)
 3. cigarette cartons=f mutton=f 4567 ==> salads=f 4561    <conf:(1)
 4. brushware=f 4518 ==> salads=f 4512    <conf:(1)
 5. brushware=f mutton=f 4495 ==> salads=f 4489    <conf:(1)
 6. cigarette cartons=f brushware=f 4481 ==> salads=f 4475    <conf:(1)
 7. cigarette cartons=f brushware=f mutton=f 4458 ==> salads=f 4452    <conf:(1)
 8. casks white wine=f 4453 ==> salads=f 4447    <conf:(1)
 9. mutton=f casks white wine=f 4430 ==> salads=f 4424    <conf:(1)
10. cigarette cartons=f casks white wine=f 4416 ==> salads=f 4410 <conf:(1)
```

If they didn't buy mutton they didn't buy salads

RULE METRICS

Confidence The percentage of times the consequent appears with antecedent.

Lift $\frac{confidence}{support}$

How much better than statistical independence.

Comes from direct marketing. If the response rate for all the data is 5% but rule finds a segment with a response rate of 20% the lift of the segment is 4.0 (20%/5%).

Leverage Based on statistical properties

Conviction Alternative measure

Support Percentage of transactions/records to which the rule applies.