Feature Selection (FS)

- Feature(attribute) analysis using mutual information

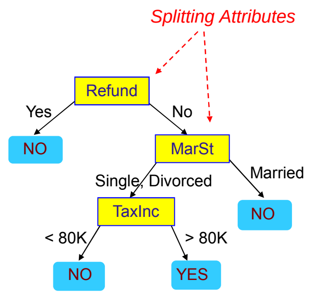
1

Decision Tree analyze features, one feature at a time

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Training Data



Model: Decision Tree

Feature Selection: Select one feature at time

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Training Data

Pick feature 'refund'
Compute its relevance to class label 'cheat'

'refund' feature is a vector over all data instances
class label is a vector over all data instances

Compute similarity(feature vector, label vector)

Repeat for next feature 'marital status'

List scores for all features
Rank a feature according to its score
Pick top 5, 10, 20, etc features for classification

Feature Selection: Select one feature at time

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Training Data

Pick feature 'refund'
Compute its relevance to class label 'cheat'

'refund' feature is a vector over all data instances
class label is a vector over all data instances

Compute similarity(feature vector, label vector)

Repeat for next feature 'marital status'

List scores for all features
Rank a feature according to its score
Pick top 5, 10, 20, etc features for classification

Feature Selection: Select one feature at time

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Training Data

Most central question:
Compute similarity(feature vector, label vector)

class label vector: discrete label values

Repeat for next feature ‘marital status’

List scores for all features
Rank a feature according to its score
Pick top 5, 10, 20, etc features for classification

Compute feature relevance to class labels

Class label vector
- Categorical values: class names

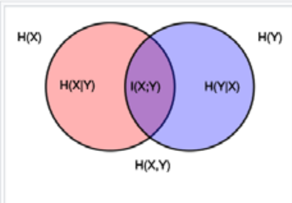
Feature Vector
- Numerical values: salary in dollars, height in inches, time in seconds, etc
- Categorical values: marriage status, job type, education, etc
- Ordinal values: grades (A-F), ranking (1-10), size(large, medium, small)

Similarity between class label vector and feature vector depends on
- Number of classes
- Feature vector value types

Compute feature relevance to class labels

Similarity between class-label-vector and feature-vector
- Number of classes = 2
  - Express class as (+1,-1)
  - Features are numerical, use Pearson correlation, t-test, Relief, sparse-coding
  - Features are categorical, num_category = 2: use Pearson correction, t-test, Relief
  - Features categorical, num_category >2: use mutual information
- Number of classes > 2
  - Features are numerical, use F-test, Relief, sparse-coding
  - Features are categorical, use mutual information

Mutual Information (information gain)



Venn diagram for various information measures associated with correlated variables X and Y. The area contained by both circles is the joint entropy $H(X,Y)$. The circle on the left (red and violet) is the individual entropy $H(X)$, with the red being the conditional entropy $H(X|Y)$. The circle on the right (blue and violet) is $H(Y)$, with the blue being $H(Y|X)$. The violet is the mutual information $I(X;Y)$.

Formally, the mutual information [1] of two discrete random variables X and Y can be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively.

Relation to other quantities [edit]

Mutual information can be equivalently expressed as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \quad = \text{Information gain} \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned}$$

where $H(X)$ and $H(Y)$ are the marginal entropies, $H(X|Y)$ and $H(Y|X)$ are the conditional entropies, $H(X,Y)$ is the joint entropy of X and Y.

Mutual Information (information gain)

Formally, the mutual information [1] of two discrete random variables X and Y can be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively.

Relation to other quantities [edit]

Mutual information can be equivalently expressed as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \quad = \text{Information gain} \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned}$$

where $H(X)$ and $H(Y)$ are the marginal entropies.

$$\begin{aligned} I(X,Y) &= \sum_{x,y} P(x,y) \log \frac{P(y|x)P(x)}{P(x)P(y)} \\ &= \sum_{x,y} P(x,y) \log \frac{P(y|x)}{P(y)} \\ &= -\sum_{x,y} P(x|y)P(y) \log P(y) + \sum_{x,y} P(y|x)P(x) \log P(y|x) \\ &= -\sum_{y} P(y) \log P(y) + \sum_{x} P(x) \sum_{y} P(y|x) \log P(y|x) \\ &= H(Y) - \sum_{x} P(x) H(Y|x) \\ &= H(Y) - H(Y|X) \end{aligned}$$

Because $I(X,Y) \geq 0$. Thus $H(Y) \geq H(Y|X)$

Information(Y) is gained [entropy(Y) is decreased] once X is specified (observed).

We seek the attribute X such that information is gained most, i.e. $H(Y) - H(Y|X) = I(X,Y)$ is maximized

Compute feature relevance to class labels

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Mutual information (refund, cheat)

Mutual information (MarStatus, cheat)

Correlation(TaxInc, cheat)

Group TaxInc into
{high:100-125K, middle: 85-95K, low: 65-75K}
Mutual information (TaxInc, cheat)

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Compute Mutual information (Marital Status=X, cheat=Y)

	Class =Yes	Class=No	
MarSt=Single	2	2	4
MarSt=Married	0	4	4
MarSt=Divorced	1	1	2
	3	7	10

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

$$\text{I(X,Y)} = \frac{2}{10}\log\frac{\frac{2}{10}}{\frac{4}{10}\frac{3}{10}} + \frac{2}{10}\log\frac{\frac{2}{10}}{\frac{4}{10}\frac{7}{10}} + \frac{4}{10}\log\frac{\frac{4}{10}}{\frac{4}{10}\frac{7}{10}} + \frac{1}{10}\log\frac{\frac{1}{10}}{\frac{2}{10}\frac{3}{10}} + \frac{1}{10}\log\frac{\frac{1}{10}}{\frac{2}{10}\frac{7}{10}} = 0.2813$$

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Compute Mutual information (Marital Status=X, cheat=Y)

	Class=Yes	Class=No
MarSt=Single	2	2
MarSt=Married	0	4
MarSt=Divorced	1	1

Mutual-info I(X,Y) = H(Y) –H(Y|X) = Info-gain

H(Y) = -0.3log(0.3)-(0.7)log(0.7) = 0.8813

Split on X=Marital Status:

H(Y|X=Single) = -(2/4)log(2/4)- (2/4)log(2/4) =1

H(Y|X=Married) =0

H(Y|X=Divorced) = -(1/2)log(1/2)- (1/2)log(1/2) =1

H(Y|X) = 0.4(1)+0.4(0)+0.2(1)=0.6

Info-Gain = H(Y) – H(Y|X) = 0.8813-0.6 = 0.2813 same as before

Compute Mutual information (refund, cheat)

Tid	Refund	Marital Status	Taxable Income	Class
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

Before Splitting:
Entropy(Parent)
= -0.3 log(0.3)-(0.7)log(0.7) = 0.8813

	Class = Yes	Class = No
Refund=Yes	0	3
Refund=No	2	4
Refund=?	1	0

Split on Refund:

Entropy(Refund=Yes) = 0

Entropy(Refund=No)
= -(2/6)log(2/6) – (4/6)log(4/6) = 0.9183

Entropy(Children)
= 0.3 (0) + 0.6 (0.9183) = 0.551

Gain = 0.9 × (0.8813 – 0.551) = 0.3303

Compute Mutual information (Taxable Income, cheat)

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	High	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Middle	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Middle	Yes
9	No	Married	Low	No
10	No	Single	Middle	Yes

Group TaxInc into
{high:100-125K,
middle: 85-95K,
low: 60-75K}

	Class=Yes	Class=No
High	0	4
Middle	3	0
Low	0	3

Mutual-info I(X,Y) = H(Y) –H(Y|X) = Info-gain
H(Y) = -0.3log(0.3)-(0.7)log(0.7) = 0.8813

Split on X =Taxable Income:
H(Y|X=High) = 0
H(Y|X=Middle) =0
H(Y|X=Low) = 0

H(Y|X)= 0.4(0)+0.3(0)+0.3(0)=0
Info-Gain = H(Y) – H(Y|X) = 0.8813-0= 0.8813

Compute feature relevance to class labels

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Mutual information is maximized

I (MarStatus, Y) = 0.2813
I (Refund, Y) = 0.3303
I (TaxInc, Y) = 0.8813

You can intuitively take X that minimize H(Y|X)
since $H(Y) - H(Y|X) = I(X,Y)$

And $H(Y)$ is fixed for different **x**

TexInc has the largest mutual information.
We should use TexInc to split the data.