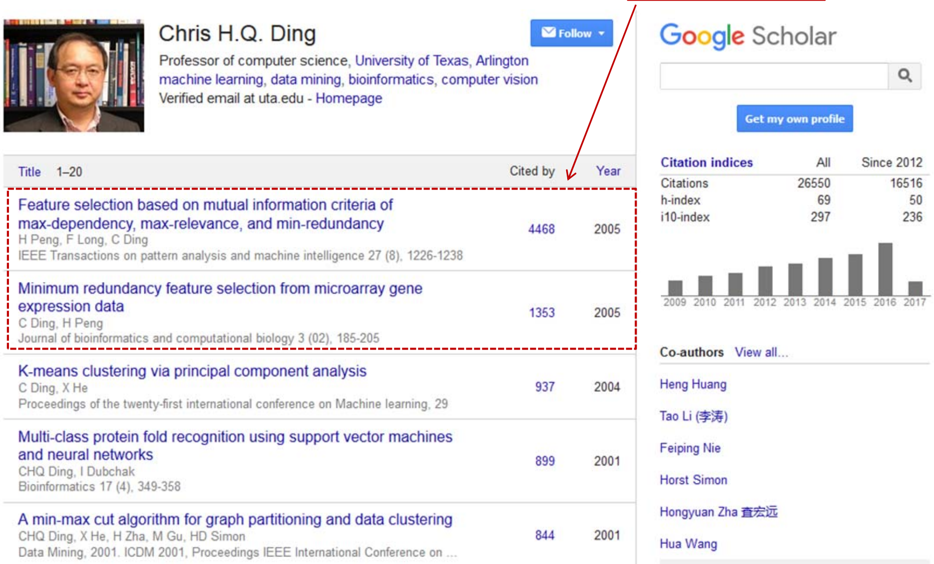
Feature Selection (FS)

- FS is one of 3 main tasks for Machine Learning
- A patient may have cancer or not (disease class label)
- An image have keywords descriptors (class labels)
- A data instance (image, patients) have class labels due to many factors/reasons.
- FS finds the most important factors for classification
  - Most relevant genes for a dieses
  - e.g. select smoking for lung cancer

Many FS methods/algorithms

- T- test, F- test, chi-statistic
- Mutual information
- ReliefF (relevant-non-relevant)
- mRMR (minimum redundancy maximum relevance, Ding & Peng,2005)
- many others

Our 2 feature selection papers cited 5821 times



Feature Selection (FS)

- FS finds the most important factors for classification

3

Feature Selection (FS)

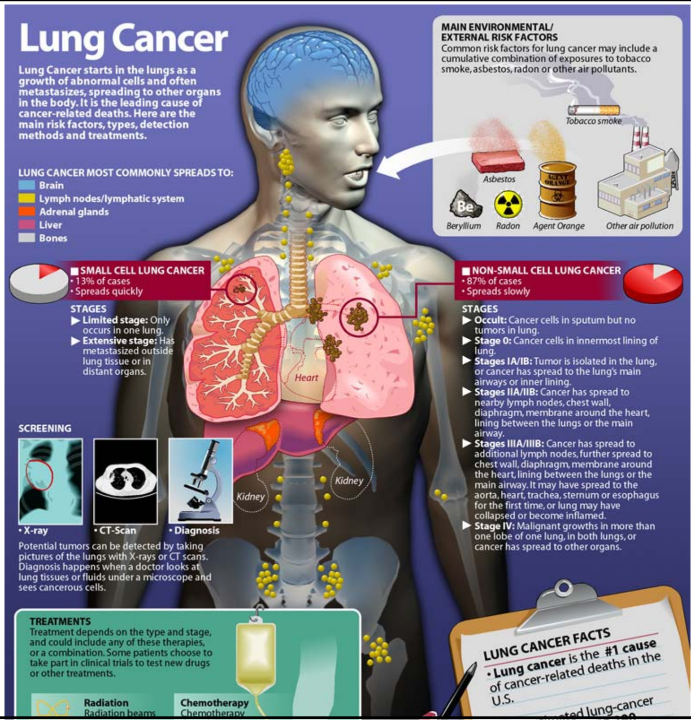
- FS finds the most important factors for classification

4

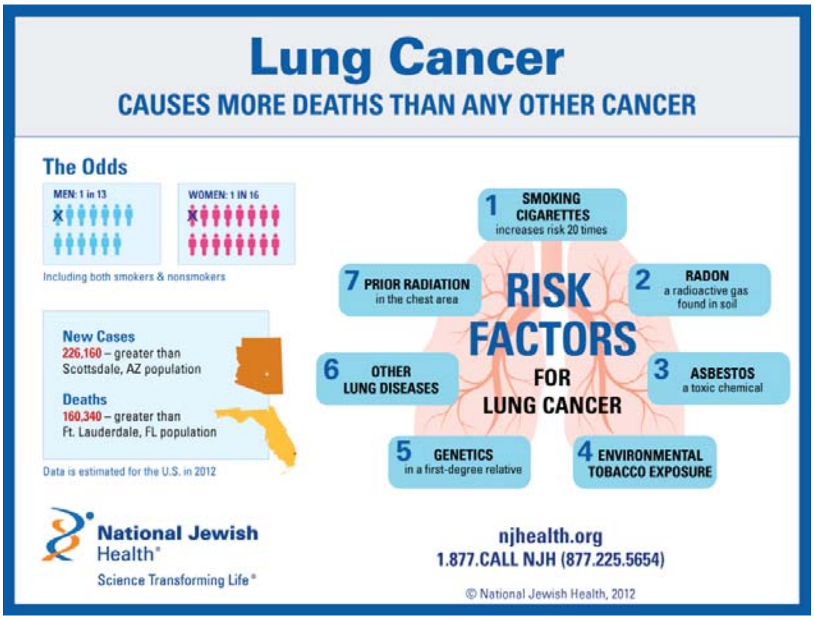
Feature Selection Example

No 1 cause of cancer-related death in US

Feature selection ➔ finding related factors



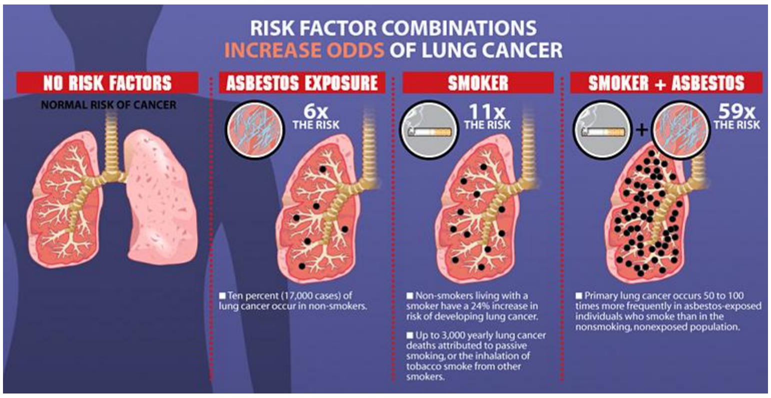
Feature Selection Example: lung cancer



Feature Selection Example: lung cancer

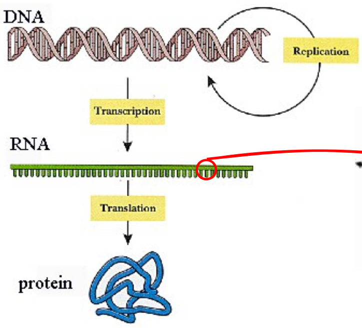
Singer Factors: smoking, expose to asbestos, etc

Multi factors: increase risk much more



7

Feature Selection Example: DNA microarray Gene Expressions



Microarray Analysis

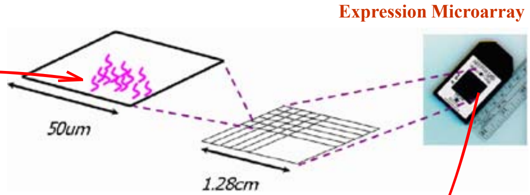


Image Courtesy of Affymetrix

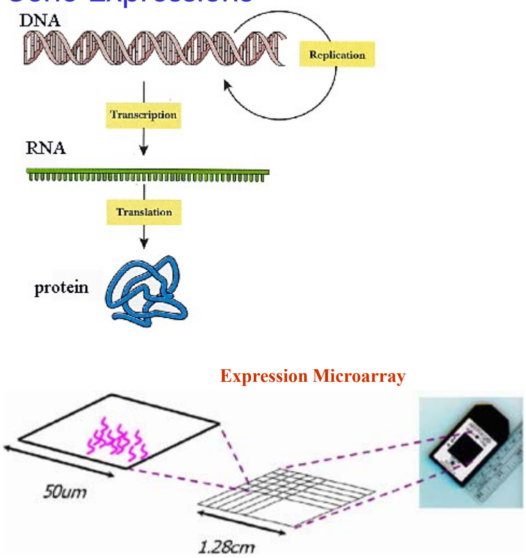
- **Analyze gene expressions of** disease types (disease diagnosis)
- **Challenge:** thousands of genes, few patients samples
- **Solution:** select several most relevant genes

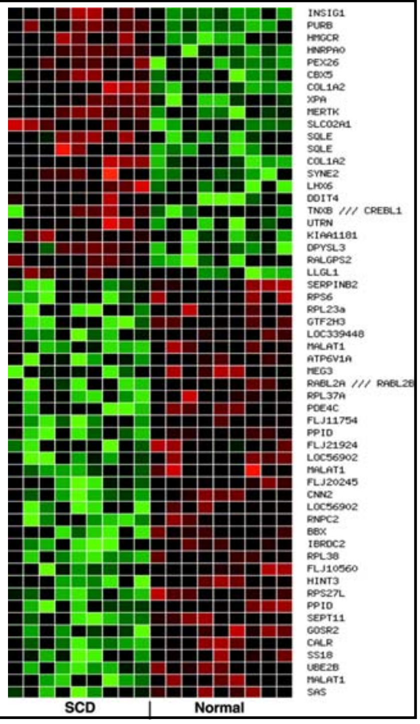
Sample \ Gene	M23197_at	U66497_at	M92287_at	...	Class
Sample 1	261	88	4778	...	ALL
Sample 2	101	74	2700	...	ALL
Sample 3	1450	34	498	...	AML
.	.	.	.	...	.

Expression Microarray Data Set

8

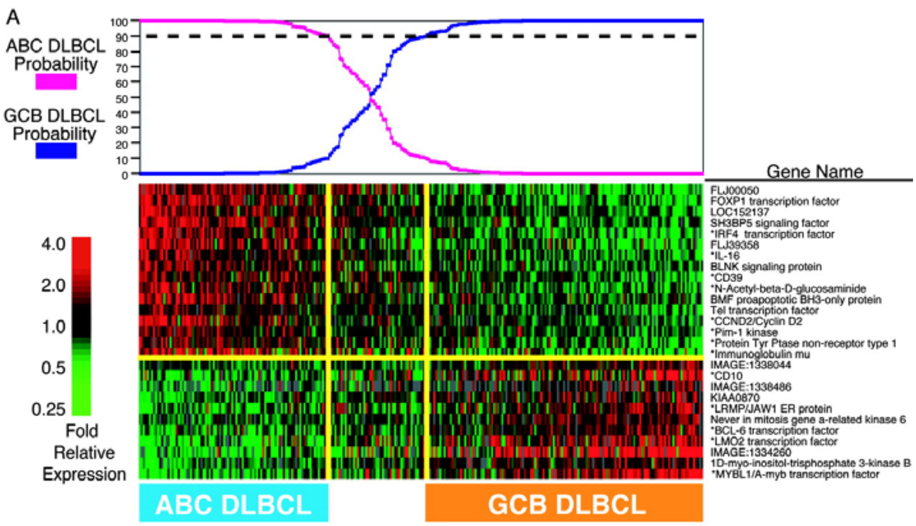
Feature Selection Example: Gene Expressions





9

Feature Selection Example: Gene Expressions pick most relevant genes for each groups: ABC vs GCB



Feature Selection Example: What are distinctive features to distinguish between men and women? old vs young?

St Marco Square, Venice

Feature Selection Example: What are distinctive features to distinguish between men and women? old vs young?

St Marco Square, Venice

Venice is sinking because too many visitors !

Feature Selection (FS)

- FS finds the most important factors for classification
- **Classifiers do feature analysis**

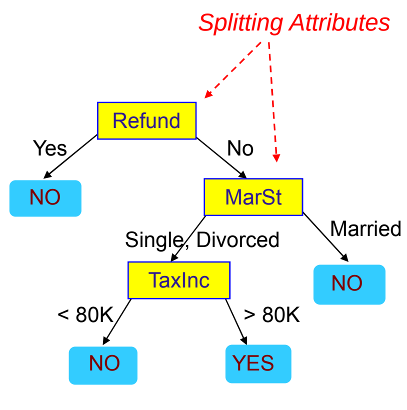
13

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

Rule-based classifier: match features to rules

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds
R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes
R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals
R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles
R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

Feature Selection (FS)

- FS finds the most important factors for classification
- Classifiers do feature analysis
  - All classifiers uses features to make prediction
- IRIS data feature selection example

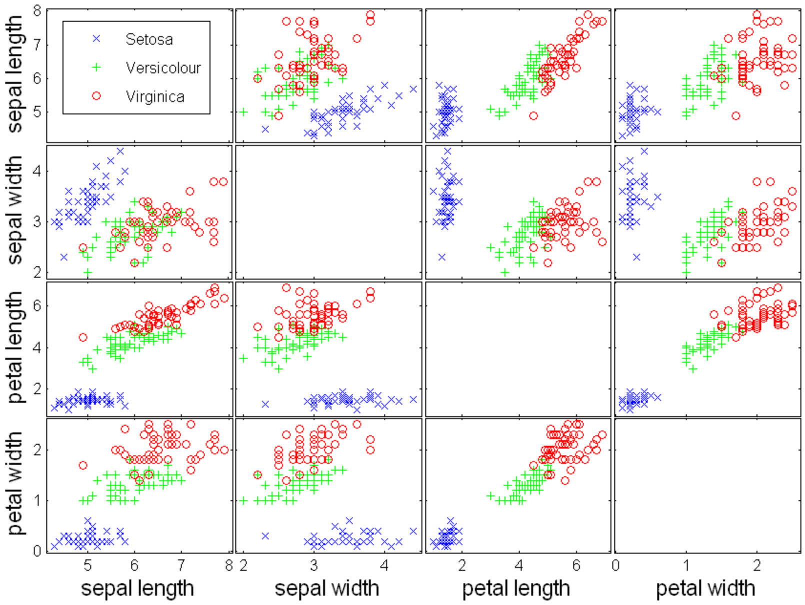
Iris Data Set

- Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Versicolour
    - Virginica
  - Four attributes/features
    - Sepal length
    - Sepal width
    - Petal length
    - Petal width

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

5.1,3.5,1.4,0.2,Iris-setosa	7.0,3.2,4.7,1.4,Iris-versicolor	6.3,3.3,6.0,2.5,Iris-virginica
4.9,3.0,1.4,0.2,Iris-setosa	6.4,3.2,4.5,1.5,Iris-versicolor	5.8,2.7,5.1,1.9,Iris-virginica
4.7,3.2,1.3,0.2,Iris-setosa	6.9,3.1,4.9,1.5,Iris-versicolor	7.1,3.0,5.9,2.1,Iris-virginica
4.6,3.1,1.5,0.2,Iris-setosa	5.5,2.3,4.0,1.3,Iris-versicolor	6.3,2.9,5.6,1.8,Iris-virginica
5.0,3.6,1.4,0.2,Iris-setosa	6.5,2.8,4.6,1.5,Iris-versicolor	6.5,3.0,5.8,2.2,Iris-virginica
5.4,3.9,1.7,0.4,Iris-setosa	5.7,2.8,4.5,1.3,Iris-versicolor	7.6,3.0,6.6,2.1,Iris-virginica
4.6,3.4,1.4,0.3,Iris-setosa	6.3,3.3,4.7,1.6,Iris-versicolor	4.9,2.5,4.5,1.7,Iris-virginica
5.0,3.4,1.5,0.2,Iris-setosa	4.9,2.4,3.3,1.0,Iris-versicolor	7.3,2.9,6.3,1.8,Iris-virginica
4.4,2.9,1.4,0.2,Iris-setosa	6.6,2.9,4.6,1.3,Iris-versicolor	6.7,2.5,5.8,1.8,Iris-virginica
4.9,3.1,1.5,0.1,Iris-setosa	5.2,2.7,3.9,1.4,Iris-versicolor	7.2,3.6,6.1,2.5,Iris-virginica
5.4,3.7,1.5,0.2,Iris-setosa	5.0,2.0,3.5,1.0,Iris-versicolor	6.5,3.2,5.1,2.0,Iris-virginica
4.8,3.4,1.6,0.2,Iris-setosa	5.9,3.0,4.2,1.5,Iris-versicolor	6.4,2.7,5.3,1.9,Iris-virginica
4.8,3.0,1.4,0.1,Iris-setosa	6.0,2.2,4.0,1.0,Iris-versicolor	6.8,3.0,5.5,2.1,Iris-virginica
4.3,3.0,1.1,0.1,Iris-setosa	6.1,2.9,4.7,1.4,Iris-versicolor	5.7,2.5,5.0,2.0,Iris-virginica
5.8,4.0,1.2,0.2,Iris-setosa	5.6,2.9,3.6,1.3,Iris-versicolor	5.8,2.8,5.1,2.4,Iris-virginica
5.7,4.4,1.5,0.4,Iris-setosa	6.7,3.1,4.4,1.4,Iris-versicolor	6.4,3.2,5.3,2.3,Iris-virginica
5.4,3.9,1.3,0.4,Iris-setosa	5.6,3.0,4.5,1.5,Iris-versicolor	6.5,3.0,5.5,1.8,Iris-virginica
5.1,3.5,1.4,0.3,Iris-setosa	5.8,2.7,4.1,1.0,Iris-versicolor	7.7,3.8,6.7,2.2,Iris-virginica
……	………	……

Scatter Plots of Iris Data



Select 1, 2, 3, 4 features on IRIS Data

Feature selection
• Motivation
• Example
• Classification of feature selection methods
Univariate methods of feature selection
Multivariate methods of feature selection
Feature extraction
Conclusions

Example: Fisher's Iris data

Complete enumeration of all combinations of features:
LOO Xval Error: Leave-one-out crossvalidation error

	Input features				LOO Xval Error	
	SL	SW	PL	PW	Dec. tree	3-NN
# inputs					100.0 %	100.0 %
1 input	x				26.7 %	28.7 %
		x			41.3 %	47.3 %
			x		6.0 %	8.0 %
				x	5.3 %	4.0 %
2 inputs	x	x			23.3 %	24.0 %
	x		x		6.7 %	5.3 %
	x			x	5.3 %	4.0 %
		x	x		6.0 %	6.0 %
		x		x	5.3 %	4.7 %
			x	x	4.7 %	5.3 %
3 inputs	x	x	x		6.7 %	7.3 %
	x	x		x	5.3 %	5.3 %
	x		x	x	4.7 %	3.3 %
		x	x	x	4.7 %	4.7 %
All inputs	x	x	x	x	4.7 %	4.7 %

- Decision tree reaches its lowest error (4.7 %) whenever PL and PW are among the inputs; it is able to choose them for decision making, more features do not harm.
- 3-NN itself does not contain any feature selection method, it uses all features available. *The lowest error is usually not achieved when using all inputs!*

Feature selection methods

- Univariate (select each feature independently of others)
  - Pearson correlation coefficient
  - T-test, f-test
  - Chi-square
  - mutual information
  - Relief, etc
- Multivariate (select a subset of features simultaneously)
  - Subset selection, forward search, floating search
  - Recursive feature elimination
  - Use a classifier’s internal structures, such as support vector machine
  - Linear combination of features, i.e., dimension reduction methods (PCA)

Feature selection methods

- Univariate (select one feature independent of others)
  - Pearson correlation coefficient
  - T-test, F-test
  - Chi-square
  - mutual information
  - Relief, etc
- Multivariate (select a subset of features simultaneously)
  - Subset selection, forward search, floating search
  - Recursive feature elimination
  - Use a classifier’s internal structures, such as support vector machine
  - Linear combination of features, i.e., dimension reduction methods (PCA)

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

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (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
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

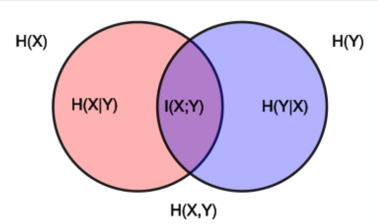
- Number of classes
- Feature vector value types

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*.

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

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}

Compute Mutual information (Taxable Income, cheat)

	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

F-test (class-label-vector, feature-vector)

– Features are numerical

Given a gene expression across n tissue samples $\mathbf{g} = (g_1, g_2, \cdots, g_n)$, the F-statistic is defined as

$$F = \left[\sum_k n_k(g_k - \bar{g})^2/(K-1)\right]/\sigma^2, \qquad (1)$$

where $\bar{g}$ is the average expression across all samples, $\bar{g}_k$ is the average within class C_k, and σ^2 is the *pooled* variance:

$$\sigma^2 = \left[\sum_k (n_k - 1)\,\sigma_k^2\right]/(n-K)$$

where n_k and σ_k are the size and variance of gene expression within class C_k. For $K = 2$,

$$F = t^2, \quad t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,\frac{\bar{g}_1 - \bar{g}_2}{\sigma}, \qquad (2)$$

F-statistic reduces to t-statistic. We pick genes with large F-values or t-values.

When gene follows the Gaussian distribution, f-value follow F(K-1,n-K) distribution. We can compute p-values and confidence levels to assess the test. This is the theory of analysis of variance (ANOVA)

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 F-test (TaxInc=X, cheat=Y)

$\mathbf{g} = (g_1, g_2, \cdots, g_n)$, the F-statistic is defined as

$$F = \left[\sum_k n_k(\bar{g}_k - \bar{g})^2/(K-1)\right]/\sigma^2, \quad (1)$$

where $\bar{g}$ is the average expression across all samples, $\bar{g}_k$ is the average within class C_k, and σ^2 is the *pooled* variance:

$$\sigma^2 = \left[\sum_k (n_k - 1)\,\sigma_k^2\right]/(n-K)$$

where n_k and σ_k are the size and variance of gene expression within class C_k.

Avg(X|Y=no) = (125+100+120+70+60+220+75)/7=110

Var(X|Y=no) = [(125-110)^2 + … + (75-110)^2]/(7-1) = 2975

Avg(X|Y=yes) = (95+85+90)/3=90

Var(X|Y=yes) = [(95-90)^2 + (85-90)^2 + (90-90)^2]/(3-1) = 25

Avg(X) = (125 + … + 90)/10=104

F = [7(110–104)^2+3(90-104)^2]/(2-1) / [(7-1)*2975+(3-1)*25]/(10-2) = 840/(17900/8)=0.3754