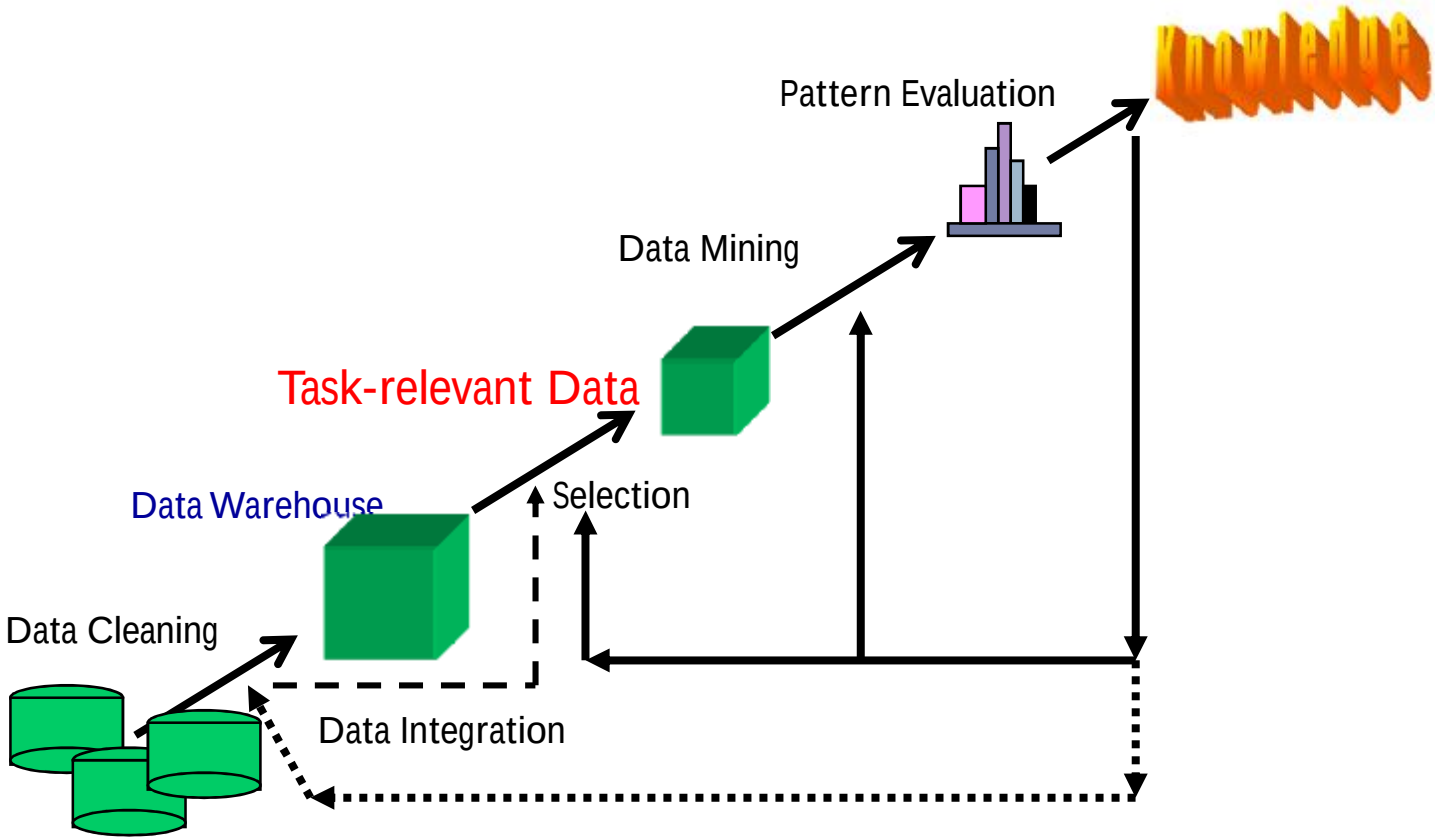
Feature Selection

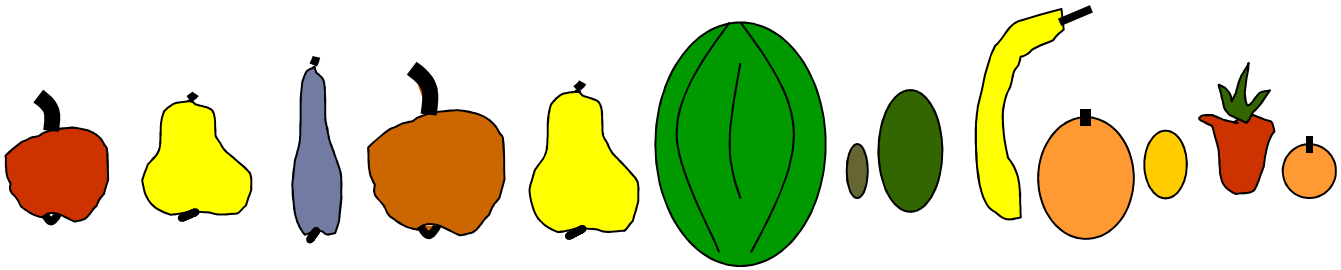
M.M. Pedram
pedram@tmu.ac.ir
Faculty of Engineering, Tarbiat Moallem University
2011

Feature Selection

- Finding the most compact and *informative set of features*, to improve the efficiency or data storage and processing.



Classify Fruits



(Øx, Øy, Øx/Øy, *curvature*, *color*, *hardness*, *weight*, *smell*, …)

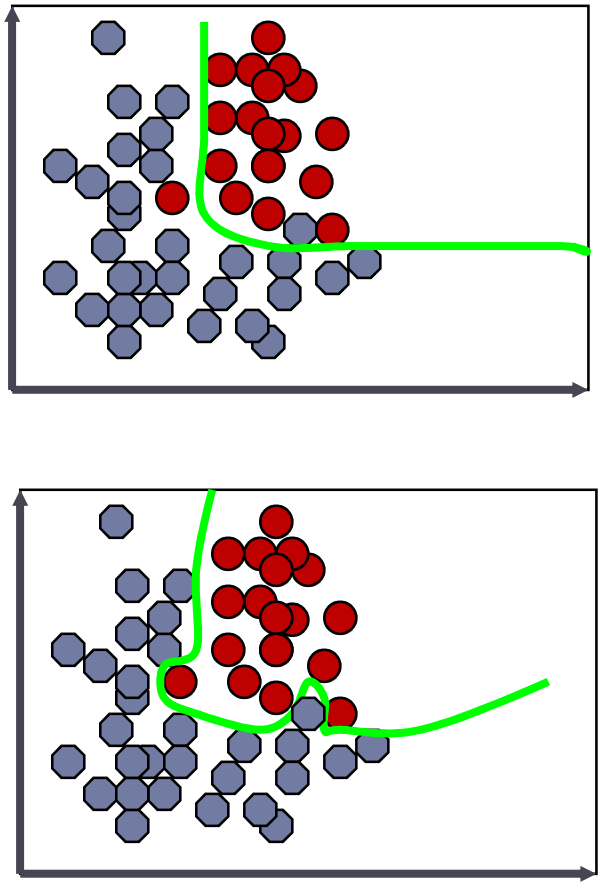
M.M. Pedram, pedram@tmu.ac.ir

Risk Functional

- A function of the parameters of the learning machine, assessing how much it is expected to fail on a given task.
  - Classification: the error rate
  - Regression: the mean square error

Overfitting

- Fit / Robustness Tradeoff



M.M. Pedram, pedram@tmu.ac.ir

Ockham's Razor

- Principle proposed by William of Ockham in the fourteenth century.
- Of two theories providing similarly good predictions, prefer *the simplest one*.
- Shave off unnecessary parameters of your models.

Why Feature Selection?

- **Lot of inputs ⇒Lots of parameters & Large input space**

  ⇒ Curse of dimensionality and risks of overfitting !

FS Methods Classification

Unsupervised	Linear	Nonlinear
Selection	Correlation between inputs	Mutual information between inputs
Projection	PCA	Kohonen maps

Supervised	Linear	Nonlinear
Selection	Correlation between inputs & Outputs	Mutual information between inputs and outputs, Greedy algorithms, Genetic algorithms
Projection	Linear Discriminant Analysis, Partial Least Squares	Projection pursuit

FS Methods Classification

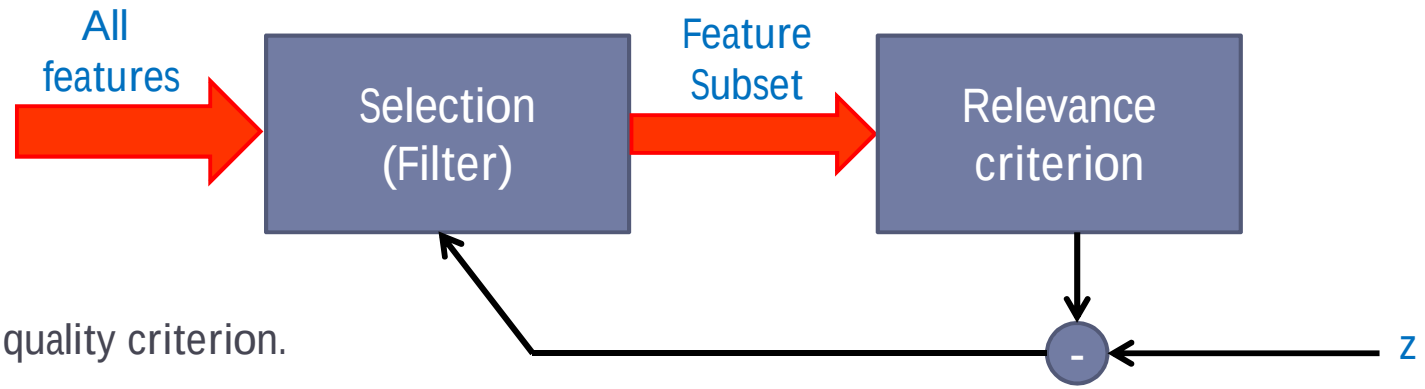
- Selection: choosing among the original features
  - easy (+)
  - interpretability of the features (+)

- Projection: creating new features from the original ones
  - more general →possibly more efficient (+)
  - more difficult (-)
  - features not interpretable (-)

M.M. Pedram, pedram@tmu.ac.ir

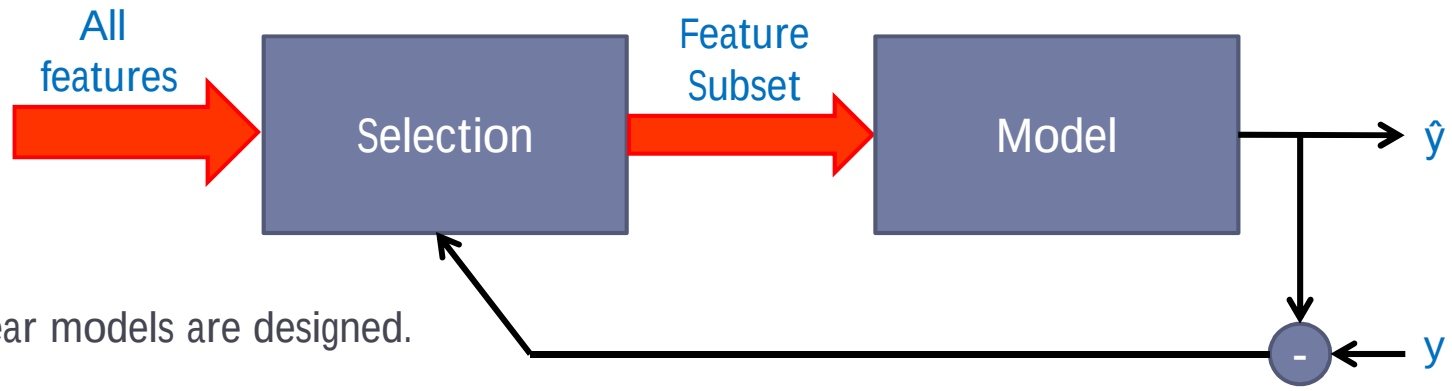
Supervised selection: filter versus wrapper

- Supervising does not necessarily mean to use the model!
- ***Filter***



  - Note the final quality criterion.

- ***Wrapper***



  - Many (non)linear models are designed.

Important Questions in FS

- Question1 : Subset relevance assessment
  - Among all 2^d-1 possible subsets, which is the best one?

- Question2 : Optimal Subset search
  - How not to consider all 2^d-1 possible subsets?

- There are two feature qualities that must be considered by FS methods: relevancy and redundancy.

Subset relevance assessment

- Relevance is difficult to define!

- Filter approach (model free):
  - a variable (or set of) is relevant if it is statistically dependent on *y*.

$$P(y|x_i) \neq P(y)$$

- Wrapper approach (uses model *f*):
  - a variable (or set of) is relevant if the model built on it shows good performances.

Subset relevance assessment

- Correlation
- Mutual information

Correlation

- Correlation, a linear filter
  - Measures linear dependencies (between -1 and +1)
  - 0 indicates no linear relation.

- Definition:
  - correlation between random variable *X* and random variable *Y*

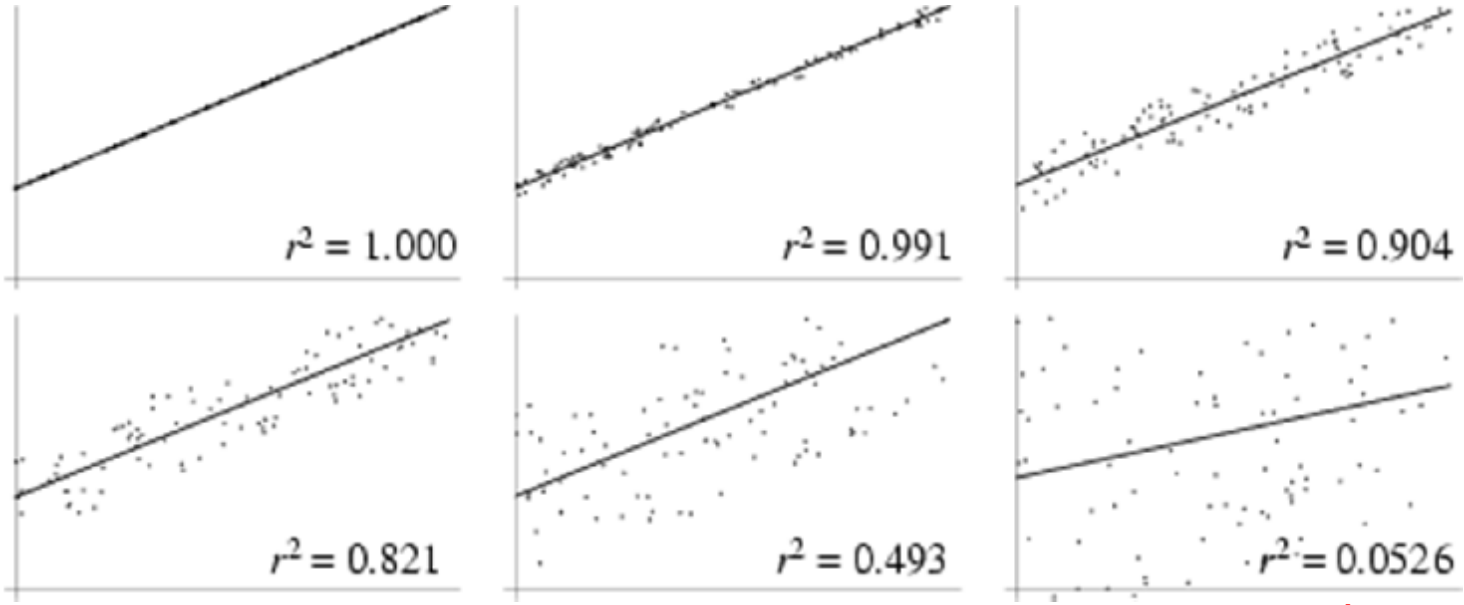
$$r = \frac{Cov(x,y)}{\sqrt{Var(x)\,Var(y)}} = \frac{E(xy) - E(x)E(y)}{\sqrt{Var(x)\,Var(y)}} = \frac{E[(x - E[x]) \cdot (y - E[y])]}{\sqrt{E\left[(x - E[x])^2\right] \cdot E\left[(y - E[y])^2\right]}}$$

  - Suppose a dataset $\{x^j, y^j\}$

$$= \frac{\sum_{j=1}^{N} \left((x^j - \bar{x}) \cdot (y^j - \bar{y})\right)}{\sqrt{\sum_{j=1}^{N} \left((x^j - \bar{x})^2\right) \cdot \sum_{j=1}^{N} \left((y^j - \bar{y})^2\right)}}$$

Correlation

Strong correlation

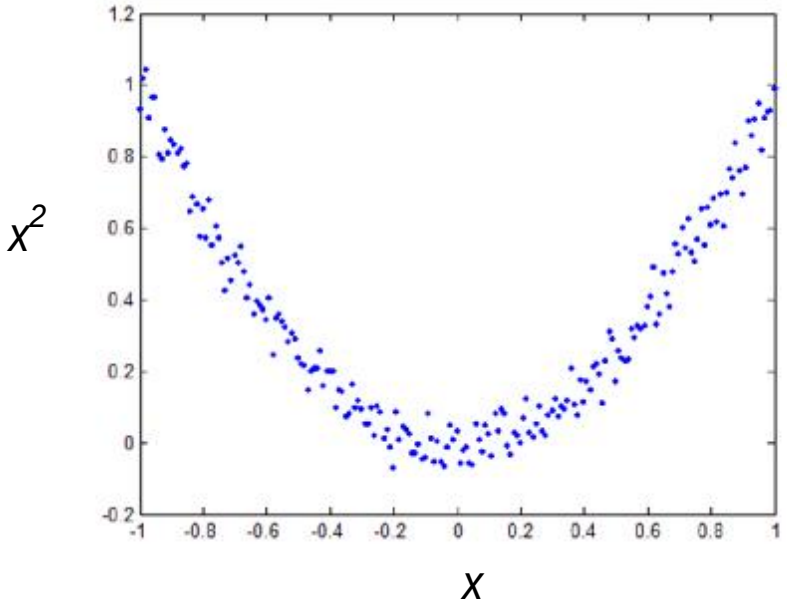


Weak correlation

M.M. Pedram, pedram@tmu.ac.ir

Correlation

- Note: Correlation does not measure nonlinear relations!

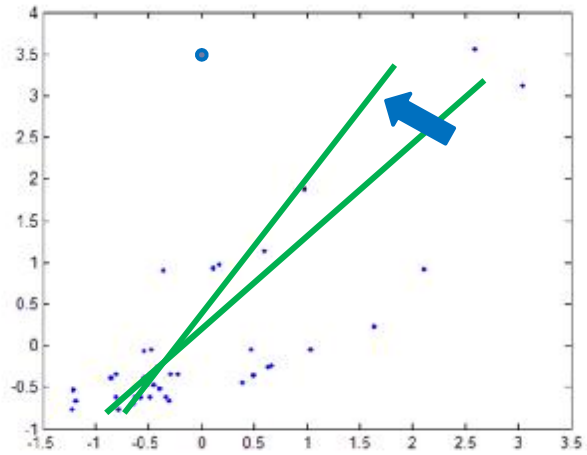


$$\rho_{xx^2} \approx 0$$

M.M. Pedram, pedram@tmu.ac.ir

Correlation

- Correlation:
  - is linear.
  - is parametric (it makes the hypothesis of a ...linear model).
  - does not *explain causality*.
  - is almost impossible to define between more than 2 variables.
  - is sensitive to outliers.



Mutual information

- Mutual information between random variable *x* and random variable *y* measures how the uncertainty on *y* is reduced when *x* is known (and vice versa).

- Relevance of a subset X_S: mutual information $MI(X_S; y)$ between this subset and the target variable *y*.

Mutual information

- Some properties:
  - If x and y are independent, $MI(y;x) = 0$
  - $MI(y;y) = H(y)$
  - $MI(y;x)$ is always non negative and less than $\min(H(y), H(x))$

Mutual information

- Nonlinear:

$$MI(X,Y) = \int P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \, dX \, dY$$

$$= KL(P(X,Y) \,||\, P(X)P(Y))$$

Rough Set

M.M. Pedram, pedram@tmu.ac.ir

Rough Sets

- Modeling imperfect knowledge.
- It can be employed to reduce the dimensionality of datasets.

Zdzislaw Pawlak: Rough Set Theory (1982)

Rough Sets

- A rough set is itself the approximation of a vague concept (set) by a pair of precise concepts, called *lower* and *upper approximations*, that are a classification of the domain of interest into disjoint categories.
- It works by exploring and exploiting the granularity structure of the data only.

Information and Decision Systems

- *Information system* : a table of data, consisting of objects (rows in the table) and attributes (columns) = database
- *Decision system* : an information system may be extended by the inclusion of decision (output) attributes.
- A decision system is *consistent* if for every set of objects whose attribute values are the same, the corresponding decision attributes are identical.

Example #1

- *Decision system* : consists of 4 conditional features (*a*, *b*, *c*, *d*), a *decision feature* (*e*), and eight objects.

$x \in \mathbb{U}$	a	b	c	d	$\Rightarrow$	e
0	S	R	T	T		R
1	R	S	S	S		T
2	T	R	R	S		S
3	S	S	R	T		T
4	S	R	T	R		S
5	T	T	R	S		S
6	T	S	S	S		T
7	R	S	S	R		S

Nomenclature

$I = (U,A)$: an information system,

U : a nonempty set of finite objects (the universe of discourse),

A : a nonempty finite set of attributes

$a : U \rightarrow V_a$: an attribute, $a \in A$.

V_a : the set of values that attribute a may take.

$A = \{C \cup D\}$: *decision systems*,

P : Feature subset, $P \subseteq A$,

C : the set of input features,

D : the set of class indexes.

$d \in D$: a class index, itself a variable
$d : U \rightarrow \{0, 1\}$ *such that for* $x \in U$, $d(x) = 1$ *if* x *has class* d and $d(x) = 0$ *otherwise.*

Indiscernibility

- With any $P \subseteq A$ there is an associated equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in U \times U \mid \forall a \in P, a(x) = a(y)\}$$

- Note that this relation corresponds to the *equivalence relation* for which two objects are equivalent if and only if they have the same vectors of attribute values for the attributes in P.

Indiscernibility

- The partition of U, determined by *IND*(*P*), is denoted U/*IND*(*P*) or *U/P*, which is simply the set of equivalence classes generated by *IND*(*P*):

$$U/IND(P) = \bigotimes\{\, U/IND(\{a\}) \mid a \in P \,\}$$

- Where

$$A \otimes B = \{X \cap Y \mid X \in A,\ Y \in B,\ X \cap Y \neq \emptyset\}$$

- The equivalence classes of the indiscernibility relation with respect to *P* are denoted $[x]_P$, $x \in U$.

M.M. Pedram, pedram@tmu.ac.ir

Indiscernibility

- Example: if *P* = *{b, c}, then objects 1, 6, and 7* are indiscernible, as are objects 0 and 4.

$x \in \mathbb{U}$	a	b	c	d	$\Rightarrow$	e
0	S	R	T	T		R
1	R	S	S	S		T
2	T	R	R	S		S
3	S	S	R	T		T
4	S	R	T	R		S
5	T	T	R	S		S
6	T	S	S	S		T
7	R	S	S	R		S

- *IND*(*P*) creates the following partition of U:

$$\begin{aligned}\mathbb{U}/IND(P) &= \mathbb{U}/IND(\{b\}) \otimes \mathbb{U}/IND(\{c\}) \\ &= \{\{0,2,4\},\{1,3,6,7\},\{5\}\} \\ &\otimes\{\{2,3,5\},\{1,6,7\},\{0,4\}\} \\ &= \{\{2\},\{0,4\},\{3\},\{1,6,7\},\{5\}\}\end{aligned}$$

M.M. Pedram, pedram@tmu.ac.ir

Lower and Upper Approximations

- Let $X \subseteq U$. X can be approximated using only the information contained within P by constructing the P-lower and P-upper approximations of the classical crisp set X:

$$\underline{P}X = \{x \mid [x]_P \subseteq X\}$$

$$\overline{P}X = \{x \mid [x]_P \cap X \neq \emptyset\}$$

- The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is called a *rough set*.

M.M. Pedram, pedram@tmu.ac.ir

Positive, Negative, and Boundary Regions

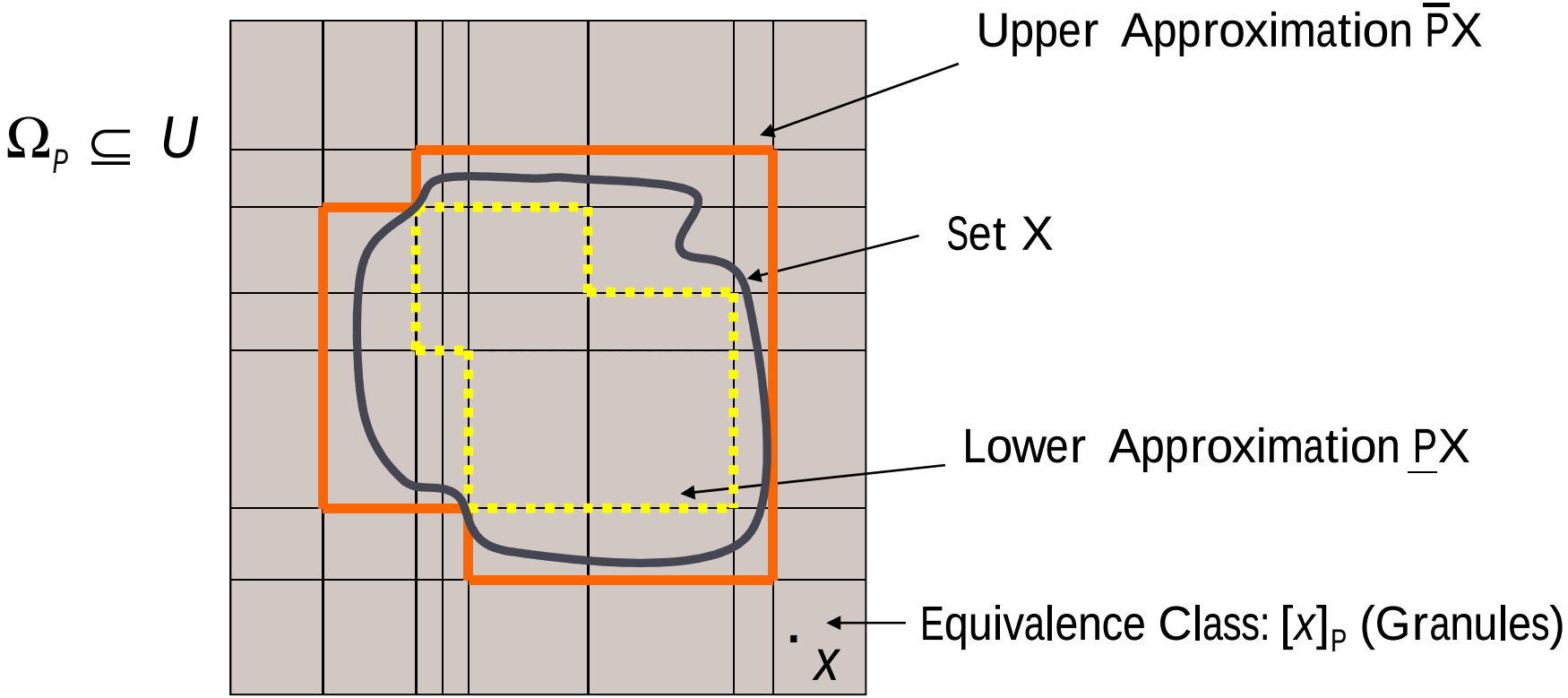
- Let *P* and *Q* be sets of attributes inducing equivalence relations over *U*, then the *positive*, *negative*, and *boundary regions* are defined as:

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}X$$

$$NEG_P(Q) = \mathbb{U} - \bigcup_{X \in \mathbb{U}/Q} \overline{P}X$$

$$BND_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \overline{P}X - \bigcup_{X \in \mathbb{U}/Q} \underline{P}X$$

Positive, Negative, and Boundary Regions



$[x]_P$: the set of all points which are *indiscernible* with point *x* in terms of feature subset P. (set of all points belonging to the same granule as of the point *x* in feature space Ω_P)

Positive, Negative, and Boundary Regions

- $POS_P(Q)$: The positive region comprises all objects of U that can be classified to classes of U/*Q* using the information contained within attributes *P*.
- $BND_P(Q)$: The boundary region is the set of objects that can possibly, but not certainly, be classified in this way.
- $NEG_P(Q)$: The negative region is the set of objects that cannot be classified to classes of U/*Q*.

Example #1

- Let $P = \{b,c\}$ and $Q = \{e\}$, *then*

S R S T T T S S

$$\mathbb{U}/IND(P) = \mathbb{U}/IND(\{b\}) \otimes \mathbb{U}/IND(\{c\}) = \{\{2\}, \{0,4\}, \{3\}, \{1,6,7\}, \{5\}\}$$

only only only

R S T

$$POS_P(Q) = \cup\{\emptyset, \{2,5\}, \{3\}\} = \{2,3,5\}$$

R S T

$$NEG_P(Q) = \mathbb{U} - \cup\{\{0,4\}, \{2,0,4,1,6,7,5\}, \{3,1,6,7\}\} \rightarrow \bigcup_{X \in \mathbb{U}/Q} \overline{P}X$$

$$= \emptyset$$

$$BND_P(Q) = \cup\{\{0,4\}, \{2,0,4,1,6,7,5\}, \{3,1,6,7\}\} - \{2,3,5\}$$

$$= \{0,1,4,6,7\}$$

$x \in \mathbb{U}$	a	b	c	d	$\Rightarrow$	e
0	S	R	T	T		R
1	R	S	S	S		T
2	T	R	R	S		S
3	S	S	R	T		T
4	S	R	T	R		S
5	T	T	R	S		S
6	T	S	S	S		T
7	R	S	S	R		S

M.M. Pedram, pedram@tmu.ac.ir

Feature Dependency

- A set of attributes Q *depends totally* on a set of attributes P, denoted $P \Rightarrow Q$, if all attribute values from Q are uniquely determined by values of attributes from P.

- If there exists a functional dependency between values of Q and P, then Q depends totally on P.

Feature Dependency and Significance

Dependency In rough set theory:

- For $P, Q \subset A$, it is said that Q *depends on* P *in a degree* k $(0 \leq k \leq 1)$, denoted $P \Rightarrow_k Q$, if

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|}$$

- $|.|$: the cardinality.

- If $k = 1$, Q depends totally on P,
  if $0 < k < 1$, Q depends partially (in a degree k) on P,
  if $k = 0$, Q does not depend on P.

Feature Dependency and Significance

- Example: the degree of dependency of attribute *{e}* on the attributes *{b,c}* is:

$$\gamma_{\{b,c\}}(\{e\}) = \frac{|POS_{\{b,c\}}(\{e\})|}{|\mathbb{U}|}$$

$$= \frac{|\{2, 3, 5\}|}{|\{0, 1, 2, 3, 4, 5, 6, 7\}|} = \frac{3}{8}$$

Feature Dependency and Significance

- Given P, Q and a feature $a \in P$, the *significance* of feature a upon Q is defined by:

$$\sigma_P(Q, a) = \gamma_P(Q) - \gamma_{P-\{a\}}(Q)$$

- If the significance of feature a is 0, then the feature a is *dispensible*.

M.M. Pedram, pedram@tmu.ac.ir

Feature Dependency and Significance

- For example #1, if *P = {a,b,c} and Q = e, then*

$$\gamma_{\{a,b,c\}}(\{e\}) = |\{2,3,5,6\}|/8 = 4/8$$
$$\gamma_{\{a,b\}}(\{e\}) = |\{2,3,5,6\}|/8 = 4/8$$
$$\gamma_{\{b,c\}}(\{e\}) = |\{2,3,5\}|/8 = 3/8$$
$$\gamma_{\{a,c\}}(\{e\}) = |\{2,3,5,6\}|/8 = 4/8$$

- And calculating the significance of the three attributes gives

$$\sigma_P(Q,a) = \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{b,c\}}(\{e\}) = 1/8$$
$$\sigma_P(Q,b) = \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{a,c\}}(\{e\}) = 0$$
$$\sigma_P(Q,c) = \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{a,b\}}(\{e\}) = 0$$

- Thus, the attribute *a* is indispensable, but attributes *b* and *c* can be dispensed.

Reducts

- For many application problems, it is often necessary to maintain a concise form of the information system.
- One way to implement this is to search for a minimal representation of the original dataset.
- For this, the concept of a *reduct is* introduced and defined as a minimal subset *R of the initial attribute set C such* that for a given set of attributes *D, the dependency does not change, i.e.*

$$\gamma_R(D) = \gamma_C(D)$$

?