What is relevant and what is redundant?

M. ROSÁRIO OLIVEIRA

CEMAT AND DEPT. MATEMÁTICA, IST-ULISBOA

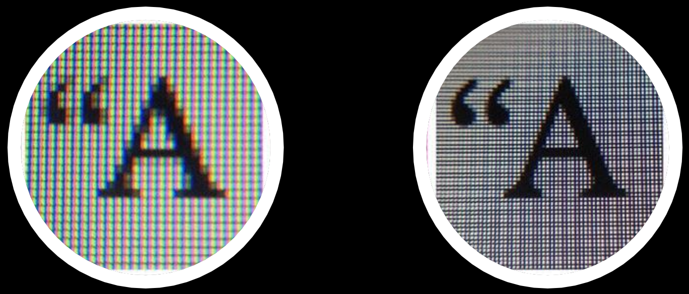
FEATURE SELECTION:

Only the "*right*" features

FEATURE SELECTION:

THE RIGHT DATA

More data is not necessarily **more information**…



Feature Selection:

- Extract from the data **useful** and **valuable** knowledge for **real problem solving**

FEATURE SELECTION:
THE RIGHT DATA

- Select a **small subset** of the original features

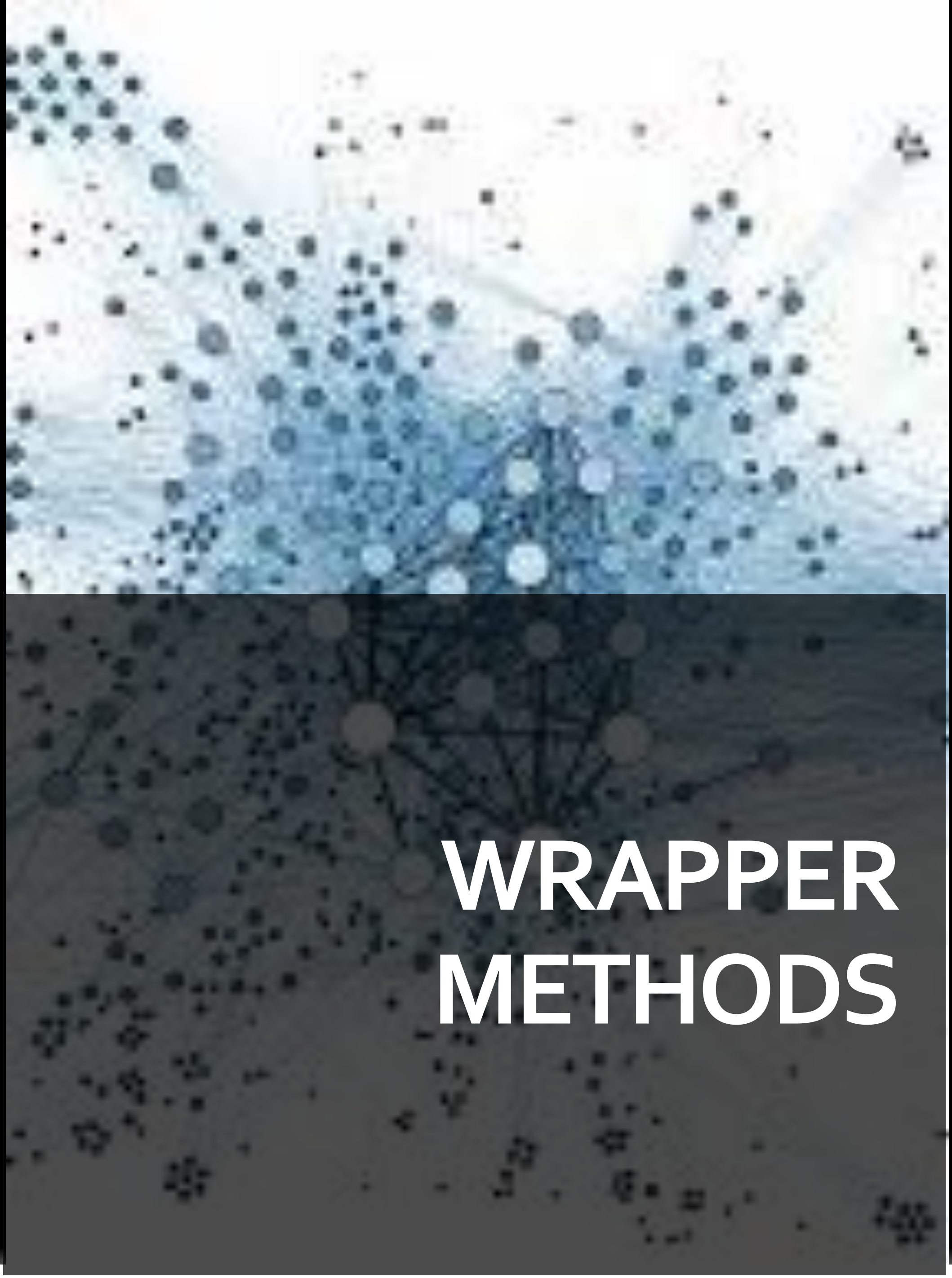
- Such that we remove **irrelevant** and **redundant** features

- In order to:
  - Reduce **computational complexity**
  - Improve model **accuracy**
  - Increase model **interpretability**

Feature Selection Methods

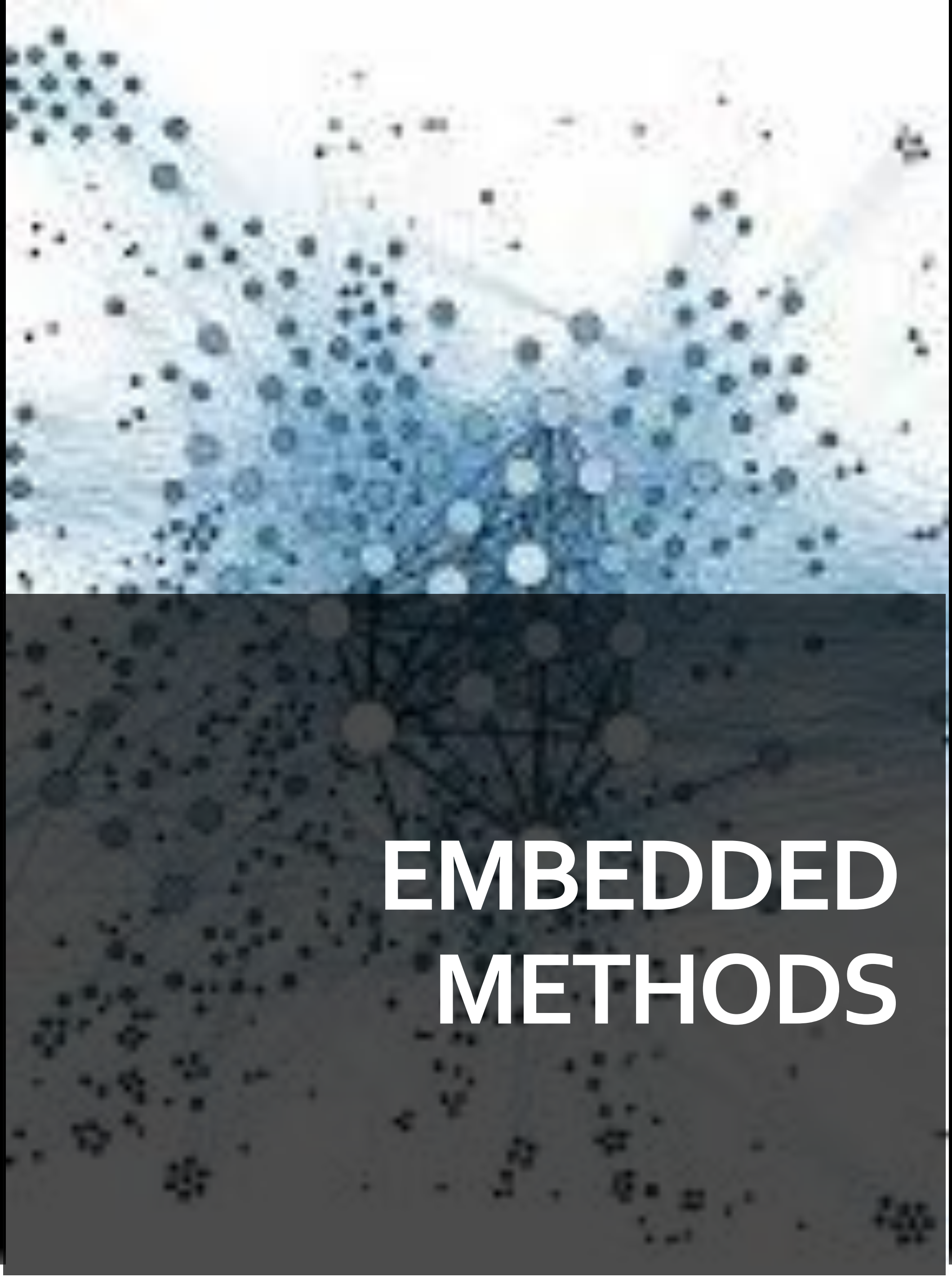
(classification)

- Classifier dependent
  - Wrapper
  - Embedded
- Classifier independent
  - Filter



WRAPPER METHODS

- **Idea:** search for feature subsets, using the classifier accuracy as the measure of utility for a candidate subset
- **Disadvantages:**
  - computational cost
  - selected features are classifier specific
- **Example:**
  - Stepwise regression



EMBEDDED METHODS

- **Idea:** Classifier estimations and feature selection are not separated and interact
- **Disadvantages:**
  - Selected features are classifier specific
  - Regularized_OF=OF+λregularization_penalty
- **Example:**
  - Regularization methods

regularization_penalty

FILTER METHODS

- **Idea:** Classifier’s estimation and feature selection are separated and depend on a specific measure of benefit

- Most popular ones: rely on **Mutual Information (MI)** and **Entropy**
- **Mutual Information:** measures **linear** and **non-linear** associations among features

- **Example:**
  - Forward feature selection methods based on MI

ENTROPY, MUTUAL INFORMATION

Entropy

Entropy

Motivated by problems in the field of telecommunications

A Mathematical Theory of Communication*

C. E. Shannon (1948)

- A measure of uncertainty
- One formula that changed the world...

Entropy
Discrete rv

$$H(\boldsymbol{X}) = -\sum_{\boldsymbol{x} \in \mathcal{X}} P(\boldsymbol{X} = \boldsymbol{x}) \ln P(\boldsymbol{X} = \boldsymbol{x}).$$

- Does not depend on the values of X, only on its prob.
- H(a)=0
- H(X)≥0, Non-negative
- H(X)=ln(n), X~Unif{$a_1,\ldots,a_n$}, is maximum

Differential Entropy
Continous rv

$$h(\boldsymbol{X}) = -\int_{\boldsymbol{x} \in \mathcal{X}} f_X(\boldsymbol{x}) \ln f_X(\boldsymbol{x}) d\boldsymbol{x}.$$

- Does not depend on the values of X, only on its pdf
- Can be negative
- h(X)=ln(a), X~Unif(0,a),
  - **a=1, h(X)=0**
  - **a<1, h(X)<0**

Mutual Information

Mutual Information
Discrete rv

$$\mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}) \ln \frac{P(\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y})}{P(\boldsymbol{X}=\boldsymbol{x})P(\boldsymbol{Y}=\boldsymbol{y})}.$$

Mutual Information
Continuous rv

$$\mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}) = \int_{y\in\mathcal{Y}}\int_{x\in\mathcal{X}} f_{X,Y}(\boldsymbol{x},\boldsymbol{y}) \ln \frac{f_{X,Y}(\boldsymbol{x},\boldsymbol{y})}{f_X(\boldsymbol{x})f_Y(\boldsymbol{y})} d\boldsymbol{x}\, d\boldsymbol{y}.$$

- Measures linear and non-linear associations between **X** and **Y**

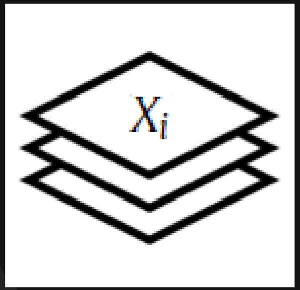
- MI(**X**,**Y**) ≥ 0
- Symmetric
- MI(**X**,**Y**)=0 iff **X** $\perp\!\!\!\perp$ **Y**
- MI(**X**,**X)=**H(**X**)

- All properties hold, except
- MI(**X**,**X)=**+∞

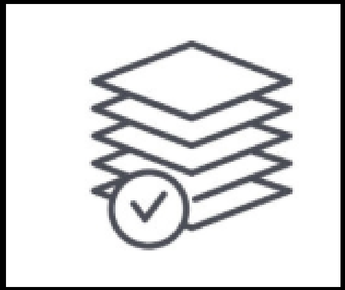
FORWARD FEATURE SELECTION

FORWARD FEATURE SELECTION

Goal: Select a **small subset** of the original features, excluding **irrelevant** and **redundant** features



$$X_j = \arg\max_{X_i \in F} \mathrm{MI}(C, \boldsymbol{S} \cup \{X_i\}).$$

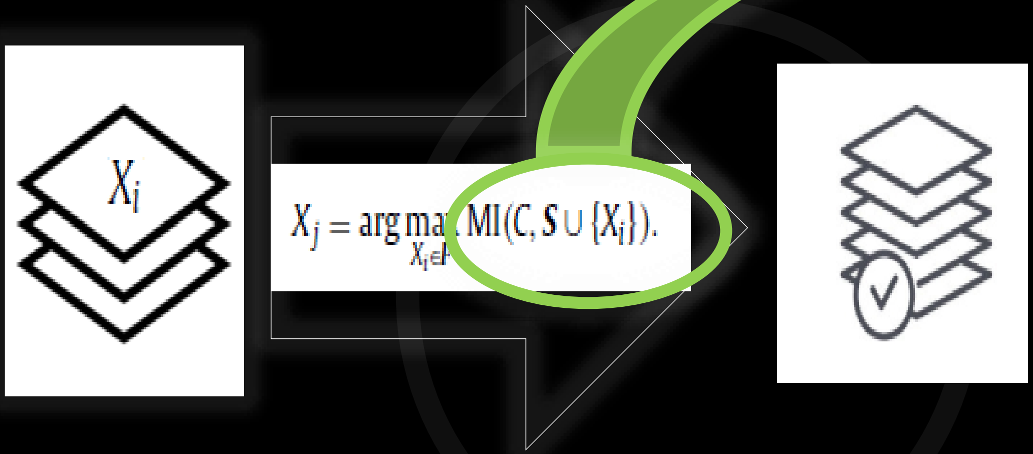


F= Candidate features

S= Selected features

C= Class-variable

FORWARD FEATURE SELECTION



$$X_j = \arg\max_{X_i \in F} \mathrm{MI}(C, S \cup \{X_i\}).$$

Problem:

$$\begin{aligned}
\mathrm{OF}(X_i) &= \cancel{\mathrm{MI}(C, S)} + \mathrm{MI}(C, X_i|S)\\
&= \cancel{\mathrm{MI}(C, S)} + \mathrm{MI}(C, X_i) - \mathrm{TMI}(C, X_i, S)\\
&= \cancel{\mathrm{MI}(C, S)} + \mathrm{MI}(C, X_i) - \mathrm{MI}(X_i, S) + \mathrm{MI}(X_i, S|C).
\end{aligned}$$

Calculation / Estimation

OBJECTIVE FUNCTIONS: INTERPRETABILITY

$$X_j = \arg\max_{X_i \in F} \mathrm{MI}(C, \boldsymbol{S} \cup \{X_i\}).$$

max

$$\mathrm{OF}(X_i) = \mathrm{MI}(C, \boldsymbol{S}) + \mathrm{MI}(C, X_i | \boldsymbol{S})$$

$$\mathrm{OF}(X_i) = H(C) - H(C | X_i, \boldsymbol{S}).$$

min

>0

$$\begin{aligned}\mathrm{OF}(X_i) &= \cancel{\mathrm{MI}(C, \boldsymbol{S})} + \mathrm{MI}(C, X_i | \boldsymbol{S}) \\ &= \cancel{\mathrm{MI}(C, \boldsymbol{S})} + \mathrm{MI}(C, X_i) - \mathrm{TMI}(C, X_i, \boldsymbol{S}) \\ &= \cancel{\mathrm{MI}(C, \boldsymbol{S})} + \mathrm{MI}(C, X_i) - \mathrm{MI}(X_i, \boldsymbol{S}) + \mathrm{MI}(X_i, \boldsymbol{S} | C).\end{aligned}$$

min

max

OBJECTIVE FUNCTIONS: INTERPRETABILITY

$$X_j = \arg\max_{X_i \in F} \mathrm{MI}(C, S \cup \{X_i\}).$$

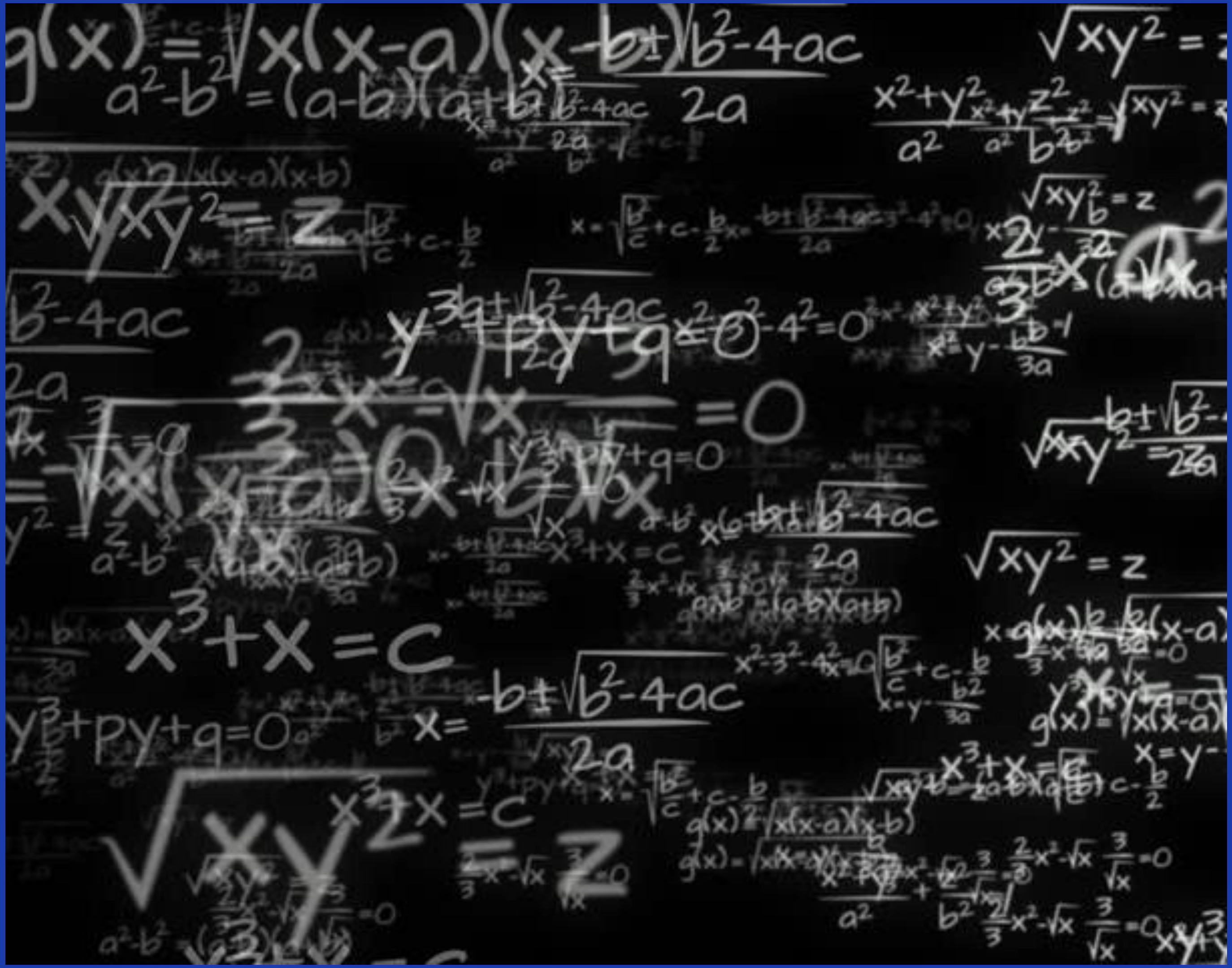
$$\begin{aligned}
\mathrm{OF}(X_i) &= \mathrm{MI}(C,S) + \mathrm{MI}(C,X_i|S) \\
&= \mathrm{MI}(C,S) + \mathrm{MI}(C,X_i) - \mathrm{TMI}(C,X_i,S) \\
&= \cancel{\mathrm{MI}(C,S)} + \mathrm{MI}(C,X_i) - \mathrm{MI}(X_i,S) + \mathrm{MI}(X_i,S|C).
\end{aligned}$$

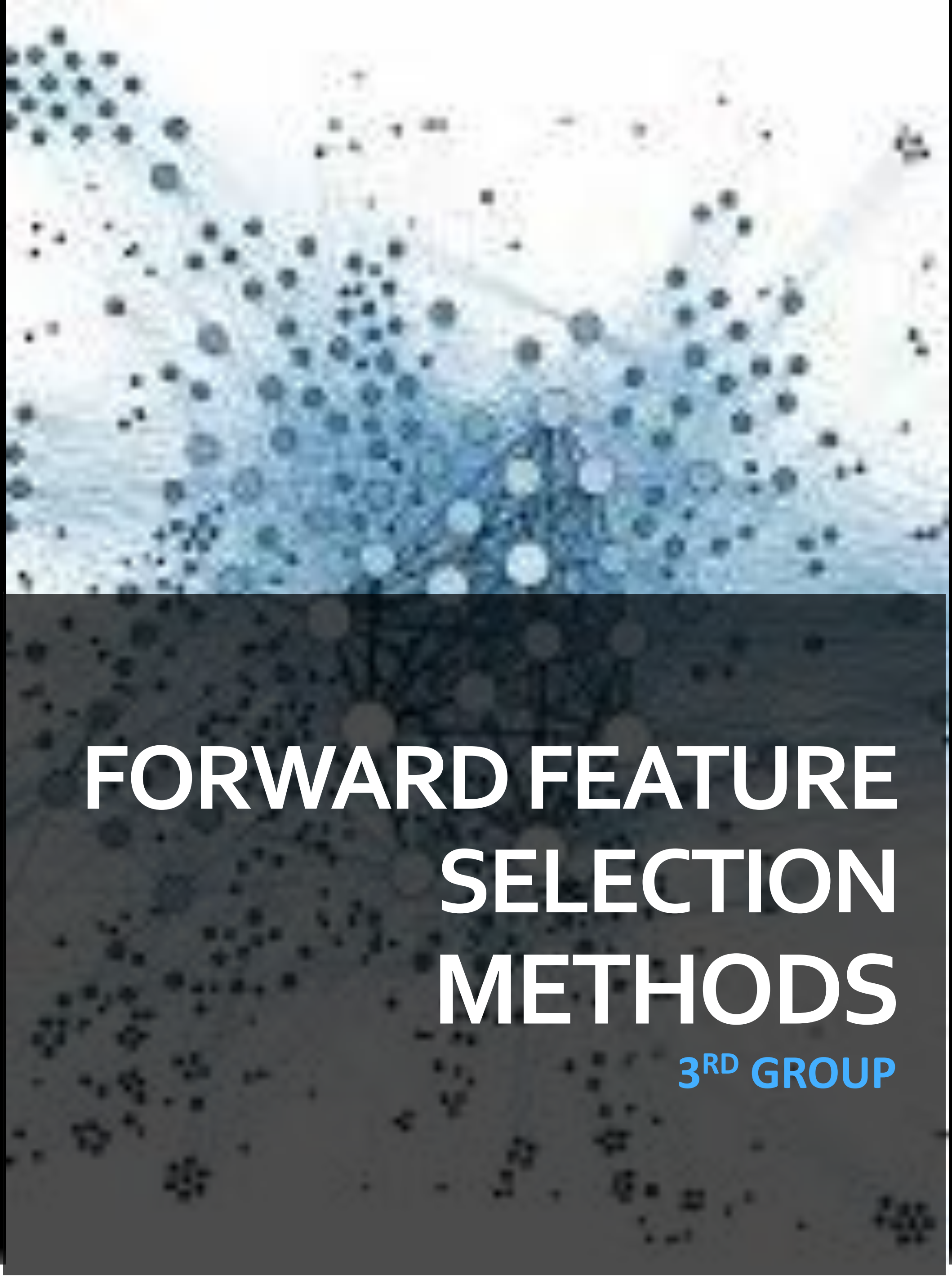
Relevance ☺

Inter-feature redundance ☹

Class-relevant Redundancy ☺

FEATURE SELECTION METHODS





FORWARD FEATURE SELECTION METHODS

3RD GROUP

$$OF(X_i) = \cancel{MI(C,S)} + MI(C,X_i) - MI(X_i,S) + MI(X_i,S|C).$$

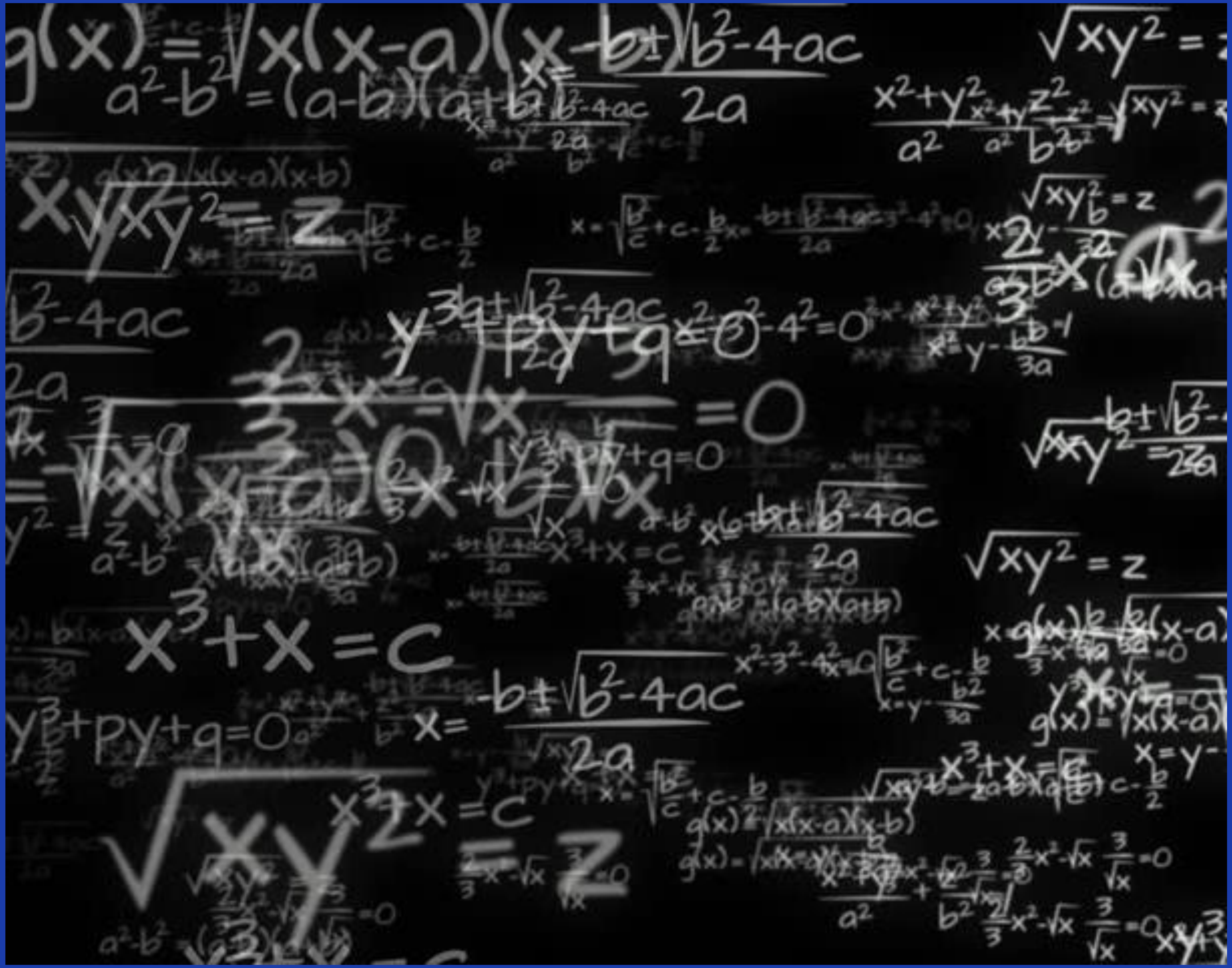
Third Group of Methods

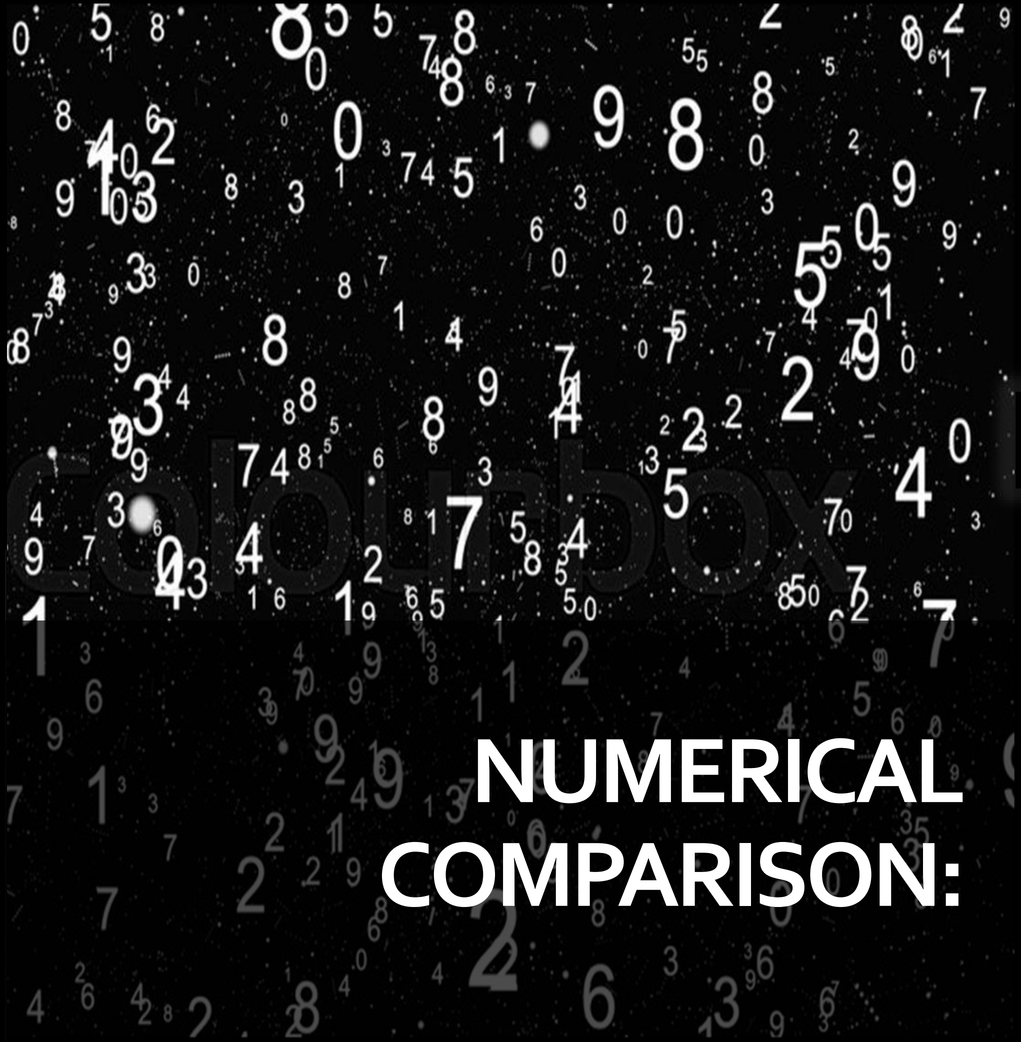
Method	Objective function evaluated at X_i
CIFE	$MI(C,X_i) - \sum_{X_s \in S}(MI(X_i,X_s) - MI(X_i,X_s|C))$
JMI	$MI(C,X_i) - \frac{1}{|S|}\sum_{X_s \in S}(MI(X_i,X_s) - MI(X_i,X_s|C))$
CMIM	$MI(C,X_i) - \max_{X_s \in S}\{MI(X_i,X_s) - MI(X_i,X_s|C)\}$
JMIM	$MI(C,X_i) - \max_{X_s \in S}\{MI(X_i,X_s) - MI(X_i,X_s|C) - MI(C,X_s)\}$

$$OF_{DMIM}(X_i) = MI(X_i,C) - \max_{X_s \in S} MI(X_i,X_s) + \max_{X_s \in S} MI(X_i,X_s|C).$$

- ***Class-relevant redundancy***: contribution of a candidate feature to the explanation of the class, when taken together with already selected features

THEORETICAL COMPARISON





NUMERICAL COMPARISON:

How comparisons are usually done:

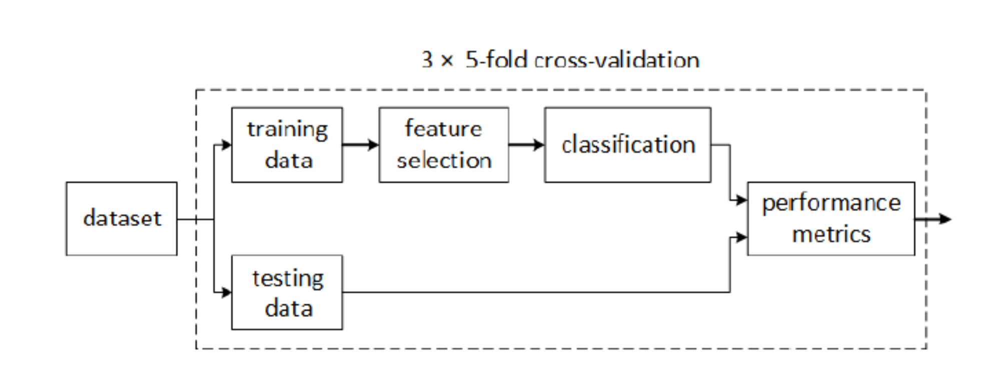
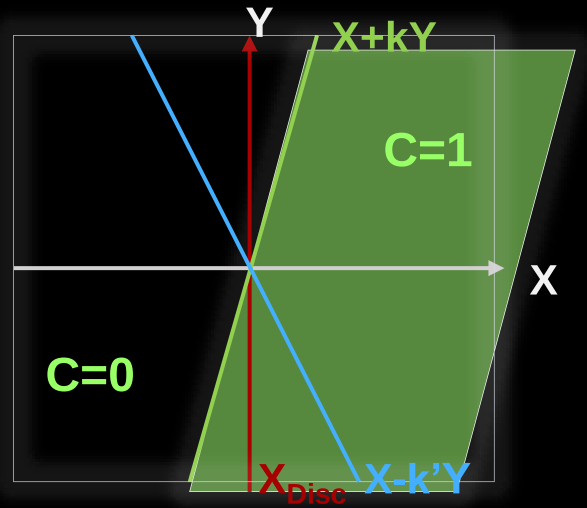


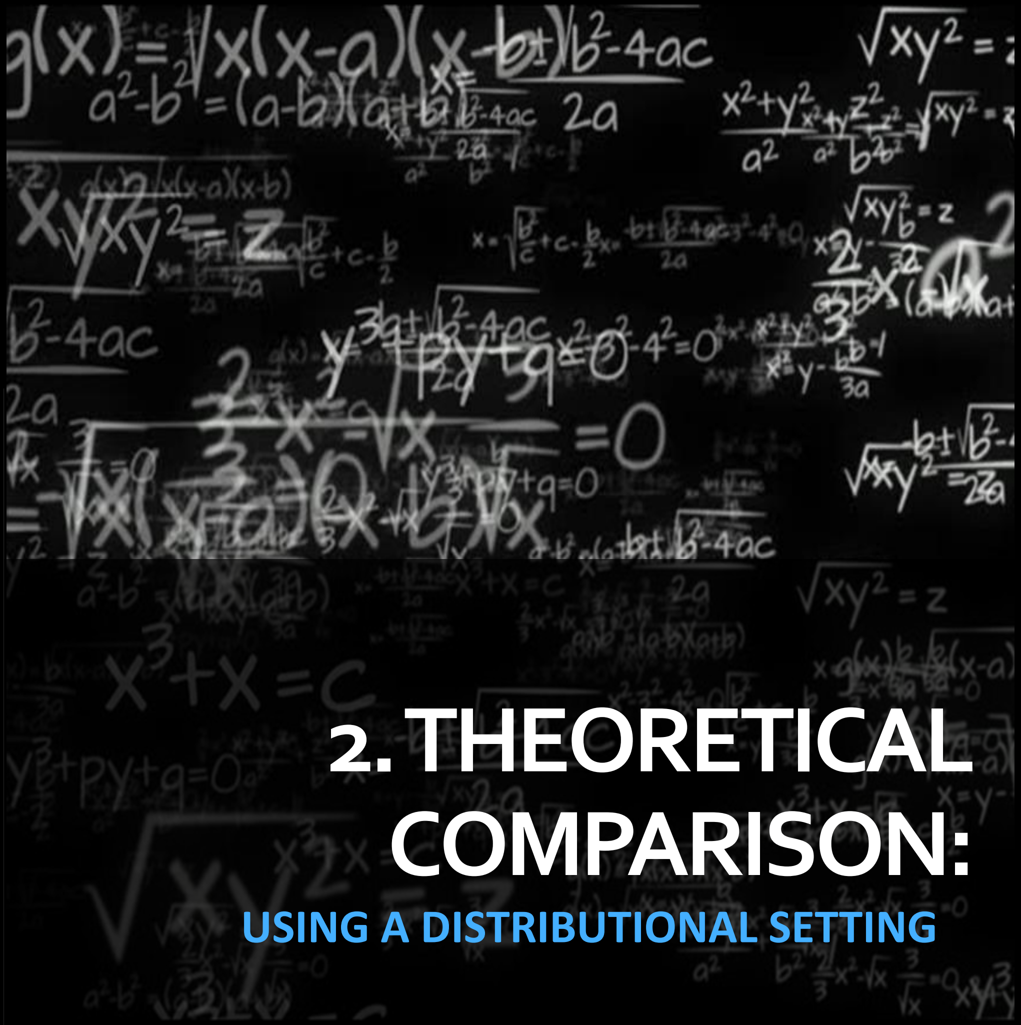
Figure (4) Evaluation process.

Theoretical Setup:

Class-Variable: $C_k = \begin{cases} 0, & X + kY < 0 \\ 1, & X + kY \geq 0 \end{cases}$

Candidate Features: X, X-k'Y, X_{Disc}, Z





2. THEORETICAL COMPARISON:

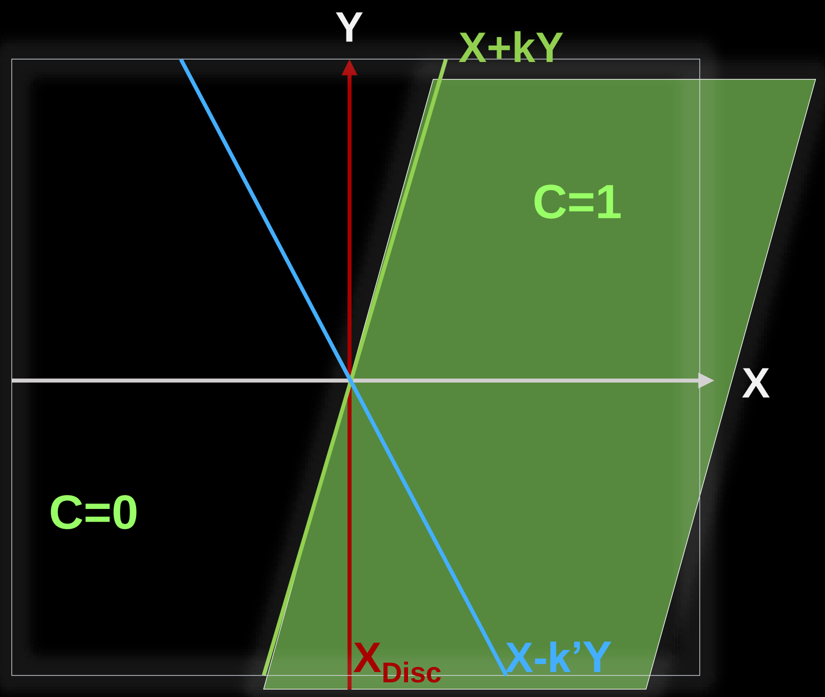
USING A DISTRIBUTIONAL SETTING

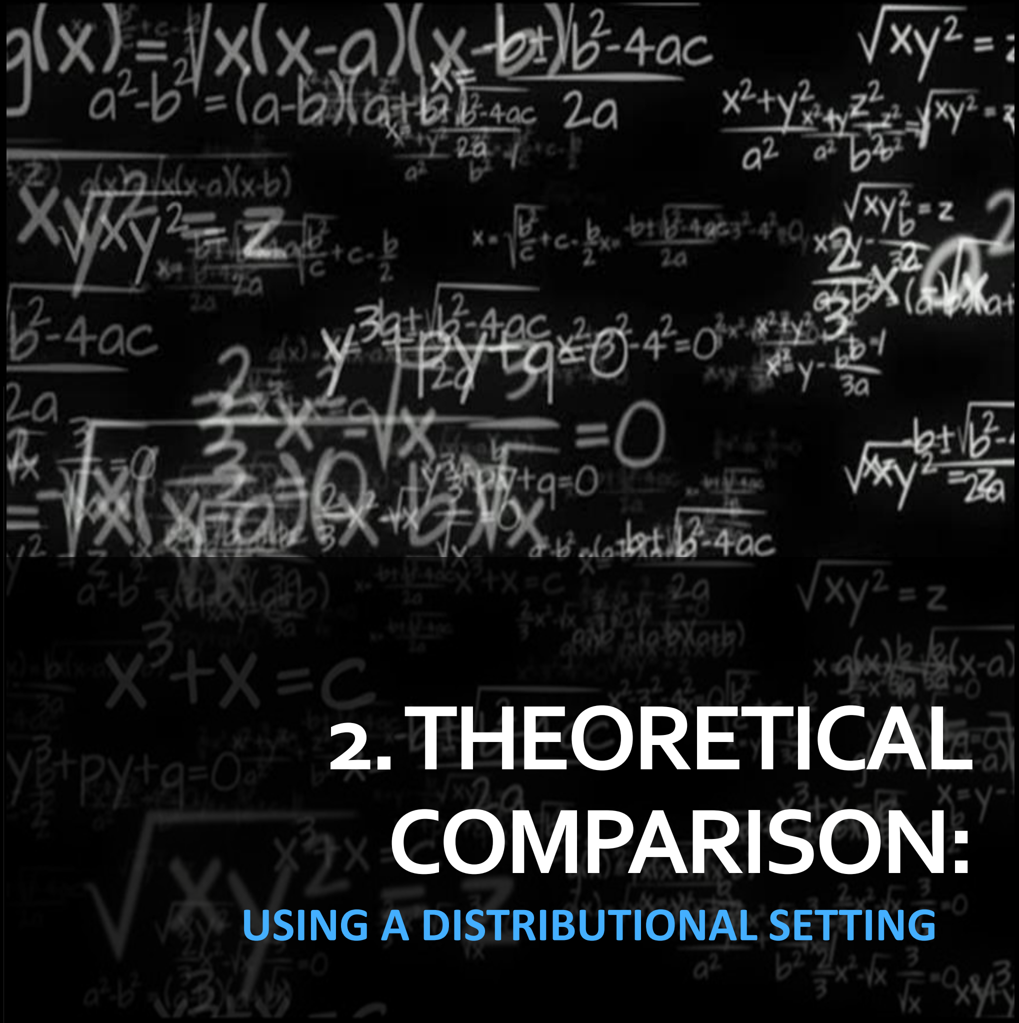
FEATURE (RELEVANCE) TYPES

- Class-Variable: $C_k = \begin{cases} 0, & X + kY < 0 \\ 1, & X + kY \geq 0 \end{cases}$'
- Candidate Features: X, X-k'Y, X_{Disc}, Z

Features Categories:

- *Irrelevant*: Z
- *Relevant*: X, X-k'Y, X_{Disc}
  - *Fully Relevant*: X-k'Y
- *Redundant*:
  - *If we chose* X then X_{Disc} is redundant
  - *If we chose* X_{Disc} then X is redundant





2. THEORETICAL COMPARISON:

USING A DISTRIBUTIONAL SETTING

Features Order: OF were calculated theoretically assuming X, Y, and Z are indep. N(0,1)

Performance Measure:

Bayes Risk and Bayes Classifier

=

min{Total Probability of Misclassifiction}

THEORETICAL COMPARISON

A	$\mathrm{MI}(C_k, A)$
X	$\frac{1}{2}\ln(2\pi e) - \frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X\mid C_k=j}(u)\ln f_{X\mid C_k=j}(u)du$
$X-k'Y$	$\frac{1}{2}\ln(2\pi e(1+k'^2)) - \frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X-k'Y\mid C_k=j}(u)\ln f_{X-k'Y\mid C_k=j}(u)du$
Z	0
X_{disc}	$2\ln(2) + \frac{\arctan k}{\pi}\ln(\frac{\arctan k}{2\pi}) + (1-\frac{\arctan k}{\pi})\ln(\frac{1}{2}-\frac{\arctan k}{2\pi})$ [b]

[a] $X\mid C_k = j \sim \mathrm{SN}(0, 1, \frac{(-1)^{j+1}}{k}),\ j = 0, 1.$
[b] $X-k'Y\mid C_k = j \sim \mathrm{SN}(0, \sqrt{1+k'^2}, (-1)^{j+1}(\frac{1-kk'}{k+k'})),\ j = 0, 1.$

Table 8
MI between pairs of input features.

A	B	$\mathrm{MI}(\cdot,\cdot)$
X	$X-k'Y$	$\frac{1}{2}\ln(1+\frac{1}{k'^2})$
X	X_{disc}	$\ln(2)$
$X-k'Y$	X_{disc}	$\frac{1}{2}\ln(2\pi e) - \frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X\mid C_{k'}=j}(u)\ln f_{X\mid C_{k'}=j}(u)du$ [a]
Z	B	$0,\ B \in \{X, X-k'Y, X_{\text{disc}}\}$

[a] $X\mid C_{k'} = j \sim \mathrm{SN}(0, 1, \frac{(-1)^{j+1}}{k'}),\ j = 0, 1.$

A	B	$\mathrm{MI}(\cdot,\cdot\mid C_k)$
X	$X-k'Y$	$\frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X\mid C_k=j}(u)\ln f_{X\mid C}$ … $\frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X-k'Y\mid C_k=j}(u)\ln$ … $(1+\ln\pi+\ln k')$
X	X_{disc}	$-\frac{\arctan k}{\pi}\ln(\frac{\arctan k}{\pi}) - (1-$ … $\frac{\arctan k}{\pi})$
$X-k'Y$	X_{disc}	$\frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X-k'Y\mid C_k=j}(u)\ln f_X$ … $h(X-k'Y\mid X_{\text{disc}}, C_k)$ [b,c]
Z	B	$0,\ B \in \{X, X-k'Y, X_{\text{disc}}\}$

[a] $X\mid C_{k'} = j \sim \mathrm{SN}(0, 1, \frac{(-1)^{j+1}}{k'}),\ j = 0, 1.$
[b] $X-k'Y\mid C_k = j \sim \mathrm{SN}(0, \sqrt{1+k'^2}, (-1)^{j+1}$ …
[c] $h(X-k'Y\mid X_{\text{disc}}, C_k)$ in (A.5).

$$\times \exp\Big(-\frac{\ldots}{2}\frac{\ldots}{1+k'^2}\Big)\int \ldots$$

$$= \frac{\sqrt{2}}{\sqrt{1+k'^2}}\int_{-\infty}^{u}\frac{1}{\sqrt{2\pi}\sqrt{1+k'^2}(k+k')}\ldots$$

$$\times\ \Phi\Big(-\frac{(1-kk')z}{(k-k')\sqrt{1+k'^2}}\Big)dz$$

A	B	$\mathrm{MI}(\cdot,\cdot)$
X	$X-k'Y$	$\frac{1}{2}\ln(1+\frac{1}{k'^2})$
X	X_{disc}	$\ln(2)$
$X-k'Y$	X_{disc}	$\frac{1}{2}\ln(2\pi e) - \frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X\mid C_{k'}=j}(u)\ln f_{X\mid C_{k'}=j}(u)du$ [a]
Z	B	$0,\ B \in \{X, X-k'Y, X_{\text{disc}}\}$

[a] $X\mid C_{k'} = j \sim \mathrm{SN}(0, 1, \frac{(-1)^{j+1}}{k'}),\ j = 0, 1.$

The desired probability density function is

$$f_{X-k'Y\mid X_{\text{disc}}=1, C_k=0}(v) = \begin{cases} \frac{2\pi}{\pi-\arctan k}\int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\phi(z)\phi(v+k'z)dz, & v \ge 0 \\ 0, & v < 0 \end{cases}$$

We finally consider $u = 0$ and $j = 0$. For $v \ge 0$,

$$P(X-k'Y \le v, X_{\text{disc}} = 0, C_k = 0)$$
$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty}\phi(w)\phi(z)dw\,dz - \int_{0}^{+\infty}\int_{-kz}^{0}\phi(w)\phi(z)dz\,dw$$
$$- \int_{-\infty}^{-\frac{v}{k'}}\int_{v+k'z}^{0}\phi(w)\phi(z)dz\,dw$$
$$= \frac{1}{2} - \int_{-\infty}^{0}\phi(z)\Big[\Phi(-kz) - \frac{1}{2}\Big]dz - \int_{-\infty}^{-\frac{v}{k'}}\Big[\frac{1}{2} - \Phi(v+k'z)\Big]\phi(z)dz$$
$$= \frac{3}{4} - \frac{1}{2}F_{\mathrm{SN}(0,1,-k)}(0) - \frac{1}{2}\Phi\Big(-\frac{v}{k'}\Big) + \int_{-\infty}^{-\frac{v}{k'}}\Phi(v+k'z)\phi(z)dz.$$

As for the case $v < 0$,

$$P(X-k'Y \le v, X_{\text{disc}} = 0, C_k = 0)$$
$$= \int_{-\infty}^{-\frac{v}{k'+k}}\int_{-\infty}^{v+k'z}\phi(w)\phi(z)dw\,dz + \int_{-\frac{v}{k'+k}}^{+\infty}\int_{-\infty}^{-kz}\phi(w)\phi(z)dz\,dw$$
$$= \int_{-\infty}^{-\frac{v}{k'+k}}\Phi(v+k'z)\phi(z)dw\,dz + \int_{-\frac{v}{k'+k}}^{+\infty}\Phi(-kz)\phi(z)dz$$
$$= \int_{-\infty}^{-\frac{v}{k'+k}}\Phi(v+k'z)\phi(z)dz + \frac{1}{2} - \frac{1}{2}F_{\mathrm{SN}(0,1,-k)}\Big(-\frac{v}{k'+k}\Big).$$

We again need to take the derivative of the two expressions with respect to v to obtain the corresponding conditional density functions. In the case of $v \ge 0$,

$$\frac{d}{dv}\Big[\frac{1}{4} + \frac{1}{2}F_{\mathrm{SN}(0,1,k)}(0) - \frac{1}{2}\Phi\Big(-\frac{v}{k'}\Big) + \int_{-\infty}^{-\frac{v}{k'}}\Phi(v+k'z)\phi(z)dz\Big]$$

$$= \int_{-\frac{v}{k'+k}}^{+\infty}\phi(z)\phi(v+k'z)dz;$$

while, for $v < 0$,

$$\frac{d}{dv}\Big[\int_{-\frac{v}{k'}}^{+\infty}\phi(z)\Phi(v+k'z)dz - \frac{1}{2}\Phi\Big(\frac{v}{k'}\Big)\Big]$$
$$= \int_{-\frac{v}{k'}}^{+\infty}\phi(z)\phi(v+k'z)dz + \frac{1}{2}\phi\Big(\frac{v}{k'}\Big)\frac{1}{k'} - \frac{1}{2}\phi\Big(\frac{v}{k'}\Big)\frac{1}{k'}$$
$$= \int_{-\frac{v}{k'}}^{+\infty}\phi(z)\phi(v+k'z)dz.$$

Note that the following important result [51, Ch. 3] was required in order to obtain both final expressions above:

$$\frac{d}{dx}\int_{a(x)}^{b(x)}g(x,y)dy = \int_{a(x)}^{b(x)}\frac{dg(x,y)}{dx}dy + g(x, b(x))b'(x) - g(x, a(x))a'(x). \quad \text{(A.4)}$$

This result was applied to $\frac{d}{dv}\int_{-\frac{v}{k'+k}}^{+\infty}\phi(z)\Phi(v+k'z)dz$, in the expression for $v \ge 0$, and also to $\frac{d}{dv}\int_{-\frac{v}{k'}}^{+\infty}\phi(z)\Phi(v+k'z)dz$, concerning the case $v < 0$.

The desired probability density function is

$$f_{X-k'Y\mid X_{\text{disc}}=1, C_k=1}(v) = \begin{cases} \frac{2\pi}{\pi-\arctan k}\int_{-\frac{v}{k'+k}}^{+\infty}\phi(z)\phi(v+k'z)dz, & v \ge 0 \\ \frac{2\pi}{\pi-\arctan k}\int_{-\frac{v}{k'}}^{+\infty}\phi(z)\phi(v+k'z)dz, & v < 0 \end{cases}$$

We now consider $u = 0$ and $j = 1$. We have to split again in two cases. For $v \ge 0$,

$$P(X-k'Y \le v, X_{\text{disc}} = 0, C_k = 1) = \int_{0}^{+\infty}\int_{-kz}^{0}\phi(w)\phi(z)dz\,dw$$
$$= \int_{0}^{+\infty}\phi(z)\Big[\frac{1}{2} - \Phi(-kz)\Big]dz$$
$$= \frac{1}{2}F_{\mathrm{SN}(0,1,-k)}(0) - \frac{1}{4}.$$

For $v < 0$,

$$P(X-k'Y \le v, X_{\text{disc}} = 0, C_k = 1)$$
$$= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\int_{-kz}^{v+k'z}\phi(w)\phi(z)dw\,dz + \int_{-\frac{v}{k'}}^{+\infty}\int_{-kz}^{0}\phi(w)\phi(z)dz\,dw$$
$$= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\phi(z)[\Phi(v+k'z) - \Phi(-kz)]dz$$
$$+ \int_{-\frac{v}{k'}}^{+\infty}\phi(z)\Big[\frac{1}{2} - \Phi(-kz)\Big]dz\,dw$$
$$= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\phi(z)\Phi(v+k'z)dz - \frac{1}{2}\Big[F_{\mathrm{SN}(0,1,-k)}\Big(-\frac{v}{k'}\Big)$$

functions. In the case of $v \ge 0$, it is simply 0 since there is no dependency on v. As for $v < 0$,

$$\frac{d}{dv}\Big[\int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\phi(z)\Phi(v+k'z)dz + \frac{1}{2}F_{\mathrm{SN}(0,1,-k)}\Big(-\frac{v}{k'+k}\Big) + \frac{1}{2}\Phi\Big(\frac{v}{k'}\Big)\Big]$$
$$= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\phi(z)\phi(v+k'z)dz + \frac{1}{2}f_{\mathrm{SN}(0,1,-k)}\Big(-\frac{v}{k'+k}\Big)\frac{1}{k'+k}$$
$$- \frac{1}{2}\phi\Big(\frac{v}{k'}\Big)\frac{1}{k'} - \frac{1}{2}f_{\mathrm{SN}(0,1,-k)}\Big(-\frac{v}{k'+k}\Big)\frac{1}{k'+k} + \frac{1}{2}\phi\Big(\frac{v}{k'}\Big)\frac{1}{k'}$$
$$= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\phi(z)\phi(v+k'z)dz.$$

Once again, (A.4) was applied, in this case to $\int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\phi(z)\Phi(v+k'z)dz$.

The desired probability density function is

$$f_{X-k'Y\mid X_{\text{disc}}=0, C_k=1}(v) = \begin{cases} 0, & v \ge 0 \\ \frac{2\pi}{\arctan k}\int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\phi(z)\phi(v+k'z)dz, & v < 0 \end{cases}$$

We now consider $u = 1$ and $j = 0$. For $v \ge 0$,

$$P(X-k'Y \le v, X_{\text{disc}} = 1, C_k = 0)$$
$$= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}}\int_{0}^{v+k'z}\phi(w)\phi(z)dw\,dz + \int_{-\frac{v}{k'+k}}^{0}\int_{0}^{-kz}\phi(w)\phi(z)dz\,dw$$
$$= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}}\phi(z)\Big[\Phi(v+k'z) - \frac{1}{2}\Big]dz$$
$$+ \int_{-\frac{v}{k'+k}}^{0}\phi(z)\Big[\Phi(-kz) - \frac{1}{2}\Big]dz\,dw$$
$$= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}}\phi(z)\Phi(v+k'z)dz - \frac{1}{2}\Big[\Phi\Big(-\frac{v}{k'+k}\Big) - \Phi\Big(-\frac{v}{k'}\Big)\Big]$$
$$+ \frac{1}{2}\Big[F_{\mathrm{SN}(0,1,-k)}(0) - F_{\mathrm{SN}(0,1,-k)}\Big(-\frac{v}{k'+k}\Big)\Big] - \frac{1}{2}\Phi\Big(\frac{v}{k'+k}\Big)$$
$$= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}}\phi(z)\Phi(v+k'z)dz - \frac{1}{2}\Phi\Big(\frac{v}{k'+k}\Big)$$
$$+ \frac{1}{2}F_{\mathrm{SN}(0,1,-k)}(0) - \frac{1}{2}F_{\mathrm{SN}(0,1,-k)}\Big(-\frac{v}{k'+k}\Big).$$

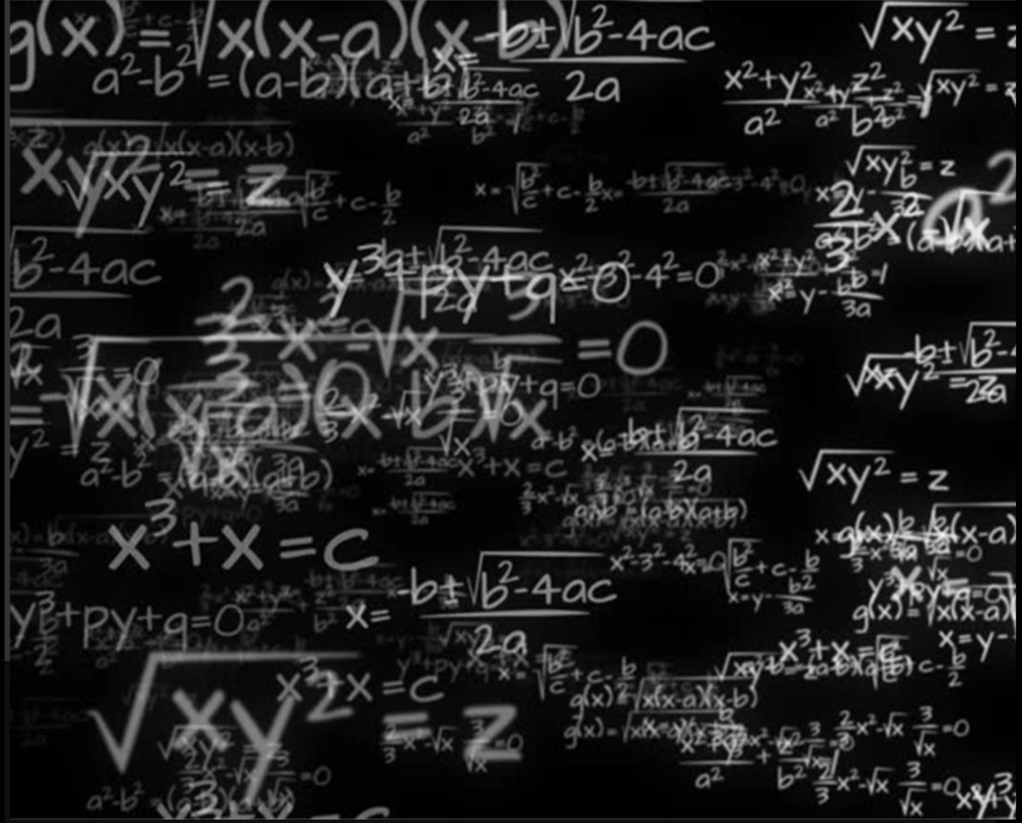
As for the case $v < 0$, $P(X-k'Y \le v, X_{\text{disc}} = 1, C_k = 0) = 0$.

We now need to take the derivative with respect to v of the expression obtained for $v \ge 0$ (the derivative of the one for $v < 0$ is 0) to obtain the corresponding conditional density functions. We have

$$\frac{d}{dv}\Big[\int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}}\phi(z)\Phi(v+k'z)dz - \frac{1}{2}\Phi\Big(-\frac{v}{k'+k}\Big) + \frac{1}{2}F_{\mathrm{SN}(0,1,-k)}(0)$$
$$- \frac{1}{2}F_{\mathrm{SN}(0,1,-k)}\Big(-\frac{v}{k'+k}\Big)\Big]$$
$$= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}}\phi(z)\phi(v+k'z)dz - \frac{1}{2}\phi\Big(-\frac{v}{k'+k}\Big)$$

$$P(X-k'Y \le v, X_{\text{disc}} = 1, C_k = \ldots$$
$$= \int_{-\frac{v}{k'+k}}^{0}\int_{-kz}^{v+k'z}\phi(w)\phi(z)dw\ldots$$
$$= \int_{-\frac{v}{k'+k}}^{0}\phi(z)[\Phi(v+k'z) - \ldots$$
$$+ \int_{0}^{+\infty}\phi(z)\Big[\Phi(v+k'z) - \ldots$$
$$= \int_{-\frac{v}{k'+k}}^{0}\phi(z)\Phi(v+k'z)dz - \ldots$$
$$- F_{\mathrm{SN}(0,1,-k)}\Big(-\frac{v}{k'+k}\Big)\Big] + \ldots$$

$$\int_{-\frac{v}{k'+k}}^{+\infty}\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz\Big)$$
$$\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz\Big)dv$$
$$\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz\Big)$$
$$\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz\Big)dv$$
$$\int_{-\infty}^{\frac{v}{k'}}\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz\Big)$$
$$\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz\Big)dv$$
$$\frac{2\pi}{\arctan k}\zeta(z,v)dz\Big)$$
$$\frac{2\pi}{\arctan k}\zeta(z,v)dz\Big)dv$$
$$\frac{2\pi}{\arctan k}\zeta(z,v)dz\Big)$$
$$\frac{2\pi}{\arctan k}\zeta(z,v)dz\Big)dv\Big).$$



CONCLUSIONS

- Theoretical framework for the comparison of feature selection methods.
- Derivation of upper and lower bounds for the target objective functions.
- Distributional setting to highlight deficiencies of feature selection methods.
- Identification of feature selection methods to be avoided and preferred.

JMI, CMIM or even better: DMIM

MIM, MIFS, mRMR, maxMIFS: Ignore complementary

Main References

Neurocomputing 513 (2022) 215–232

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Feature selection using Decomposed Mutual Information Maximization

Francisco Macedo [a], Rui Valadas [b,c,*], Eunice Carrasquinha [a,b], M. Rosário Oliveira [a], António Pacheco [a]

Theoretical foundations of forward feature selection methods based on mutual information

Francisco Macedo [a,b], M. Rosário Oliveira [a,*], António Pacheco [a], Rui Valadas [c]

[a] CEMAT and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa 1049–001, Portugal
[b] EPF Lausanne, SB-MATHICSE-ANCHP, Station 8, Lausanne CH-1015, Switzerland
[c] IT and Departament of Electrical and Computer Engineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa 1049–001, Portugal

Theoretical evaluation of feature selection methods based on mutual information

Cláudia Pascoal [a], M. Rosário Oliveira [a,*], António Pacheco [a], Rui Valadas [b]

[a] CEMAT and Dep. Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[b] IT and Dep. Electrical and Computer Engineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal