Feature Selection via ℓ_1-penalized Squared-loss Mutual Information

Wittawat Jitkrittum
(10M38450)

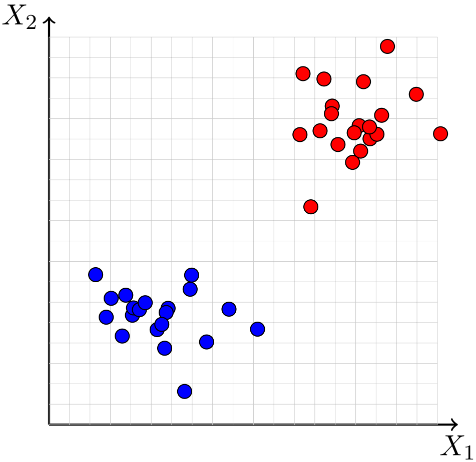
Sugiyama Lab.
Department of Computer Science
Tokyo Institute of Technology

7 February 2012

Outline

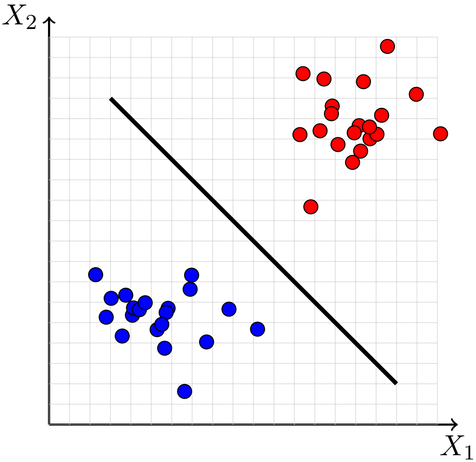
Outline

Example of Feature Selection



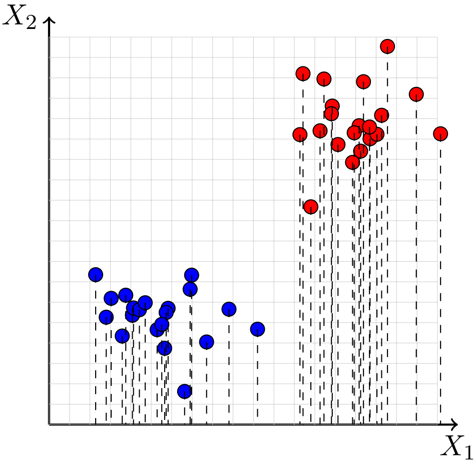
- binary classification
- 2 features: (X_1, X_2)
- Linearly separable
- But, X_1 and X_2 are redundant. Let's choose X_1.
- X_1 alone can distinguish the 2 classes.

Example of Feature Selection



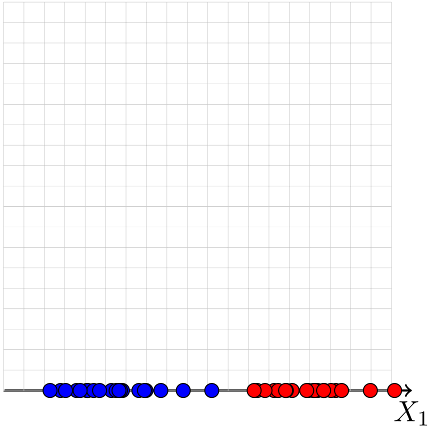
- binary classification
- 2 features: (X_1, X_2)
- Linearly separable
- But, X_1 and X_2 are redundant. Let's choose X_1.
- X_1 alone can distinguish the 2 classes.

Example of Feature Selection



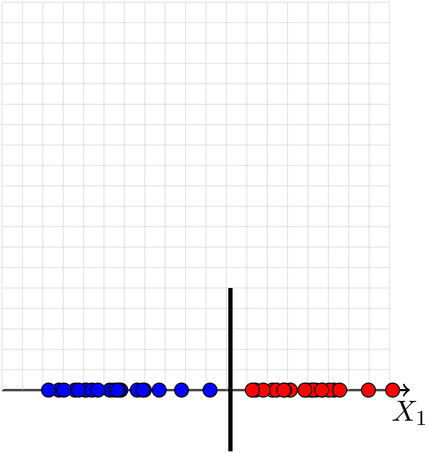
- binary classification
- 2 features: (X_1, X_2)
- Linearly separable
- But, X_1 and X_2 are redundant. Let's choose X_1.
- X_1 alone can distinguish the 2 classes.

Example of Feature Selection



- binary classification
- 2 features: (X_1, X_2)
- Linearly separable
- But, X_1 and X_2 are redundant. Let's choose X_1.
- X_1 alone can distinguish the 2 classes.

Example of Feature Selection



- binary classification
- 2 features: (X_1, X_2)
- Linearly separable
- But, X_1 and X_2 are redundant. Let's choose X_1.
- X_1 alone can distinguish the 2 classes.

Feature Selection

What :

- Given an input $\mathbf{X} \in \mathbb{R}^{m \times n}$ and n-dimensional output vector $\mathbf{Y}$
  - m features (dimensions)
  - n observations (sample size)

select k features $(k < m)$ in $\mathbf{X}$ which can explain $\mathbf{Y}$ well.

Why :

- Reduces data collection cost
- Reduces computation required to train a predictor.
- Facilitates model interpretation

Example : document classification

- Using bag-of-words representation, feature selection can be used to understand which words can explain different categories.

Outline

Outline

Feature Ranking

Definitions

- $X := (X_1, \ldots, X_m)$: input variables
- Y: output variable
- f: feature quality measure e.g., correlation

Procedure:

- Ranks $\{X_i\}_{i=1}^{m}$ in descending order of $f(X_i, Y)$.
- Select top k features.

Advantage:

- Simple & fast

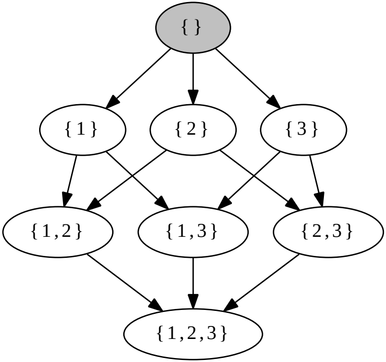
Disadvantage:

- Not consider feature redundancy

Feature redundancy

Features are redundant if they are similar (previous example).

Forward Search



- $\mathcal{X}$: set of features
- k: desired number of features

Procedure:

- Start from $\mathcal{X} = \emptyset$.
- Add a feature to $\mathcal{X}$ until $|\mathcal{X}| = k$.

Advantage:

- Consider feature redundancy

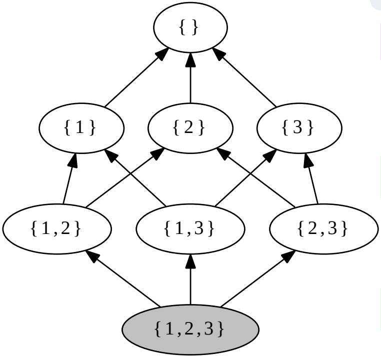
Disadvantage:

- Not consider feature interaction

Feature interaction

Interacting features are individually weak, but strong when combined (e.g. XOR problem).

Backward Search



- $\mathcal{X}$: set of features
- k: desired number of features

Procedure:

- Start from ($\mathcal{X}$ = all features).
- Remove a feature until $|\mathcal{X}| = k$.

Advantages:

- Considers redundancy
- Considers interaction

Disadvantage:

- $O(m^2)$ (m = number of features)

ℓ_1-penalized Feature Weight Learning

$$\begin{array}{ll} \underset{\boldsymbol{w}}{\text{maximize}} & f(\mathrm{diag}(\boldsymbol{w})\mathbf{X}, \mathbf{Y}) \\ \text{subject to} & \|\boldsymbol{w}\|_1 \leq z \end{array}$$

- f: feature quality measure
- $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{m \times n}$ (m features $\times$ n samples)
- $\mathrm{diag}(\boldsymbol{w})\mathbf{X} = (\mathrm{diag}(\boldsymbol{w})\boldsymbol{x}_1, \ldots, \mathrm{diag}(\boldsymbol{w})\boldsymbol{x}_n) \in \mathbb{R}^{m \times n}$
- $\mathrm{diag}(\boldsymbol{w})\boldsymbol{x} = (w_1 x_1, \ldots, w_m x_m)^T$

- w_j: weight of the j^{th} feature
- If $z > 0$ is sufficiently small, obtained $\widehat{\boldsymbol{w}}$ becomes sparse [Tibshirani, 1996].
- $\widehat{w}_j = 0 \Rightarrow j^{th}$ feature is not necessary
- A k-feature subset can be obtained by tuning z.

Comparison of Optimization Strategies

- k: number of desired features
- m: number of total features

	Ranking	**Forward**	**Backward**	**Exhaustive**	ℓ_1
Optimization	discrete	discrete	discrete	discrete	cont.
Search Complexity	m	km	m^2	2^m	m
Consider Redundancy	×	△	○	◎	○
Consider Interaction	×	×	○	◎	○

Advantages of ℓ_1:

- Considers all features at the same time
  - considers redundancy and interaction
- Low computational complexity
- We use ℓ_1 as the optimization strategy

Outline

Pearson Correlation (PC)

- For binary or continuous Y,

$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

- For categorical Y [Hall, 2000],

$$\rho_c(X, Y) = \sum_{c=1}^{C} p(Y = c)|\rho(X, B_c)|.$$

  - $B_c =$ binary variable taking 1 when $Y = c$

X, Y: univariate random variables

Advantage :

- Computationally efficient

Disadvantage :

- Detects only linear dependency

Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2005]

$$\text{HSIC}(X, Y) = \frac{1}{(n-1)^2} \operatorname{tr}(KHLH)$$

- $(K)_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2{\sigma_x}^2}\right)$
- $(L)_{i,j} = l(\boldsymbol{y}_i, \boldsymbol{y}_j) = \exp\left(-\frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{2{\sigma_y}^2}\right)$
- $H = I - \mathbf{1}\mathbf{1}^T/n$

- Non-linear extension of Pearson correlation (PC)
- Measures infinite-order moment (kernel tricks)
- HSIC$(X, Y) = 0 \Leftrightarrow X$ and Y are independent.

Advantages

- Considers non-linear dependency

Disadvantage

- No model selection criterion for σ_x and σ_y
  - Popular heuristic is $\sigma_x = \operatorname{median}(\{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|\}_{i<j})$

Mutual Information (MI) [Cover and Thomas, 1991]

$$I(X,Y) = \iint \log\left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right)p(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{x}d\boldsymbol{y}$$

- Well-known information-theoretic measure
- $I(X,Y)=0 \Leftrightarrow X$ and Y are independent.

Advantages:

- Considers non-linear dependency
- Model selection e.g., estimator called MLMI (Maximum Likelihood MI) [Suzuki et al., 2008]

Disadvantage

- Costly to estimate (due to $\log$)

Squared-loss Mutual Information (SMI) [Suzuki et al., 2009]

$$I_s(X,Y) = \frac{1}{2}\iint \left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1\right)^2 p(\boldsymbol{x})p(\boldsymbol{y})\, d\boldsymbol{x}d\boldsymbol{y}$$

- Same family as MI (f-divergence).
- $I_s(X,Y) = 0 \Leftrightarrow X$ and Y are independent.

Advantages:

- Considers non-linear dependency
- Estimator LSMI (Least-Squares MI) has a model selection criterion.
- LSMI can be computed analytically.

4 Feature Quality Measures

	PC	HSIC	MI	SMI
Non-linear Dependency	×	○	○	○
Model Selection	not needed	×	○	○
Computational Efficiency	◎	○	×	△

- PC: cannot handle non-linear dependency.
- HSIC: no model selection
- MI: costly to estimate
- SMI: good balance of all properties. (◡̈)
- We use SMI as the feature quality measure.

Summary of Optimization Strategies and Measures

	Rank.	Forward	Backward	Exhaustive	ℓ_1
PC	◯	×	×	×	×
HSIC	-	◯	◯	×	△
MI	◯	◯	◯	×	-
SMI	◯	◯	◯	×	-

◯, △ method exists , × unreasonable, impractical , - not exist

- PC : Goodness of a subset $\mathcal{X}$ is $\sum_{i \in \mathcal{X}} \rho(X_i, Y)$.
  - Forward, backward and ℓ_1 give the same solution.
- After an extensive research, we propose to use ℓ_1+SMI .
  - Referred to as ℓ_1-LSMI (LSMI = an estimator of SMI) .

Outline

ℓ_1-LSMI (proposed method)

$$\begin{aligned} \underset{\boldsymbol{w}\in\mathbb{R}^m}{\text{maximize}} \quad & \widehat{I}_s(\mathrm{diag}(\boldsymbol{w})\mathbf{X}, \mathbf{Y}) \\ \text{subject to} \quad & \mathbf{1}^T\boldsymbol{w} \leq z \\ & \boldsymbol{w} \geq \mathbf{0}, \end{aligned}$$

- $\boldsymbol{w} \geq \mathbf{0}$ is imposed to narrow search space (signs do not matter).
- $s(z)$: number of obtained features using z
- $s(z)$ tends to increase as z increases.
- To find a k-feature subset:
  1. $z \leftarrow$ small value
  2. Repeat until $s(z) > k$
     - i $z \leftarrow 2z$
     - ii Solve ℓ_1-LSMI's problem with gradient projection
     - iii **if** $s(z) = k$ **return** obtained features
  3. $z_h \leftarrow z$
  4. $z_l \leftarrow z_h/2$
  5. Find $z \in (z_l, z_h)$ with a binary search so that $s(z) = k$.
- Repeat with different initial $\boldsymbol{w}$.

Outline

Outline

3 Toy Datasets

(1) and-or $(k = 4, m = 10)$

- $Y = (X_1 \wedge X_2) \vee (X_3 \wedge X_4)$
- $X_1, \ldots, X_7 \sim$ Bernoulli(0.5)
- $X_8, \ldots, X_{10} = Y$ with 0.2 chance of bit flip
- **Characteristics:** feature redundancy, weak interaction

(2) quad $(k = 2, m = 10)$

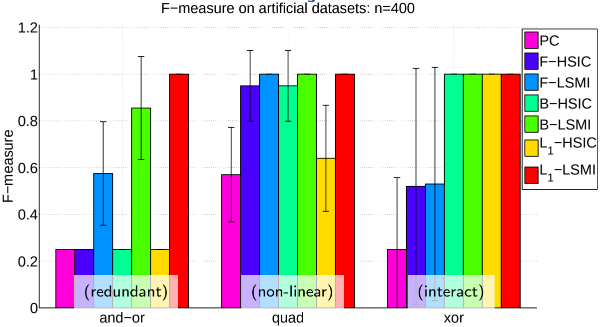
- $Y = \frac{X_1^2 + X_2}{0.5 + (X_2 + 1.5)^2} + 0.1\epsilon$
- $X_1, \ldots, X_8, \epsilon \sim \mathcal{N}(0, 1)$
- $X_9 \sim 0.5X_1 + \mathcal{U}(-1, 1)$
- $X_{10} \sim 0.5X_2 + \mathcal{U}(-1, 1)$
- **Characteristic:** non-linear dependency

(3) xor $(k = 2, m = 10)$

- $Y = \text{xor}(X_1, X_2)$
- $X_1, \ldots, X_5 \sim$ Bernoulli(0.5)
- $X_6, \ldots, X_{10} \sim$ Bernoulli(0.75)
- **Characteristic:** feature interaction

- m: #total features
- k: #features to select
- $X \sim$ Bernoulli(p) $\Rightarrow$ binary variable with $P(X = 1) = p$

Results on the 3 Toy Datasets



- 50 trials, $n = 400$
- F-measure (F)
- $F = 2PR/(P + R)$
- $0 \leq F \leq 1$
- $F = 1 \Leftrightarrow$ only and all true features are selected.

- **PC**, **F-HSIC**, **F-LSMI** cannot handle interacting features.
  - Simultaneous consideration of features is necessary.
- Inappropriate σ_x (Gaussian width) makes **F-HSIC**, **B-HSIC**, **ℓ_1-HSIC** fail sometimes in quad problem.
- **B-LSMI** sometimes greedily keeps some of redundant features in and-or problem.
- **ℓ_1-LSMI** works well in all cases.

Outline

SVC/SVR Test Errors on Medium-Dimensional Real Data

Dataset	m	n	k	PC	ℓ_1-HSIC	ℓ_1-LSMI	mRMR	Relief
abalone (R)	8	400	4	**1.63 (0.9)**	**1.65 (0.9)**	**1.60 (0.8)**	**1.64 (0.8)**	**1.58 (0.8)**
bcancer (C2)	9	277	4	0.24 (0.0)	**0.23 (0.0)**	**0.23 (0.0)**	0.25 (0.0)	0.26 (0.0)
glass (C6)	9	214	4	**0.29 (0.0)**	0.30 (0.0)	0.30 (0.0)	0.30 (0.0)	0.31 (0.0)
housing (R)	13	400	4	4.03 (0.2)	**3.95 (0.2)**	**3.91 (0.2)**	**3.97 (0.2)**	4.10 (0.2)
vowel (C11)	13	400	4	**0.20 (0.0)**	**0.20 (0.0)**	0.21 (0.0)	**0.20 (0.0)**	0.21 (0.0)
wine (C3)	13	178	4	0.03 (0.0)	**0.03 (0.0)**	**0.03 (0.0)**	**0.03 (0.0)**	0.03 (0.0)
image (C2)	18	400	4	0.10 (0.0)	0.13 (0.0)	0.06 (0.0)	0.14 (0.0)	**0.05 (0.0)**
segment (C7)	18	400	4	0.19 (0.0)	0.11 (0.0)	**0.05 (0.0)**	**0.05 (0.0)**	0.13 (0.0)
vehicle (C4)	18	400	4	0.32 (0.0)	0.34 (0.0)	**0.27 (0.0)**	0.39 (0.1)	0.32 (0.0)
german (C2)	20	400	4	**0.24 (0.0)**	0.25 (0.0)	**0.25 (0.0)**	0.25 (0.0)	0.26 (0.0)
cpuact (R)	21	400	4	**0.25 (0.0)**	0.54 (0.3)	**0.25 (0.2)**	**0.23 (0.1)**	0.37 (0.1)
ionosphere (C2)	33	351	4	0.07 (0.0)	0.07 (0.0)	**0.07 (0.0)**	0.09 (0.0)	**0.07 (0.0)**
satimage (C6)	36	400	10	0.22 (0.0)	0.14 (0.0)	**0.13 (0.0)**	0.14 (0.0)	0.16 (0.0)
spectf (C2)	44	267	10	0.19 (0.0)	0.19 (0.0)	**0.17 (0.0)**	0.18 (0.0)	0.18 (0.0)
senseval2 (C3)	50	400	10	**0.18 (0.0)**	0.19 (0.0)	**0.18 (0.0)**	0.18 (0.0)	0.21 (0.0)
speech (C2)	50	400	10	0.01 (0.0)	**0.01 (0.0)**	0.01 (0.0)	0.02 (0.0)	0.03 (0.0)
sonar (C2)	60	208	10	0.23 (0.0)	0.21 (0.0)	**0.16 (0.0)**	**0.18 (0.0)**	0.19 (0.0)
msd (R)	90	400	10	**0.95 (0.1)**	**0.94 (0.1)**	**0.93 (0.1)**	0.97 (0.1)	**0.96 (0.1)**
musk1 (C2)	166	400	20	0.19 (0.0)	0.16 (0.0)	0.16 (0.0)	**0.15 (0.0)**	0.19 (0.0)
musk2 (C2)	166	400	20	0.09 (0.0)	**0.09 (0.0)**	**0.08 (0.0)**	0.09 (0.0)	**0.09 (0.0)**
ctslices (R)	384	400	20	0.82 (0.1)	0.65 (0.0)	**0.38 (0.0)**	**0.45 (0.0)**	0.56 (0.0)
isolet (R)	617	400	20	5.92 (0.3)	5.85 (0.4)	**5.30 (0.4)**	**5.39 (0.4)**	6.27 (0.3)
Top Count				7	8	**17**	10	5

- classification error/mean squared error (SD)
- Paired t-test with 5% significance level. 50 trials.

SVC/SVR Test Errors on High-Dimensional Real Data

Dataset	m	n	PC	ℓ_1-HSIC	ℓ_1-LSMI	mRMR	Relief
warp. (C10)	2429	210	0.062 (0.00)	0.052 (0.01)	**0.031 (0.01)**	0.033 (0.00)	0.043 (0.00)
BASE. (C2)	4862	400	0.120 (0.03)	**0.082 (0.02)**	0.120 (0.03)	0.094 (0.02)	0.270 (0.10)
TOX. (C4)	5748	171	0.370 (0.00)	0.280 (0.02)	**0.150 (0.06)**	0.260 (0.00)	0.310 (0.00)
CLL. (C3)	11349	111	**0.110 (0.00)**	0.120 (0.01)	0.130 (0.01)	0.140 (0.00)	0.260 (0.00)
SMK. (C2)	19993	187	0.240 (0.00)	**0.200 (0.02)**	0.220 (0.01)	0.220 (0.00)	0.250 (0.00)

- 10 trials
- Select $k = 20$ features

Discussion :

- ℓ_1-LSMI and ℓ_1-HSIC perform well.
- Results of Relief and PC suggest high-dimensional data have redundant features i.e., TOX.
- ℓ_1-LSMI performs well on TOX.

Conclusions

	Ranking	**Forward**	**Backward**	**Exhaustive**	ℓ_1
Optimization	discrete	discrete	discrete	discrete	cont.
Search Complexity	m	km	m^2	2^m	m
Consider Redundancy	×	△	○	◎	○
Consider Interaction	×	×	○	◎	○

	PC	**HSIC**	**MI**	**SMI**
Non-linear Dependency	×	○	○	○
Model Selection	not needed	×	○	○
Computational Efficiency	◎	○	×	△

- Extensively studied combinations of optimizations and measures.
- Proposed ℓ_1-LSMI $= \ell_1 +$ SMI.
- Demonstrated that ℓ_1-LSMI works well on real datasets.
- To present at IBISML and submit a journal to IEICE.

LSMI

- $I_s(X,Y) = \frac{1}{2}\iint \left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1\right)^2 p(\boldsymbol{x})p(\boldsymbol{y})\, d\boldsymbol{x}d\boldsymbol{y}$
- Directly model $\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$ with

  $g(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{G} := \{\boldsymbol{\alpha}^T\boldsymbol{\varphi}(\boldsymbol{x},\boldsymbol{y}) \mid \boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_b)^T \in \mathbb{R}^b\}$
- Find $\boldsymbol{\alpha}$ which minimizes the squared error $J(\boldsymbol{\alpha})$.

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\iint \left(g(\boldsymbol{x},\boldsymbol{y}) - g^*(\boldsymbol{x},\boldsymbol{y})\right)^2 p(\boldsymbol{x})p(\boldsymbol{y})\, d\boldsymbol{x}d\boldsymbol{y}$$
$$= \frac{1}{2}\iint g(\boldsymbol{x},\boldsymbol{y})^2 p(\boldsymbol{x})p(\boldsymbol{y})\, d\boldsymbol{x}d\boldsymbol{y} - \iint g(\boldsymbol{x},\boldsymbol{y})p(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{x}d\boldsymbol{y} + C$$

$$J(\boldsymbol{\alpha}) \approx \widehat{J}(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T\widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^T\boldsymbol{\alpha}$$
$$\widehat{\boldsymbol{H}} := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\boldsymbol{\varphi}(\boldsymbol{x}_i,\boldsymbol{y}_{i'})\boldsymbol{\varphi}(\boldsymbol{x}_i,\boldsymbol{y}_{i'})^T$$
$$\widehat{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varphi}(\boldsymbol{x}_i,\boldsymbol{y}_i)$$

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}\boldsymbol{\alpha}^T \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^T \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha}$$

$\widehat{\boldsymbol{\alpha}}$ can be solved analytically with

$$\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_{b\times b}\right)^{-1} \widehat{\boldsymbol{h}}$$

$\widehat{\boldsymbol{\alpha}}$ can be used to estimate the SMI by

$$\widehat{I}_s = \frac{1}{2}\widehat{\boldsymbol{h}}^T \widehat{\boldsymbol{\alpha}} - \frac{1}{2}$$

- $\widehat{I}_s$ is called Least-Squares Mutual Information (LSMI)
- Basis functions are defined by the product kernel:

$$\begin{aligned} \varphi_l(\boldsymbol{x}, \boldsymbol{y}) &= \phi_l^x(\boldsymbol{x}) \phi_l^y(\boldsymbol{y}) \\ &= \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{c}(l)}\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{y}_{\boldsymbol{c}(l)}\|^2}{2\sigma^2}\right) \end{aligned}$$

where $\boldsymbol{c} \subseteq \{1, \ldots, n\}$ is the list of b indices of observations chosen as Gaussian centers.

Delta Kernel for Classification Task

Delta kernel is used on $\mathbf{Y}$ for classification task.

$$\phi_l^y(y) = \delta(y, y_{\boldsymbol{c}(l)})$$

$$\delta(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$$

So that,

$$\begin{aligned} \varphi_l(\boldsymbol{x}, y) &= \phi_l^x(\boldsymbol{x}) \phi_l^y(y) \\ &= \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{c}(l)}\|^2}{2\sigma^2} \right) \delta(y, y_{\boldsymbol{c}(l)}) \end{aligned}$$

Model Selection by Cross Validation

Cross validation is available for the SMI estimator for selecting (σ, λ).

- Divide $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ into K disjoint subsets $\{\mathcal{S}_k\}_{k=1}^K$
- Calculate $\widehat{\boldsymbol{H}}_{\mathcal{S}_k}$ and $\widehat{\boldsymbol{h}}_{\mathcal{S}_k}$ with $\mathcal{S}_k$
- Calculate $\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_{-k}}$ with $\{\mathcal{S}_j\}_{j \neq k}$
- Choose (σ, λ) which minimizes

$$\widehat{J}^{(K-CV)} := \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{2} \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_{-k}}^T \widehat{\boldsymbol{H}}_{\mathcal{S}_k} \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_{-k}} - \widehat{\boldsymbol{h}}_{\mathcal{S}_k}^T \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_{-k}} \right)$$

Sample Application: Document Classification

- $\mathbf{X}$: term-document matrix.
- $\mathbf{Y}$: document categories

- Use bag-of-words representation

Term \ Doc	Doc_1	Doc_2	$\cdots$
approach	1.0	3.0	$\cdots$
binary	0.0	2.0	$\cdots$
block	2.0	0.0	$\cdots$
common	0.0	1.0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$
Category	sport	science	$\cdots$

- Feature selection can be used to understand which words can explain different categories.

Feature Interaction

Feature interaction

Features are interacting if they can explain the output in presence of each other, even though each feature may not be explanatory.

X_1	X_2	$Y = \text{xor}(X_1, X_2)$
0	0	0
0	1	1
1	0	1
1	1	0

- $X_1 = 0 \Rightarrow Y$ can be 0 or 1.
- $X_1 = 1 \Rightarrow Y$ can be 0 or 1.
- Same for X_2.
- Neither X_1 nor X_2 can explain Y.
- But, X_1 and X_2 together can explain.

- All features need to be considered simultaneously.

Feature Interaction

Feature interaction

Features are interacting if they can explain the output in presence of each other, even though each feature may not be explanatory.

X_1	X_2	$Y = \text{xor}(X_1, X_2)$
0	0	0
0	1	1
1	0	1
1	1	0

- $X_1 = 0 \Rightarrow Y$ can be 0 or 1.
- $X_1 = 1 \Rightarrow Y$ can be 0 or 1.
- Same for X_2.
- Neither X_1 nor X_2 can explain Y.
- But, X_1 and X_2 together can explain.

- All features need to be considered simultaneously.

Feature Interaction

Feature interaction

Features are interacting if they can explain the output in presence of each other, even though each feature may not be explanatory.

X_1	X_2	$Y = \text{xor}(X_1, X_2)$
0	0	0
0	1	1
1	0	1
1	1	0

- $X_1 = 0 \Rightarrow Y$ can be 0 or 1.
- $X_1 = 1 \Rightarrow Y$ can be 0 or 1.
- Same for X_2.
- Neither X_1 nor X_2 can explain Y.
- But, X_1 and X_2 together can explain.

- All features need to be considered simultaneously.

Feature Interaction

Feature interaction

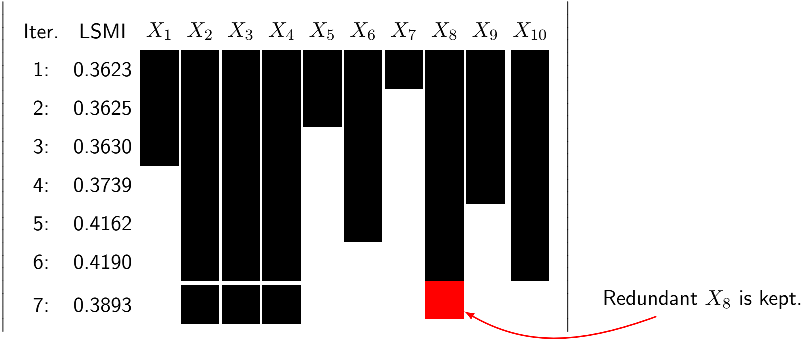
Features are interacting if they can explain the output in presence of each other, even though each feature may not be explanatory.

X_1	X_2	$Y = \text{xor}(X_1, X_2)$
0	0	0
0	1	1
1	0	1
1	1	0

- $X_1 = 0 \Rightarrow Y$ can be 0 or 1.
- $X_1 = 1 \Rightarrow Y$ can be 0 or 1.
- Same for X_2.
- Neither X_1 nor X_2 can explain Y.
- But, X_1 and X_2 together can explain.

- All features need to be considered simultaneously.

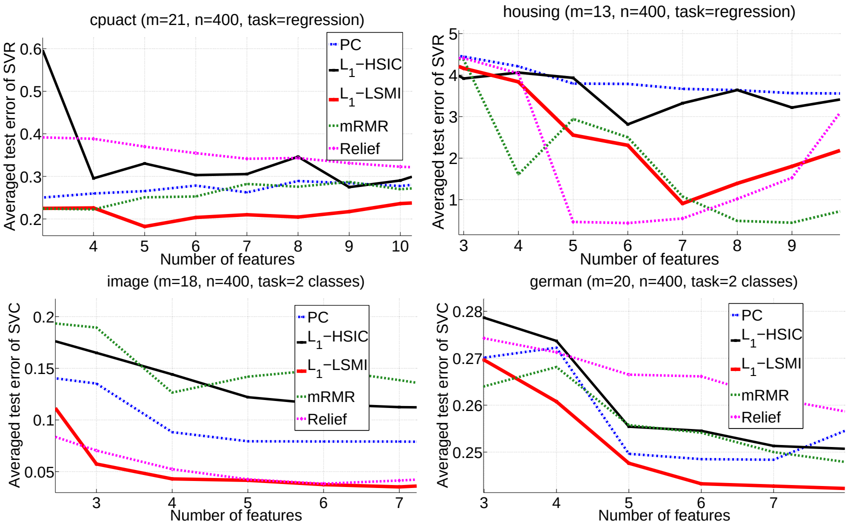
Results on and-or of B-LSMI



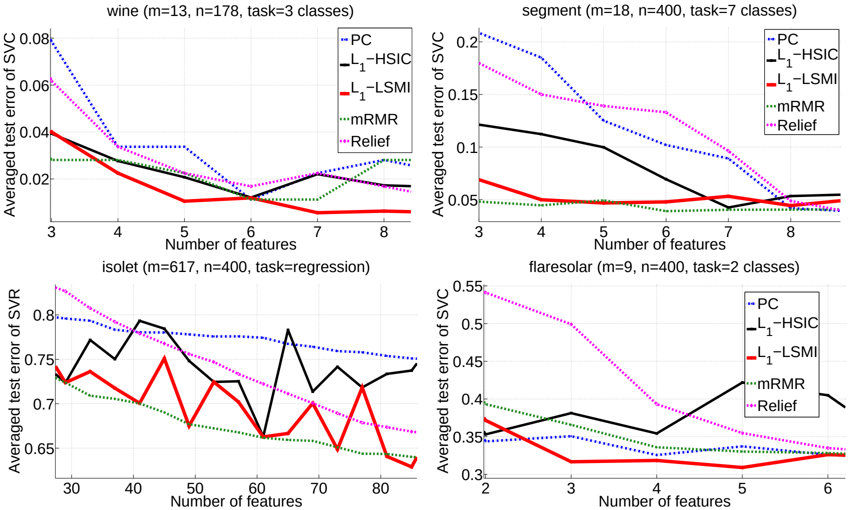
and-or $(k = 4, m = 10)$

- $Y = (X_1 \wedge X_2) \vee (X_3 \wedge X_4)$
- $X_1, \ldots, X_7 \sim$ Bernoulli(0.5)
- $X_8, \ldots, X_{10} = Y$ with 0.2 chance of bit flip
- **Characteristics:** feature redundancy, weak interaction

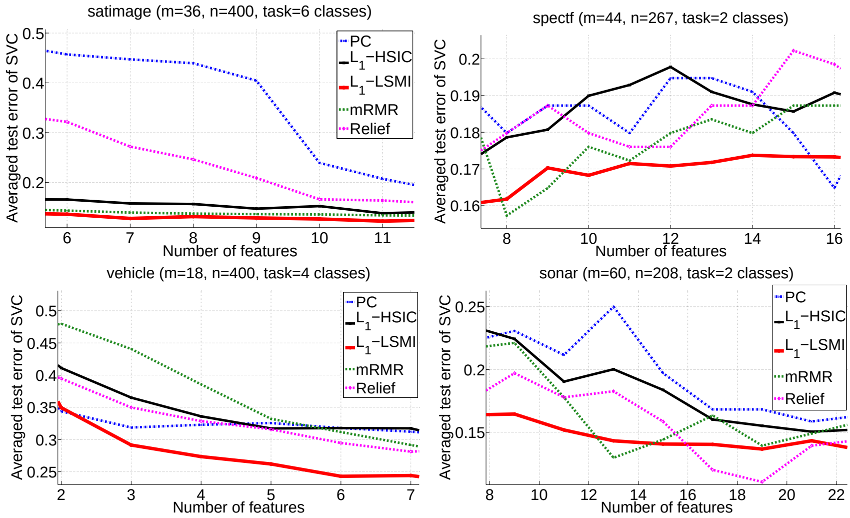
SVC/SVR CV Errors on Real Data I



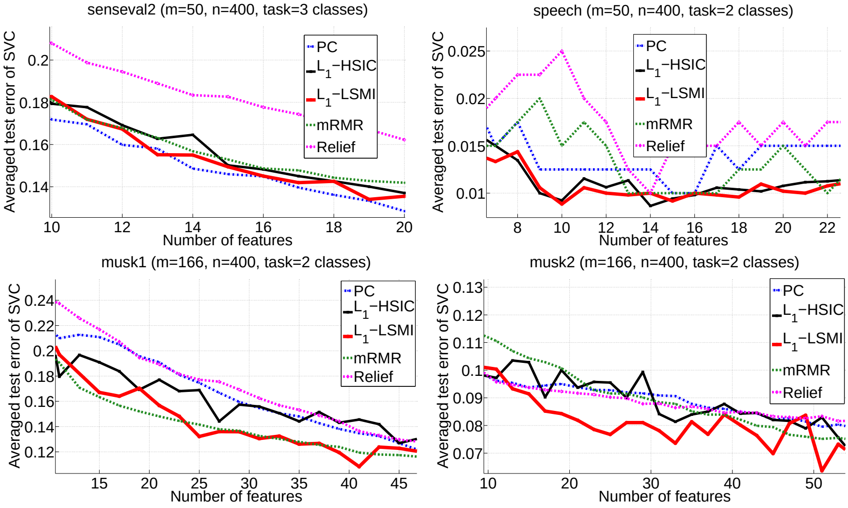
SVC/SVR CV Errors on Real Data II



SVC/SVR CV Errors on Real Data III



SVC/SVR CV Errors on Real Data IV



F-measure of the Selected Features

Precision

$P = (\#$ correctly selected features) / ($\#$ selected features)

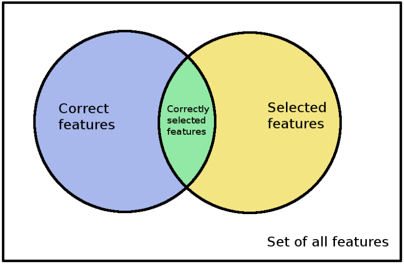
Recall

$R = (\#$ correctly selected features) / ($\#$ correct features)

F-measure

$F = 2PR/(P + R)$

- $0 \leq F \leq 1$



Gradient Projection Algorithm to Solve ℓ_1-LSMI

Require: $\boldsymbol{w}_0$ (initial point), z (ℓ_1 ball's radius)

1: **for** $t = 0 \to t_{max} - 1$ **do** // t_{max} `denotes the maximum number of iterations`

2: $s \leftarrow 1/\sqrt{t}$ // `step size`

3: $\boldsymbol{w}_{t+1} \leftarrow \pi_z(\boldsymbol{w}_t + s\nabla\widehat{I}_s(\mathrm{diag}(\boldsymbol{w}_t)\mathbf{X}, \mathbf{Y}))$ // π_z `is a projection operator onto the positive` ℓ_1 `ball with radius` z

4: **if** $\|\boldsymbol{w}_{t+1}\|_0 \leq 1$
or $\|\nabla\widehat{I}_s(\mathrm{diag}(\boldsymbol{w}_{t+1})\mathbf{X}, \mathbf{Y})\|_2 < \tau_{opt}$
or $|\widehat{I}_s(\mathrm{diag}(\boldsymbol{w}_{t+1})\mathbf{X}, \mathbf{Y}) - \widehat{I}_s(\mathrm{diag}(\boldsymbol{w}_t)\mathbf{X}, \mathbf{Y})| < \tau_{prog}$ **then**

5: break

6: **end if**

7: **end for**

8: **return** $\boldsymbol{w}_{t+1}$ // `set of selected features` $\mathcal{X}_z$ `can be determined by inspecting` $\boldsymbol{w}_{t+1}$

Minimal Redundancy Maximal Relevance (mRMR)

- Let $X = (X_1, \ldots, X_m)$ denote input variables.
- mRMR [Peng et al., 2005] uses mutual information [Cover and Thomas, 1991] to measure the dependency.

$$I(X, Y) = \iint \log\left(\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right) p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{x} d\boldsymbol{y}$$

- The optimization problem of mRMR is

$$\begin{aligned} \underset{\mathcal{I} \subset \{1, \ldots, m\}}{\text{maximize}} \quad & \overbrace{\frac{1}{k} \sum_{i \in \mathcal{I}} I(X_i, Y)}^{\text{relevance part}} - \overbrace{\frac{1}{k^2} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I}} I(X_i, X_j)}^{\text{pairwise redundancy constraint}} \\ \text{subject to} \quad & |\mathcal{I}| = k. \end{aligned}$$

- Still, mRMR cannot find interacting features since features are considered univariately.

Relief

- Relief [Kira and Rendell, 1992] is an iterative, distance-based, feature ranking algorithm.
- Relief disregards redundancy of features.

Set feature weights $\boldsymbol{w} \leftarrow \mathbf{0}_m$
for $i = 1 \rightarrow n$ **do**
  $\boldsymbol{s} \leftarrow$ Near-Hit of $\boldsymbol{x}_i$ // Nearest instance which has the same class as $\boldsymbol{x}_i$
  $\boldsymbol{d} \leftarrow$ Near-Miss of $\boldsymbol{x}_i$ // Nearest instance which has a different class to $\boldsymbol{x}_i$
  for $j = 1 \rightarrow m$ **do** // for each feature j
    $w_j \leftarrow w_j - (x_j - s_j)^2/n + (x_j - d_j)^2/n$
  end for
end for
Rank features in descending order of w_j.

LSMI Values in Andor Problem

Feature indices				LSMI
1	2	3	4	**0.4958**
1	2	3	8	0.3654
1	2	3	9	0.3806
1	2	3	10	0.3571
1	2	4	8	0.3764
1	2	4	9	0.3843
1	2	4	10	0.3724
1	2	8	9	0.3459
1	2	8	10	0.3302
1	2	9	10	0.3355
1	3	4	8	0.3822
1	3	4	9	0.3761
1	3	4	10	0.3915
1	3	8	9	0.3249
1	3	8	10	0.3303
1	3	9	10	0.3334
1	4	8	9	0.3423

Feature indices				LSMI
1	4	8	10	0.3354
1	4	9	10	0.3410
2	3	4	8	0.3666
2	3	4	9	0.3817
2	3	4	10	0.3903
2	3	8	9	0.3407
2	3	8	10	0.3120
2	3	9	10	0.3217
2	4	8	9	0.3403
2	4	8	10	0.3277
2	4	9	10	0.3281
3	4	8	9	0.3556
3	4	8	10	0.3487
3	4	9	10	0.3533
1	8	9	10	0.3299
2	8	9	10	0.3335
3	8	9	10	0.3031
4	8	9	10	0.3346

All possible 35 four-feature subsets of $\{X_1, \ldots, X_4\} \cup \{X_8, \ldots, X_{10}\}$ in andor dataset, and their corresponding values of LSMI to the output $Y = (X_1 \wedge X_2) \vee (X_3 \wedge X_4)$

Illustration of the Search for a k-feature Subset

Illustration of the Search for a k-feature Subset

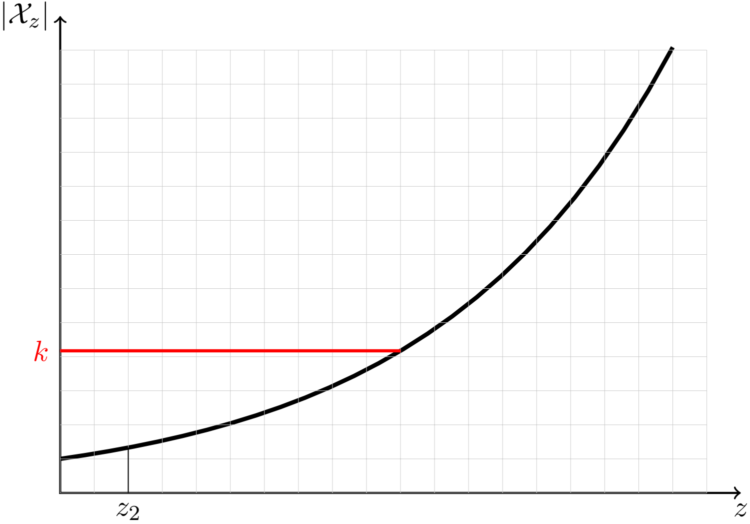


Illustration of the Search for a k-feature Subset

Illustration of the Search for a k-feature Subset

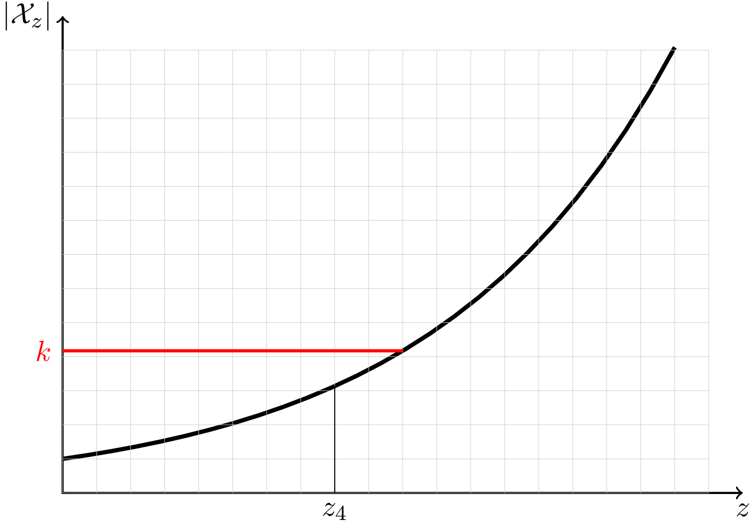


Illustration of the Search for a k-feature Subset

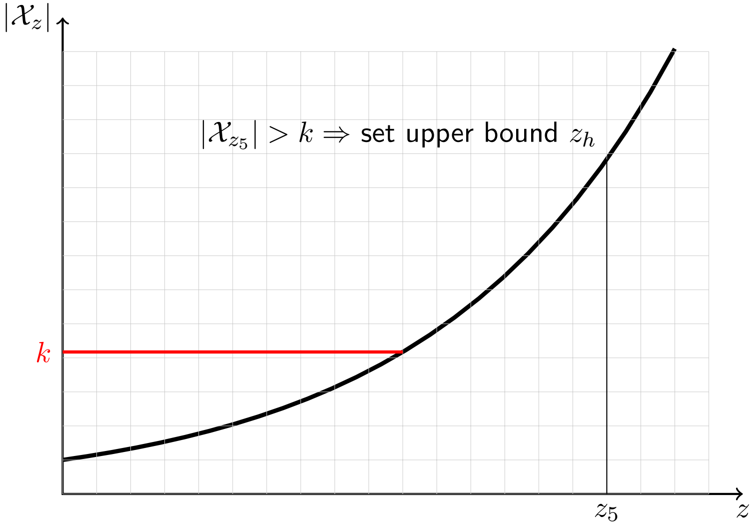


Illustration of the Search for a k-feature Subset

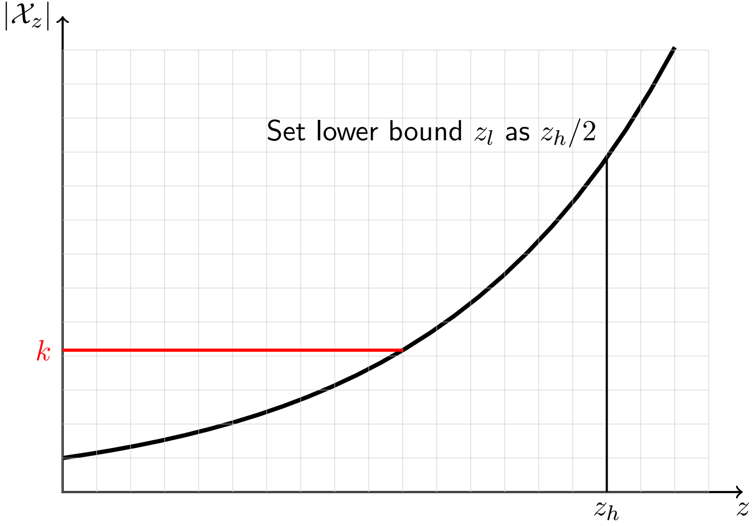


Illustration of the Search for a k-feature Subset

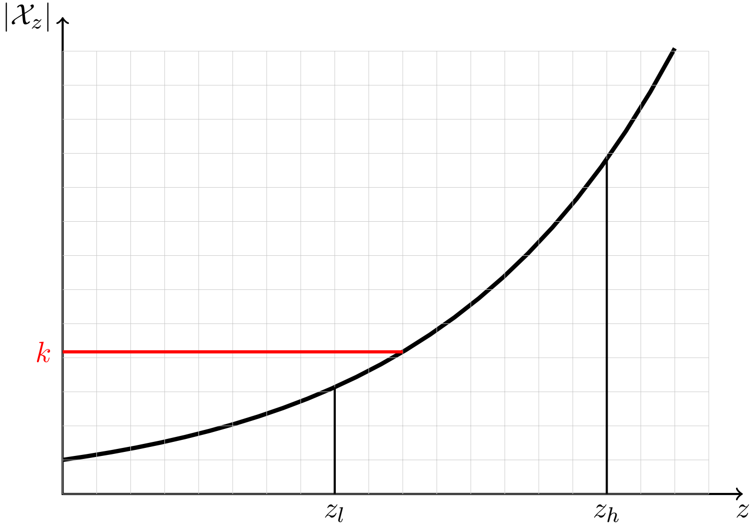


Illustration of the Search for a k-feature Subset

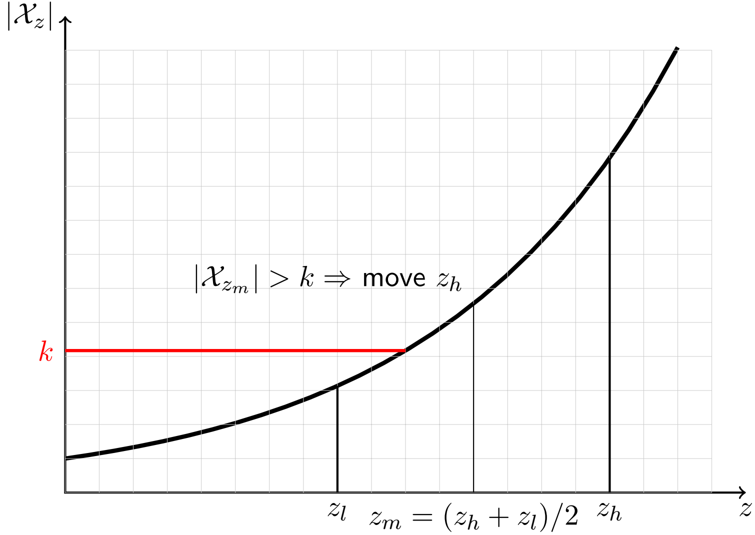
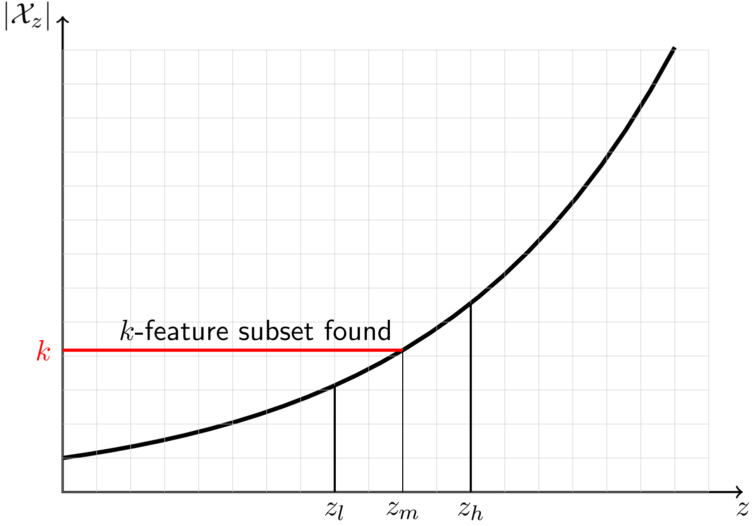


Illustration of the Search for a k-feature Subset

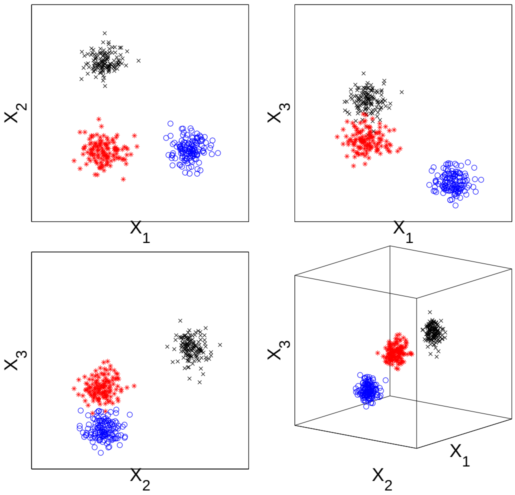
Illustration of the Search for a k-feature Subset



Toy Dataset: 3clusters

3clusters

- $\{X_1, X_2\}$ give a perfect separability, and are regarded as the true features
- X_3 is redundant.
- $X_4, \ldots, X_{10} \sim \mathcal{U}(0, 1)$
- **Characteristic:** feature redundancy



Summary of Real Datasets

Dataset	m	n	Task	Class balance (%)
BASEHOCK	4862	1993	B	49.9/50.1
CLL_SUB_111	11340	111	M3	9.9/44.1/45.9
SMK_CAN_187	19993	187	B	48.1/51.9
TOX_171	5748	171	M4	26.3/26.3/22.8/24.6
abalone	8	4177	R	-
bcancer	9	277	B	70.8/29.2
cpuact	21	3000	R	-
ctslices	384	53500	R	-
flaresolar	9	1066	B	44.7/55.3
german	20	1000	B	70.0/30.0
glass	9	214	M6	32.7/35.5/7.9/6.1/4.2/13.6
housing	13	506	R	-
image	18	1155	B	42.9/57.1
ionosphere	33	351	B	64.1/35.9
isolet	617	6238	R	-
msd	90	10000	R	-
musk1	166	476	B	56.5/43.5
musk2	166	6598	B	84.6/15.4
satimage	36	6435	M6	23.8/10.9/21.1/9.7/11.0/23.4
segment	18	2310	M7	14.3% per class
senseval2	50	534	M3	33.3% per class
sonar	60	208	B	46.6/53.4
spectf	44	267	B	20.6/79.4
speech	50	400	B	50.0/50.0
vehicle	18	846	M4	25.1/25.7/25.8/23.5
vowel	13	990	M11	9.1% per class
warpPIE10P	2420	210	M10	10% per class
wine	13	178	M3	33.1/39.9/27.0

References I

Cover, T. M. and Thomas, J. A. (1991).
Elements of information theory.
Wiley-Interscience, New York, NY, USA.

Gretton, A., Bousquet, O., Smola, A. J., and Schölkopf, B. (2005).
Measuring statistical dependence with hilbert-schmidt norms.
In Jain, S., Simon, H.-U., and Tomita, E., editors, *ALT*, volume 3734 of *Lecture Notes in Computer Science*, pages 63–77. Springer.

References II

Hall, M. A. (2000).
Correlation-based feature selection for discrete and numeric class machine learning.
In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 359–366, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kira, K. and Rendell, L. A. (1992).
The feature selection problem: Traditional methods and a new algorithm.
In *AAAI*, pages 129–134, Cambridge, MA, USA. AAAI Press and MIT Press.

References III

Peng, H., Long, F., and Ding, C. (2005).
Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 27:1226–1238.

Suzuki, T., Sugiyama, M., Kanamori, T., and Sese, J. (2009).
Mutual information estimation reveals global associations between stimuli and biological processes.
BMC Bioinformatics, 10(S-1).

References IV

Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. (2008).
Approximating mutual information by maximum likelihood density ratio estimation.
In Saeys, Y., Liu, H., Inza, I., Wehenkel, L., and de Peer, Y. V., editors, *Proceedings of ECML-PKDD2008 Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery 2008 (FSDM2008)*, volume 4 of *JMLR Workshop and Conference Proceedings*, pages 5–20, Antwerp, Belgium.

Tibshirani, R. (1996).
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society (Series B), 58:267–288.