Aux 3-{4}-5. Introducción a la Minería de Datos

Gastón L'Huillier[1,2], Richard Weber[2]
glhuilli@dcc.uchile.cl

[1]Departamento de Ciencias de la Computación
Universidad de Chile

[2]Departamento de Ingeniería Industrial
Universidad de Chile

2010

Auxiliares 3, 4 y 5

Selección de atributos (aux3)

- Coeficiente de correlación
- Ganancia de información (*Information Gain*)
- Wrapper Methods (*Forward Selection* y *Backward Elimination*)
- Embedded Methods (Arboles de Decisión)

Enunciado tarea 1 - Introducción a RapidMiner5.0 (aux4)

- Enunciado tarea 1
- Cargar datos, Re-codificar variables y conceptos de operadores básicos para pre-procesamiento de datos

Extracción de atributos (aux5)

- *Principal Component Analysis* (PCA)
- *Independent Component Analysis* (ICA)
- Aplicación de técnicas de selección y extracción de atributos en RapidMiner5.0

Recuerdo: Aprendizaje Supervisado

Dataset

$$\mathcal{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,A} \\ x_{2,1} & \cdots & x_{2,A} \\ \vdots & \vdots & \vdots \\ x_{N,1} & \cdots & x_{N,A} \end{pmatrix}, \mathcal{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

$$\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N} \text{ , } \mathbf{x_i} = (x_{i,1}, \ldots, x_{i,A}) \text{ , } \forall i \in \{1, \ldots, N\}$$

$$\mathcal{Y} = (y_1, \ldots, y_N)^T$$

Probabilidad que cliente$_j$ pague el credito

$$y = f(\mathcal{X})$$

$$\Rightarrow y_i = P(\text{ pague el credito } |\mathbf{x}_i)$$

Recuerdo: Aprendizaje No Supervisado

Dataset

$$\mathcal{X} = \begin{pmatrix} x_{1,1} & \dots & x_{1,A} \\ x_{2,1} & \dots & x_{2,A} \\ \vdots & \vdots & \vdots \\ x_{N,1} & \dots & x_{N,A} \end{pmatrix}, k$$

$$\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N} \ , \ \mathbf{x_i} = (x_{i,1}, \dots, x_{i,A}) \ , \ \forall i \in \{1, \dots, N\}$$

$$k \in \mathbb{N}_+$$

Segmento del cliente$_j$

$$\{\mathbf{c}_i\}_{i=1}^{k} \leftarrow f(\mathcal{X}, k)$$

$$\Rightarrow y_j = \arg\min_i d(\mathbf{x}_j, \mathbf{c}_i) \ , \ \forall i \in \{1, \dots, k\}$$

Selección y Extracción de atributos

Motivación

- Por cada objeto lo puedo representar por muchos atributos ($\mathbf{x} = \{x_1, \ldots, x_A\}$).
- El conjunto de atributos puede no entregarme información válida.
- Cada atributo del conjunto de atributos puede entregarme la misma información.
- Alta dimensionalidad ($A >> 0$) puede hacer el modelo más complejo, y dificil de explicar.
- Restricciones de procesamiento (hardware) ante una cantidad muy grande de atributos.
- etc...

Limpieza y Transformación de Datos

Algunas etapas

- Visualización y exploración de datos[a].
- Imputación de datos (fill($x_{i,j}$))
- Eliminación de outliers
- Transformación de atributos ($T(\mathbf{x}_{\cdot,j})$)
- Normalización y estandarización de atributos
- Selección de observaciones (*Sampling*) ($\mathbf{x}_i$)
- Selección de atributos ($\mathbf{x}_{\cdot,j}$)

[a] No necesariamente se ejecutan en este orden, es un proceso iterativo.

Algunas referencias

- [Myatt, 2006, Myatt and Johnson, 2009]

Selección y Extracción de Atributos

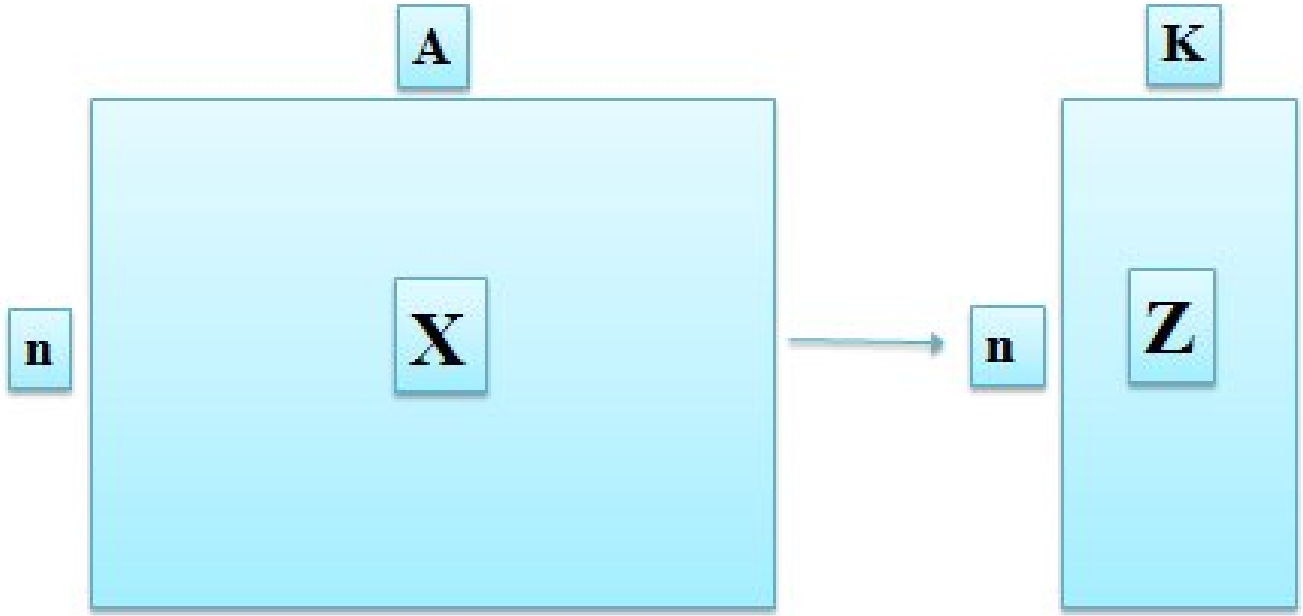

Figura: Reducción de dimensionalidad

Selección de atributos - Correlación

Coeficiente de correlación [Guyon and Elisseeff, 2003]

$$\text{Corr}(\mathbf{x}_{\cdot,j}) = \frac{cov(\mathbf{x}_{\cdot,j}, \mathcal{Y})}{\sqrt{var(\mathbf{x}_{\cdot,j})var(\mathcal{Y})}} \tag{1}$$

$$\text{Corr}(\mathbf{x}_{\cdot,j}) = \frac{\sum_{i=1}^{N}(x_{i,j} - \bar{x_i})(y_j - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_{i,j} - \bar{x_i})^2 \sum_{i=1}^{N}(y_j - \bar{y})^2}}$$

- Detecta sólo dependencias lineales.
- Alta correlación $\Rightarrow \text{Corr}(\mathbf{x}_{\cdot,j}) > 0{,}75$.

Selección de atributos - Ganancia de Información

Entropía de una variable

$$\mathrm{H}(\mathbf{x}_{\cdot,j}) = -\sum_{x_j \in \mathcal{X}_j} p(x_j) log(p(x_j)) \text{ , } \mathcal{X}_j \text{ es el espacio de valores de } \mathbf{x}_{\cdot,j}$$

Entropía Condicional

$$H(\mathbf{x}_{\cdot,j}|\mathcal{Y}) = -\sum_{x_j \in \mathcal{X}_j} \sum_{y \in \mathcal{Y}} p(x_j, y) log\left(\frac{p(x_j, y)}{p(x_j)}\right)$$

Ganancia de Información [Guyon and Elisseeff, 2003]

$$\mathrm{I}(j) = H(\mathbf{x}_{\cdot,j}) - H(\mathbf{x}_{\cdot,j}|\mathcal{Y})$$

- Calcular ganancia de información $\forall \mathbf{x}_{\cdot,j} \in X$, listar en orden decreciente y seleccionar los top k atributos.

Selección de Atributos - Wrapper & Embedded Methods

Wrapper Methods

- Se evalúa un modelo en base a un conjunto de atributos construido iterativamente.
- *Forward Selection*: Comienzo con un conjunto vacío, y se van agregando atributos de manera incremental.
- *Backward Elimination*: Comienzo con todo el conjunto de atributos, y se van eliminando iterativamente.
- *Random Sampling*: Se considera la evaluación de conjuntos aleatorios de atributos.
- Problema de optimización combiatorial NP-hard (*set-covering*)

Embedded Methods

- El mismo modelo se preocupa de una selección intrínseca de atributos.
- e.g. Árboles de decisión (lo veremos en + adelante).

Extracción de Atributos - PCA

Análisis de Componentes Principales

- Determinar conjunto de atributos $\mathcal{K}$ del conjunto original $\mathcal{X}_A = \{\mathbf{x}_{\cdot,j}\}_{j=1}^{A}$, donde $|\mathcal{K}| \leq |\mathcal{X}_A| = A$
- Cada atributo $k \in K$ es generado en base a una combinación lineal no correlacionada (ortonormal) de los atributos originales.
- Maximizar la varianza de los datos en cada componente principal.
- Todas las componentes son independientes entre ellas (ortogonales).
- La primera componente principal es aquella que representa la máxima variabilidad con respecto a los atributos originales.
- Las siguientes componentes están ordenadas según la ortogonalidad de la componente principal anterior.

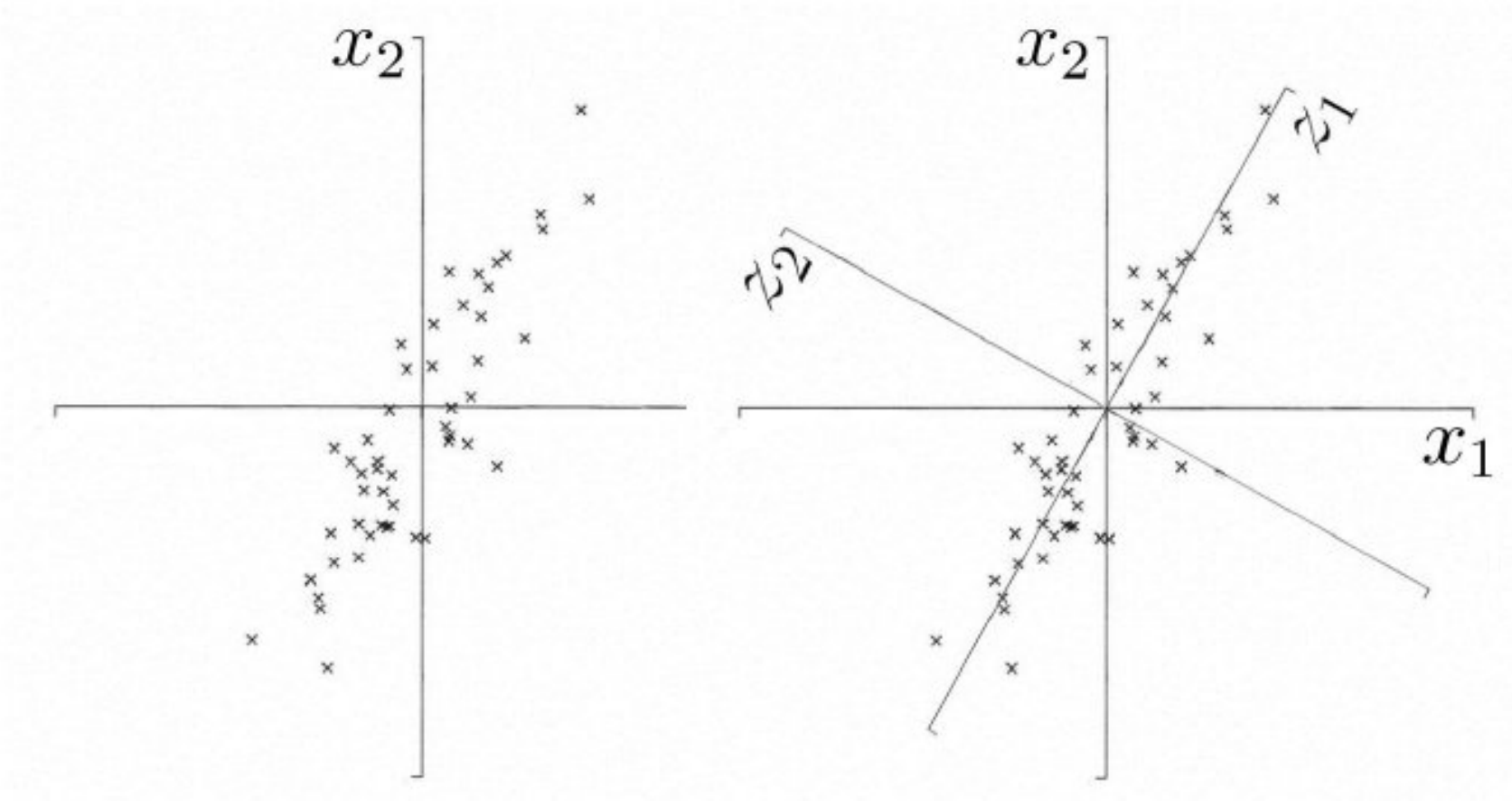

Figura: PCA

Extracción de Atributos - PCA

Análisis de Componentes Principales

$$\begin{aligned} \max_{\mathbf{a}_k} \quad & var(\mathbf{z}_k) \\ \text{subject to} \quad & cov(\mathbf{z}_k, \mathbf{z}_l) = 0,\ k > l \geq 0 \\ & \mathbf{a}_k^T \cdot \mathbf{a}_k = 1, \mathbf{a}_k \cdot \mathbf{a}_l^T = 0, \forall k, l \in K \text{ and } \forall k \neq l \end{aligned}$$

PCA

- Dado el conjunto de observaciones $\{\mathbf{x}_i\}_{i=1}^N$, el objetivo es determinar la componente principal $\mathbf{z}_k$, estimando la combinación lineal $\mathbf{a}_k = (a_{1,k}, \ldots, a_{N,k})$, donde $\mathbf{z}_k = \mathbf{a}_k^T \cdot \mathbf{x}_i$
- Donde $var(\cdot)$ y $cov(\cdot)$ son las funciones de varianza y covarianza respectivamente.

Extracción de Atributos - ICA

Independent Component Analysis

- Técnica que permite extraer factores latentes que entreguen una representación “relevante” del conjunto de información inicial.
- Al igual que PCA, asume combinaciones lineales entre los atributos originales.
- Se diferencia en que se desea minimizar la información mutua entre los atributos.

$$\mathrm{I}(j) = H(\mathbf{x}_{\cdot,j}) - H(\mathbf{x}_{\cdot,j}|\mathcal{Y})$$

- En ICA además se busca independencia estadística entre los atributos extraídos

$$f_{1,2,\ldots,k}(z_1, \ldots, z_k) = f_1(z_1) \cdots f_k(z_k)$$

ICA vs PCA

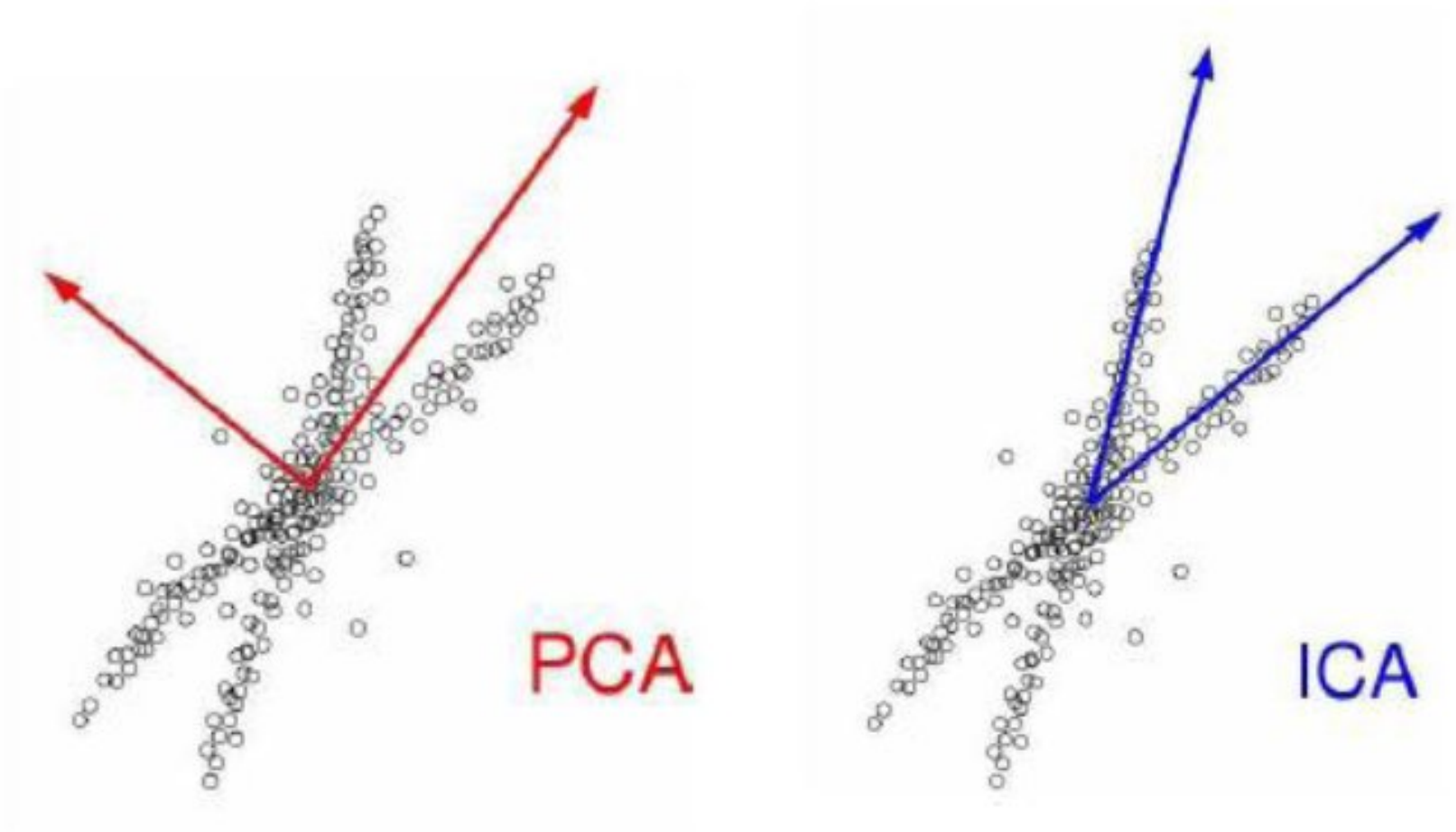

Figura: ICA vs PCA

RapidMiner v5.0

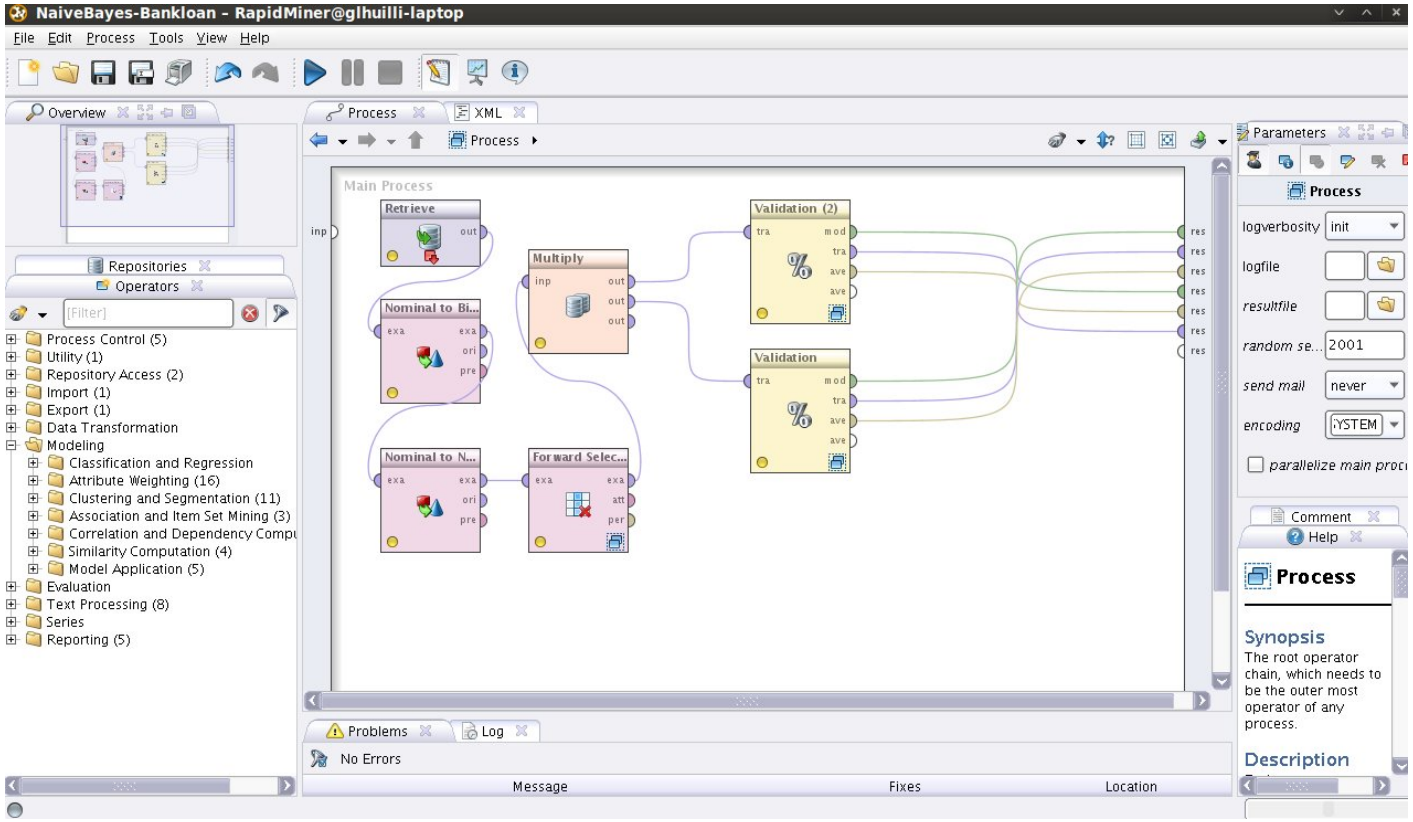

Figura: Software RapidMiner

`http://rapid-i.com/content/view/167/82/` (v5.0 32bits)

References I


Guyon, I. and Elisseeff, A. (2003).
An introduction to variable and feature selection.
J. Mach. Learn. Res., 3:1157–1182.


Myatt, G. J. (2006).
Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining.
Wiley-Interscience.


Myatt, G. J. and Johnson, W. P. (2009).
Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications.
Wiley Publishing.