Aprendizado de Máquina

Preparação de Dados

Eduardo R. Hruschka

Agenda

- Pré-processamento de dados
- Preparação de dados para métodos não supervisionados
- Preparação de dados para métodos supervisionados
    - Filtros
    - Wrappers

*“**Data is the new oil**. It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.” (Clive Humby)*

Pré-processamento

- Dados reais em geral são:

  - incompletos: faltam valores e/ou atributos (Salário = “ “).
  - ruidosos: contêm erros e/ou outliers (Salário = -150).
  - inconsistentes: apresentam discrepâncias em códigos e nomes (1→A, 2→B, 3→C).

- Problemas causados por humanos, softwares, problemas de hardware, dados de diferentes fontes etc.

- Lembrar do princípio geral GIGO.

- Preparar os dados pode consumir mais de 80% do esforço de modelagem.

Atividades comuns

- Limpeza de dados: lidar com valores ausentes, dados ruidosos, outliers, inconsistências etc.
- Integração de dados de múltiplos bancos de dados e arquivos.
- Transformação de dados: normalização e agregação.
- Redução de dados: menos dados com mesmos resultados analíticos (seleção de atributos e de amostras).
- Discretização: compactar informação.
- Nunca deixar de fazer análise exploratória (médias, medianas, variâncias, min, max, gráficos etc).

Valores Ausentes

- Ocorrência comum:
  - Mau funcionamento de dispositivos de coleta de dados;
  - Dado omitido pela fonte de informação numa pesquisa;
  - Falha na digitação ou na composição da base;

- Formas de eliminação de valores ausentes:
  - Eliminar registros/atributos com valores ausentes;
    - Perda de dados pode ser considerável.

  - Preenchimento de valores (imputação)
    - Por uma constante. Ex.: Média/Moda do atributo
      - Desconsidera a relação entre atributos da base de dados
    - Por valores que tentem preservar as relações entre atributos da base de dados
      - Uso de um algoritmo de aprendizado.

Noção intuitiva sobre padrões de ausência

- Completamente aleatória (*Missing Completely at Random – MCAR*)
- Aleatória (*Missing at Random – MAR*)
  - Ausência de valor num atributo depende de valores de outro(s) atributo(s)
- Não aleatória (*Missing not at Random – MNAR*)
  - Ausência de um valor num atributo relacionada a uma condição envolvendo o próprio valor do atributo

Exemplo - pressão arterial de pacientes:
X: Medidas em Janeiro
Y: Medidas em Fevereiro:
Completo: de todos pacientes
MCAR: de pacientes escolhidos ao acaso
MAR: de pacientes com pressão < 140 em Janeiro
MNAR: registro de medidas maiores que 140

X (Jan)	*Y (Fev)*			
	All	MCAR	MAR	MNAR
169	148	148	148	148
126	123	-	-	-
132	149	-	-	149
160	169	-	169	169
105	138	-	-	-
116	102	-	-	-
133	150	-	-	150
109	96	-	-	-
106	148	-	-	148
176	137	-	137	-
128	155	-	-	155
131	131	-	-	-
130	101	101	-	-
145	155	-	155	155
136	140	-	-	-
146	134	-	134	-
111	129	-	-	-
97	85	85	-	-
134	124	124	-	-
153	112	-	112	-
137	122	122	-	-

Duas abordagens para avaliar imputações

a) **Predição**:

- Comparar valor imputado com valor conhecido;
- Qual métrica poderia ser usada?
- Viável em aplicações práticas?
- Avaliação em dados completos diminui a informação disponível pra avaliação da ferramenta de imputação.

b) **Modelagem**:

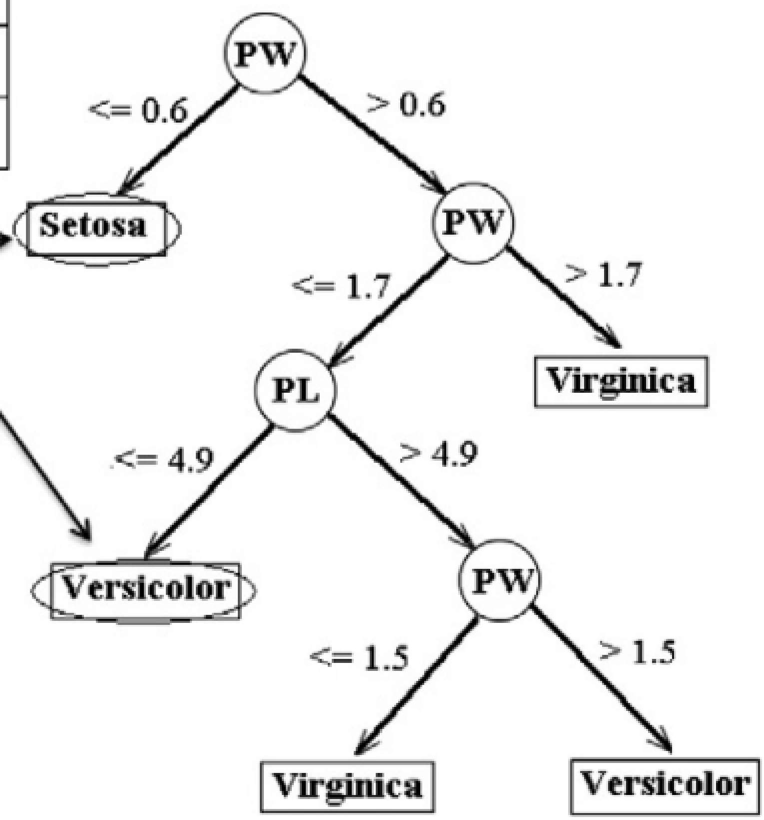
- Minimizar a influência na classificação, nas partições etc.
- Vejamos um exemplo ilustrativo...

Cuidado ao avaliar algoritmos de imputação:

- É preciso considerar a tarefa de modelagem;
- Imputação causa a falsa sensação de que valor passa a ser conhecido;

ID	SL	SW	PL	PW	CLASS
44	5	3.5	1.6	0.6	Setosa
151	5	3.5	1.6	?	Setosa

Imputation Method A => [5.0 3.5 1.6 **0.2**]
Imputation Method B => [5.0 3.5 1.6 **0.601**]



Abordagem simples para *clustering*

- Utilizar distância tolerante a ausentes;

- Exemplo para distância euclidiana:

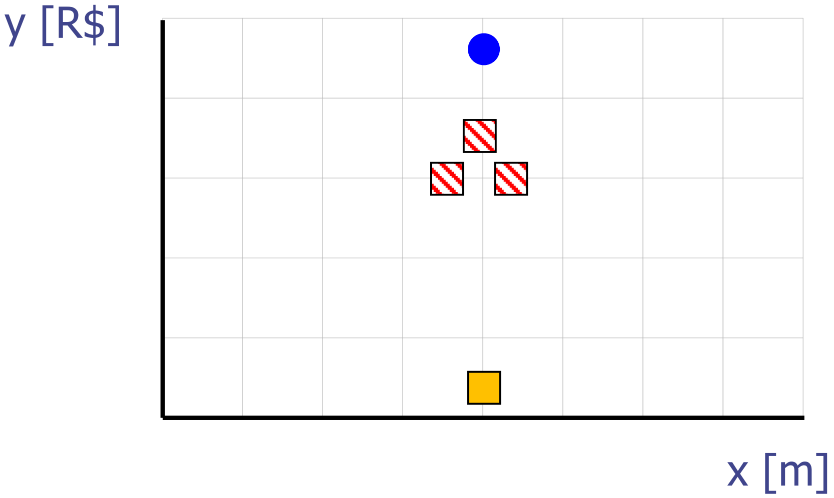
Obj. /Atrib.	a_1	a_2	A_3	a_4
x_1	2	-1	?	0
x_2	7	0	-4	8
x_3	?	3	5	2
x_4	?	10	?	5

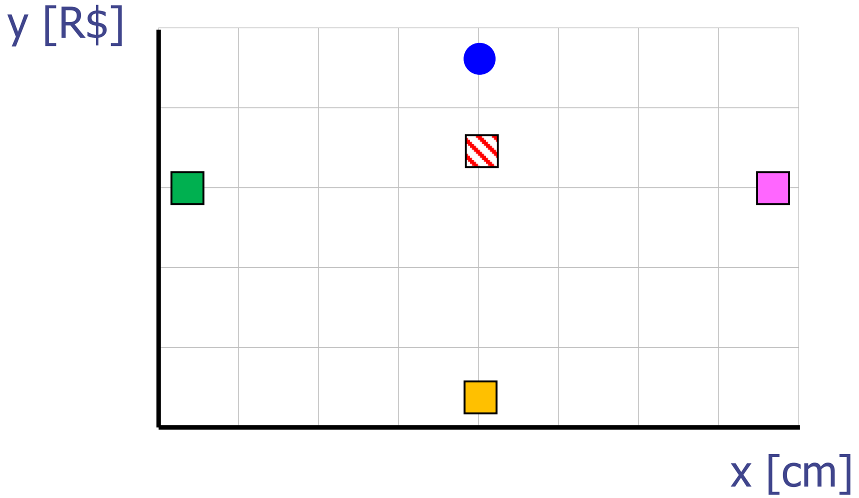
Exercício: calcule todas as demais distâncias.

Agenda

- Pré-processamento de dados
- Preparação de dados para métodos não supervisionados
- Preparação de dados para métodos supervisionados
  - Filtros
  - Wrappers

Preparação para aprendizado não supervisionado





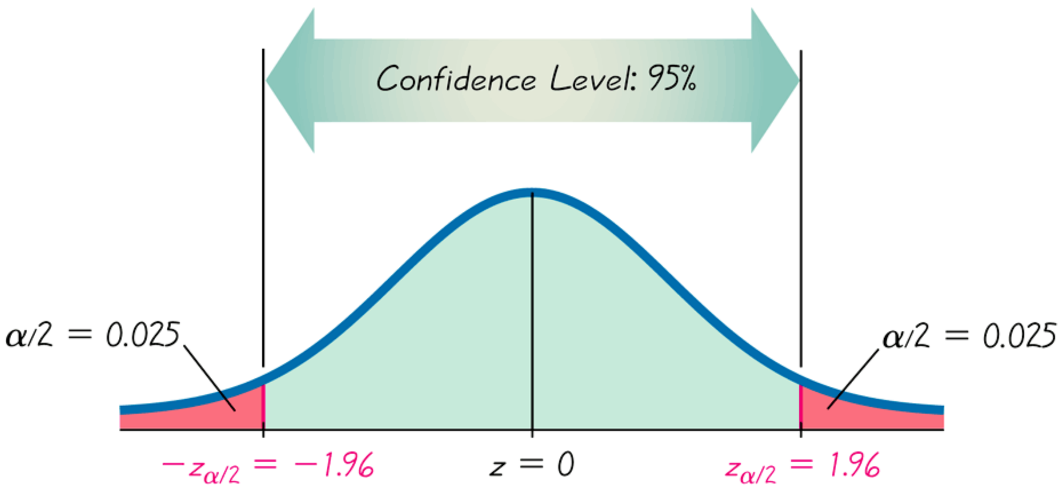
- Pode-se lidar com tais problemas por meio do que usualmente se denomina **normalização**.

- Vamos rever as formas de normalização mais comuns.

Normalizações comuns

- Reescala Linear [0,1]: $$l_{ij} = \frac{x_{ij} - \min(\mathbf{a}_j)}{\max(\mathbf{a}_j) - \min(\mathbf{a}_j)}$$

- Padronização por escore z: $$z_{ij} = \frac{x_{ij} - \mu_{\mathbf{a}_j}}{\sigma_{\mathbf{a}_j}}$$



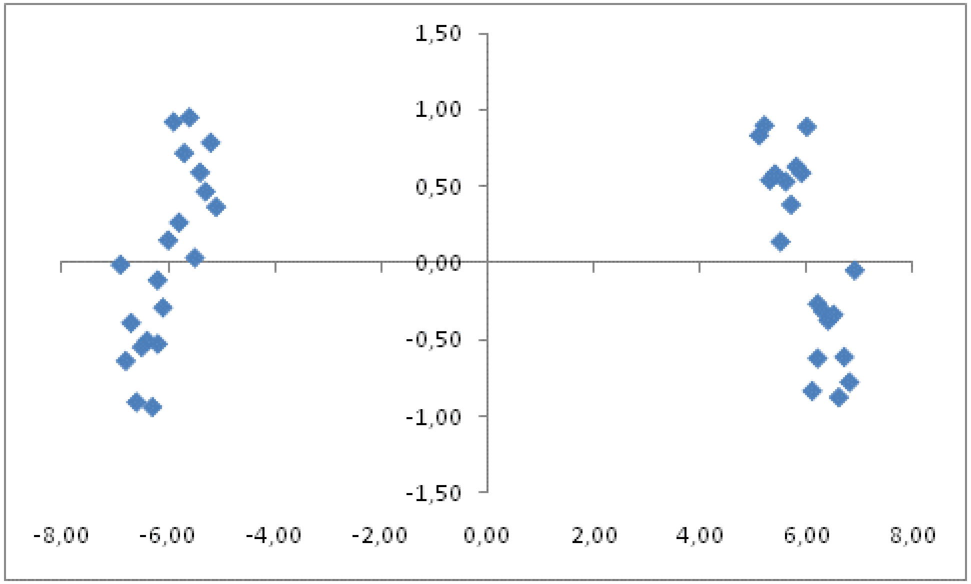
N(0,1) se atributo possui dist. Normal

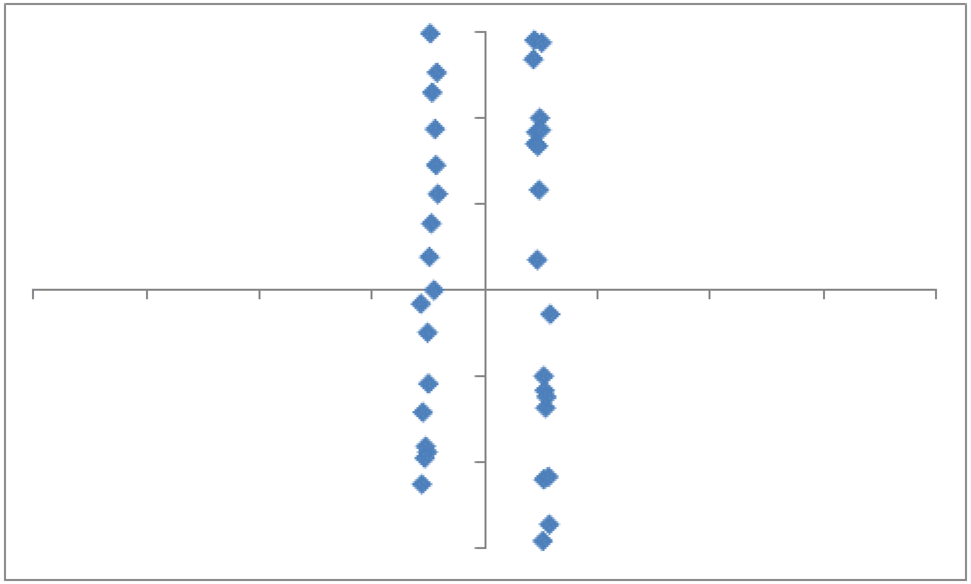
Triola, Notas de Aula, Copyright © 2004 Pearson Education

Normalização é sempre apropriada?

➢ *escore z*

(efeito semelhante para linear [0,1])





Discussão

- Atributos com escala mais ampla / maior variabilidade tendem a ter maior peso nos cálculos de distâncias;
  - Isso representa uma forma de pré-**ponderação** dos dados;
- Normalização busca eliminar esse efeito, presumindo-o ser artificial:
  - Simples consequência do uso de unidades de medida específicas;
  - Porém, impõe uma (contra) ponderação aos dados originais;
  - Introduz distorções se (ao menos parte das) diferentes variabilidades originais refletiam corretamente a natureza do problema;
- Agrupamento de dados é considerada uma área muito desafiadora.

Como lidar com atributos discretos?

	Sexo	País	Estado Civil	Comprar
$\mathbf{x}_1$	M	França	solteiro	Sim
$\mathbf{x}_2$	M	China	separado	Sim
$\mathbf{x}_3$	F	França	solteiro	Sim
$\mathbf{x}_4$	F	Inglaterra	casado	Sim
$\mathbf{x}_5$	F	França	solteiro	Não
$\mathbf{x}_6$	M	Alemanha	viúvo	Não
$\mathbf{x}_7$	M	Brasil	casado	Não
$\mathbf{x}_8$	F	Alemanha	casado	Não
$\mathbf{x}_9$	M	Inglaterra	solteiro	Não
$\mathbf{x}_{10}$	M	Argentina	casado	Não

Motivação:

$d(\mathbf{x}_1,\mathbf{x}_6)=?$

$d(\mathbf{x}_1,\mathbf{x}_7)=?$

Atributos binários

- Calcular a distância entre $\mathbf{x}_1 = [1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 0]$ e
  $\mathbf{x}_2 = [0\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 0\ 0]$
- Usando uma tabela de contingências temos:

		Objeto $\mathbf{x}_j$		
Objeto $\mathbf{x}_i$		1	0	Total
	1	n_{11}	n_{10}	$n_{11}+n_{10}$
	0	n_{01}	n_{00}	$n_{01}+n_{00}$
	Total	$n_{11}+n_{01}$	$n_{10}+n_{00}$	n

$$S^{SM}_{(\mathbf{x}_i,\mathbf{x}_j)} = \frac{n_{11}+n_{00}}{n_{11}+n_{00}+n_{10}+n_{01}} = \frac{n_{11}+n_{00}}{n}$$

Coeficiente de Casamento Simples (Zubin, 1938)

$$1 - S^{SM}_{(\mathbf{x}_i,\mathbf{x}_j)} = \frac{n_{10}+n_{01}}{n} = \frac{d^{\mathrm{Hamming}}_{(\mathbf{x}_i,\mathbf{x}_j)}}{n}$$

Atributos assimétricos

- **Atributos simétricos**: valores igualmente importantes
  - Exemplo típico $\rightarrow$ Sexo (M ou F)
- **Atributos assimétricos**: valores com importâncias distintas – presença de um efeito é mais importante do que sua ausência.
  - Exemplo: sejam 3 objetos que apresentam (1) ou não (0) dez sintomas para uma determinada doença

$$\mathbf{x}_1 = [1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1]$$

$$\mathbf{x}_2 = [1\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 0\ 0]$$

$$\mathbf{x}_3 = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$$

$S^{SM}(\mathbf{x}_1,\mathbf{x}_2) = 0.5;$

$S^{SM}(\mathbf{x}_1,\mathbf{x}_3) = 0.5;$

- Conclusão?

- Para atributos assimétricos, pode-se usar, por exemplo, o *Coeficiente de Jaccard* (1908):

$$S_{(\mathbf{x}_i,\mathbf{x}_j)}^{Jaccard} = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

  - Focada nos *casamentos* do tipo 1-1
  - Despreza *casamentos* do tipo 0-0

- Existem outras medidas similares na literatura, mas CCS e Jaccard são as mais utilizadas.
  - vide (Kaufman & Rousseeuw, 2005)

Exemplo:

p = [1 0 0 0 0 0 0 0 0 0]

q = [0 0 0 0 0 0 1 0 0 1]

n_{01} = 2 (número de atributos em que **p** = 0 e **q** = 1)
n_{10} = 1 (número de atributos em que **p** = 1 e **q** = 0)
n_{00} = 7 (número de atributos em que **p** = 0 e **q** = 0)
n_{11} = 0 (número de atributos em que **p** = 1 e **q** = 1)

$$CCS = (n_{11} + n_{00})/(n_{01} + n_{10} + n_{11} + n_{00})$$
$$= (0+7) / (2+1+0+7) = 0.7$$

$$J = n_{11} / (n_{01} + n_{10} + n_{11}) = 0 / (2 + 1 + 0) = 0$$

Atributos ordinais

Ex.: Gravidade de um efeito: {nula, baixa, média, alta}

- Ordem dos valores é importante

- Normalizar e então utilizar medidas de (dis)similaridade para valores contínuos (p. ex. Euclidiana, cosseno etc.):

    - {1, 2, 3, 4} $\rightarrow$ (rank – 1) / (número de valores – 1)

    - {0, 1/3, 2/3, 1}

➢ Abordagem comum

Atributos de várias naturezas misturados

Método de Gower (1971):

$$s_{(\mathbf{x}_i,\mathbf{x}_j)} = \frac{1}{n}\sum_{k=1}^{n} s_{ijk} \longrightarrow d_{(\mathbf{x}_i,\mathbf{x}_j)} = 1 - S_{(\mathbf{x}_i,\mathbf{x}_j)}$$

Para atributos nominais / binários:

$$\begin{cases} (x_{ik} = x_{jk}) \Rightarrow s_{ijk} = 1; \\ (x_{ik} \neq x_{jk}) \Rightarrow s_{ijk} = 0; \end{cases}$$

Para atributos ordinais ou contínuos:

$$s_{ijk} = 1 - \left| x_{ik} - x_{jk} \right| / R_k \qquad R_k = \max_m \mathbf{x}_{mk} - \min_m \mathbf{x}_{mk}$$

R_k= faixa de observações do *k-ésimo* atributo *(termo de normalização)*

Sumário

- Diferentes medidas de dis(similaridade) afetam a formação (indução) dos *clusters*
  - Como selecionar a medida de (dis)similaridade?
  - Devemos padronizar? Caso afirmativo, como?
- Infelizmente, não há respostas definitivas e globais.
- Análise de agrupamento de dados é, em essência, um processo subjetivo, dependente do problema
- Lembrem: **análise exploratória de dados**!

Agenda

- Pré-processamento de dados
- Preparação de dados para métodos não supervisionados
- Preparação de dados para métodos supervisionados
  - Filtros
  - Wrappers

Preparação para métodos supervisionados

Além das técnicas mencionadas anteriormente é comum realizar seleção de atributos (*feature selection*):

- Subconjunto mínimo de atributos tal que a distribuição de probabilidades para diferentes classes seja parecida à distribuição original (com todos os atributos);
- Facilita interpretação dos modelos obtidos;
- Reduz custo computacional de armazenamento (sistemas produtivos) e de inferência.

Referências bibliográficas:
- Guyon, I., Elisseeff, A., An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, 2003.
- Liu, H., Yu, L., Toward Integrating Feature Selection Algorithms for Classification and Clustering, IEEE Transactions on Knowledge and Data Engineering, 17(3), 1-12, 2005.

Complexidade e estratégias

- Otimização combinatória: existem 2^n subconjuntos possíveis de "*n*" atributos;
  - Busca exaustiva é usualmente inviável;
  - Diversas estratégias de busca:
    - Seleção *forward*;
    - Eliminação *backward;*
    - Bidirecional.
- Como parar a busca?
- Como avaliar os subconjuntos de atributos?

Abordagens

(i) Incorporados (*embedded)*: a seleção de atributos é intrínseca ao próprio método (e.g. C4.5, 1R).

(ii) Filtragem (*filters):* selecionar atributos de acordo com características dos dados que presumivelmente influenciam a eficácia do algoritmo de aprendizado. Independem do algoritmo de mineração a ser usado.

(iii) Empacotamento (*wrappers*): subconjunto de atributos selecionados é avaliado por meio do próprio algoritmo de aprendizado.

- Em geral fornecem melhores resultados do que a *filtragem*, mas são computacionalmente mais caros. Atributos selecionados podem não ser apropriados para algoritmos de aprendizado diferentes daquele usado para avaliar os subconjuntos de atributos .

(iv) Métodos Híbridos (*hybrid approaches*): procuram combinar as vantagens oferecidas pelos modelos (i)-(iii). Filtragem por correlação linear e modelagem não linear.

Exemplos de filtros

- Usar o critério do ganho de informação (árvores);

- Escore de Fisher (Duda & Hart, Pattern Classification and Scene Analysis, Wiley, 1973):

- Considerando um problema formado por duas classes, representadas aqui por (+,-), para cada atributo $j=1,...,n$ calcular:

$$w_j = \frac{(\mu_j^+ - \mu_j^-)^2}{(\sigma_j^+)^2 + (\sigma_j^-)^2}$$

- Presume-se que a qualidade de cada atributo (w_j) pode ser avaliada individualmente, sem levar em conta as interações entre atributos.

- Pode-se lidar com múltiplas classes de maneira análoga;
- Considerando cada classe *i* e atributo *j* temos:

$$\mu_{j,i} = \frac{1}{|C_i|} \sum_{x \in C_i} x_j$$

- A média total para *j* é definida como:

$$\mu_j = \frac{1}{m} \sum_{x} x_j$$

- Usando as duas equações acima pode-se definir a dispersão entre classes para o atributo *j* como:

$$B_j = \sum_{i=1}^{C} |C_i| (\mu_{j,i} - \mu_j)^2$$

- Em função de B_j podemos usar a seguinte função de escore:

$$B_{dispersão,j} = \frac{B_j}{\sum_{i=1}^{C} \sigma_{ji}}$$

Chai et al., An Evaluation of Gene Selection Methods for Multi-class Microarray Data Classification, Proc. European Workshop on Data Mining and Text Mining in Bioinformatics, 2004.

Exemplos de wrapper – Naive Bayes

- Atributos irrelevantes e redundantes podem comprometer acurácia de classificação;

- Selecionar atributos com base no desempenho do classificador NB. Pode-se sumarizar o NBW como segue:

1) Construir um classificador NB para cada atributo X_i (i = 1,..., n). Escolher X_i para o qual o NB apresenta a melhor acurácia e inseri-lo em A_S = {atributos selecionados};

2) Para todo $X_i \notin A_S$ construir um NB formado por $\{X_i\} \cup A_S$. Escolher o melhor classificador dentre os disponíveis e verificar se é melhor do que o obtido anteriormente:

a) SE sim, ENTÃO atualizar A_S, inserindo o atributo adicional e repetindo o passo 2;

b) SE não, ENTÃO parar e usar o classificador obtido anteriormente.

Complexidade do *wrapper*

- NB possui complexidade de tempo linear com o número de exemplos e de atributos;
- Constante de tempo do NB também é baixa (computar frequências relativas e/ou densidades);
- Algoritmo NB é facilmente paralelizável;
- O que dizer sobre o NBW?
  - Teoria: $O(2^n)$, onde n é o número de atributos;
  - Busca gulosa *poda* o espaço de busca do problema de otimização combinatória: $O(n + (n-1) + ... + 1) = O(n^2)$
  - Por exemplo, para n=100 temos: 1.2×10^{30} versus 10^4 avaliações de classificadores diferentes para escolher o melhor.

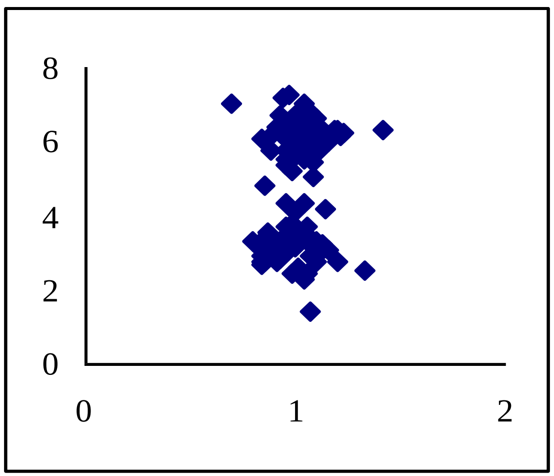
Comparando técnicas

- NÃO se pode selecionar atributos no conjunto completo de dados disponíveis e então rodar a validação cruzada apenas com os atributos selecionados (e.g., via filtros);
- Queremos estimar a capacidade de generalização do modelo: validação cruzada;
- Separar dados em conjuntos de *treinamento* e de *teste*;
- Executar <u>validação cruzada</u> no conjunto de treinamento pra selecionar atributos;
- A partir dos atributos selecionados, construir o classificador no conjunto de treinamento e avaliá-lo no conjunto de teste;
- Classificador que vai pra produção: construir com todos os dados disponíveis e parâmetros aprendidos na validação cruzada.

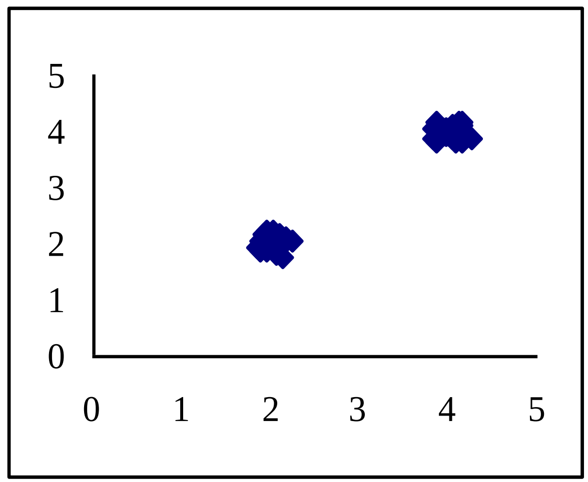
Reunanen, J., Overfitting in Making Comparisons Between Variable Selection Methods, Journal of Machine Learning Research (3), pp. 1371-1382, 2003.

Como selecionar atributos para *clustering*?

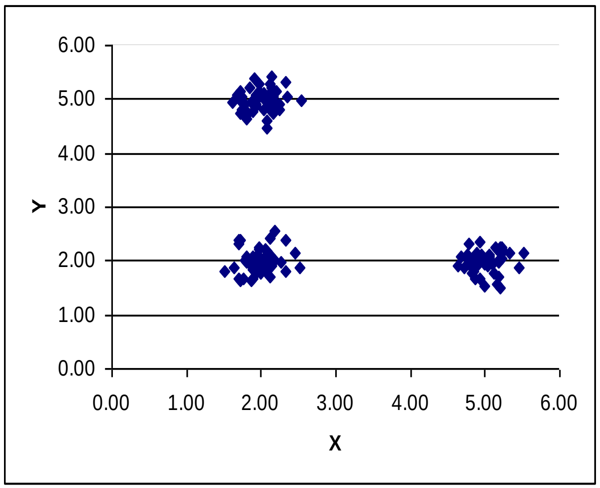
- Informação da classe não está disponível;
- Número de *clusters* e de atributos está intimamente relacionado;
- Problema se torna muito difícil quando *k* é desconhecido a priori;
- Vejamos alguns exemplos:



(a) “x” irrelevante.



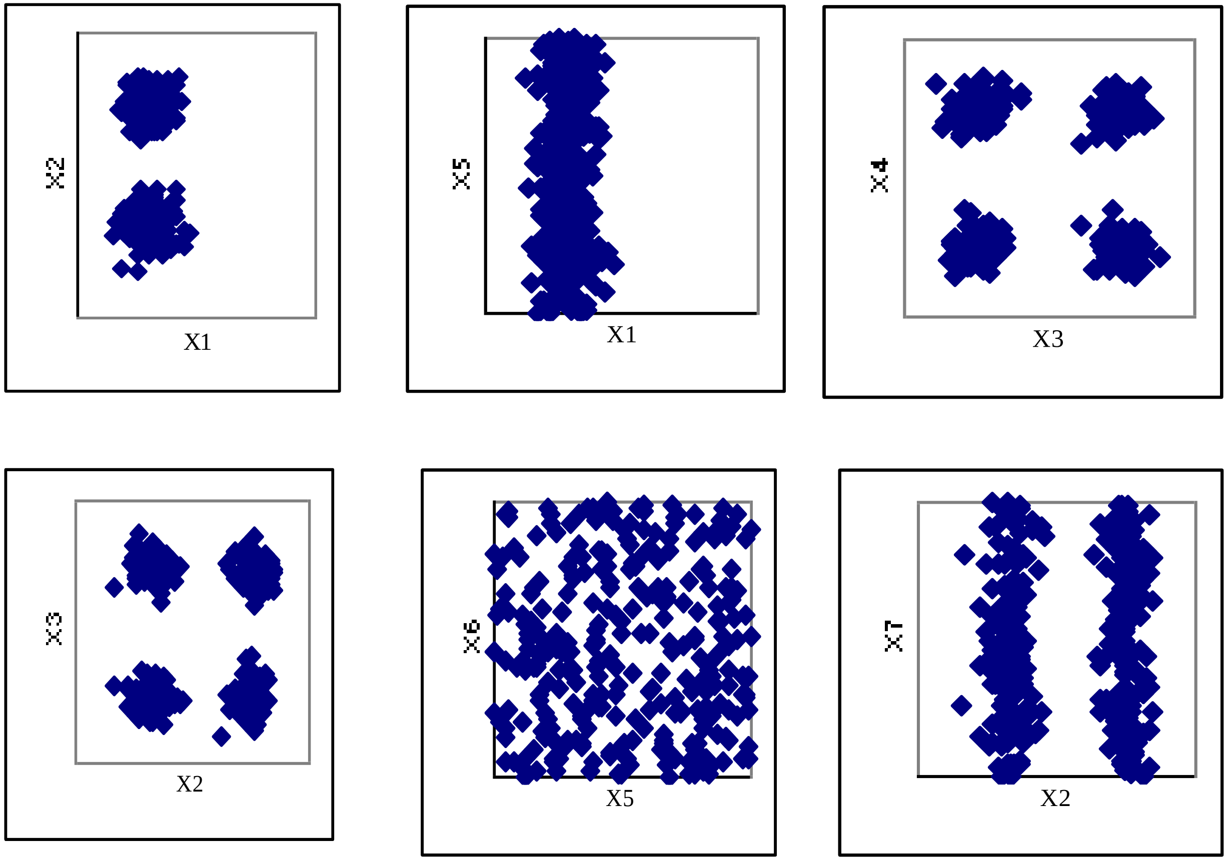
(b) Atributos redundantes.



(c) *k* depende dos selecionados.

→ O que pode acontecer em bases com mais do que dois atributos?

Consideremos 6 atributos (X1, X2,..., X6):



Quantos *clusters naturais* existem nesta base de dados?

Possíveis abordagens

- Filtros;
- Métodos baseados em empacotamento:
  - Como estabelecer critérios de validade?
  - Como comparar partições formadas por diferentes quantidades de grupos e de atributos selecionados?
  - Exemplo: combinar k-means com Naive Bayes.
- Métodos híbridos (empacotar + filtrar);
- Problema pouco estudado.

Agenda

- Pré-processamento de dados
- Preparação de dados para métodos não supervisionados
- Preparação de dados para métodos supervisionados
    - Filtros
    - Wrappers

Tocando em frente

- Aprofundar conhecimento em técnicas específicas
- Aprendizado semi-supervisionado
- Aprendizado ativo
- Aprendizado por reforço
- Fluxos de dados
- Redes complexas
- Modelos gráficos probabilísticos
- Deep learning
- Algoritmos evolutivos
- Sistemas de recomendação
- Processamento paralelo e distribuído
- Bancos de dados