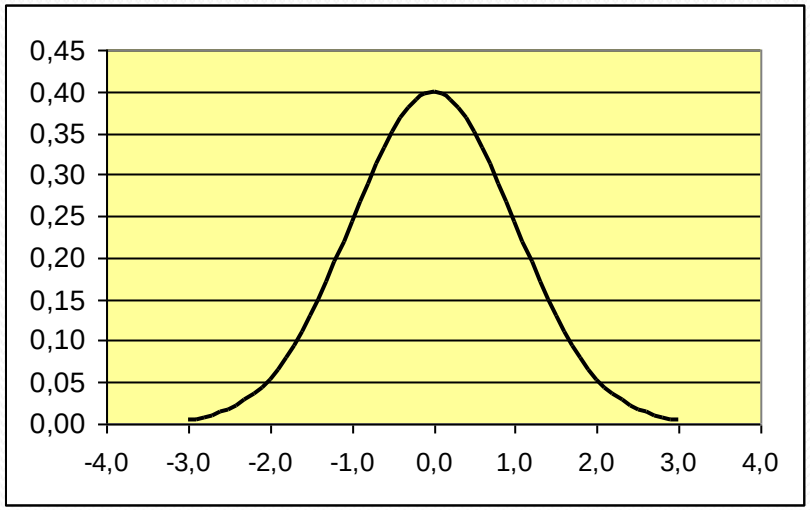
MODELAGEM E PREPARAÇÃO DE DADOS PARA APRENDIZADO DE MÁQUINA:

Redução de dimensionalidade e Seleção de atributos

Professor:

Luis E. Zárate

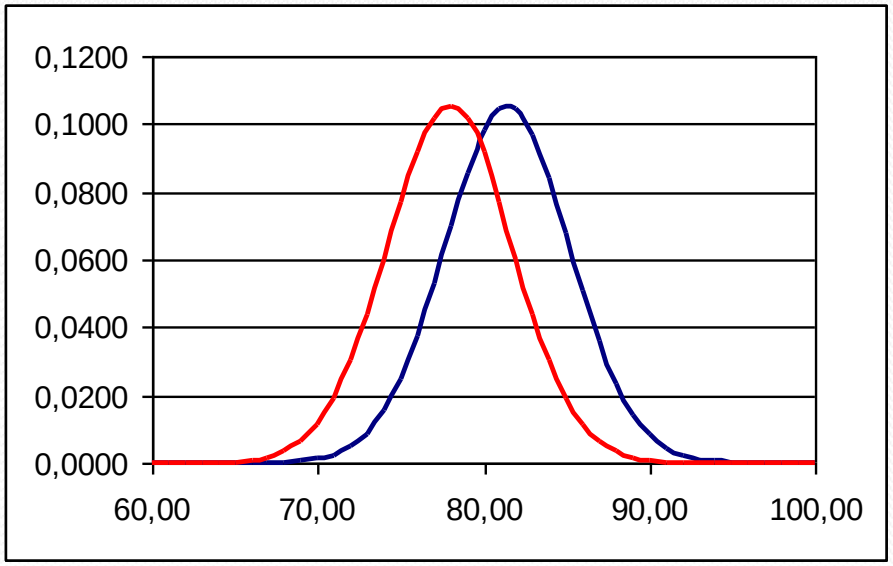
1) Eliminação pela Análise da Média e da Variância - atributo isolado



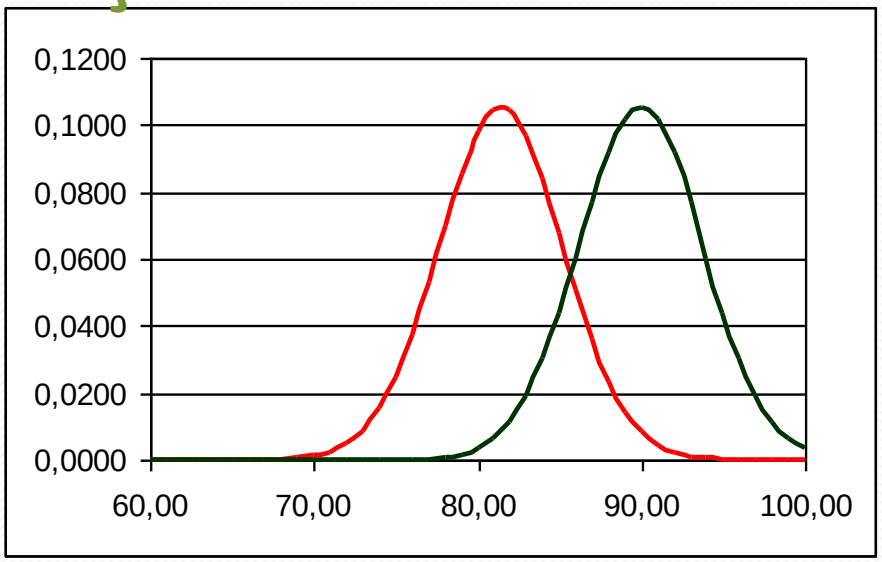
Caso os dados apresentem uma Distribuição Normal, a relevância dos atributos pode ser avaliada pela variância

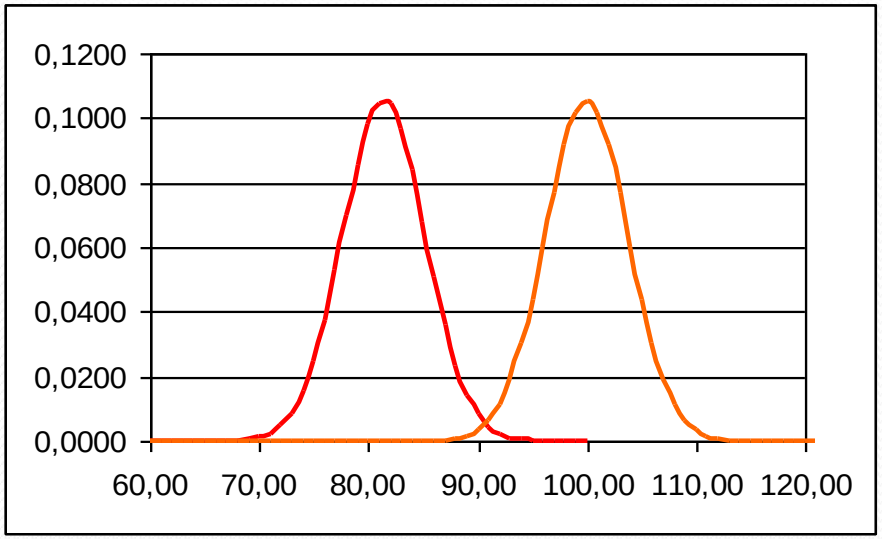
Idade	Peso (Kg)	Altura (cm)	Pescoco (cm)	Peito (cm)	Abdom (cm)
61,00	74,13	171,45	36	91,6	81,8
61,00	89,00	170,18	37,4	105,3	99,7
61,00	89,88	167,01	38,4	104,8	98,3
62,00	83,75	181,61	35,5	97,6	91,5
62,00	100,63	176,53	40,5	111,5	104,2
62,00	108,00	186,06	41,4	112,3	104,8
64,00	75,13	170,82	38,1	97,1	89
64,00	80,00	167,01	36,5	104,3	90,9
64,00	95,13	184,79	39,3	103,1	97,8
66,00	85,63	175,90	37,4	102,7	98,6
67,00	81,88	172,09	38,4	97,7	95,8
67,00	83,50	171,45	36,5	98,9	89,7
69,00	88,88	173,99	38,7	102	95
70,00	85,38	177,80	38,7	101,8	94,9
72,00	78,88	170,82	37,7	97,5	88,1
72,00	84,00	175,90	38,5	101,4	99,8
81,00	80,63	178,44	37,8	96,4	95,4
66,18	**86,14**	**174,81**	**38,05**	**101,53**	**95,02**
5,35	**8,74**	**5,62**	**1,51**	**5,30**	**5,99**

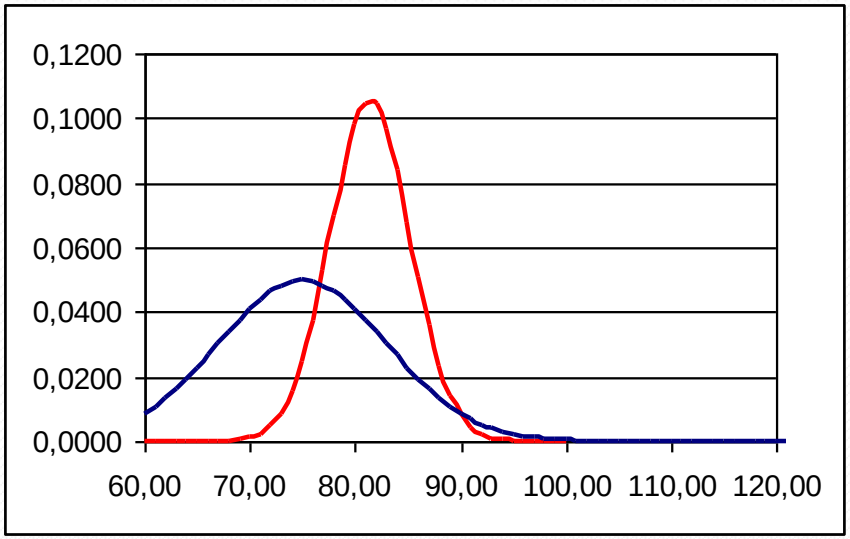
2) Eliminação pela Análise da Média e da Variância - atributo isolado para classificação

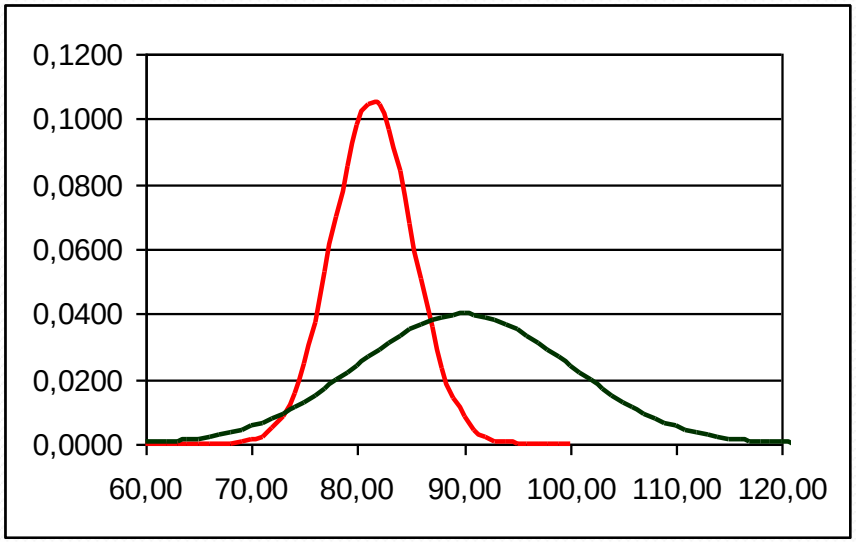


- As médias e a variância podem indicar se um atributo é relevante na distinção de duas classes ou não.









- Esta técnica deve ser aplicada com atenção, desde que pode correlacionar atributos às classes.

3) Eliminação pela Análise da Média e da Variância para atributo isolado para classificação

Considerando dois atributos X1 e X2, onde X2 representa uma classificação A ou B e X1 contendo n1 exemplos da classe A e n2 exemplos da classe B.

É possível determinar se o atributo X1 contribui para a distinção das classes. Ou seja se as médias estão próximas ou muito distantes.

$$\Delta(A-B) = \sqrt{\frac{Var(A)}{n_1} + \frac{Var(B)}{n_2}}$$

$$TESTE : \left| \frac{média(A) - média(B)}{\Delta(A-B)} \right| > Limiar$$

4) Eliminação pela Análise da Média e da Variância para dois atributos para Classificação

Considere dois atributos X1 e X2, onde X2 representa uma classificação: A ou B. X1 contêm n1 exemplos da classe A e n2 exemplos da classe B.

$$\Delta(A-B) = \sqrt{\frac{0{,}0233}{3} + \frac{0{,}0633}{3}}$$

$$\Delta(A-B) = 0{,}4677$$

X1	X2		
0,3	A	**Média_A**	0,4667
0,6	A	**Média_B**	0,4333
0,5	A	**DesPad_A**	0,1528
0,2	B	**DesPad_B**	0,2517
0,7	B	**Variância_A**	0,0233
0,4	B	**Variância_B**	0,0633

$$TESTE: \left|\frac{0{,}4667 - 0{,}4333}{0{,}4677}\right| > 0{,}5$$

$$\left|0{,}0714\right| < 0{,}5$$

X1 poderia ser eliminado, pois a distância entre as médias é menor que o Limiar

5) Eliminação pela Análise da Média e da Variância para Classificação N-dimensional

Considerando três atributos X1, X2 e X3, onde X3 representa uma classificação: A ou B. Observe que este método avalia cada atributo independente dos outros.

X1	X2	X3	Para Atributo X1		Para atributo X2	
0,3	0,7	A	**Média_A**	0,4667	**Média_A**	0,6000
0,2	0,9	B	**Média_B**	0,4333	**Média_B**	0,8333
0,6	0,6	A	**DesPad_A**	0,1528	**DesPad_A**	0,1000
0,5	0,5	A	**DesPad_B**	0,2517	**DesPad_B**	0,1155
0,7	0,7	B	**Variância_A**	0,0233	**Variância_A**	0,0100
0,4	0,9	B	**Variância_B**	0,0633	**Variância_B**	0,0133

$$\Delta(X1A - X1B) = \sqrt{\frac{0,0233}{3} + \frac{0,0633}{3}}$$

$$\Delta(X1A - X1B) = 0,4677$$

$$\Delta(X2A - X2B) = \sqrt{\frac{0,0100}{3} + \frac{0,0133}{3}}$$

$$\Delta(X2A - X2B) = 0,0875$$

$$TESTE: \left|\frac{0,4667 - 0,4333}{0,4677}\right| > 0,5$$

$$|0,0714| < 0,5$$

$$TESTE: \left|\frac{0,6 - 0,8333}{0,0875}\right| > 0,5$$

$$|2,6667| > 0,5$$

X1 poderia ser eliminado, pois a distância entre as médias é menor que o Limiar e X2 tem potencial para ser um atributo de distinção entre as duas classes.

Quando existem K classes, K comparações pairwise podem ser efetuadas comparando cada classe com seu complemento. Ao final, se uma característica é significativa em todas as comparações pairwise, esta deverá ser mantida.

Para 3 classes: A, B e C

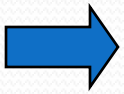
(A,B)	**(B,C)**	**(C,D)**	**…**	**(N-1,N)**
(A,C)	**(B,D)**	**…**	**…**	
(A,D)	**…**	**(C,N)**		
…	**(B,N)**			
(A,N)				

5) Eliminação pela Análise da Média e da Variância para Classificação N-dimensional

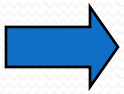
A técnica anterior testa os atributos separadamente. Se os atributos são examinados coletivamente, pode haver informação adicional de suas características.

Dada a base de dados de duas classes distintas

x1	x2	...	xN	Classe
...	...	...	...	A
...	...	...	...	B
...	...	...	...	B
...	...	...	...	A



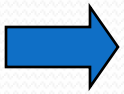
x1	x2	...	XN	Classe
...	...	...	...	A
...	...	...	...	A
...	...	...	...	A



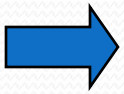
x1	x2	...	xN	Classe
...	...	...	...	B
...	...	...	...	B
...	...	...	...	B

Por exemplo:

X1	X2	Classe
0,3	0,7	A
0,2	0,9	B
0,6	0,6	A
0,5	0,5	A
0,7	0,7	B
0,4	0,9	B



x1	x2	Classe
0,3	0,7	A
0,6	0,6	A
0,5	0,5	A



x1	x2	Classe
0,2	0,9	B
0,7	0,7	B
0,4	0,9	B

Calcular as média de cada atributo para cada classe:

Médias da classe A =

m1	m2	m3	...	mN

Médias da classe B =

m1	m2	m3	...	mN

M A =

0,4667	0,6000

M B =

0,4333	0,8333

Calcular as matrizes de COVARIÂNCIAS:

$$C_{i,j} = \frac{1}{n}\sum_{k=1}^{n}[x(k,i) - m(i)][x(k,j) - m(j)]$$

Cov (A) =

0,0233	**-0,0100**
-0,0100	**0,0100**

Cov (B) =

0,0633	**-0,02670**
-0,0267	**0,0133**

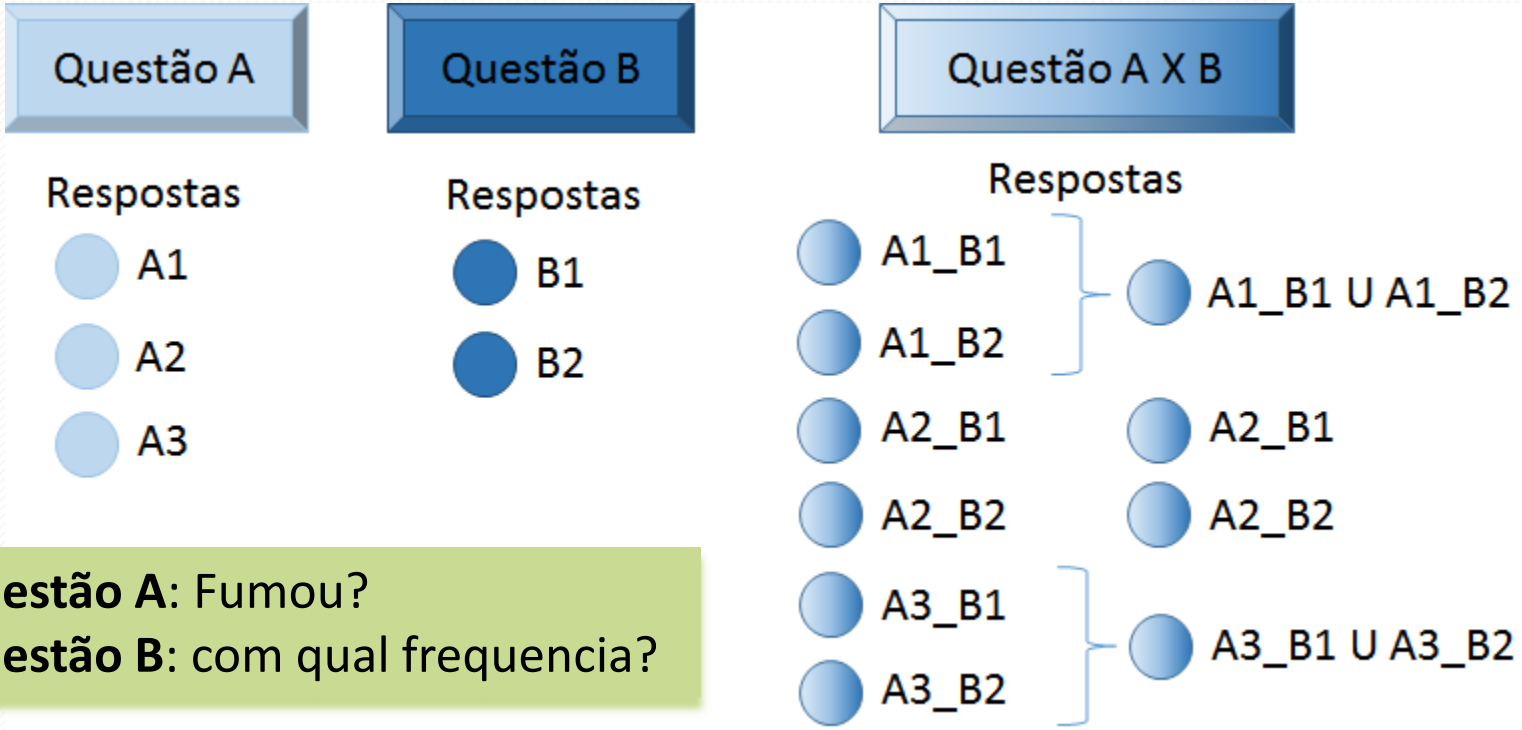
Uma medida heurística (HM) para filtrar características que separam duas classes é definida como:

$$HM = (M_A - M_B)(C_A + C_B)^{-1}(M_A - M_B)^T$$

Para o exemplo dado: **HM = 6,1551**

11) Redução da dimensionalidade pela Junção de Atributos

- Em dados nominais, a junção de questões altamente relacionadas pode trazer uma redução significativa na dimensionalidade da base, sem grandes perdas de informação.
- A junção pode ser um simples produto cartesiano das opções de resposta das questões, ou uma re-codificação deste, criando novas opções



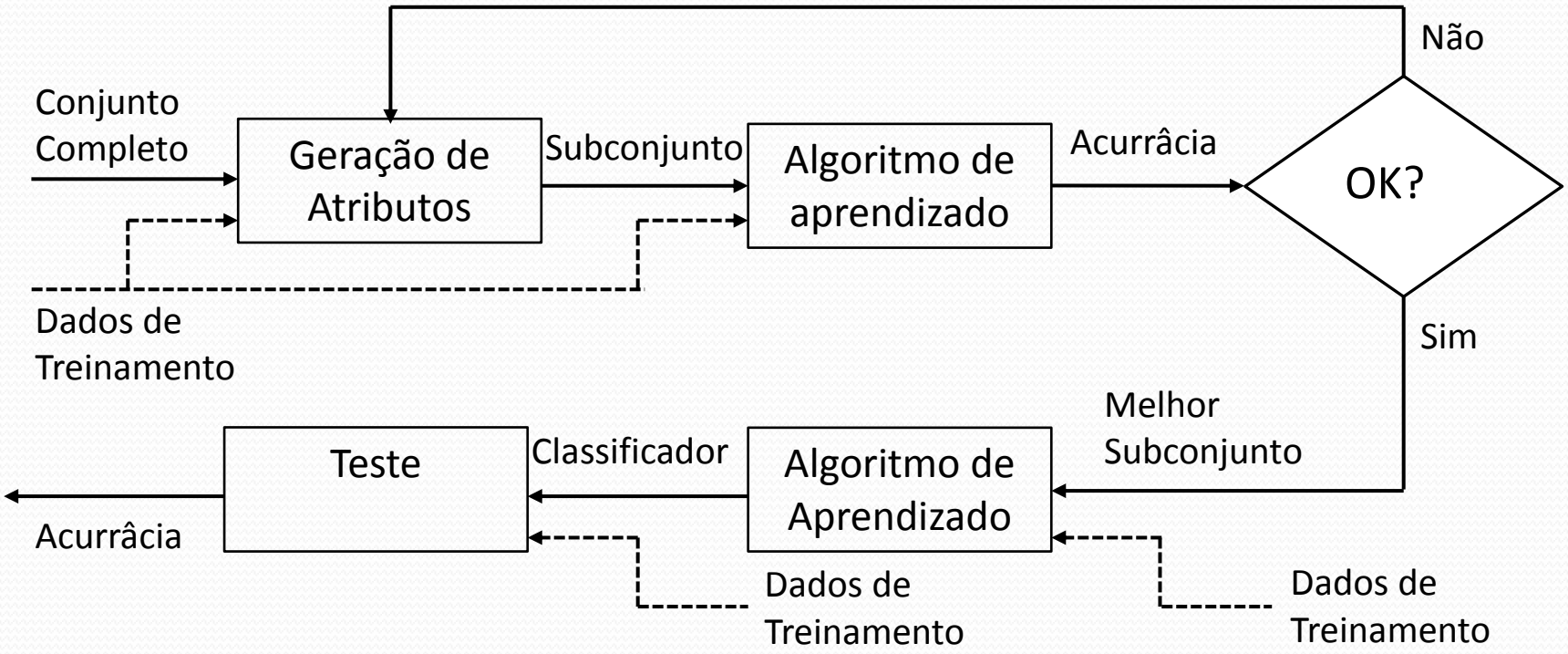
Questão A: Fumou?
Questão B: com qual frequencia?

13) Seleção de características via Filtro e Wrapper

Método baseado em filtros: Este procedimento de seleção de subconjunto de atributos é independente do algoritmo de aprendizado e é geralmente uma etapa do pré-processamento. Este é um processo mais rápido para o aprendizado, porém é possível que pelo critério utilizado pode resultar num subconjunto que pode não trabalhar muito bem no algoritmo de aprendizado.

Método baseado em Wrapper: A seleção do subconjunto de atributos é baseado no algoritmo de aprendizado utilizado para treinamento do subconjunto escolhido. O procedimento é comparar o modelo atual em relação ao(s) modelo(s) anterior(es).

Seleção de características via Modelo Wrapper



Obrigado

Professor:

Luis E. Zárate