Conceitos Adicionais e Métricas

- Nesta apresentação são descritos conceitos adicionais em AM, incluindo métricas que são obtidas a partir do conjunto de exemplos, do classificador induzido ou de regras individuais do classificado, caso ele seja simbólico

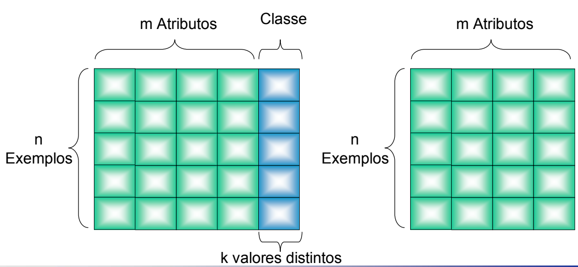
José Augusto Baranauskas
Departamento de Física e Matemática – FFCLRP-USP
augusto@usp.br
http://dfm.ffclrp.usp.br/~augusto

Métricas

- Algumas métricas são específicas de um conjunto particular de exemplos (ou seja, são independentes do classificador induzido)
  - Distribuição de classes
    - Classe minoritária
    - Classe majoritária
  - Prevalência de classe
  - Erro majoritário
- Outras métricas dependem tanto do conjunto de exemplos como do classificador induzido
  - Taxa de erro, precisão, ...
- Há ainda métricas específicas para regras, caso o classificador induzido seja simbólico

Conjunto de Exemplos (*Dataset*)

- No Aprendizado Supervisionado, cada exemplo é rotulado segundo sua classe
- No Aprendizado Não Supervisionado, cada exemplo não possui classe associada



Formato Padrão

	X_1	X_2	...	X_m	Y
z_1	x_{11}	x_{12}	...	x_{1m}	y_1
z_2	x_{21}	x_{22}	...	x_{2m}	y_2
...	...	...	...	...	...
z_n	x_{n1}	x_{n2}	...	x_{nm}	y_n

Formato Padrão

Conjunto de Exemplos (Dataset)

	X_1	X_2	...	X_m	Y
z_1	x_{11}	x_{12}	...	x_{1m}	y_1
z_2	x_{21}	x_{22}	...	x_{2m}	y_2
...	...	...	...	...	...
z_n	x_{n1}	x_{n2}	...	x_{nm}	y_n

Formato Padrão

Exemplo

	X_1	X_2	...	X_m	Y
z_1	x_{11}	x_{12}	...	x_{1m}	y_1
z_2	x_{21}	x_{22}	...	x_{2m}	y_2
...	...	...	...	...	...
z_n	x_{n1}	x_{n2}	...	x_{nm}	y_n

Formato Padrão

Atributo

	X_1	X_2	…	X_m	Y
z_1	x_{11}	x_{12}	…	x_{1m}	y_1
z_2	x_{21}	x_{22}	…	x_{2m}	y_2
…	…	…	…	…	…
z_n	x_{n1}	x_{n2}	…	x_{nm}	y_n

Formato Padrão

Classe

	X_1	X_2	…	X_m	Y
z_1	x_{11}	x_{12}	…	x_{1m}	y_1
z_2	x_{21}	x_{22}	…	x_{2m}	y_2
…	…	…	…	…	…
z_n	x_{n1}	x_{n2}	…	x_{nm}	y_n

Formato Padrão

- No formato padrão um conjunto **T** com **n** exemplos e **m** atributos a linha **i** refere-se ao **i**-ésimo exemplo (**i** = 1, 2, ..., **n**) e a entrada x_{ij} refere-se ao valor do **j**-ésimo (**j** = 1, 2, ..., **m**) atributo X_j do exemplo **i**

$$
\begin{array}{ccccccc}
\vec{z}_1 & = & (\vec{x}_1, y_1) & = & (\mathbf{x}_1, y_1) & = & (x_{11}, x_{12}, x_{13}, \cdots, x_{1m}, y_1) \\
\vec{z}_2 & = & (\vec{x}_2, y_2) & = & (\mathbf{x}_2, y_2) & = & (x_{21}, x_{22}, x_{23}, \cdots, x_{2m}, y_2) \\
\vdots & = & \vdots & = & \vdots & = & \vdots \\
\vec{z}_i & = & (\vec{x}_i, y_i) & = & (\mathbf{x}_i, y_i) & = & (x_{i1}, x_{i2}, x_{i3}, \cdots, x_{im}, y_i) \\
\vdots & = & \vdots & = & \vdots & = & \vdots \\
\vec{z}_n & = & (\vec{x}_n, y_n) & = & (\mathbf{x}_n, y_n) & = & (x_{n1}, x_{n2}, x_{n3}, \cdots, x_{nm}, y_n)
\end{array}
$$

Formato Padrão

- Como pode ser notado, exemplos são tuplas
  - $\vec{z}_i = (\vec{x}_i, y_i) = (x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$
  - também denotados por $z_i = (x_i, y_i)$ onde fica subentendido o fato que tanto z_i como x_i são vetores
- A última coluna, $y_i = f(x_i)$, é a função que tenta-se predizer a partir dos atributos
- Observa-se que cada x_i é um elemento do conjunto $dom(X_1) \times dom(X_2) \times \ldots \times dom(X_m)$, onde $dom(X_j)$ é o domínio do atributo X_j e
- A última coluna
  - y_i pertence a uma das k classes, isto é, $y_i \in \{C_1, C_2, \ldots, C_k\}$ no caso de classificação
  - y_i é um número real ($y_i \in \Re$) no caso de regressão

Exemplo de um Dataset para Classificação

- Dez exemplos (n=10)
- Rótulo Classe (discreto)
- Três classes (k=3):
  - C_1=amigo; C_2=chato; C_3=inimigo
- Três atributos (m=3):
  - Cabeça (nominal)
    - dom(Cabeça) = {redonda, triangular, quadrada}
  - Peso (contínuo)
    - dom(Peso) = $\{\forall w : w \in \Re^+\}$
  - Sorri (nominal)
    - dom(Sorri) = {sim, não}
- Atributo dependente (classe) é *categórico*

Cabeça	Peso	Sorri	Classe
redonda	10.0	não	amigo
triangular	12.0	sim	amigo
redonda	5.6	sim	amigo
quadrada	11.0	não	chato
quadrada	10.0	sim	amigo
triangular	5.5	não	inimigo
redonda	5.7	sim	chato
quadrada	15.3	sim	chato
quadrada	10.2	sim	amigo
redonda	5.0	não	inimigo

Exemplo de um Dataset para Regressão

- Dez exemplos (n=10)
- Rótulo Preço (contínuo)
  - dom(Preço) = $\{\forall p : p \in \Re^+\}$
- Três atributos (m=3):
  - Cabeça (nominal)
    - dom(Cabeça) = {redonda, triangular, quadrada}
  - Peso (contínuo)
    - dom(Peso) = $\{\forall w : w \in \Re^+\}$
  - Sorri (nominal)
    - dom(Sorri) = {sim, não}
- Atributo dependente (preço) é *numérico*

Cabeça	Peso	Sorri	Preço
redonda	10.0	não	3500
triangular	12.0	sim	2500
redonda	5.6	sim	10000
quadrada	11.0	não	3000
quadrada	10.0	sim	4000
triangular	5.5	não	12000
redonda	5.7	sim	15000
quadrada	15.3	sim	5000
quadrada	10.2	sim	7000
redonda	5.0	não	11000

Distribuição de Classes

- Em classificação, dado um conjunto T com n exemplos é possível calcular sua **distribuição de classes**
- Para cada classe C_j em T sua distribuição distr(C_j) é calculada como sendo o número de exemplos em T que possuem classe C_j dividido pelo número total de exemplos n, ou seja, a proporção de exemplos em cada classe, dada por:

$$\mathrm{distr}(C_j) = \frac{1}{n}\sum_{i=1}^{n} \| y_i = C_j \|$$

- A classe com a maior distribuição de exemplos é denominada **majoritária** ou **prevalente**
- A classe com a menor distribuição de exemplos é denominada **minoritária**

Distribuição de Classes

- Dez exemplos (n=10)
- Três classes (k=3):
  - C_1=amigo; C_2=chato; C_3=inimigo
- Distribuição de classes:
  - distr(amigo) = 5/10 = 50%
  - distr(chato) = 3/10 = 30%
  - distr(inimigo) = 2/10 = 20%
- Classe amigo é a classe majoritária
- Classe inimigo é a classe minoritária

Cabeça	Peso	Sorri	Classe
redonda	10.0	não	amigo
triangular	12.0	sim	amigo
redonda	5.6	sim	amigo
quadrada	11.0	não	chato
quadrada	10.0	sim	amigo
triangular	5.5	não	inimigo
redonda	5.7	sim	chato
quadrada	15.3	sim	chato
quadrada	10.2	sim	amigo
redonda	5.0	não	inimigo

Distribuição de Classes

- Como outro exemplo, assuma um conjunto com 100 exemplos possui 60 exemplos da classe C_1, 15 exemplos da classe C_2 e 25 exemplos da classe C_3 então sua distribuição de classes é
  - distr(C_1,C_2,C_3) = (0.60, 0.15, 0.25) =
  - = (60.00%, 15.00%, 25.00%)
- A classe C_1 é a classe majoritária
- A classe C_2 é a classe minoritária

Erro Majoritário

- O **erro majoritário** de um conjunto T é definido como 1 menos a distribuição da classe majoritária, ou seja:

$$\text{maj-err}(T) = 1 - \underset{i=1,\cdots,k}{máx}\{\mathrm{distr}(\mathrm{C_i})\}$$

- O erro majoritário de um conjunto de exemplos é ***independente*** do algoritmo de aprendizado
- Ele fornece um limiar máximo abaixo do qual o erro de um classificador deve ficar

Erro Majoritário

- Dez exemplos (n=10)
- Três classes (k=3):
  - C_1=amigo; C_2=chato; C_3=inimigo
- Distribuição de classes:
  - distr(amigo) = 5/10 = 50%
  - distr(chato) = 3/10 = 30%
  - distr(inimigo) = 2/10 = 20%
- Classe amigo é a classe majoritária
- Classe inimigo é a classe minoritária
- Erro majoritário = 1-5/10 = 50%

Cabeça	Peso	Sorri	Classe
redonda	10.0	não	amigo
triangular	12.0	sim	amigo
redonda	5.6	sim	amigo
quadrada	11.0	não	chato
quadrada	10.0	sim	amigo
triangular	5.5	não	inimigo
redonda	5.7	sim	chato
quadrada	15.3	sim	chato
quadrada	10.2	sim	amigo
redonda	5.0	não	inimigo

Erro Majoritário

- Considerando novamente o exemplo em que
  - distr(C_1,C_2,C_3) = (0.60, 0.15, 0.25) =
  - = (60.00%, 15.00%, 25.00%)
- Neste caso, o erro majoritário é
  - maj-err(T) = 1 – 0.60 = 40.00%

Exercício:
Indique a distribuição de classes, as classes majoritária e minoritária e o erro majoritário

Nome	Dá luz	Põe ovos	Voa	Vive na água	Tem pernas	Classe
humano	sim	não	não	não	sim	mamíferos
píton	não	sim	não	não	não	répteis
salmão	não	sim	não	sim	não	peixes
baleia	sim	não	não	sim	não	mamíferos
sapo	não	sim	não	eventual/e	sim	anfíbios
komodo	não	sim	não	não	sim	répteis
morcego	sim	não	sim	não	sim	mamíferos
pombo	não	sim	sim	não	sim	pássaros
gato	sim	não	não	não	sim	mamíferos
tubarão	sim	não	não	sim	não	peixes
tartaruga	não	sim	não	eventual/e	sim	répteis
pingüim	não	sim	não	eventual/e	sim	pássaros
porco-espinho	sim	não	não	não	sim	mamíferos
enguia	não	sim	não	sim	não	peixes
salamandra	não	sim	não	eventual/e	sim	anfíbios
monstro gila	não	sim	não	não	sim	répteis
ornitorrinco	não	sim	não	não	sim	mamíferos
coruja	não	sim	sim	não	sim	pássaros
golfinho	sim	não	não	sim	não	mamíferos
águia	não	sim	sim	não	sim	pássaros

- Distribuição de classes
  - distr(mamíferos) = 7/20 = 35.00%
  - distr(répteis) = 4/20 = 20.00%
  - distr(peixes) = 3/20 = 15.00%
  - distr(anfíbios) = 2/20 = 10.00%
  - distr(pássaros) = 4/20 = 20.00%
- Classe majoritária:
  - mamíferos
- Classe minoritária:
  - anfíbios
- Maj-err
  - 1 – 0.35 = 65.00%

Prevalência de Classe

- Um ponto muito importante em AM refere-se ao desbalanceamento de classes em um conjunto de exemplos
- Por exemplo, suponha um conjunto de exemplos T com a seguinte distribuição de classes dist(C_1, C_2, C_3) = (99.00%, 0.25%, 0.75%), com **prevalência da classe** C_1
- Um classificador simples que classifique sempre novos exemplos como pertencentes à classe majoritária C_1 teria uma precisão de 99.00% (maj-err(T) = 1.00%)
- Isto pode ser indesejável quando as classes minoritárias são aquelas que possuem uma informação muito importante, por exemplo, supondo C_1: paciente normal, C_2: paciente com doença A e C_3: paciente com doença B

Prevalência de Classe

- É importante estar ciente, quando se trabalha com conjuntos de exemplos desbalanceados, que é desejável utilizar uma medida de desempenho diferente da precisão
- Isto deve-se ao fato que a maioria dos sistemas de aprendizado é projetada para otimizar a precisão
- Com isso, normalmente os algoritmos apresentam um desempenho ruim se o conjunto de treinamento encontra-se fortemente desbalanceado, pois os classificadores induzidos tendem a ser altamente precisos nos exemplos da classe majoritária, mas freqüentemente classificam incorretamente exemplos das classes minoritárias
- Algumas técnicas foram desenvolvidas para lidar com esse problema, tais como a introdução de custos de classificação incorreta (explicada mais adiante), a remoção de exemplos redundantes ou prejudiciais ou ainda a detecção de exemplos de borda e com ruído

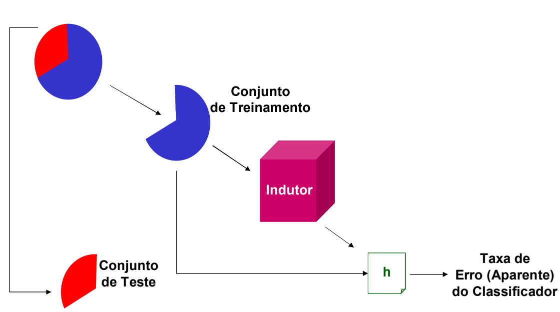
Conjuntos de Treinamento e Teste

- Usualmente, um conjunto de exemplos é dividido em dois subconjuntos disjuntos:
  - **conjunto de treinamento** que é usado para o aprendizado do conceito e o
  - **conjunto de teste** usado para medir o grau de efetividade do conceito aprendido
- Os subconjuntos são disjuntos para assegurar que as medidas obtidas utilizando o conjunto de teste sejam de um conjunto diferente do usado para realizar o aprendizado, tornando a medida estatisticamente válida

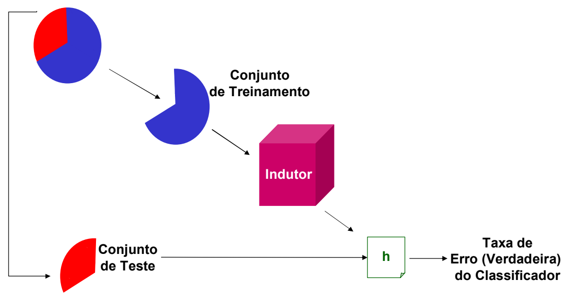
Conjuntos de Treinamento e Teste

- Após induzir uma hipótese, é possível avaliá-la no conjunto de treinamento bem como no conjunto de teste
- É usual denominar as medidas de desempenho de um classificador efetuadas sobre o conjunto de treinamento como **aparentes** (também conhecidas como medidas de **resubstituição**) e as medidas efetuadas sobre o conjunto de teste como medidas **reais** (ou **verdadeiras**)
- Por exemplo, caso a medida seja o **erro**, pode-se ter o **erro aparente** e o **erro verdadeiro**
- Para a maioria das hipóteses, a medida aparente é um estimador ruim do seu desempenho futuro, uma vez que ela tem a tendência de possuir um *bias* otimista
  - Em geral, o erro calculado sobre o conjunto de exemplos de treinamento (erro aparente) é menor que o erro calculado sobre o conjunto de exemplos de teste (erro verdadeiro)

Erro Aparente



Erro Verdadeiro



Erro e Precisão

- Principais fatores de erro:
  - Qualidade (representatividade) da informação dos atributos
  - Adaptação do algoritmo de aprendizado aos exemplos
  - Distribuição dos exemplos futuros
  - Quantidade de exemplos

Erro e Precisão

- Recordando a notação adotada
  - Exemplo: par (x,y) = (x,f(x))
  - f(·) é desconhecida
  - Atributos: x
  - Classe (rotulada ou atribuída pelo processo no mundo real): y = f(x)
  - Classificador (hipótese): h(·)
  - Classe do exemplo x (atribuída pelo classificador): ŷ = h(x)
  - *n* é o número de exemplos

Erro e Precisão: Classificação

- Classificação
  - err(h) = taxa de erro (*error rate*) da hipótese h
  - acc(h) = precisão (*accuracy*) da hipótese h

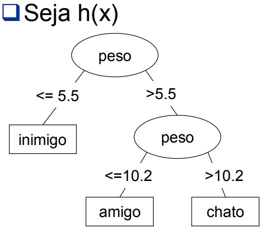
$$err(h) = \frac{1}{n}\sum_{i=1}^{n} \| y_i \neq h(x_i) \| = \frac{1}{n}\sum_{i=1}^{n} \| y_i \neq \hat{y}_i \|$$

$$acc(h) = \frac{1}{n}\sum_{i=1}^{n} \| y_i = h(x_i) \| = \frac{1}{n}\sum_{i=1}^{n} \| y_i = \hat{y}_i \| = 1 - err(h)$$

- O operador || *E* || retorna:
  - 1 se *E* é verdadeiro
  - 0 se *E* é falso
- É óbvio que:
  - 0 ≤ err(h) ≤ 1 (ou, equivalentemente, 0% ≤ err(h) ≤ 100%)
  - 0 ≤ acc(h) ≤ 1 (ou, equivalentemente, 0% ≤ acc(h) ≤ 100%)

Erro e Precisão: Classificação

- Seja h(x)



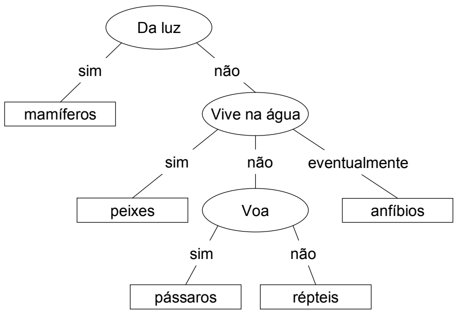
Cabeça X_1	Peso X_2	Sorri X_3	Classe Y=f(x)	Predita Ŷ=h(x)
redonda	10.0	não	amigo	amigo
triangular	12.0	sim	amigo	chato
redonda	5.6	sim	amigo	amigo
quadrada	11.0	não	chato	chato
quadrada	10.0	sim	amigo	amigo
triangular	5.5	não	inimigo	inimigo
redonda	5.7	sim	chato	amigo
quadrada	15.3	sim	chato	chato
quadrada	10.2	sim	amigo	amigo
redonda	5.0	não	inimigo	inimigo

- err(h) = 2/10 = 20%
- acc(h) = 1-2/10 = 80%

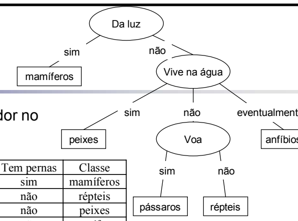
Exercício

Nome	Dá a luz	Põe ovos	Voa	Vive na água	Tem Pernas	Classe
humano	sim	não	não	não	sim	mamíferos
píton	não	sim	não	não	não	répteis
salmão	não	sim	não	sim	não	peixes
baleia	sim	não	não	sim	não	mamíferos
sapo	não	sim	não	eventualmente	sim	anfíbios
dragão de komodo	não	sim	não	não	sim	répteis
morcego	sim	não	sim	não	sim	mamíferos
pombo	não	sim	sim	não	sim	pássaros
gato	sim	não	não	não	sim	mamíferos
tubarão	sim	não	não	sim	não	peixes
tartaruga	não	sim	não	eventualmente	sim	répteis
pinguim	não	sim	não	eventualmente	sim	pássaros
porco-espinho	sim	não	não	não	sim	mamíferos
enguia	não	sim	não	sim	não	peixes
salamandra	não	sim	não	eventualmente	sim	anfíbios
monstro gila	não	sim	não	não	sim	répteis
ornitorrinco	não	sim	não	não	sim	mamíferos
coruja	não	sim	sim	não	sim	pássaros
golfinho	sim	não	não	sim	não	mamíferos
águia	não	sim	sim	não	sim	pássaros

Exercício



Exercício



Calcule o erro e a precisão do classificador no conjunto de exemplos abaixo

Nome	Dá luz	Põe ovos	Voa	Vive na água	Tem pernas	Classe
humano	sim	não	não	não	sim	mamíferos
píton	não	sim	não	não	não	répteis
salmão	não	sim	não	sim	não	peixes
baleia	sim	não	não	sim	não	mamíferos
sapo	não	sim	não	eventual/e	sim	anfíbios
komodo	não	sim	não	não	sim	répteis
morcego	sim	não	sim	não	sim	mamíferos
pombo	não	sim	sim	não	sim	pássaros
gato	sim	não	não	não	sim	mamíferos
tubarão	sim	não	não	sim	não	peixes
tartaruga	não	sim	não	eventual/e	sim	répteis
pingüim	não	sim	não	eventual/e	sim	pássaros
porco-espinho	sim	não	não	não	sim	mamíferos
enguia	não	sim	não	sim	não	peixes
salamandra	não	sim	não	eventual/e	sim	anfíbios
monstro gila	não	sim	não	não	sim	répteis
ornitorrinco	não	sim	não	não	sim	mamíferos
coruja	não	sim	sim	não	sim	pássaros
golfinho	sim	não	não	sim	não	mamíferos
águia	não	sim	sim	não	sim	pássaros

$$err(h) = \frac{1}{n}\sum_{i=1}^{n} \| y_i \neq h(x_i) \|$$

$$acc(h) = 1 - err(h)$$

Solução

$$err(h) = \frac{4}{20} = 20.00\%$$

$$acc(h) = \frac{16}{20} = 80.00\%$$

Nome	Dá luz	Põe ovos	Voa	Vive na água	Tem Pernas	Classe	Ŷ=h(x)
humano	sim	não	não	não	sim	mamíferos	**mamíferos**
píton	não	sim	não	não	não	répteis	**répteis**
salmão	não	sim	não	sim	não	peixes	**peixes**
baleia	sim	não	não	sim	não	mamíferos	**mamíferos**
sapo	não	sim	não	eventualmente	sim	anfíbios	**anfíbios**
dragão de komodo	não	sim	não	não	sim	répteis	**répteis**
morcego	sim	não	sim	não	sim	mamíferos	**mamíferos**
pombo	não	sim	sim	não	sim	pássaros	**pássaros**
gato	sim	não	não	não	sim	mamíferos	**mamíferos**
tubarão	sim	não	não	sim	não	peixes	**mamíferos**
tartaruga	não	sim	não	eventualmente	sim	répteis	**anfíbios**
pingüim	não	sim	não	eventualmente	sim	pássaros	**anfíbios**
porco-espinho	sim	não	não	não	sim	mamíferos	**mamíferos**
enguia	não	sim	não	sim	não	peixes	**peixes**
salamandra	não	sim	não	eventualmente	sim	anfíbios	**anfíbios**
monstro gila	não	sim	não	não	sim	répteis	**répteis**
ornitorrinco	não	sim	não	não	sim	mamíferos	**répteis**
coruja	não	sim	sim	não	sim	pássaros	**pássaros**
golfinho	sim	não	não	sim	não	mamíferos	**mamíferos**
águia	não	sim	sim	não	sim	pássaros	**pássaros**

Erro: Regressão

- mse-err(h) = erro médio quadrático (*mean squared error*)

$$\text{mse-err}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2$$

- rmse-err(h) = raiz do erro médio quadrático (*root mean squared error*)

$$\text{rmse-err}(h) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2}$$

- mad-err(h) = distância/erro absoluta(o) média(o) (*mean absolute distance/error*)

$$\text{mad-err}(h) = \frac{1}{n}\sum_{i=1}^{n}|y_i - h(x_i)|$$

Erro: Regressão

- rse-err(h) = erro relativo quadrático (*relative squared error*)

$$\text{rse-err}(h) = \frac{\sum_{i=1}^{n}(y_i - h(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}; \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- rrse-err(h) = raiz do erro relativo quadrático (*root relative squared error*)

$$\text{rrse-err}(h) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - h(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- rae-err(h) = erro absoluto relativo (*relative absolute error*)

$$\text{rae-err}(h) = \frac{\sum_{i=1}^{n}|y_i - h(x_i)|}{\sum_{i=1}^{n}|y_i - \bar{y}|}$$

- Coeficiente de correlação

$$\text{corr}(h) = \frac{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(h(x_i) - \bar{h})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \times \dfrac{\sum_{i=1}^{n}(h(x_i) - \bar{h})^2}{n-1}}}$$

Exemplo: *Xiphopenaeus kroyeri*

- Coleta de camarões “sete-barbas” na baía da Ubatuba
- Previsão para a população de um mês baseado na população coletada em meses anteriores
  - mse-err(h) = 379.66
  - mad-err(h) = 19.00
  - rmse-err(h) = 19.46
  - rse-err(h) = 0.3785
  - rrse-err(h) = 0.6152
  - rae-err(h) = 0.6196

Valor Correto (y)	Valor Predito h(x)
486	503
600	575
535	520

Matriz de Confusão

- A **matriz de confusão** de uma hipótese h oferece uma medida efetiva do modelo de classificação, ao mostrar o número de classificações corretas *versus* as classificações preditas para cada classe, sobre um conjunto de exemplos T
- As entradas da matriz são representadas por $M(C_i,C_j)$ indicando o número de exemplos de T que são da classe C_i mas que foram classificados pela hipótese h como sendo da classe C_j

$$M(C_i,C_j) = \sum_{\{\forall(x,y)\in T:\, y=C_i\}} \| h(x) = C_j \|$$

Matriz de Confusão

- O número de acertos, para cada classe, se localiza na diagonal principal $M(C_i,C_i)$ da matriz
- Os demais elementos $M(C_i,C_j)$, para $i \neq j$, representam erros na classificação
- A matriz de confusão de um classificador ideal possui todos os elementos fora da diagonal iguais a zero uma vez que ele não comete erros

$$\begin{array}{c|cccc}
\text{Classe} & \text{predita } C_1 & \text{predita } C_2 & \cdots & \text{predita } C_k \\
\hline
\text{verdadeira } C_1 & M(C_1,C_1) & M(C_1,C_2) & \cdots & M(C_1,C_k) \\
\text{verdadeira } C_2 & M(C_2,C_1) & M(C_2,C_2) & \cdots & M(C_2,C_k) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{verdadeira } C_k & M(C_k,C_1) & M(C_k,C_2) & \cdots & M(C_k,C_k)
\end{array}$$

Matriz de Confusão

- h: if X_1= a and X_2 = s then classe = + else classe = -

	Atributos				
Exemplo	X_1	X_2	X_3	Classe (Y)	h
z_1	a	s	2	+	+
z_2	a	s	1	-	+
z_3	b	n	1	+	-
z_4	b	s	2	-	-
z_5	c	n	2	+	-

Classe	Predita +	Predita -
Verdadeira +	1	2
Verdadeira -	1	1

Exercício: Qual a matriz de confusão de *h*?

Nome	Dá luz	Põe ovos	Voa	Vive na água	Tem Pernas	Classe	*h*
humano	sim	não	não	não	sim	mamíferos	**mamíferos**
píton	não	sim	não	não	não	répteis	**répteis**
salmão	não	sim	não	sim	não	peixes	**peixes**
baleia	sim	não	não	sim	não	mamíferos	**mamíferos**
sapo	não	sim	não	eventualmente	sim	anfíbios	**anfíbios**
dragão de komodo	não	sim	não	não	sim	répteis	**répteis**
morcego	sim	não	sim	não	sim	mamíferos	**mamíferos**
pombo	não	sim	sim	não	sim	pássaros	**pássaros**
gato	sim	não	não	não	sim	mamíferos	**mamíferos**
tubarão	sim	não	não	sim	não	peixes	**mamíferos**
tartaruga	não	sim	não	eventualmente	sim	répteis	**anfíbios**
pingüim	não	sim	não	eventualmente	sim	pássaros	**anfíbios**
porco-espinho	sim	não	não	não	sim	mamíferos	**mamíferos**
enguia	não	sim	não	sim	não	peixes	**peixes**
salamandra	não	sim	não	eventualmente	sim	anfíbios	**anfíbios**
monstro gila	não	sim	não	não	sim	répteis	**répteis**
ornitorrinco	não	sim	não	não	sim	mamíferos	**répteis**
coruja	não	sim	sim	não	sim	pássaros	**pássaros**
golfinho	sim	não	não	sim	não	mamíferos	**mamíferos**
águia	não	sim	sim	não	sim	pássaros	**pássaros**

Solução

Classe	*h*
mamíferos	**mamíferos**
répteis	**répteis**
peixes	**peixes**
mamíferos	**mamíferos**
anfíbios	**anfíbios**
répteis	**répteis**
mamíferos	**mamíferos**
pássaros	**pássaros**
mamíferos	**mamíferos**
peixes	**mamíferos**
répteis	**anfíbios**
pássaros	**anfíbios**
mamíferos	**mamíferos**
peixes	**peixes**
anfíbios	**anfíbios**
répteis	**répteis**
mamíferos	**répteis**
pássaros	**pássaros**
mamíferos	**mamíferos**
pássaros	**pássaros**

		Classe Predita por *h*				
		anfíbios	peixes	répteis	pássaros	mamíferos
Classe Verdadeira	anfíbios	2	0	0	0	0
	peixes	0	2	0	0	1
	répteis	1	0	3	0	0
	pássaros	1	0	0	3	0
	mamíferos	0	0	1	0	6

Matriz de Confusão

		Classe predita por h				
		C_1	C_2	...	C_k	$M(C_i,*) = \sum_{j=1}^{k} M(C_i,C_j)$
Classe Verdadeira	C_1	$M(C_1,C_1)$	$M(C_1,C_2)$	...	$M(C_1,C_k)$	$M(C_1,*)$
	C_2	$M(C_2,C_1)$	$M(C_2,C_2)$	...	$M(C_2,C_k)$	$M(C_2,*)$
	...	...	...	...	...	...
	C_k	$M(C_k,C_1)$	$M(C_k,C_2)$	...	$M(C_k,C_k)$	$M(C_k,*)$
		$M(*,C_1)$	$M(*,C_2)$	...	$M(*,C_k)$	n

$$M(*,C_j) = \sum_{i=1}^{k} M(C_i,C_j) \qquad n = \sum_{i=1}^{k} M(C_i,*) = \sum_{i=1}^{k} M(*,C_i)$$

Índice Kappa

- Índice Kappa: medida de concordância
  - Mede a fração de concordância observada entre as classes preditas por h e as classes verdadeiras
  - kappa <= 0: indica ausência de concordância
  - θ_o: concordância total observada
  - θ_e: concordância esperada pelo simples acaso

$$\text{kappa} = \frac{\theta_o - \theta_e}{1 - \theta_e}$$

$$\theta_o = \frac{1}{n}\sum_{i=1}^{k} M(C_i, C_i)$$

$$\text{kappa} \begin{cases} \text{mínimo} < 0 \\ \text{máximo} = 1 \end{cases}$$

$$\theta_e = \frac{1}{n^2}\sum_{i=1}^{k} M(C_i, *) \times M(*, C_i)$$

Exemplo

κ=0.74 err=20.00%	anfíbios	peixes	répteis	pássaros	mamíferos	$M(C_i,*)$
anfíbios	2	0	0	0	0	2
peixes	0	2	0	0	1	3
répteis	1	0	3	0	0	4
pássaros	1	0	0	3	0	4
mamíferos	0	0	1	0	6	7
$M(*,C_i)$	4	2	4	3	7	20

κ=1.00 err=0.00%	anfíbios	peixes	répteis	pássaros	mamíferos	$M(C_i,*)$
anfíbios	2	0	0	0	0	2
peixes	0	3	0	0	0	3
répteis	0	0	4	0	0	4
pássaros	0	0	0	4	0	4
mamíferos	0	0	0	0	7	7
$M(*,C_i)$	2	3	4	4	7	20

κ=0.81 err=15.00%	anfíbios	peixes	répteis	pássaros	mamíferos	$M(C_i,*)$
anfíbios	2	0	0	0	0	2
peixes	0	3	0	0	0	3
répteis	0	0	4	0	0	4
pássaros	0	0	0	4	0	4
mamíferos	0	0	3	0	4	7
$M(*,C_i)$	2	3	7	4	4	20

κ= –0.22 err=100.00%	anfíbios	peixes	répteis	pássaros	mamíferos	$M(C_i,*)$
anfíbios	0	0	0	0	2	2
peixes	3	0	0	0	0	3
répteis	0	0	0	4	0	4
pássaros	0	0	4	0	0	4
mamíferos	0	7	0	0	0	7
$M(*,C_i)$	3	7	4	4	2	20

Matriz de Confusão para 2 Classes

- Com apenas duas classes, as escolhas estão estruturadas para predizer a ocorrência (**positivo** ou **+**) ou não (**negativo** ou **-**) de um evento simples
- Neste caso, os dois erros possíveis são denominados **falso positivo** (F_P) e **falso negativo** (F_N) e a matriz de confusão resume-se a:
  - T_P é o número de exemplos positivos classificados corretamente
    - Verdadeiro Positivo (True Positive)
  - F_N é o número de exemplos positivos classificados incorretamente (como sendo negativos)
    - Falso Negativo (False Negative)
  - T_N é o número de exemplos negativos classificados corretamente
    - Falso Positivo (False Positive)
  - F_P é o número de exemplos negativos classificados incorretamente (como sendo positivos)
    - Verdadeiro Negativo (True Negative)
  - Total de $n=(T_P+F_N+F_P+T_N)$ exemplos

Classe	Predita C+	Predita C-
Verdadeira C+	T_P	F_N
Verdadeira C-	F_P	T_N

Matriz de Confusão para 2 Classes

Classe	Predita C+	Predita C-	Taxa de Erro da Classe	Taxa de Erro Total
Verdadeira C+	T_P	F_N	$\frac{F_N}{T_P + F_N}$	$\frac{F_P + F_N}{n}$
Verdadeira C-	F_P	T_N	$\frac{F_P}{F_P + T_N}$	

T_P = Verdadeiro Positivo (True Positive)
F_N = Falso Negativo (False Negative)
F_P = Falso Positivo (False Positive)
T_N = Verdadeiro Negativo (True Negative)
n = $T_P + F_N + F_P + T_N$

Exercício

- Qual a matriz de confusão do classificador *h* para os exemplos fornecidos (disease.arff)?

ParedeCelular	Núcleo	Cauda	Cor	Doente (classe)	h (predita)
thin	2	2	light	+	+
thin	1	2	light	-	-
thin	2	2	dark	+	+
thin	2	1	light	-	+
thick	2	2	light	+	+
thick	1	1	light	-	-
thick	2	2	dark	+	+
thin	1	1	dark	-	-
thick	1	1	dark	+	-
thick	2	1	dark	+	+

Classe	Predita +	Predita -	Taxa de Erro da Classe
Verdadeira +	5	1	16.66%
Verdadeira -	1	3	25.00%

Métricas Derivadas da Matriz de Confusão para 2 Classes

- Confiabilidade positiva

$$\text{prel}(h) = \frac{T_P}{T_P + F_P}$$

- Confiabilidade negativa

$$\text{nrel}(h) = \frac{T_N}{T_N + F_N}$$

- Suporte

$$\text{sup}(h) = \frac{T_P}{n}$$

- Sensitividade (recall)

$$\text{sens}(h) = \frac{T_P}{T_P + F_N}$$

- Especificidade

$$\text{spec}(h) = \frac{T_N}{F_P + T_N}$$

- Precisão total

$$\text{tacc}(h) = \frac{T_P + T_N}{n}$$

- Cobertura

$$\text{cov}(h) = \frac{T_P + F_P}{n}$$

- Medida-F

$$F-\text{measure}(h) = \frac{2}{\frac{1}{prel(h)} + \frac{1}{sens(h)}}$$

Métricas Derivadas da Matriz de Confusão para 2 Classes em IR

- Em Recuperação de Informação (*Information Retrieval*):
  - A confiabilidade positiva é denominada precisão (*precision*)
  - A sensitividade é denominada *recall*
- F-measure (média harmônica de *precision* e *recall*) também é conhecida como F_1-measure

$$F-\text{measure}(h)=\frac{2}{\frac{1}{prel(h)}+\frac{1}{sens(h)}}=\frac{2\times prel(h)\times sens(h)}{prel(h)+sens(h)}$$

- A fórmula geral de F_α-measure é

$$F_\alpha-\text{measure}(h)=\frac{\alpha+1}{\frac{1}{prel(h)}+\frac{\alpha}{sens(h)}}=\frac{(\alpha+1)\times prel(h)\times sens(h)}{\alpha\times prel(h)+sens(h)}$$

- Dois valores comuns para α são
  - α=2, que pondera *recall* duas vezes mais do que *precision*
  - α=0.5, que pondera *precision* duas vezes mais do que *recall*

Métricas Derivadas da Matriz de Confusão para 2 Classes

- Assumindo um conjunto de exemplos sobre pacientes no qual há duas classes:
  - Pacientes com gripe (classe positiva)
  - Pacientes sem gripe (classe negativa)
- Após induzir um classificador h:
  - A confiabilidade positiva é a proporção (probabilidade) que, caso um paciente seja rotulado por h como tendo gripe, que ele realmente tenha gripe
  - A confiabilidade negativa é a proporção (probabilidade) que, caso um paciente seja rotulado por h como não tendo gripe, que ele realmente não tenha gripe
  - A porcentagem de pacientes com gripe e rotulados como tendo gripe é o suporte
  - A porcentagem de pacientes rotulados como tendo gripe é a cobertura
  - Sensitividade indica a proporção de pacientes com gripe que são classificados como tais
    - Sensitividade de 100% significa que todos os pacientes com gripe são classificados como tais
  - Especificidade indica a proporção de pacientes sem gripe que são classificados como tais
    - Especificidade de 100% significa que todos os pacientes sem gripe são classificados como tais

Matriz de Confusão para 2 Classes

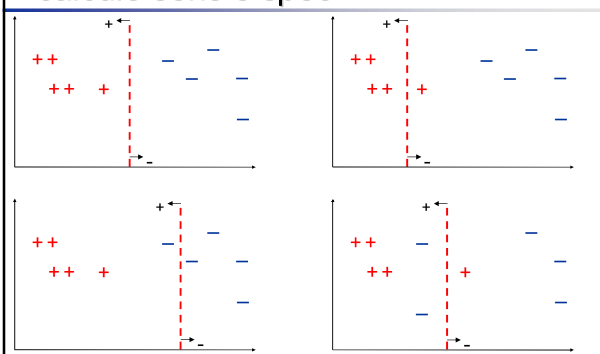
4	1
2	3

- prel = 4/6 = 0.67
- nrel = 3/4 = 0.75
- sup = 4/10 = 0.40
- sens = 4/5 = 0.80
- spec = 3/5 = 0.60
- tacc = 7/10 = 0.70
- cov = 6/10 = 0.60
- F-measure = 0.73
- Kappa = 0.40

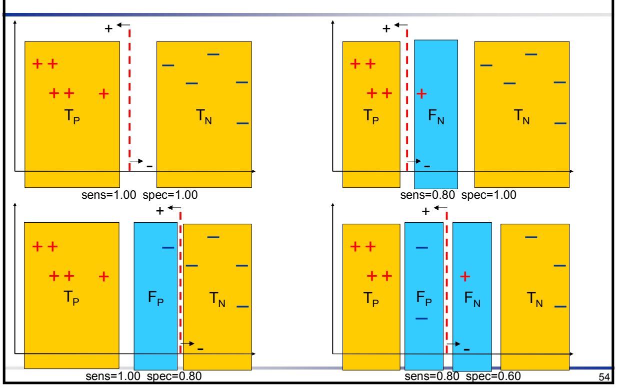
5	0
1	4

- prel = 5/6 = 0.83
- nrel = 4/4 = 1.00
- sup = 5/10 = 0.50
- sens = 5/5 = 1.00
- spec = 4/5 = 0.80
- tacc = 9/10 = 0.90
- cov = 6/10 = 0.60
- F-measure = 0.91
- Kappa = 0.80

Exercício: Indique T_P, F_N, F_P, T_N e calcule sens e spec



Solução



Custos de Erros

- Medir adequadamente o desempenho de classificadores, através da taxa de erro (ou precisão) assume um papel importante em AM, uma vez que o objetivo consiste em construir classificadores com baixa taxa de erro em novos exemplos
- Entretanto, ainda considerando o problema anterior contendo duas classes, se o custo de ter falsos positivos e falsos negativos não é o mesmo, então outras medidas de desempenho devem ser usadas
- Uma alternativa natural, quando cada tipo de classificação incorreta possui um custo diferente ou mesmo quando existe prevalência de classes, consiste em associar um custo para cada tipo de erro

Custos de Erros

- O custo cost(C_i,C_j) é um número que representa uma penalidade aplicada quando o classificador faz um erro ao rotular exemplos, cuja classe verdadeira é C_i, como pertencentes à classe C_j, onde i,j = 1, 2, ..., k e k é o número de classes
- Assim, cost(C_i,C_i) = 0, uma vez que não constitui um erro e cost(C_i,C_j) > 0, i ≠ j
- Em geral, os indutores assumem que cost(C_i,C_j)=1, i≠j, caso esses valores não sejam definidos explicitamente

Custos de Erros

- No cálculo utilizando custos, os erros são convertidos em custos pela multiplicação do erro pelo custo correspondente, calculados utilizando-se

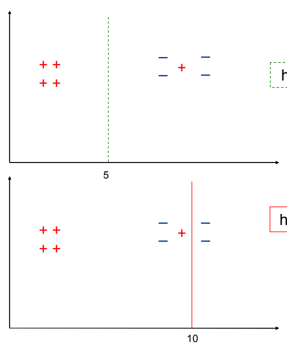
$$\mathrm{err-cost}(h) = \frac{1}{n}\sum_{i=1}^{n} \| y_i \neq h(x_i) \| \times \mathrm{cost}(y_i, h(x_i))$$

- É também possível obter os custos através da matriz de confusão utilizando-se

$$\mathrm{err-cost}(h) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{k} M(C_i, C_j) \times \mathrm{cost}(C_i, C_j)$$

- Assim, ao invés de projetar um algoritmo que minimize a taxa de erro, o objetivo poderia ser minimizar custos de classificação incorreta

Custos de Erros



cost(+,-) = 10
cost(-,+) = 2

h_1: if X1<5 then classe = + else classe = -

err-cost(h_1) = 10

h_2: if X1<10 then classe = + else classe = -

err-cost(h_2) = 4

Cobertura

- Seja regra L → R
  - Exemplos que satisfazem a parte L da regra são **cobertos** pela regra (ou a regra **dispara** para esses exemplos)
  - Exemplos que satisfazem tanto a condição L como a conclusão R são **cobertos corretamente** pela regra
  - Exemplos satisfazendo a condição L mas não a conclusão R são **cobertos incorretamente** pela regra
  - Exemplos que não satisfazem a condição L **não são cobertos** pela regra

Exemplos satisfazendo...	são...
¬L	Não cobertos pela regra
L	Cobertos pela regra
L ∧ R	Cobertos corretamente pela regra
L ∧ ¬R	Cobertos incorretamente pela regra

Cobertura: Exemplo

- if X_1= a and X_2 = s then classe = +

	Atributos				
Exemplo	X_1	X_2	X_3	Classe	Cobertura
z_1	a	s	2	+	Coberto (corretamente)
z_2	a	s	1	-	Coberto (incorretamente)
z_3	b	n	1	+	Não coberto
z_4	b	s	2	-	Não coberto
z_5	c	n	2	*	Não coberto

Matriz de Contingência

- A matriz de confusão é aplicada ao classificador visto como uma caixa-preta, ou seja, o classificador pode ser simbólico ou não para se calcular essa matriz
- Já a **matriz de contingência** é calculada para cada regra, exigindo, desta forma, que o classificador seja simbólico
- Considerando cada regra no formato L → R, sua correspondente matriz de contingência é dada ao lado

$$\begin{array}{c|cc|c} & L & \bar{L} & \\ \hline R & lr & \bar{l}r & r \\ \bar{R} & l\bar{r} & \bar{l}\bar{r} & \bar{r} \\ \hline & l & \bar{l} & n \end{array}$$

Matriz de Contingência

- Nesta tabela, L denota o conjunto de exemplos para os quais a condição da regra é verdadeira e seu complemento ¬L denota o conjunto de exemplos para os quais a condição da regra é falsa e analogamente para R e ¬R
- LR denota o conjunto de exemplos L ∩ R no qual ambos L e R são verdadeiros, L¬R denota o conjunto de exemplos L ∩ ¬R no qual L é verdadeiro e R é falso e assim por diante

	L	$\bar{L}$	
R	lr	$\bar{l}r$	r
$\bar{R}$	$l\bar{r}$	$\bar{l}\bar{r}$	$\bar{r}$
	l	$\bar{l}$	n

Matriz de Contingência

- Por generalidade, denota-se a cardinalidade de um conjunto **A** por **a**, ou seja, **a = |A|**
- Assim, **l** denota o número de exemplos no conjunto **L**, ou seja, l = |L|, **r** denota o número de exemplos no conjunto **R**, ou seja r = |R|, **lr** denota o número de exemplos no conjunto LR com lr = |LR| e assim por diante

	L	$\bar{L}$	
R	lr	$\bar{l}r$	r
$\bar{R}$	$l\bar{r}$	$\bar{l}\bar{r}$	$\bar{r}$
	l	$\bar{l}$	n

Matriz de Contingência

- A freqüência relativa |**A**|/**n** = **a**/**n** associada ao subconjunto **A** é denotada por p(**A**), onde **A** é um subconjunto dos n exemplos
- Dessa forma, a freqüência relativa é usada como uma estimativa de probabilidade
- A notação p(**A**|**B**) segue sua definição habitual em probabilidade, dada pela equação seguinte, onde **A** e **B** são ambos subconjuntos dos **n** exemplos

$$p(A \mid B) = \frac{p(AB)}{p(B)} = \frac{\frac{|AB|}{n}}{\frac{|B|}{n}} = \frac{\frac{ab}{n}}{\frac{b}{n}} = \frac{ab}{b}$$

Matriz de Contingência: Exemplo

if X_1 = a and X_2 = s then classe = +

Exemplo	Atributos X_1	X_2	X_3	Classe	Cobertura
E_1	a	s	2	+	Coberto (corretamente)
E_2	a	s	1	-	Coberto (incorretamente)
E_3	b	n	1	+	Não coberto
E_4	b	s	2	-	Não coberto
E_5	c	n	2	*	Não coberto

	L	¬L	
R	1 (E_1)	1 (E_3)	2
¬R	1 (E_2)	2 (E_4,E_5)	3
	2	3	5

Métricas Derivadas da Matriz de Contingência

- Confiabilidade positiva $\mathrm{prel}(L \rightarrow R) = p(R \mid L) = \frac{lr}{l}$
- Confiabilidade negativa $\mathrm{nrel}(L \rightarrow R) = p(\bar{R} \mid \bar{L}) = \frac{\bar{l}\bar{r}}{\bar{l}}$
- Suporte $\sup(L \rightarrow R) = p(LR) = \frac{lr}{n}$
- Sensitividade $\mathrm{sens}(L \rightarrow R) = p(L \mid R) = \frac{lr}{r}$

Métricas Derivadas da Matriz de Contingência

- Especificidade $\mathrm{spec}(L \rightarrow R) = p(\bar{L} \mid \bar{R}) = \frac{\bar{l}\bar{r}}{\bar{r}}$
- Precisão total $\mathrm{tacc}(L \rightarrow R) = p(LR) + p(\bar{L}\bar{R}) = \frac{lr + \bar{l}\bar{r}}{n}$
- Cobertura $\mathrm{cov}(L \rightarrow R) = p(L) = \frac{l}{n}$
- Novidade $\mathrm{nov}(L \rightarrow R) = p(LR) - p(L)p(R) = \frac{lr}{n} - \frac{l \cdot r}{n^2}$
- Satisfação $\mathrm{sat}(L \rightarrow R) = \frac{p(\bar{R}) - p(\bar{R} \mid L)}{p(\bar{R})} = 1 - \frac{n \cdot l\bar{r}}{l \cdot \bar{r}}$

Slides baseados no Capítulo 4 do livro:

Rezende, S.O. (ed).
Sistemas Inteligentes, Manole, 2003,
ISBN 85-204-1683-7

Material elaborado por
José Augusto Baranauskas
Revisão 2007