Faculdade IMPACTA TECNOLOGIA

Pós Graduação em BI com Big Data

Web Mining

Classificação

Prof. MSc. Fernando Sousa

Bibliografia

- Bibliografia básica para esta aula
  - Liu B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. 1st ed. 2007. Corr. 2nd printing edition. Berlin ; New York: Springer; 2009. 552 p – Cap 3 e 6.
  - Weiss SM, Indurkhya N, Zhang T. Fundamentals of Predictive Text Mining. 2010 edition. London ; New York: Springer; 2010. 226 p. - Cap. 3
  - Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. 1 edition. New York: Cambridge University Press; 2008. 506 p. – Cap. 13 e 14

Bibliografia

- Bibliografia complementar
  - Salton G, McGill MJ. Introduction to Modern Information Retrieval. New York, NY, USA: McGraw-Hill, Inc.; 1986.
  - Sousa FS. Análise Comparativa de Métodos de Recuperação de Informação para Categorização de Conteúdos Web Relacionados à Saúde [Dissertação de Mestrado]. [São Paulo]: Universidade Federal de São Paulo (UNIFESP); 2011.

Introdução

- Pré-processamento
  - Como transformar dados textuais em numéricos
  - Encontrar tokens e termos
  - Remover palavras desnecessárias
  - Gerar o vetor de características
- Após o pré-processamento utilizam-se os vetores de características para classificação de textos

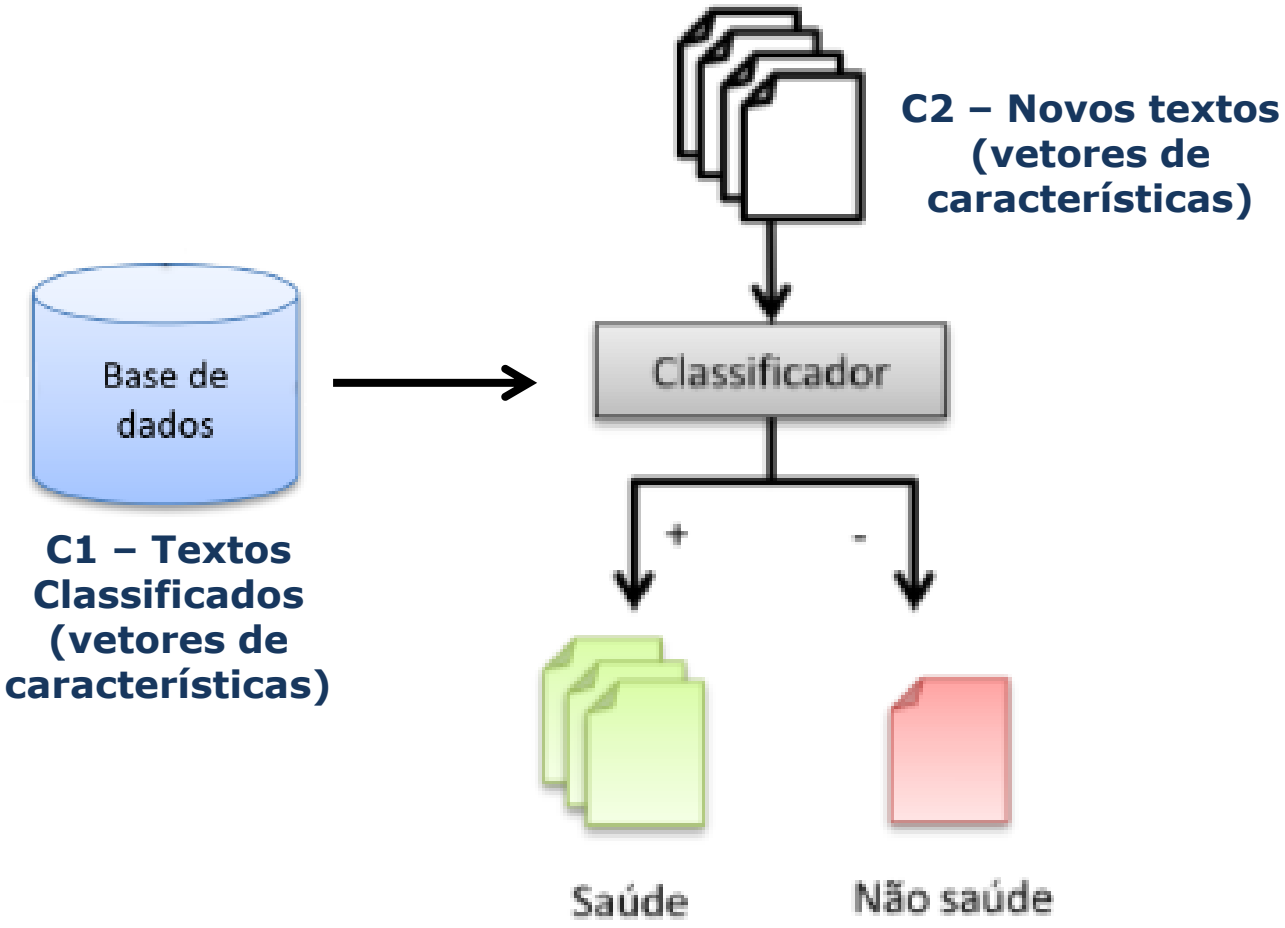
Faculdade IMPACTA TECNOLOGIA

Introdução

Classificação de textos

- Problema:
  - Existe um conjunto de textos C1 de uma experiência passada (já classificado – sabe-se a que categoria pertence)
  - Existe um conjunto de novos textos C2 que ainda não sabe-se a categoria
  - A tarefa de classificação é atribuir uma categoria para os exemplos de C2 baseado dos exemplos de C1

Classificação de textos



Classificação de textos

- Exemplos
  - Classificar e-mails (suporte de hardware, suporte de software, vendas...)
  - Separar spams de e-mails seguros
  - Classificar conteúdos da web em diferentes categorias (ciência, esportes, saúde, economia,...)
  - Classificar tweets como positivos ou negativos (análise de sentimento)

Classificação de textos

- Como o algoritmo funciona?
  - Ele encontra **padrões** existentes nos textos já classificados
  - Este padrão será utilizado para analisar os novos textos e prever sua categoria
    - Se um determinado padrão ocorre para um conjunto de textos de uma mesma categoria, um novo exemplo que contém este mesmo padrão provavelmente pertence a mesma categoria

Classificação de textos

- Um padrão pode ser um conjunto de palavras que ocorrem frequentemente em um conjunto de textos de um mesmo assunto
  - Textos de esporte costumam ter palavras como: “time”, “jogador”, “gol”, “pontos”, “resultado”
  - Textos de economia costumam ter palavras como: “índice”, “juros”, “inflação”, “dinheiro”, “pontos”
- Repare que uma mesma palavra pode ser frequente em assuntos diferentes
  - O conjunto de palavras e de um texto suas frequências definem o padrão

Classificação de textos

- Por exemplo:
  - Baseado em um conjunto de textos já classificado (conjunto de treinamento), sabe-se que as palavras **ganhar** e **presente** ocorrem com frequência e-mails do tipo spam
  - Define-se a regra: e-mails recebidos com as palavras **ganhar** e **presente** serão classificados como spam
    - **Esta é uma regra simples apenas ilustrativa. Não é otimizada para classificadores de padrões**

Classificação de textos

- A regra anterior pode ser colocada em um vetor de características
  - Executar as tarefas de pré-processamento
  - Se o texto contem a palavra **presente**, o valor dessa característica será 1; 0 caso contrário
  - Se o texto contem a palavra **ganhar**, o valor dessa característica será 1; 0 caso contrário

Classificação de textos

C1

Texto	presente	...	ganhar	...
T1	1	...	1	...
T2	0	...	0	...
T3	1	...	0	...
T4	1	...	1	...

C2

Texto	presente	...	ganhar	...
T5	0	...	0	...
T6	1	...	1	...
T7	1	...	1	...
T8	0	...	1	...

Classificação de textos

- Na tarefa de classificação coloca-se um último atributo no vetor de características, que será a categoria de um texto
  - No exemplo, span (1) ou não (0)
- A definição da categoria do conjunto de treinamento é feita manualmente
  - A classificação manual é necessária para “treinar” o classificador

Classificação de textos

C1

Texto	presente	...	ganhar	...	span?
T1	1	...	1	...	1
T2	0	...	0	...	0
T3	1	...	0	...	1
T4	1	...	1	...	0

C2

Texto	presente	...	ganhar	...	span?
T5	0	...	0	...	?
T6	1	...	1	...	
T7	1	...	1	...	?
T8	0	...	1	...	?

Classificação de textos

- Os valores dos atributos do vetor de características não precisam ser necessariamente 0 ou 1
  - Pode-se utilizar a contagem de ocorrência dos termos (to, tf, tf.idf)
  - Pode-se utilizar valores categóricos
    - Sim ou verdadeiro para a palavra que ocorre
    - Não ou falso para a palavra que não ocorre

Classificação de textos

- Os algoritmos de data mining são utilizados para realizar a classificação dos textos (classificador)
  - Textos estão representados de forma numérica.
  - Existe uma característica especial que determina a categoria de um documento
- O objetivo do algoritmo é encontrar a categoria mais provável para um novo documento, baseado nos padrões do conjunto de treinamento (o que ele já conhece)
  - Aprendizado supervisionado
  - Um conjunto de textos já foi classificado manualmente para servir de guia para o classificador

Classificação de textos

- Os padrões dos textos podem ser um pouco nebulosos e difíceis de encontrar.
  - Por exemplo, em C1 não há nenhum texto onde apareça a palavra ganhar (1) e não apareça a palavra presente (0)
    - E se aparecer um novo texto com este padrão (T8 de C2)? Spam ou não?
  - Ou ainda, se houver 2 textos com as duas palavras, mas um classificado como como spam e outro não (T1 e T4), qual categoria deve ser associada?

Classificação de textos

C1

Texto	presente	...	ganhar	...	span?
T1	1	...	1	...	1
T2	0	...	0	...	0
T3	1	...	0	...	1
T4	1	...	1	...	0

C2

Texto	presente	...	ganhar	...	span?
T5	0	...	0	...	?
T6	1	...	1	...	
T7	1	...	1	...	?
T8	0	...	1	...	?

Classificação de textos

- Duas palavras não são suficientes para determinar a categoria de um documento
  - Existem milhares de palavras no vocabulário de um idioma.
- A classificação deve considerar todas as palavras presentes nos textos
  - Entretanto, muitas palavras também pode não ser bom (ler sobre seleção de atributos)
  - Ex.: stopwords não são boas palavras para a classificação, e são removidas do vetor de características
- Além disso, a classificação de um novo documento não é feita baseada em apenas um texto já classificado, mas sim em uma combinação deles

Modelos

- Baseado no vetor de características dos documentos já classificados, o algoritmo de data mining cria um modelo
  - Modelo é uma função matemática criada a partir dos padrões encontrados no vetor de características
    - O conjunto de textos utilizado para criar o modelo recebe o nome de Conjunto de treinamento
  - Este modelo é capaz de inferir a categoria de um texto, baseado nos padrões conhecidos a priori

Faculdade IMPACTA TECNOLOGIA

Modelos

- Relembrando, os padrões podem ser um pouco nebulosos
  - Pode aparecer um novo padrão que não estava no conjunto te treinamento
- O algoritmo de classificação não pode simplesmente procurar por um padrão exatamente igual ao novo exemplo
  - É bem provável que não exista

Modelos

- É necessário que o algoritmo faça uma generalização do que aprendeu no conjunto já classificado
  - Para um novo texto ele encontrará o padrão que mais se aproxima
  - Por exemplo, calcular a similaridade entre os novos documentos e os documentos já classificados
- Esta pode parecer uma tarefa complexa, mas existem técnicas matemáticas e computacionais que simplificam o cálculo da similaridade
- O que os algoritmos de data mining fazem é criar um modelo matemático com os dados de treinamento

Algoritmos de Data Mining

- Exemplos de algoritmos
  - Naive Bayes
  - Vizinhos mais próximos
  - Redes neurais
  - Árvores de decisão
- Depois da criação do modelo, o próximo passo é a avaliação

Modelagem em Python

- Utilizando o conhecimento adquirido nas aulas de Data Mining, utilize Árvores de Decisão para classificar a base de notícias
  - Você deve alterar o a leitura dos arquivos para armazenar também a categoria de cada arquivo
- Encontre a acurácia do classificador utilizando 30% da base para testes

Modelagem com Rapid Miner

- Crie um novo processo para gerar o vetor de características da base de notícias
  - Na seleção das pastas, separe cada uma das categorias de noticias
  - Neste processo também será importante selecionar a opção add meta information, para manter as categorias
- Procure o operador Select Attributes na árvore Blending -> Attributes -> Selection
  - Vamos remover os atributos que não são os termos, mantendo a categoria

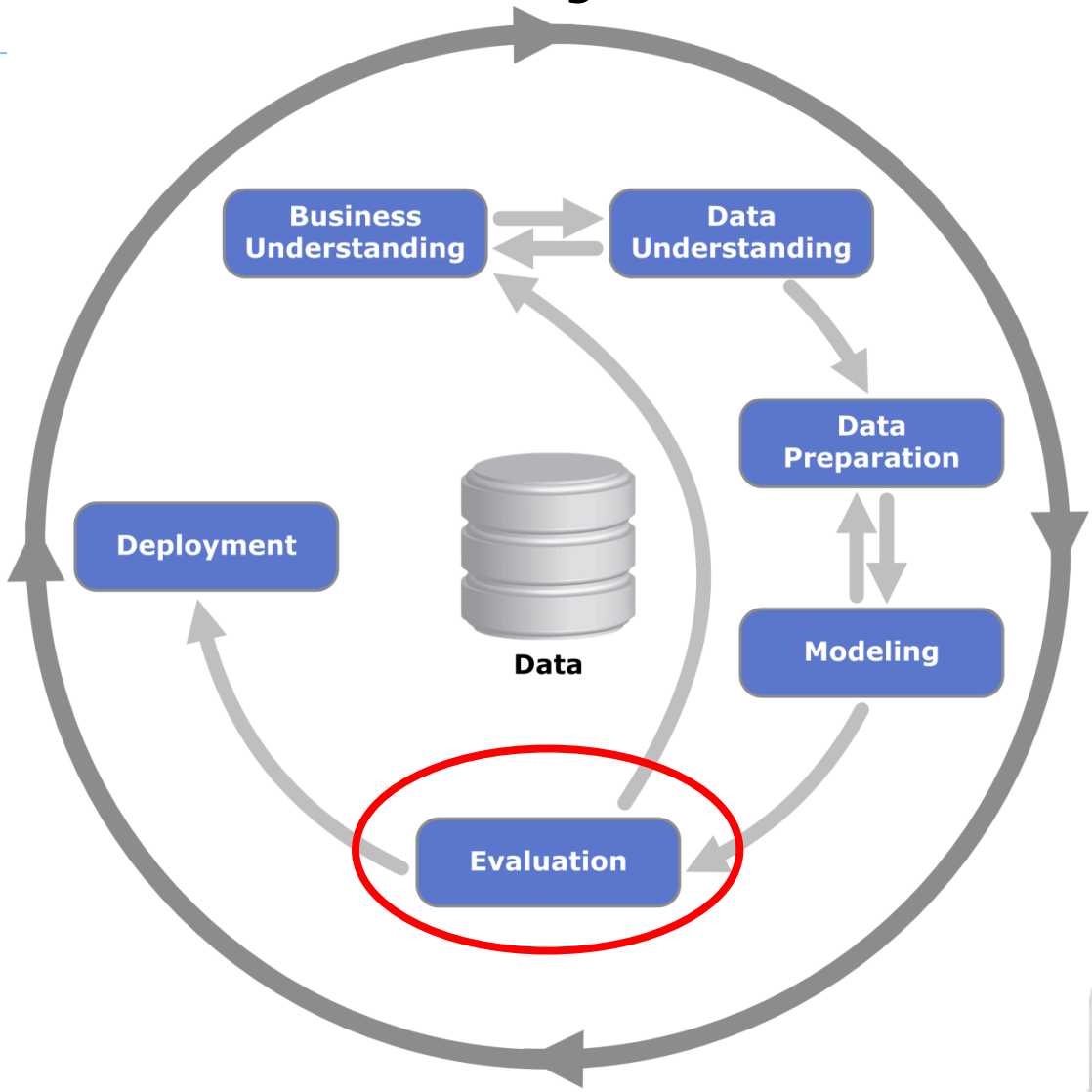
Modelagem com Rapid Miner

- Configures os seguintes parâmetros:
  - attribute filter type: subset
    - Escolher um subgrupo de atributos
  - attributes:
    - Clique no botão Select Attributes
    - Coloque os seguintes atributos do lado direito:
      - Content-Type, Description, Keywords, Language, Robots, Title, metadata_date, metadata_file, metadata_path
      - Se não encontrá-los na lista da esquerda, basta escrever no campo Selected Attributes e clicar no botão +
  - invert selection: marcado
    - Ou seja, os atributos selecionados anteriormente serão removidos, e não selecionados
  - include special attributes: marcado

Modelagem com Rapid Miner

- Repare que a coluna da categoria está verde
  - É um atributo especial (label)
  - O RapidMiner identifica este atributo especial e separa dos atributos utilizados para criar o modelo

Validação e Avaliação



Validação e Avaliação

- Depois que o modelo é criado é necessário avaliar o desempenho de acerto do classificador
- É necessário que o classificador tenha um bom desempenho nesse quesito para a aplicação (deployment)
- A medida de acurácia é a principal medida para verificar o quanto o classificador está acertando
  - Número de textos classificador corretamente dividido pela quantidade de textos

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$

- A medida de erro (1 - acurácia) também pode ser utilizada

Métodos de Avaliação

- Usualmente o conjunto de textos é dividido em dois subconjuntos: treinamento e testes
- O conjunto de treinamento é utilizado para o classificador "aprender" os padrões existentes nos textos (modelagem)
- O conjunto de testes é utilizados para avaliar o desempenho do classificador
  - Textos que o classificador ainda não conhece
- Todos os textos devem estar classificados
  - Treinamento: o classificador tem que saber qual é a categoria de um padrão
  - Testes: a categoria inferida pelo classificador é comparada com a categoria original

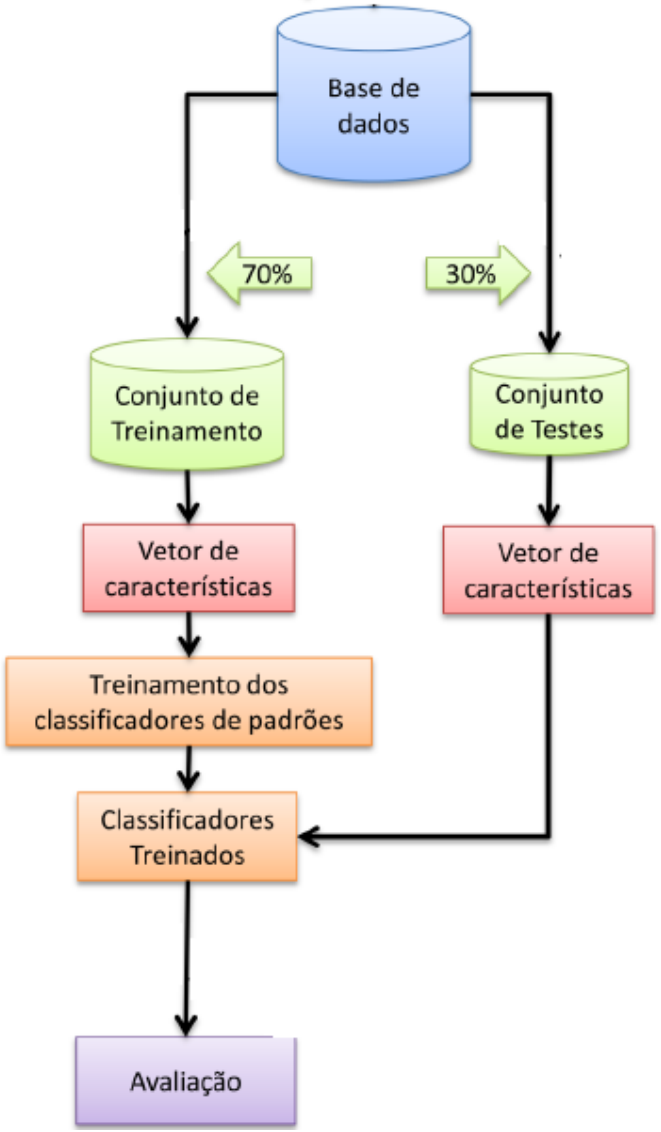
Métodos de Avaliação

- Nenhum texto utilizado no treinamento deve ser utilizado nos testes
  - O classificador já conhece o padrão dos textos de treinamento
  - Se o mesmo texto for utilizado para testá-lo, o resultado será enviesado (a acurácia será maior)
- Veremos a seguir os principais métodos de avaliação do classificador

Métodos de Avaliação - Holdout

- Os textos disponíveis são divididos em dois conjuntos disjuntos
  - Conjunto de treinamento
  - Conjunto de testes (*holdout set*)
  - Nenhum texto pertencerá aos dois conjuntos
- Muito utilizado quando o conjunto é grande
- É usual dividir o conjunto total ao meio (50% treinamento e 50% testes) ou 2/3 para treinamento e 1/3 para testes

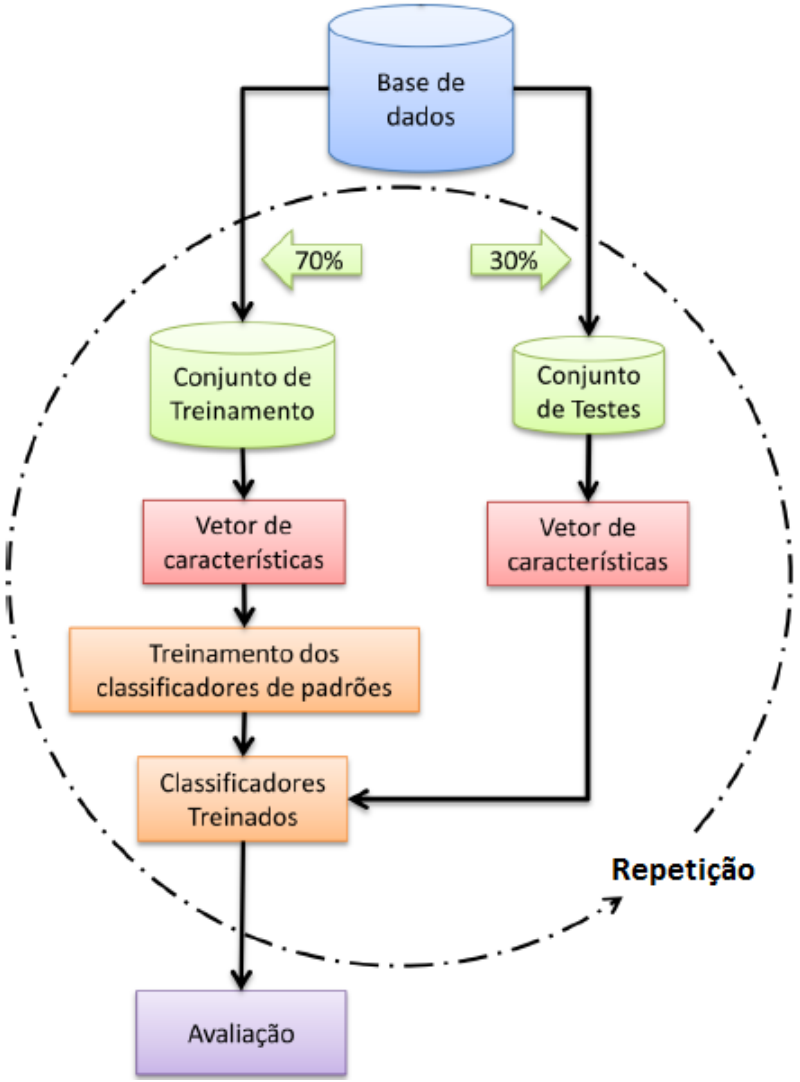
Métodos de Avaliação - Holdout



Métodos de Avaliação – Conj. Aleatórios

- O conjunto total disponível é muito pequeno para se confiar no desempenho do classificador (sobram poucos textos para os testes)
- Cria-se então *n* conjuntos aleatórios de treinamento e testes para avaliar o desempenho *n* vezes
- O desempenho final será a média do desempenho dos n conjuntos

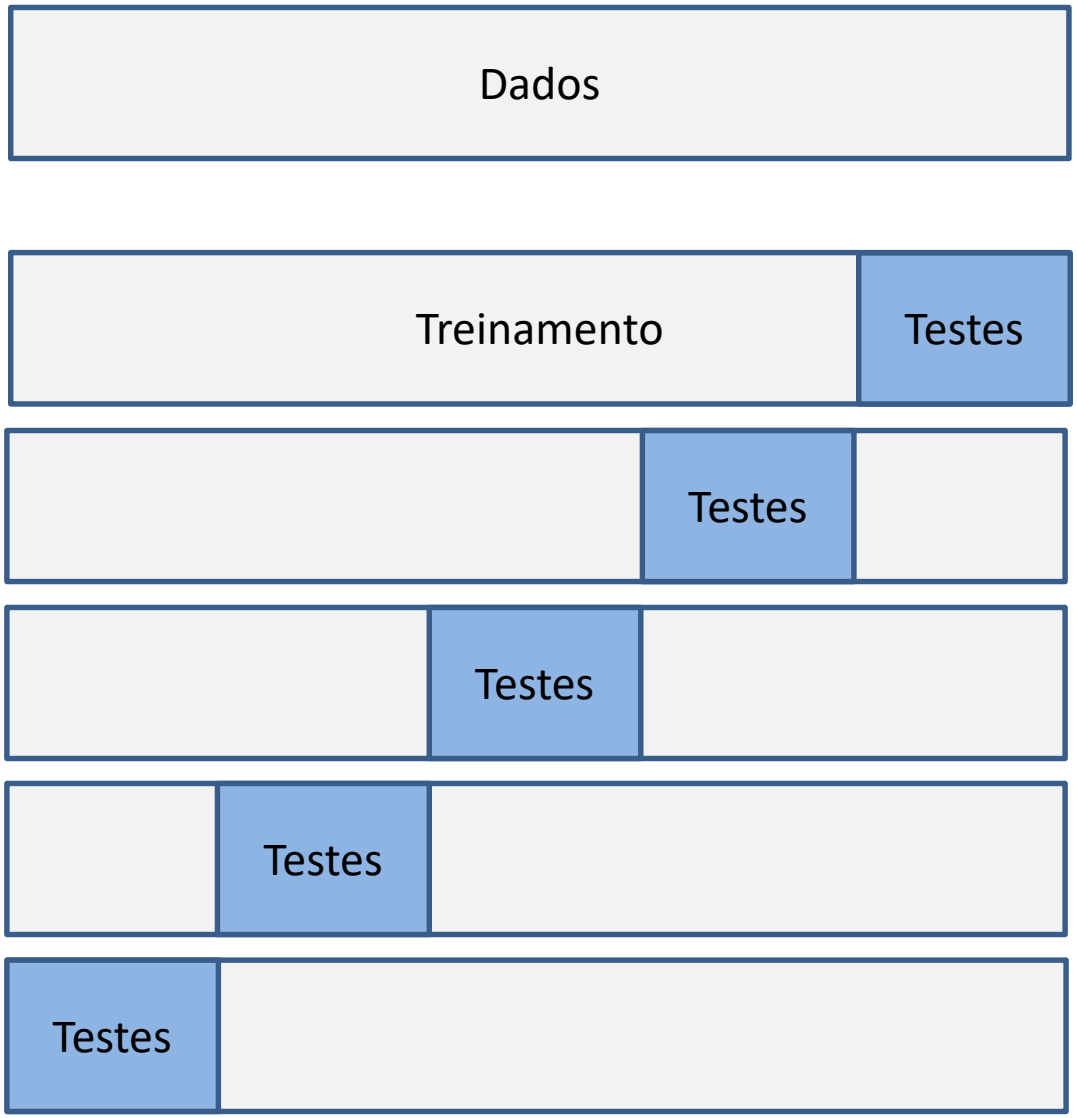
Métodos de Avaliação – Conj. Aleatórios



Métodos de Avaliação – Cross-validation

- Também utilizado para conjuntos pequenos
- O conjunto total é dividido em n conjuntos disjuntos de mesmo tamanho (*n-fold cross-validation*)
- n-1 conjuntos são utilizados para treinar o classificador, e o conjunto restante é utilizado para testar
- Este processo e executado n vezes, gerando n medidas de desempenho. O desempenho final será a média
- n= 10 e n = 5 são comuns
  - 10-fold cross-validation e 5-fold cross-validation

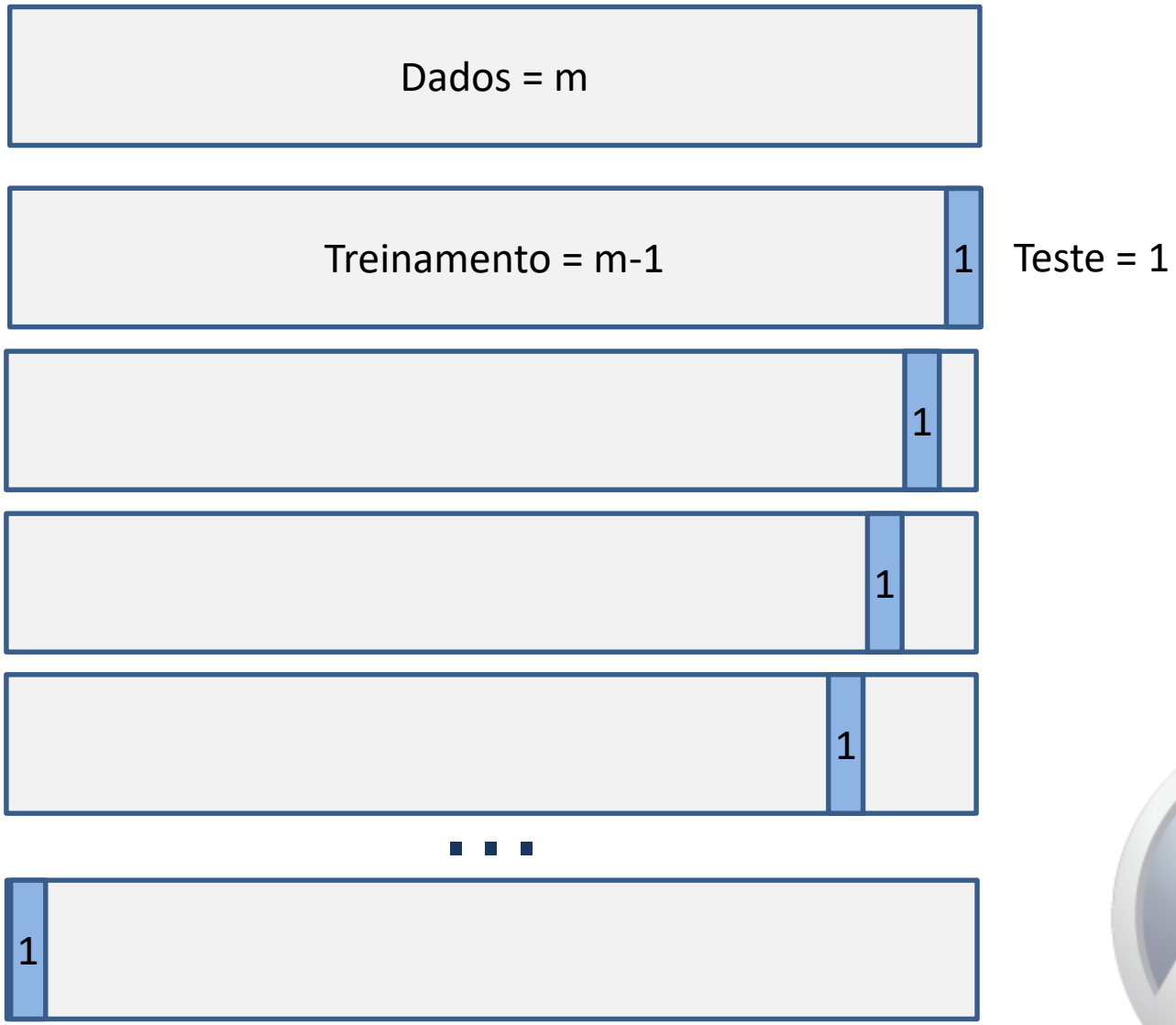
Métodos de Avaliação – Cross-validation



Métodos de Avaliação – Leave-one-out

- Caso especial do Cross-validation onde o conjunto de testes é composto por apenas um texto
  - Se o conjunto total tem m textos, serão criados m conjuntos diferentes
  - m-1 conjuntos são utilizados para treinar e o conjunto restante utilizado para testar

Métodos de Avaliação – Leave-one-out



Medidas de avaliação

- Precisão (*precision*) e revocação (*recall*) são duas medidas clássicas para avaliar o desempenho do classificador
- Em classificação de textos é comum que o problema tenha uma categoria alvo
  - Páginas web que são de saúde
  - Opiniões positivas no twitter
  - Detecção de spam em e-mail
- A categoria alvo é a categoria **positiva**
- Todas as outras categorias do conjunto são englobadas como uma única categoria, **negativa**

Medidas de avaliação

- A precisão avalia o quão precisa é a classificação dos texto de categoria positiva
  - Ou seja, quantos textos classificados como positivo realmente são positivos
- A revocação avalia o quão completa é a classificação dos textos da categoria positiva
  - Ou seja, dos textos que eram para ser classificados como positivo, quantos foram classificados como positivo
- A entendimento dessas medidas fica mais claro com a matriz de confusão

Matriz de confusão

- A matriz de confusão, ou matriz de contingência é uma tabela que contém informações sobre as categorias reais e as preditas pelos classificadores. Ela contém:
  - O número de classificações **corretas** dos exemplos **positivos** (verdadeiro positivo - VP)
  - O número de classificações **incorretas** dos exemplos **positivos** (falso positivo – FP)
  - O número de classificações **incorretas** dos exemplos **negativos** (falso negativo – FN)
  - O número de classificações **corretas** dos exemplos **negativos** (verdadeiro negativo– VN)

Matriz de confusão

	Classificados +	Classificados -
Original +	VP	FN
Original -	FP	VN

- A precisão e a revocação são calculadas com base na matriz de confusão

$$Precisão = \frac{VP}{VP + FP} \qquad Revocação = \frac{VP}{VP + FN}$$

- Precisão: número de textos positivos classificados corretamente (VP) dividido pelo número total de textos classificados como positivos (VP + FP)
- Revocação: número de textos positivos classificados corretamente (VP) dividido pelo total de textos que realmente são da categoria positiva

Exemplo

- Um conjunto de testes tem 100 textos da classe positiva e 1000 textos da classe negativa. A matriz de confusão após a utilização do classificador é a seguinte:

	Classificados +	Classificados -
Original +	1	99
Original -	0	1000

- Qual a precisão e revocação deste classificador?

Liu B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. 1st ed. 2007. Corr. 2nd printing edition. Berlin ; New York: Springer; 2009. 552 p – Cap 3

Exemplo

- Um conjunto de testes tem 100 textos da classe positiva e 1000 textos da classe negativa. A matriz de confusão após a utilização do classificador é a seguinte:

	Classificados +	Classificados -
Original +	1	99
Original -	0	1000

- Qual a precisão e revocação deste classificador?

$$Precisão = \frac{VP}{VP + FP} = \frac{1}{1 + 0} = 1$$

$$Revocação = \frac{VP}{VP + FN} = \frac{1}{1 + 99} = 0.01$$

Liu B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. 1st ed. 2007. Corr. 2nd printing edition. Berlin ; New York: Springer; 2009. 552 p – Cap 3

F-Score

- Em aplicações reais é comum que um valor de precisão alto seja atingido em detrimento da revocação, e vice-versa, como no exemplo anterior
  - A medida que será mais importante dependerá da aplicação
- É possível utilizar uma única medida chamada F-score
  - Média harmônica entre precisão e revocação

$$F = \frac{2\,pr}{p + r}$$

- Para o F-Score ser alto, tanto a precisão quanto a revocação devem ser altos

Múltiplas categorias

- Existem situações em que o problema contém mais de duas categorias
  - Classificar notícias em esportes, saúde, educação, etc
  - Classificar os e-mails recebidos por uma empresa
- A mesma abordagem citada anteriormente pode ser utilizada
- A diferença estará na matriz de confusão
  - Ela terá n linhas e n colunas
- Para calcular a precisão e revocação de uma categoria, esta será a categoria positiva e todas as outras a categoria negativa

Múltiplas categorias

		Classificação			
		Esportes	Agricultura	Saude	revocação
Original	Esportes	26	0	4	86.66%
	Agricultura	3	23	5	74.19%
	Saude	2	4	31	83.78%
	precisão	83.87%	85.18%	77.50%	

- Exemplo para a categoria esportes:
  - Precisão = 26/(26+3+2) = 83,87%
  - Revocação = 26 / (26 + 0 + 4) = 86,67%
  - F-score = = 2 * 0,8387 * 0,8667/(0,8387 + 0,8667) = 85,25%

Avaliação com Rapid Miner

- Procure o operador Cross Validation na árvore Validation
- Ligue a saída do Selected Attributes na entrada do Cross Validation
- Cross Validation é um subprocesso. Clique duas vezes para abri-lo
- Este subprocesso tem 2 partes: treinamento e testes

Avaliação com Rapid Miner

- Em treinamento (training) vamos colocar o algoritmo de data mining
- Procure a extensão do Weka no Marketplace e instale
- Depois de instalado, procure o operador W-J48 na árvore Extensions -> Weka -> Modeling -> Predictive
  - J48 é a implementação em java do C4.5
  - Coloque este operador no espaço de treinamento (training) do subprocesso de Cross Validation
  - Ligue a entrada do treinamento do subprocesso Cross Validation (tra) na entrada do W-J48 (tra)
  - Ligue a saída do W-J48 (mod) na saída do treinamento do subprocesso Cross Validation (mod)

Avaliação com Rapid Miner

- Agora vamos aplicar o modelo na parte de testes (testing) e calcular a acurácia
- Procure o operador Apply Model na árvore Scoring e coloque no espaço de testes (testing) do subprocesso de Cross Validation
- Ligue a entrada de testes mod na entrada mod do Apply Model
- Procure

Faculdade
IMPACTA
TECNOLOGIA