Avaliação do classificador

Prof. Clodoaldo A. M. Lima & Sarajane M. Peres

Treinamento e Teste

- O desempenho de um classificador pode ser medido por meio da taxa de erro:
  - A taxa de erro de erro é a proporção de erros obtidos sobre um conjunto completo de instancias.

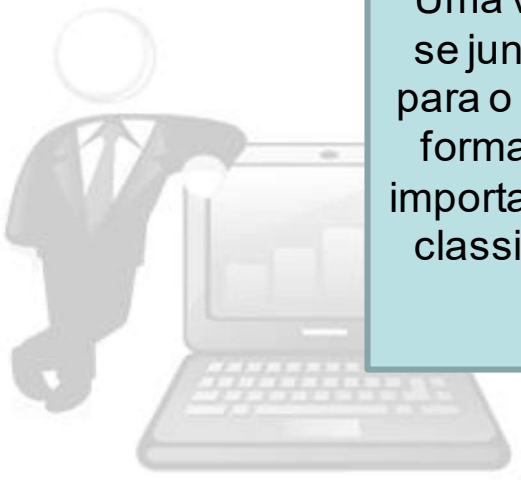
> O classificador prediz a classe de cada instância; se ela é correta, é contada como um “sucesso”; se não, é contada como um “erro”.

  - O que interessa é o desempenho do classificador mediante “novos” dados, e não sobre os dados velhos (usados no processo de treinamento).

Treinamento, validação e teste

- Frequentemente é útil dividir o conjunto de dados disponíveis em três partes, para três diferentes propósitos:
  - Conjunto de treinamento: usado por um ou mais métodos de aprendizado para construir o classificador.
  - Conjunto de validação: usado para otimizar os parâmetros do classificador, ou para selecionar um em particular.
  - Conjunto de teste: usado para clacular a taxa de erro final do modelo já otimizado.

Uma vez que a taxa de erro foi determinada, os dados de testes podem se juntar aos dados de treinamento para produzir um novo classificador para o uso real. Não há problema nisso quando usado apenas como uma forma de maximizar o classificador que será usado na prática. O que é importante é que a taxa de erro não seja calculada com base nesse último classificador gerado. Além disso, o mesmo pode ser feito com os dados de validação. (Witten & Frank, 2005)



Matriz de Confusão

- Oferece uma medida da eficácia do modelo de classificação,mostrando o número de classificações corretas *versus o* número de classificação prevista para cada classe.

Classe	C_1 Prevista	C_2 Prevista	⋯	C_k Prevista
C_1 Real	$M(C_1,C_1)$	$M(C_1,C_2)$	⋯	$M(C_1,C_k)$
C_2 Real	$M(C_2,C_1)$	$M(C_2,C_2)$	⋱	$M(C_2,C_k)$
⋮	⋮	⋮	⋯	⋮
C_k Real	$M(C_k,C_1)$	$M(C_k,C_2)$	⋯	$M(C_k,C_k)$

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

Matriz de Confusão para duas classes

Classe	prevista C_+	prevista C_-	Taxa de erro da classe	Taxa de erro total
real C_+	T_p	F_n	$F_n / (T_p + F_n)$	$(F_p + F_n) / n$
real C_-	F_p	T_n	$F_p / (F_p + T_n)$	

TP = True Positive (verdadeiro positivo)

FN = False Negative(falso negativo)

FP = False Positive (falso positivo)

TN = True Negative (verdadeiro negativo)

n = (TP+FN+FP+TN)

Matriz de Confusão para duas classes

- Outras métricas derivadas da tabela anterior:

C_+ Predictive Value = $T_p / (T_p + F_p)$
C_- Predictive Value = $T_n / (T_n + F_n)$
True C_+ Rate ou Sesitivity y ou Recall = $T_p / (T_p + F_n)$
True C_- Rate ou Specifity = $T_n / (F_p + T_n)$
Precision = $(T_p + T_n) / n$

Avaliação do classificador

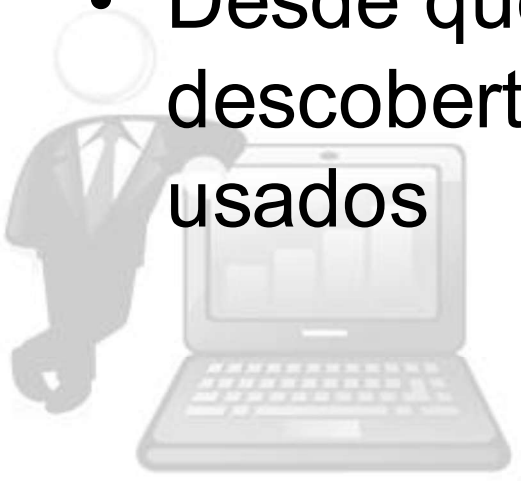
- Para estimar o erro verdadeiro de um classificador, a amostra para teste deve ser aleatoriamente escolhida
- Amostras não devem ser pré-selecionadas de nenhuma maneira
- Para problemas reais, tem-se uma amostra de uma única população, de tamanho *n,* e a tarefa é estimar o erro verdadeiro para essa população

Métodos para estimar o erro verdadeiro de um classificador

- Resubstitution
- Random
- Holdout
- r-fold cross-validation
- r-fold stratified cross-validation
- Leave-one-out
- Bootstrap

Resubstitution

- Gera o classificador e testa a sua performance com o mesmo conjunto de dados
- Os desempenhos computados com este método são otimistas e tem grande bias
- Desde que o bias da resubstitution foi descoberto, os métodos de cross-validation são usados



Holdout
(Witten & Frank, 2005)

- Estratégia para teste de classificador que reserva um certo montante de dados para treino e o restante para teste (podendo ainda usar parte para validação).

- Comumente esta estratégia uma 1/3 dos dados dados para teste e o restante para treinamento, escolhido randomicamente.

- É interessante assegurar que a amostragem randômica seja feita de tal maneira que garanta que cada classe é apropriadamente representada tanto no conjunto de treinamento quanto no conjunto de teste. Este procedimento é chamado de *estratificação* (*holdout estratificado*).

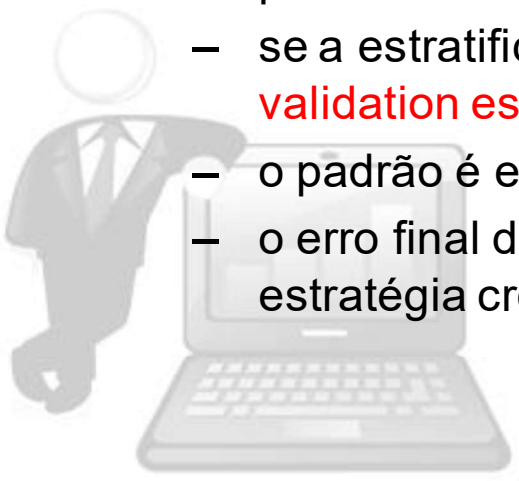
- Também é útil, para amenizar tendências, repetir todo o processo de treino e teste várias vezes com diferentes amostragens randômicas (holdout repetitivo/iterativo).

Random

- l classificadores, l<<n, são induzidos de cada conjunto de treinamento
- O erro é a média dos erros dos classificadores medidos por conjuntos de treinamentos gerados aleatória e independentemente
- Pode produzir estimativas melhores que o holdout

Cross Validation
(Witten & Frank, 2005)

- Trata-se de uma estratégia para lidar com um montante de dados limitado.
- Nesta estratégia decide-se um numero fixos de folds, ou partições dos dados. Supondo que sejam usados três folds (3-fold cross validation):
  - o conjunto de dado é dividido em três partições de tamanhos aproximadamente iguais e, de maneira rotativa, cada uma delas é usada para teste enquanto as duas restantes são usadas para treinamento.
  - ou seja: use 2/3 para treinamento e 1/3 para teste e repita o processo três vezes, tal que, no fim, cada instância tenha sido usadas exatamente uma vez para teste.
  - se a estratificação é adotada, então o procedimento se chama 3-fold cross validation estratificado (aconselhável).
  - o padrão é executar o 10-fold cross validation, 10 vezes.
  - o erro final do classificador é a média dos erros obtidos em cada iteração da estratégia cross-validation

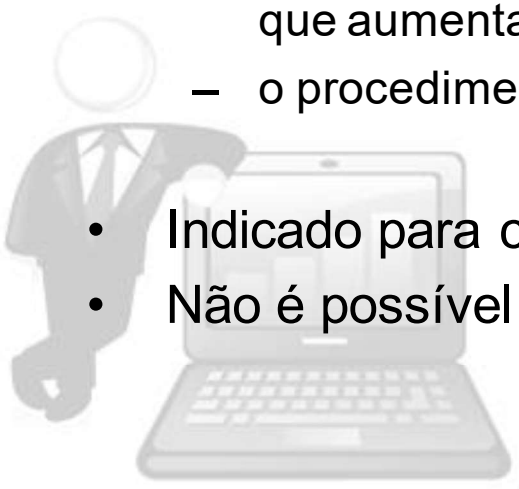


Leave-one-out
(Witten & Frank, 2005)

- Leave-one-out cross-validation é um n-fold cross-validation, onde n é o número de instâncias no conjunto de dados.
- A avaliação é sobre a corretude de classificação da instância em teste – um ou zero para sucesso ou falha, respectivamente.
- Os resultados de todas as n avaliações, uma para cada instância do conjunto de dados, são analisados via média, e tal média representa o erro final estimado.

- Motivações:
  - o maior número possível de dados é usado para treinamento em cada caso, o que aumenta as chance do classificador alcançar acuidade.
  - o procedimento é determinístico.

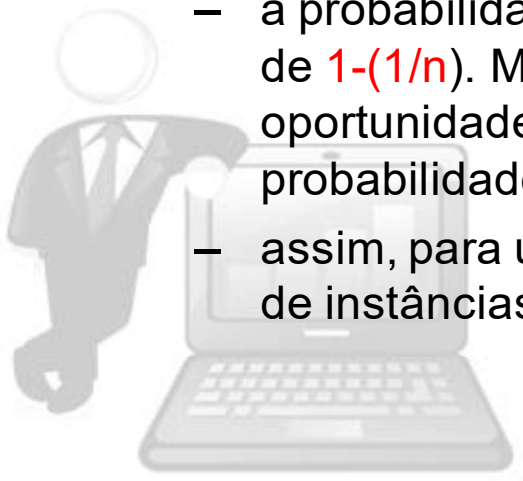
- Indicado para conjunto de dados pequenos.
- Não é possível aplicar qualquer procedimento de estratificação.



Boostrap
(Witten & Frank, 2005)

- Baseado em um procedimento estatístico de amostragem com reposição.
- Uma instância não é retirada do conjunto de dados original quando ela é escolhida para compor o conjunto de treinamento.
  - Ou seja, a mesma instância pode ser selecionada várias vezes durante o procedimento de amostragem.
- As instâncias do conjunto original que não foram escolhidas para compor o conjunto de treinamento, comporão o conjunto de teste.

- O 0,632 bootstrap:
  - a probabilidade de uma instância ser escolhida é 1/n. E de não ser escolhida é de 1-(1/n). Multiplicando essas probabilidades de acordo com o número de oportunidades de escolha (n), tem-se $(1 - (1/n))^n \sim e^{-1} = 0,368$ como a probabilidade de uma instância não ser escolhida.
  - assim, para um conjunto de dados grande, o conjunto de testes conterá 36,8% de instâncias e o conjunto de treinamento, 63,2% delas.



Boostrap
(Witten & Frank, 2005)

- A medida de erro obtida é uma estimativa pessimista do erro verdadeiro porque o conjunto de treinamento, embora tenha tamanho n, contém somente 63% das instâncias, o que não é grande coisa se comparado com os 90% usados no 10-fold cross-validation.
- Para compensar isso, pode-se combinar o erro do conjunto de teste com o erro de resubstituição (estimativa otimista).

- O boostrap combina da seguinte forma:
  - erro = 0,632 * erro de teste + 0.368 * erro de treinamento

- O procedimento deve ser repetido várias vezes, e uma média de erro final deve ser encontrada.

O bootstrap é o procedimento mais indicado para estimar erro para conjuntos de dados muito pequenos.

Parâmetros dos estimadores

M.

	holdout	random	leave – one – out	r - fold cv	r - fold strat cv	bootstrap
Train size	pn	t	$n-1$	$n(r-1)/r$	$n(r-1)/r$	n
Test size	$(1-p)n$	$n-t$	1	n/r	n/r	$n-t$
Iterations	1	$I << n$	n	r	r	200
Replacement	*no*	*no*	*no*	*no*	*no*	*yes*
Class Prevalence	*no*	*no*	*no*	*no*	*yes*	*yes/no*

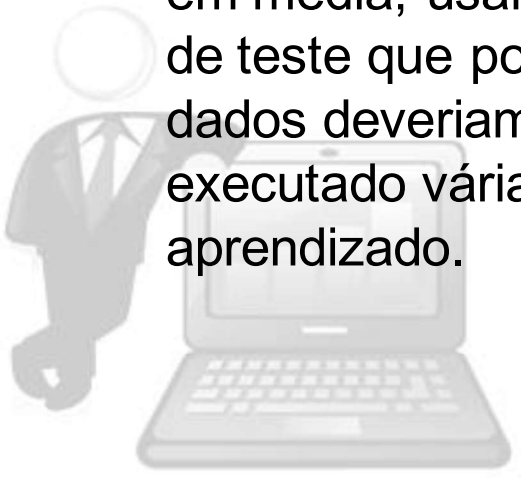
Comparando métodos

(Witten & Frank, 2005)

- Encontrar a taxa de erro para as técnicas comparadas e escolher aquela com a menor taxa é a forma mais simples de comparação e, pode ser adequada para problemas isolados.

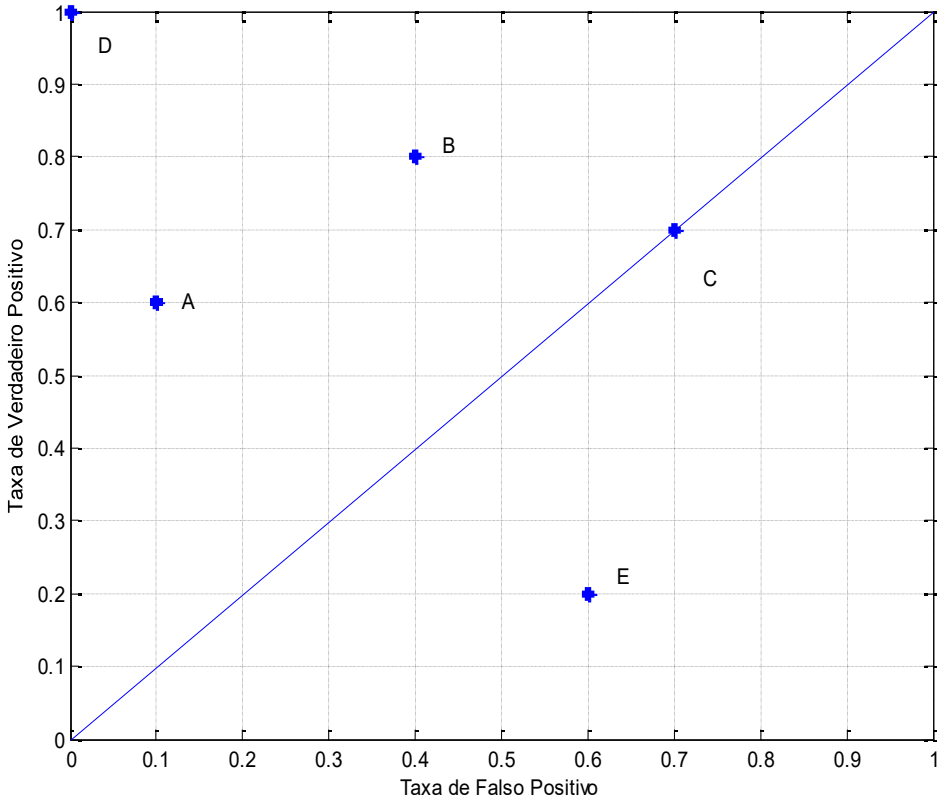
- Se um novo algoritmo é proposto, seus proponentes devem mostrar que ele melhora o estado da arte para o problema em estudo e demonstrar que a melhora observada não é apenas um efeito de sorte do processo de estimativa do erro.

- O objetivo é determinar se um esquema é melhor ou pior do que o outro, em média, usando todas as possibilidades de conjuntos de treinamento e de teste que podem ser criados a partir do domínio. Todos os conjuntos de dados deveriam ser do mesmo tamanho e o experimento deveria ser executado várias vezes, com diferentes tamanhos, para obter uma curva de aprendizado.



Avaliação dos classificadores

- Gráfico ROC com cinco classificadores discretos.
  - A é dito um classificador "conservador", B é o inverso de E, D é um classificador perfeito e C é dito aleatório.



Avaliação dos classificadores

- Considere as seguintes saídas de um classificador:

z	y	L (-0.7)	L(-0.6)	L(0.8)	L(0.9)	L(1.0)	L(>1.0)
-0.7	-1	1	-1	-1	-1	-1	-1
-0.6	-1	1	1	-1	-1	-1	-1
0.8	1	1	1	1	-1	-1	-1
0.9	1	1	1	1	1	-1	-1
1.0	1	1	1	1	1	1	-1

- -0,7: TP = 1 FP = 1 ponto (1 ; 1)
- -0,6: TP = 1 FP = 0,5 ponto (0.5 ; 1)
- 0,8: TP = 1 FP = 0 ponto (0 ; 1)
- 0,9: TP = 0,66 FP = 0 ponto (0 ; 0,66)
- 1,0: TP = 0,33 FP = 0 ponto (0 ; 0,33)