Tópicos Especiais em Computação I

Avaliação de modelos

Patrícia Lucas

Bacharelado em Sistemas de Informação
IFNMG - Campus Salinas

Salinas
Março 2021

INSTITUTO FEDERAL
Norte de Minas Gerais
Campus Salinas

Versão 1.2021

Referência

Avaliação de modelos

Capítulo 9: Avaliação de modelos preditivos.

Inteligência Artificial: Uma abordagem de aprendizado de máquina. Katti Faceli...[et al.]. - Rio de Janeiro: LTC, 2011.

Visão geral

Avaliação de modelos

Uma característica particular dos algoritmos de aprendizagem de máquina é a necessidade de experimentação.

A validação de qualquer técnica de aprendizagem de máquina geralmente envolve a realização de experimentos controlados, em que se demonstre a sua efetividade na solução de diversos problemas, representados por seus conjuntos de dados associados.

É recomendável seguir procedimentos que garantam a corretude, a validade e a reprodutibilidade dos experimentos realizados e, mais importante, das conclusões obtidas a partir de seus resultados.

Visão geral

Avaliação de modelos

A avaliação experimental de um algoritmo de aprendizagem de máquina pode ser realizada sobre diversos aspectos: acurácia do modelo, compreensibilidade do conhecimento extraído, tempo de aprendizado, etc.

Nessa aula, vamos concentrar a discussão em medidas relacionadas ao desempenho obtido nas predições realizadas.

- Métricas de erro para problemas de classificação.
- Métricas de erro para problemas de regressão.
- Amostragem.

Métricas de erro para classificação

Taxa de erro de classificadores:

$$err(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i)) \tag{1}$$

Em que:

- $I(a) = 1$, se a é verdadeiro e 0, caso contrário.

- Dado um conjunto de dados contendo n objetos, sobre o qual a avaliação será realizada, essa taxa equivale à proporção de exemplos desse conjunto classificados incorretamente por $\hat{f}$ e é obtida pela comparação da classe conhecida de x_i, y_i, com a classe predita, $\hat{f}(x_i)$.

- A taxa de erro varia entre 0 e 1 e valores próximos de 0 são considerados melhores.

Métricas de erro para classificação

Avaliação de modelos

Acurácia de classificadores:

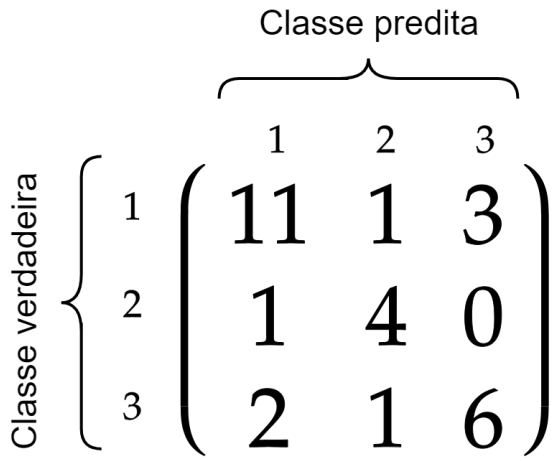
$$ac(\hat{f}) = 1 - err(\hat{f}) \tag{2}$$

- Nesse caso, valores próximos de 1 são considerados melhores.

Métricas de erro para classificação

Avaliação de modelos

Matriz de confusão: ilustra o número de predições corretas e incorretas em cada classe. Exemplo para um problema com 3 classes:



Temos que:

- 11 amostras da classe 1 foram classificadas corretamente, 1 incorretamente como classe 2 e 3 incorretamente como classe 3.
- 4 amostras da classe 2 foram classificadas corretamente e 1 incorretamente como classe 1.
- 6 amostras da classe 3 foram classificadas corretamente, 1 incorretamente como classe 2 e 2 incorretamente como classe 1.

Métricas de erro para classificação

Matriz de confusão para duas classes:

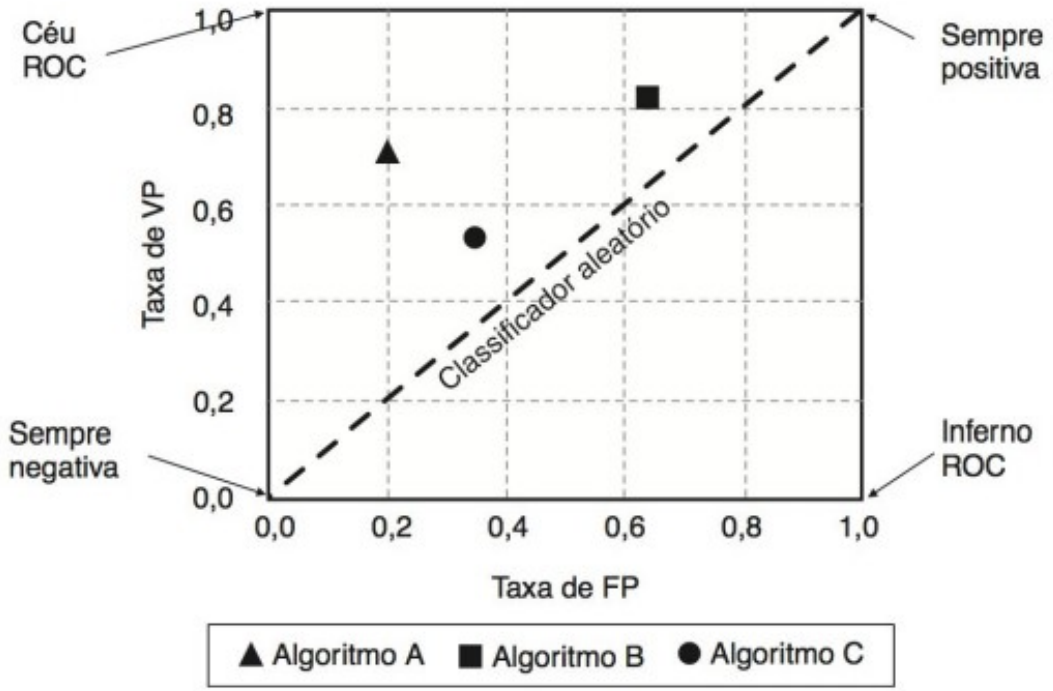
		Classe predita	
		+	-
Classe verdadeira	+	VP	FN
	-	FP	VN

- VP: corresponde ao número de exemplos da classe positiva classificados corretamente.
- VN: corresponde ao número de exemplos da classe negativa classificados corretamente.
- FP: corresponde ao número de exemplos cuja classe verdadeira é negativa mas que foram classificados incorretamente como pertencendo à classe positiva.
- FN: corresponde ao número de exemplos originalmente à classe positiva que foram classificados incorretamente como pertencendo à classe negativa.

Métricas de erro para classificação

Avaliação de modelos

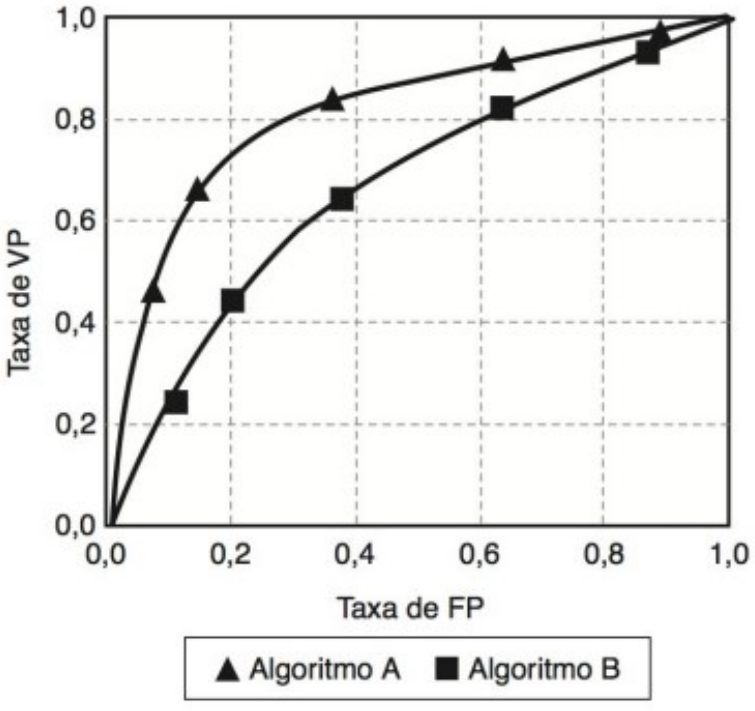
Análise ROC: o gráfico ROC é um gráfico bidimensional plotado em um espaço denominado espaço ROC, com eixos X e Y representando as medidas TFP e TVP, respectivamente.



Métricas de erro para classificação

Avaliação de modelos

Curva ROC



Área abaixo da curva ROC (AUC): medida que produz valores entre 0 e 1, sendo que os mais próximos de 1 são considerados melhores.

Métricas de erro para regressão

Avaliação de modelos

Para problemas de regressão o erro de $\hat{f}$ pode ser calculado pela distância entre o valor y_i conhecido e aquele predito pelo modelo $\hat{f}(x_i)$.

As medidas de erro mais utilizadas são o erro quadrático médio (MSE - *mean squared error*) e a distância absoluta média (MAD - *mean absolute distance*):

$$MSE(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{3}$$

$$MAD(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{f}(x_i)| \tag{4}$$

O MSE e MAD são sempre positivos e para ambas as medidas, valores mais baixos correspondem a melhores modelos.

Amostragem

Avaliação de modelos

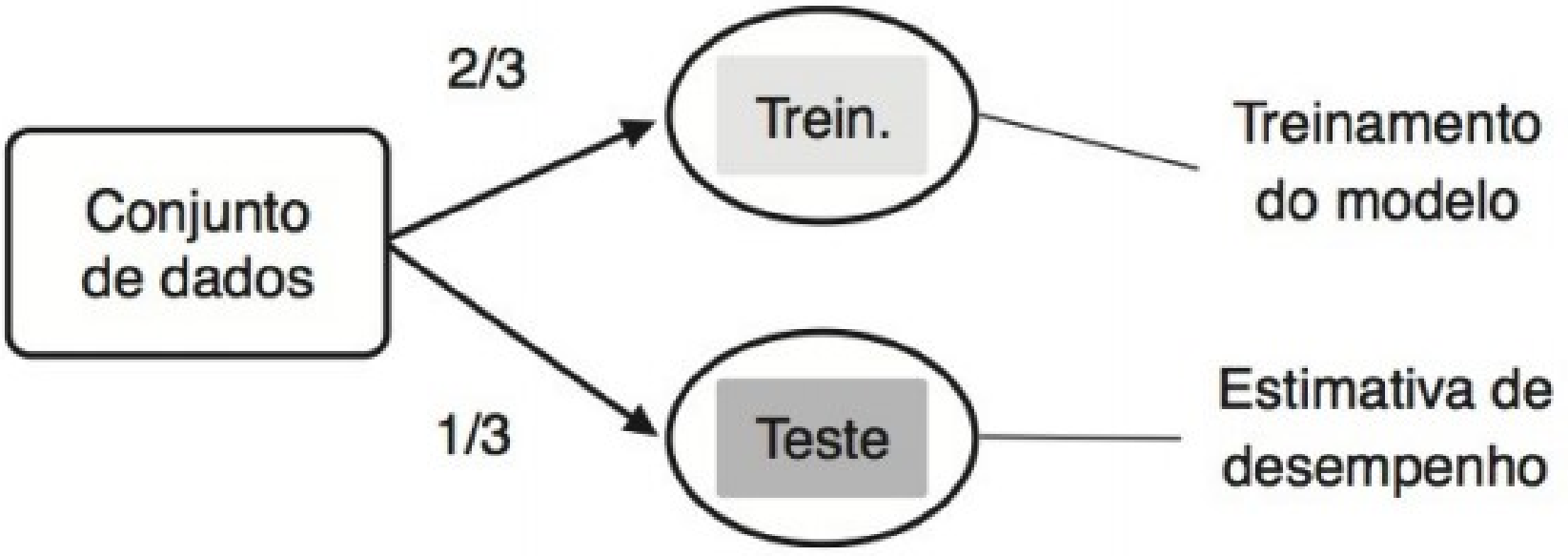
Para se obter estimativas de desempenho preditivo mais confiáveis devemos definir um subconjunto de dados de treinamento e um subconjunto de dados de teste.

O subconjunto de treinamento é usado na etapa de ajuste do modelo, enquanto que o subconjunto de teste é usado para simular a apresentação de objetos novos ao preditor.

Esses subconjuntos devem ser disjuntos para assegurar que as medidas de desempenho sejam obtidas a partir de um conjunto de exemplos diferente daquele usado no aprendizado.

No caso de métodos de amostragem que envolvem média de desempenho, deve-se reportar também os valores de desvio padrão associados. Um alto desvio padrão indica uma alta variabilidade nos resultados, ou seja, uma instabilidade.

Métodos de amostragem *holdout*:

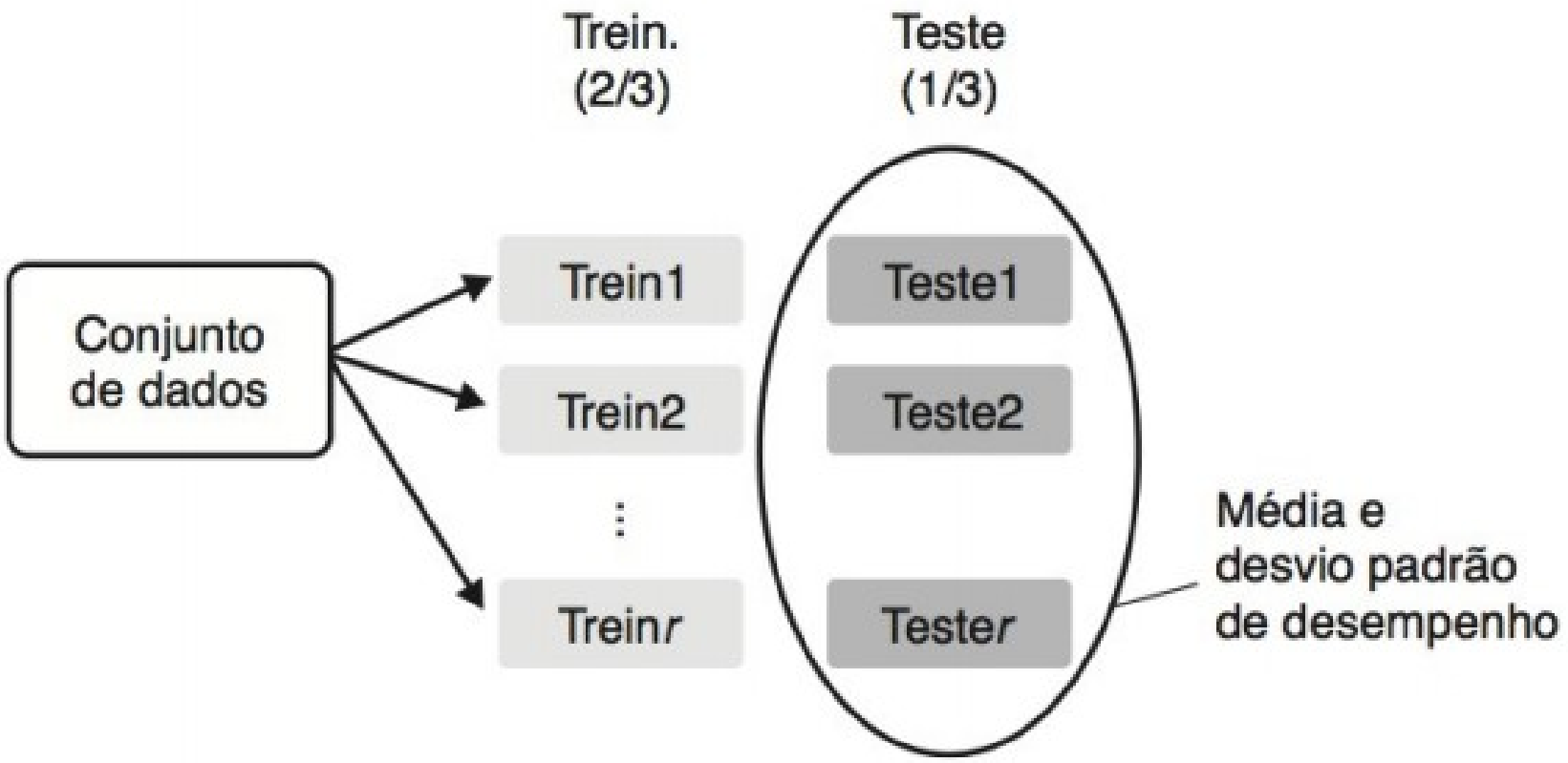


Problema: o *holdout* não permite avaliar o quanto o desempenho de uma técnica varia quando diferentes combinações de objetos são apresentados em seu treinamento.

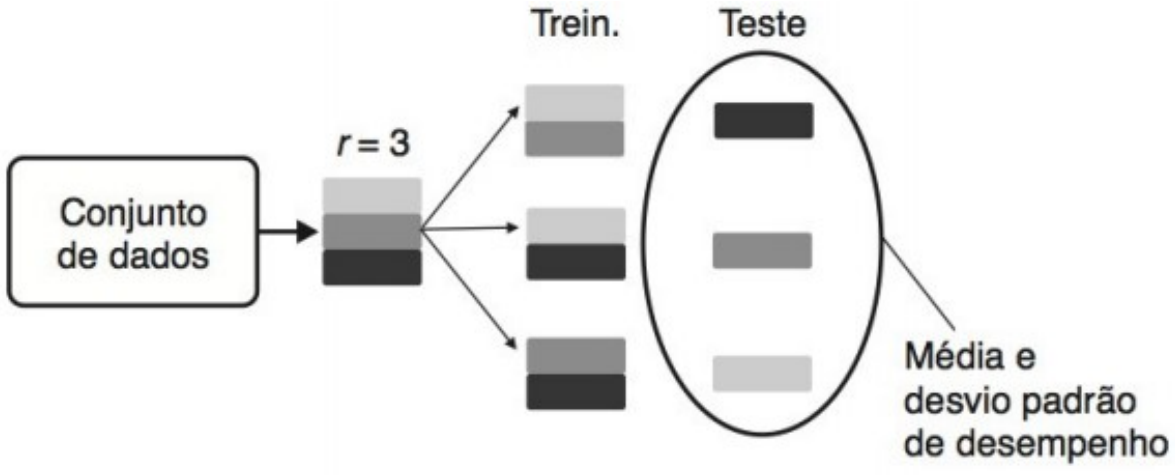
Amostragem

Avaliação de modelos

Métodos de amostragem aleatória:



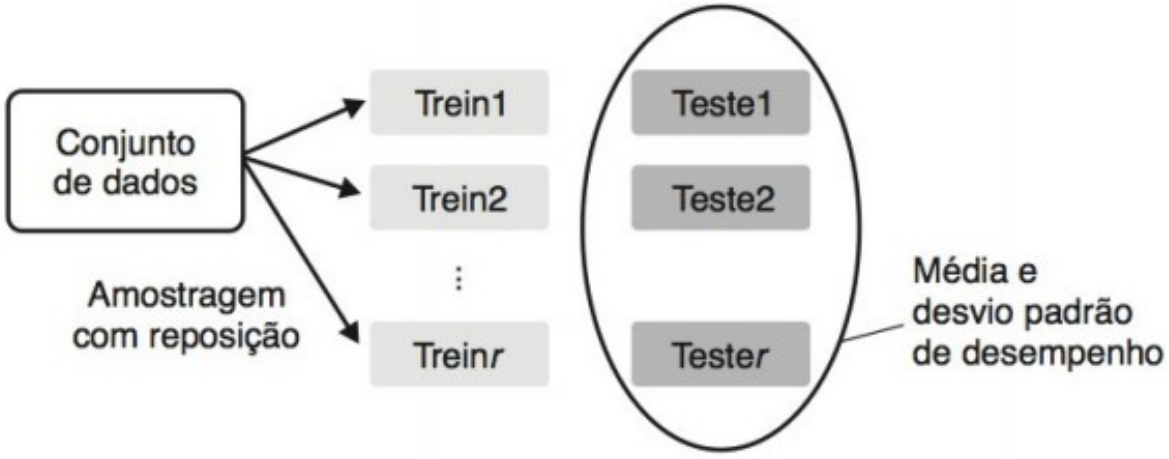
Métodos de amostragem por validação cruzada:



O conjunto de dados é dividido em r subconjuntos de tamanho aproximadamente igual. O modelo é treinado usando $r - 1$ exemplos e o restante é usado para teste. Esse processo é repetido r vezes, utilizado em cada ciclo uma partição diferente para teste.

O desempenho final do preditor é a média dos desempenhos observados sobre cada subconjunto de teste.

Métodos de amostragem *bootstrap*:



Nesse método, *r* subconjuntos de treinamento são gerados pela amostragem aleatória de exemplos do conjunto original, com repetição. Os exemplos não selecionados compõem os subconjuntos de teste.

O desempenho final do preditor é a média dos desempenhos observados sobre cada subconjunto de teste.