Validação Cruzada

EST0133 - Introdução à Modelagem de Big Data

Marcus Nunes
https://introbigdata.org/
https://marcusnunes.me/

Universidade Federal do Rio Grande do Norte

Introdução

Introdução

- Os métodos que veremos a partir de agora são computacionalmente complexos
- Eles envolvem passos extras que normalmente não são utilizados em métodos paramétricos de ajuste de modelos
- Os dois conceitos principais que nos permitirão avaliar se nossos modelos foram bem ajustados são
  - Divisão dos dados em conjuntos de treino e teste
  - Validação cruzada

Conjuntos de Treino e Teste

Conjuntos de Treino e Teste

- A partir de hoje, veremos alguns algoritmos para classificação e predição de dados
- Estes algoritmos são ferramentas importantes para descobrir padrões
- Eles permitem que generalizemos comportamentos presentes nos dados

Conjuntos de Treino e Teste

- Classificação ou Regressão?
- A resposta é categórica ou numérica?
- Resposta categorizada: classificação
- Resposta numérica: regressão

Conjuntos de Treino e Teste

- Aprendizagem Supervisionada: existe um conjunto de treino no qual o algoritmo se baseia para encontrar as relações entre os dados, com os valores da variável resposta bem definidos
- Aprendizagem Não-Supervisionada: não existe um conjunto de treino no qual o algoritmo se baseia para encontrar as relações entre os dados, sem os valores da variável resposta bem definidos

Conjuntos de Treino e Teste

- Ao aplicarmos um algoritmo de aprendizagem supervisionada, necessitamos ser capazes de avaliar o quão bom (ou ruim) é o nosso método
- A maneira mais comum de fazer isto é através de divisão dos dados originais em dois conjuntos:
  - Conjunto de treino
  - Conjunto de teste

Conjuntos de Treino e Teste

- O conjunto de treino é aquele no qual aplicamos o algoritmo, informando a resposta correta para o algoritmo
- Em geral, utilizamos de 50% a 80% dos dados originais no conjunto de treino
- O algoritmo, então, se ajusta de modo a prever com a maior exatidão possível os outputs que nos interessam

Conjuntos de Treino e Teste

- O conjunto de teste é aquele que utilizamos para prever resultados e verificar como o algoritmo se comportaria em dados reais
- Em geral, utilizamos de 50% a 20% dos dados originais no conjunto de teste (o percentual de dados deste conjunto depende do percentual utilizado no conjunto de treino, pois um é complementar do outro)
- Assim, podemos avaliar o quão bem o algoritmo está conseguindo prever novos resultados que não estavam no conjunto original
- É como se simulássemos a coleta de novos dados

Métricas de Avaliação

Métricas de Avaliação

- Para avaliar a eficiência do algoritmo de classificação, utilizamos medidas como sensitividade e especificidade
- Sensitividade é a razão entre o número de positivos encontrados pelo modelo pelo total de positivos nos dados
- Especificidade é a razão entre o número de negativos encontrados pelo modelo pelo total de negativos nos dados

Métricas de Avaliação

		Referência	
		Positivo	Negativo
Predição	Positivo	a	b
	Negativo	c	d

- Acurácia: $p_0 = \frac{a+d}{a+b+c+d}$
- Sensitividade: $\frac{a}{a+c}$
- Especificidade: $\frac{d}{b+d}$
- Curva ROC: Sensitividade (Verdadeiros Positivos) vs. 1-Especificidade (Falsos Positivos)

Métricas de Avaliação

		Referência	
		Positivo	Negativo
Predição	Positivo	a	b
	Negativo	c	d

- Kappa:

$$p_{\text{Sim}} = \frac{a+b}{a+b+c+b} \times \frac{a+c}{a+b+c+b}$$

$$p_{\text{Não}} = \frac{c+d}{a+b+c+b} \times \frac{b+d}{a+b+c+b}$$

$$p_e = p_{\text{Sim}} + p_{\text{Não}}$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

Métricas de Avaliação

- Para avaliar a eficiência do algoritmo de regressão, utilizamos medidas como erro quadrático médio, erro absoluto médio e coeficiente de determinação
- Suponha que temos uma amostra de tamanho n e observações y_i, $i = 1, \cdots, n$
- Suponha que possuímos uma forma de estimar os valores de y_i e chamamos estas estimações de $\widehat{y}_i$

Métricas de Avaliação

- A fórmula do erro quadrático médio é dada por

$$\text{EQM} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Normalmente, a estatística que se usa é a raiz do erro quadrático médio, dada por

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$$

Métricas de Avaliação

- O erro absoluto médio é dado por

$$\text{EAM} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_i|$$

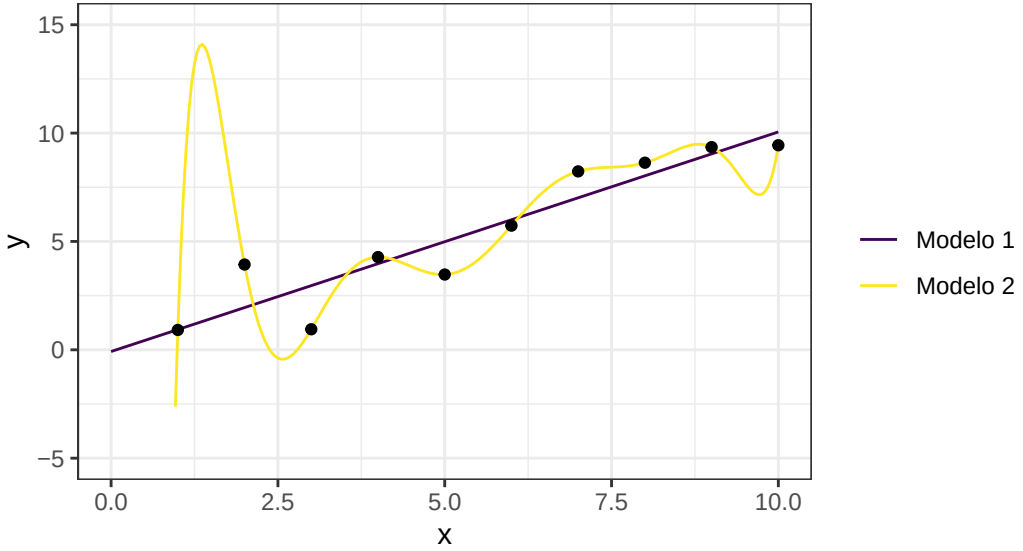
- Podemos calcular o coeficiente de determinação como

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Métricas de Avaliação

- O ideal é que as medidas de ajuste do modelo sejam similares nos conjuntos de treino e teste
- Fazemos isso para verificar que não houve sobreajuste (*overfitting*) nos dados
- Se o modelo está sobreajustado, isto significa que ele se ajusta muito bem aos dados originais, mas não é um modelo que é generalizável

Métricas de Avaliação



Validação Cruzada

Validação Cruzada

- Dividir os dados apenas uma vez em treino e teste pode, ainda assim, gerar vícios
- Então por que não realizar esta divisão mais de uma vez?
- Desta forma, a ocorrência de eventos anômalos fica diluída

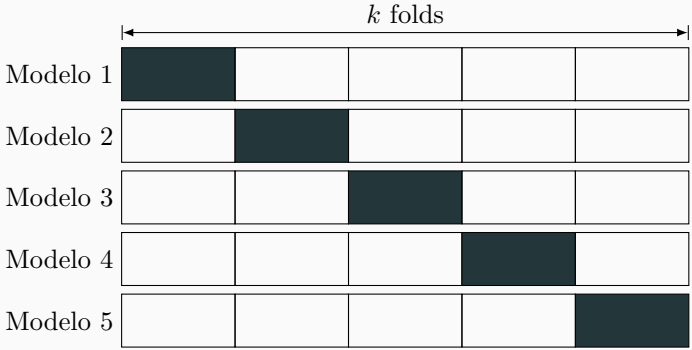
Validação Cruzada

- Validação Cruzada é um método de reamostragem (*resampling*)
- Baseia-se na ideia de tomar diversas amostras aleatórias da mesma população
- Estas amostras aleatórias são todas tomadas a partir de uma amostra que já obtivemos

Validação Cruzada

- O procedimento geral para realizar a validação cruzada é o seguinte:
- Crie k partições aleatórias do conjunto de dados com o mesmo tamanho aproximado
- Para $j = 1, \cdots, k$, faça
  1. Treine o modelo em todos os blocos, exceto o bloco j
  2. Teste o modelo no bloco j
  3. Estime o erro em cada bloco j
- Calcule a média dos erros

Validação Cruzada

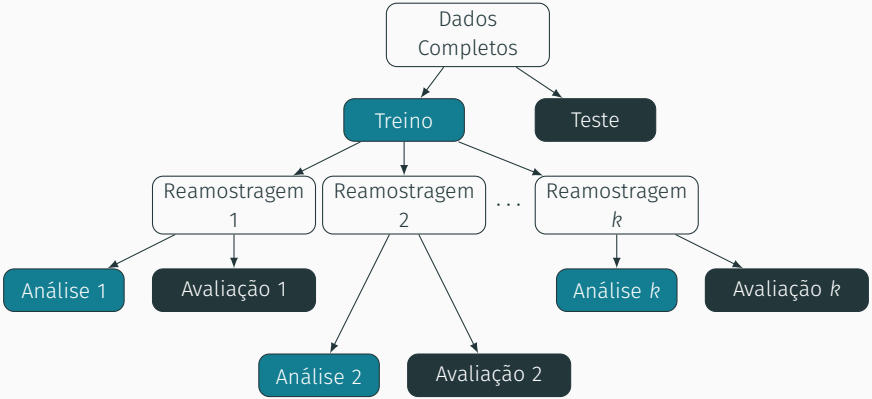


Processo Completo

Processo Completo

- A divisão dos dados em treino e teste ocorre em toda análise que formos realizar
- Encontramos o melhor modelo no conjunto de treino através da validação cruzada
- Verificamos o resultado obtido no conjunto de teste
- Ou seja, os dois métodos se complementam

Processo Completo



Validação Cruzada

EST0133 - Introdução à Modelagem de Big Data

Marcus Nunes
https://introbigdata.org/
https://marcusnunes.me/

Universidade Federal do Rio Grande do Norte