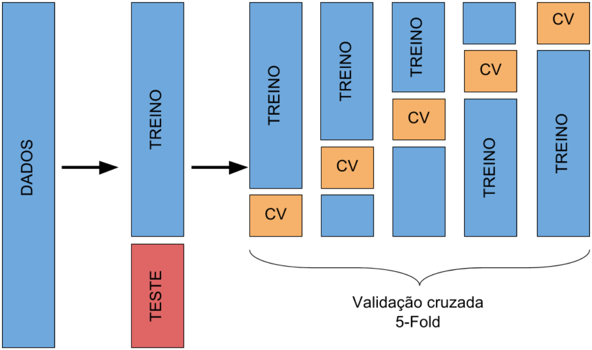
Métodos de reamostragem

Profs.: Eduardo Vargas Ferreira
Walmes Marques Zeviani

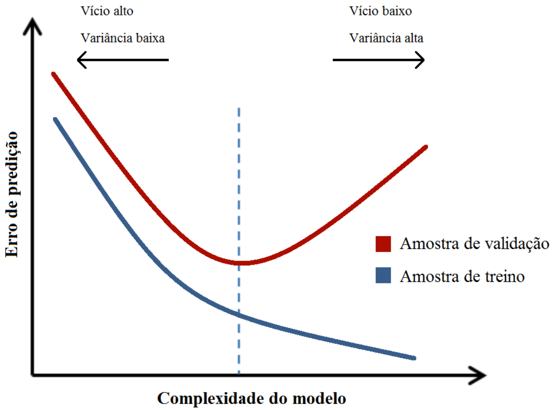
Validação cruzada e bootstrap

- Nesta seção vamos discutir dois métodos de reamostragem: validação cruzada e bootstrap;
- Tais métodos reajustam o modelo de interesse a partir de amostras dos dados de treino, a fim de obter informações adicionais sobre o modelo;
- Por exemplo, fornecem estimativas do erro de predição da amostra de teste, o vício e desvio padrão das estimativas dos parâmetros;
- Lembrando:
  - ⋆ **Erro de validação:** média do erro resultante da predição de uma nova observação (que não fazia parte dos dados de treino);
  - ⋆ **Erro do treino:** é calculado mediante aplicação do método estatístico nos dados de treino.

Separação dos dados

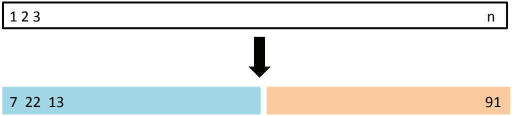


Dados de treino *x* Dados de validação



Validação cruzada holdout

- Nesta abordagem dividimos os dados em apenas duas partes: **treinamento** e **validação**;



- O modelo é ajustado utilizando os dados de treinamento, e com o modelo obtido fazemos predições nas observações dos dados de validação;

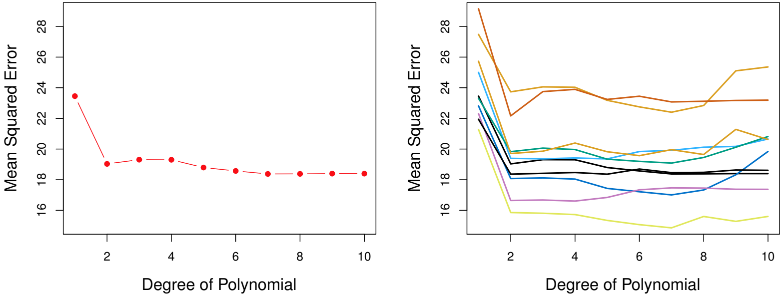
- O erro resultante dos dados de validação fornece uma estimativa do **erro do teste**, baseando-se em determinado indicador;

- Tipicamente, utilizamos o **erro quadrado médio** para variáveis quantitativas e **taxa de classificação incorreta** para as qualitativas;

Exemplo: `Auto` data set

- Queremos comparar polinômios de diferentes ordens em uma regressão linear;

- Separamos aleatoriamente as 392 observações em duas amostras: treinamento (com 196 dados) e validação (196 dados);



- O gráfico da esquerda temos **divisão única** e da direita **divisão múltipla**.

Inconvenientes desta abordagem de validação

- A estimativa do erro, dependendo do conjunto de treinamento e validação, pode ser altamente variável;

- Lembre-se, nesta abordagem apenas uma parte da amostra treino (aquela não utilizada no treinamento) é utilizada para prever o erro;

- Esse fato sugere que a estimativa do erro do teste será superestimada, pois o tamanho da amostra será pequeno com relação aos dados totais;

> Menos dados $\Rightarrow$ geralmente menos informação $\Rightarrow$ maior variabilidade

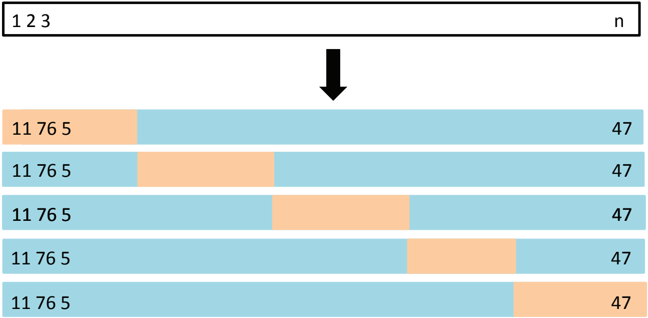
- Para minimizar esses problemas, outras abordagens de validação cruzada surgiram, uma delas é por **k-fold**;

Validação cruzada por $k - fold$

- **k-fold** é amplamente usado para estimar o erro do teste;
- As estimativas podem ser usadas para seleção do melhor modelo, e fornecer uma ideia do erro do teste para o modelo escolhido;
- O método consiste em dividir os dados em k partes iguais. Ajusta-se o modelo com $k - 1$ partes, e uma é destinada para às predições;
- Isto é feito para cada parte $k = 1, \ldots, K$, e em seguida os resultados são combinados;

1	2	3	4	5
Validation	Train	Train	Train	Train

Validação cruzada 5 – *fold*



Validação cruzada por $k - fold$

- Sejam as K partes denotadas por $C_1, C_2, \ldots, C_K$, em que C_k representa o índice da k-ésima parte;

- Considere que temos n_k observações na parte k (se n é múltiplo de K, então $n_k = n/K$);

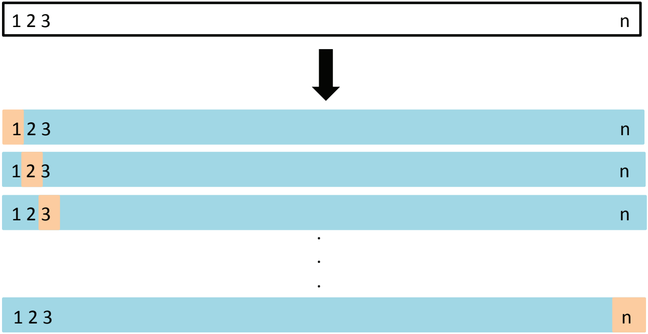
- Calcule:

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} EQM_k.$$

- O $EQM_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, e $\hat{y}_i$ é o valor ajustado da observação i, obtido dos dados com a parte k removida;

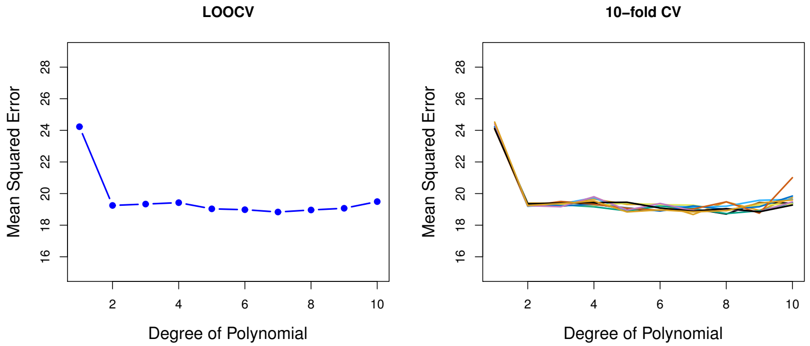
- Um caso particular é quando $K = n$ gerando o método **leave-one-out cross-validation** (LOOCV).

Leave-one-out cross-validation



Exemplo: `Auto` data set

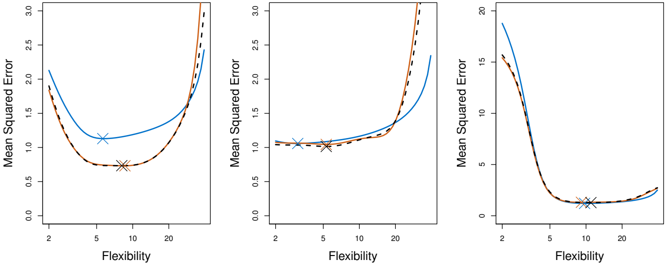
- O gráfico da direita apresenta 9 diferentes validações cruzadas 10 − *fold*. Em cada uma temos uma nova partição dos dados;



- Note que a variabilidade é menor quando comparado com a abordagem **holdout** (slide 6);

Exemplo simulado

- O gráfico abaixo apresenta a verdadeira curva do MSE em azul, a estimativa LOOCV pontilhada e $10 - $ *fold* em laranja;



- No primeiro gráfico as curvas estimadas têm o comportamento correto, todavia subestima o erro quadrado médio;

- Da mesma forma, no gráfico central, as estimativas estão muito próximas do valor real quando os graus de flexibilidade do modelo são baixos.

Bias-Variance Trade-Off

- Um bom desempenho de um método em um conjunto de teste requer um baixo erro quadrático médio. Porém, note que

$$\mathrm{E}\left[y_0 - h(\boldsymbol{x}_0)\right]^2 = \mathrm{V}ar\left[h(\boldsymbol{x}_0)\right] + \left[\mathrm{Vício}(h(\boldsymbol{x}_0)\right]^2 + \mathrm{V}ar(\varepsilon). \tag{1}$$

Variância

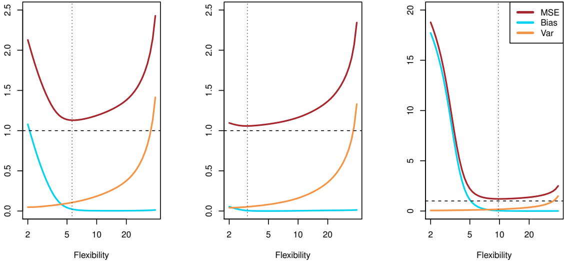
- ⋆ Refere-se ao quanto $h(\boldsymbol{x}_0)$ muda quando a estimamos utilizando diferentes dados de treino;
- ⋆ Idealmente as estimativas de $h(\boldsymbol{x}_0)$ não deveriam mudar muito entre os conjuntos;
- ⋆ Em geral, quanto mais flexível o modelo, maior a variância.

Vício

- ⋆ Refere-se ao erro de aproximar um problema real (extremamente complicado) por uma função simples;
- ⋆ Em geral, quanto mais simples o modelo, maior o vício.

Bias-Variance Trade-Off

- Os três gráficos abaixo ilustram a Equação (1):



- ⋆ A curva azul representa o $[\mathrm{Vício}(h(\boldsymbol{x}_0))]^2$, para diferentes níveis de flexibilidade;
- ⋆ A curva laranja a $\mathrm{V}ar\,[h(\boldsymbol{x}_0)]$;
- ⋆ A linha pontilhada $\mathrm{V}ar(\varepsilon)$, o erro irredutível;
- ⋆ Finalmente, a linha vermelha representa o EQM do teste.

CV em problemas de classificação

- Sejam as K partes denotadas por $C_1, C_2, \ldots, C_K$, em que C_k representa o índice da k-ésima parte;

- Considere que temos n_k observações na parte k (se n é múltiplo de K, então $n_k = n/K$);

- Calcule:

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} Err_k, \quad \text{com} \quad Err_k = \sum_{i \in C_k} \mathbb{1}(y_i \neq \hat{y}_i)/n_k.$$

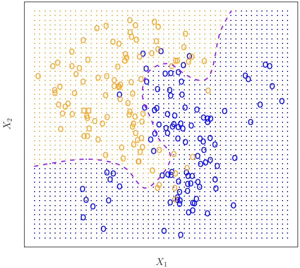
- O desvio padrão estimado do CV_k é

$$\widehat{DP}(CV_k) = \sqrt{\sum_{k=1}^{K} \left(Err_k - \overline{Err}_k\right)^2 /(K-1)}$$

- Que é uma estimativa útil, mas não muito válida (existe correlação entre os desvios, pois compartilham parte da amostra de treino).

Exemplo simulado

- Os dados consistem em 100 observações;
- Classificadas em um dos 2 grupos (indicados em azul e laranja);
- Ajustamos quatro modelos de regressão logística aos dados;
- P. ex., um modelo quadrático fica:

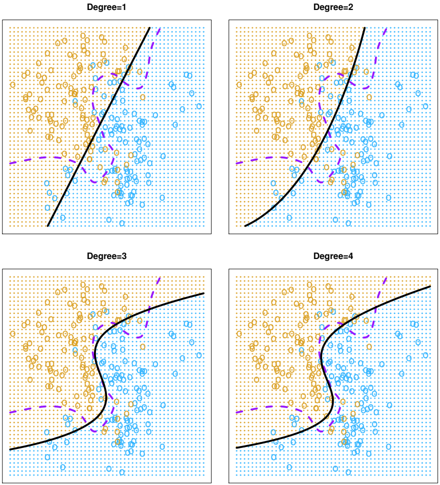


$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$$

- A verdadeira taxa de erro do teste é 0.133 (são dados simulados!).

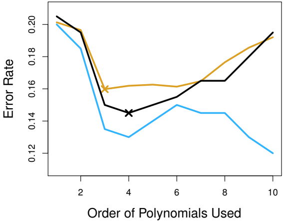
Exemplo simulado

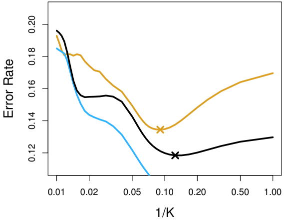
- A taxa de erro do teste para os quatro ajustes são, respectivamente, 0.201, 0.197, 0.160, e 0.162.



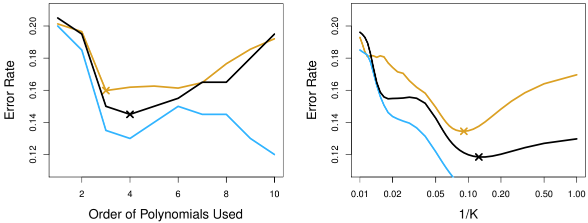
Exemplo simulado

- Na prática, em dados reais, a verdadeira taxa de erro do teste são desconhecidas;
- Então, como decidir entre os modelos propostos? **Validação cruzada**;
- Abaixo temos a taxa de erro por validação cruzada $10 - fold$ em preto, o verdadeiro erro do teste em marrom e o erro do treinamento em azul;
- Na esquerda temos o classificador por regressão logística. E na direita, utilizamos *KNN* (nas abscissas temos o inverso do número de vizinhos).





Exemplo simulado



- Analisando os gráficos percebemos que a taxa de erro do treinamento decresce com o aumento da flexibilidade do modelo;
- Na esquerda, a taxa de erro do teste apresenta uma forma de U. Semelhante à validação cruzada $10 -$ *fold* (com uma boa aproximação);
- Na direita, a taxa do erro por validação cruzada também apresenta um mínimo muito próximo do obtido com os dados de teste.

CV: certo e errado

- Considere um classificador aplicado aos dados de duas classes:

  1. Começando com 500000 preditoras e amostra de tamanho 50, filtramos as 100 preditoras que têm maior correlação nas classes;

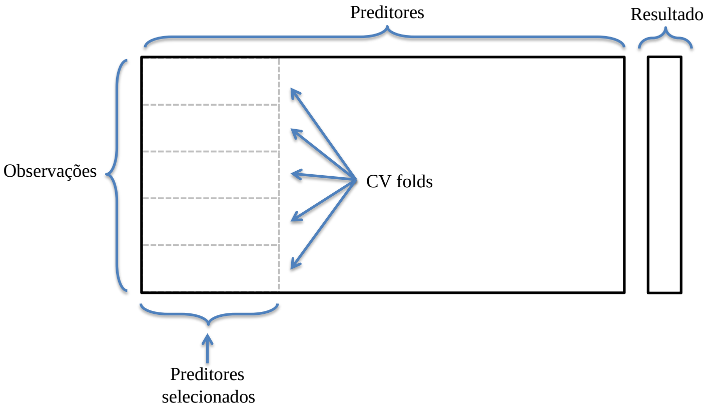
  2. Aplicamos um classificador (e.g. regressão logística) utilizando somente as 100 preditoras.

- Como podemos estimar o desempenho do teste para este classificador? **Validação cruzada**

- Podemos aplicar validação cruzada no Passo 2, esquecendo o Passo 1 (não incorporando o fato de termos eliminado 4900 preditoras)? **Não!**

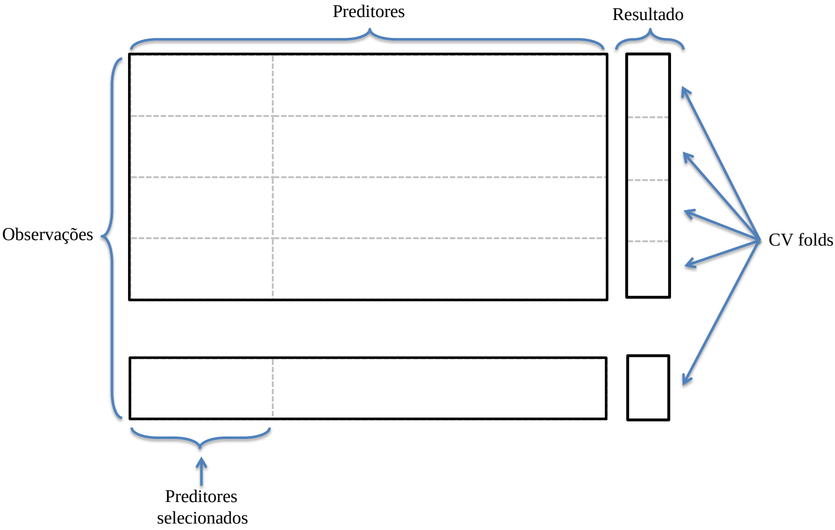
  - ⋆ Isso seria ignorar o fato de que no Passo 1 o procedimento já viu os rótulos de treinamento, e aprendeu com isso.

Errado!



Preditores
Resultado
Observações
CV folds
Preditores
selecionados

Certo!



Preditores
Resultado
Observações
CV folds
Preditores
selecionados

Bootstrap

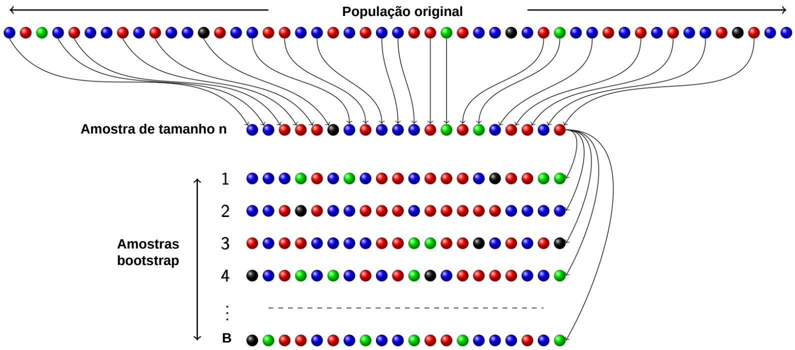
- **Bootstrap** é uma ferramenta utilizada para quantificar a incerteza associada a determinado estimador ou método de aprendizagem;

- Por exemplo, pode ser usado para estimar o erro padrão dos coeficientes do ajuste de uma regressão linear (ou seus intervalos de confiança);

- Na verdade, o poder do bootstrap reside no fato de ser aplicável em uma vasta gama de métodos estatísticos;

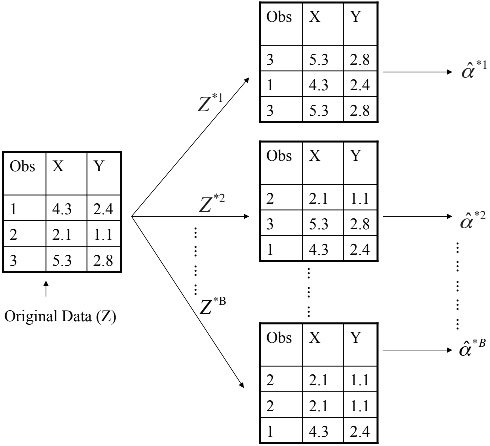
- Incluindo alguns para os quais uma medida de variação é difícil de obter (e não é fornecido automaticamente pelo software);

Ideia básica do bootstrap



Exemplo com somente 3 observações

- Cada conjunto de dados bootstrap contém n observações amostradas **com reposição** dos dados originais.



Bootstrap estimando o erro de predição

- Na validação cruzada, o K- ésimo *fold* de validação é distinto dos demais $k-1$ *folds* usados no treinamento;

1	2	3	4	5
Validation	Train	Train	Train	Train

- Não há **overlap** entre os dados de treino e validação. O que é crucial para seu sucesso. Queremos uma ideia sobre os dados de teste (novos dados);
- Para estimar o erro utilizando bootstrap, podemos pensar em separar cada amostra bootstrap para treinamento e a original como validação;
- Entretanto, tais amostras apresentam um significativo *overlap* com os dados originais (cerca de 2/3);
- Este fato causa sérios problemas de subestimação do verdadeiro erro de predição (podendo ser parcialmente solucionado - não veremos no curso);
- Então, validação cruzada apresenta uma abordagem mais simples e atrativa para estimar o erro de predição (Keep it simple!).