Exercício 2

Aluno: Deivison Venicio Souza

Uma preocupação constante na aplicação de métodos de aprendizado de máquinas é garantir que o modelo final faça boas generalizações. Para tanto, uma estratégia comum é a divisão de dados em conjuntos de treino (com fração de treino e fração de avaliação) e teste.

Suponha que tenhamos dois modelo em competição:

$$h_1(x) = b_0 + b_1 x,$$
$$h_2(x) = b_0 + b_1 x + b_2 x^2.$$

Considerando a função custo:

$$J(y_i, h(\mathbf{x})) = \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{h}(x_i)]^2$$

Descreva como decidir sobre o melhor modelo, detalhando onde cada conjunto de dados será utilizado e como. Utilize o método de 5-fold.

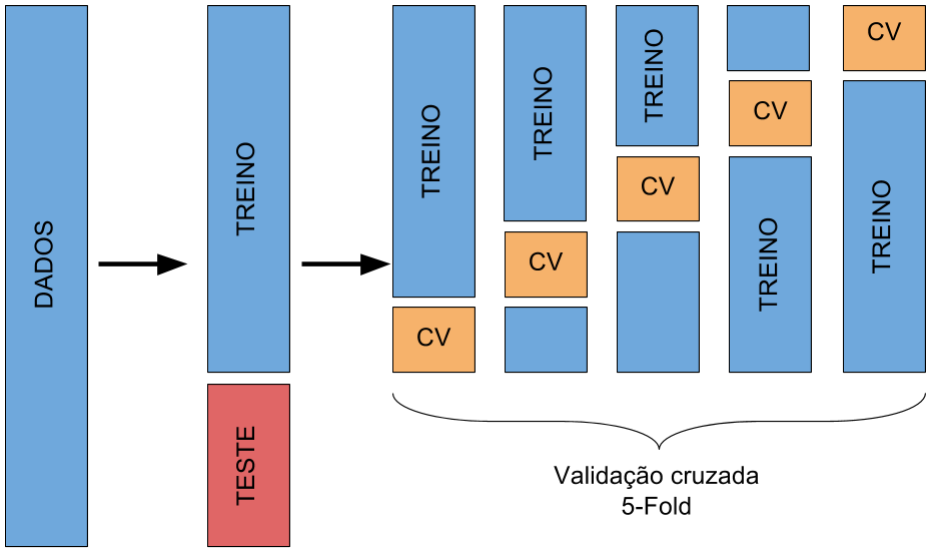

Figura 1: Validação cruzada 5-fold

R: Uma abordagem comum para estimar o desempenho esperado de um modelo preditivo é implementar algum método de reamostragem (resampling method) dos dados originais. A literatura reporta várias técnicas de reamostragem dos dados originais: a) k-fold cross-validation, leave-one-out cross-validation, repeated k-fold cross-validation, bootstrap. A figura acima retrata o k-fold cross-validation, em que o valor k=5.

Para avaliar o desempenho dos modelos, ou seja, aquele com melhor capacidade de generalização sobre dados não usados no processo de aprendizagem, poderíamos seguir os seguintes passos:

- Inicialmente, realizaria a divisão do conjunto de dados em: conjunto de treinamento (aprendizagem) e conjunto de teste. Uma proporção bastante usada é dividir 80% (treino) e 20% (teste). Porém, não é regra geral e dependerá da quantidade de dados disponível;

- Em seguida, ambos os modelos seriam aprendidos usando o conjunto de treinamento. Para tanto, seria utilizado o método de reamostragem 5-fold CV. Usando a abordagem 5-fold CV ter-se-ia os seguintes procedimentos:

a) Inicialmente, o conjunto de treinamento seria dividido em 5 subgrupos (*k=5*) aproximadamente iguais (N/5) e distintos garantindo, assim, que cada subgrupo seja um bom representante do todo;
b) Em seguida, os k-1 subgrupos seriam usados para treinar ou aprender os modelos, e o subgrupo mantido de fora (*hold-out set*) seria usado para obter estimativas imparciais do desempenho. O processo continuaria iterativamente, até que cada subgrupo seja usado exatamente uma vez para validar um modelo aprendido;
c) Ao final do procedimento 5-fold CV ter-se-ia 5 estimativas da capacidade de generalização (medida por meio da função perda quadrática) para ambos os modelos. O desempenho final dos modelos seria obtido pela média aritmética das estimativas obtidas em cada subgrupo da validação.

Finalizado o procedimento de treinamento do modelo com uso do método 5-fold CV podemos decidir pelo melhor modelo observando aquele com melhor capacidade de generalização, ou com menor valor para a função custo. Em geral, as métricas utilizadas para tanto são RMSE (do inglês, *Root Mean Square Error*), rRMSE (do inglês, *Relative Root Mean Square Error*) e R^2 (do inglês, *R-squared*).

Enfim, supondo que o melhor foi h1(x) = b0 + b1x deve-se validar seu poder preditivo sobre o conjunto de teste (não utilizado na fase de treinamento do modelo). Finalmente, um modelo final poderia ser treinado com todo conjunto de dados.