1 Seleção e Avaliação de Modelos

1. Descreva o procedimento geral de holdout visando medir a capacidade de generalização do modelo. Ilustre com a regressão logística. Mostre as funções de custo no treino e no teste.

2. O que é o erro de classificação 0/1? Defina matematicamente.

3. O que é estratificação nos conjuntos de treino e teste?

4. Descreva o procedimento geral para a obtenção de hiperparâmetros ótimos. Defina os custos de treino, validação e teste.

5. Descreva o procedimento de divisão treino/teste por *K-fold*.

6. Descreva o procedimento de divisão treino/teste por *leave-one-out*.

7. Descreva o procedimento de divisão treino/teste por *bootstrap*.

8. Explique como a curva do custo no treino e na validação para diferentes valores de um hiperparâmetro (como o de regularização) permite o diagnóstico de *under/overfitting*. Mostre graficamente.

9. Explique como a curva de aprendizado em função do tamanho do treino permite o diagnóstico de *under/overfitting*. Mostre graficamente.

10. O que é análise de erro em um modelo de *machine learning*?

11. Explique porque a acurácia não é uma boa medida em problemas com classes desbalanceadas.

12. Descreva matematicamente (fórmula) e conceitualmente (significado intuitivo) o conceito de *precision*.

13. Descreva matematicamente (fórmula) e conceitualmente (significado intuitivo) o conceito de *recall*.

14. Descreva como é construída e qual a utilidade de uma cuva de *precision/recall*.

15. Defina matematicamente o F_1-*score* e explique sua importância.

16. A afirmação "O vencedor não é o melhor algoritmo, mas sim quem tem mais dados" é clássica em *big data*, mas ela pressupõe duas condições. Quais são elas?