2ª Lista de Exercícios - EST5534 Aprendizado de Máquina Estatístico

---

Exercício 1. Obtenha um classificador de Bayes ingênuo (Naive Bayes) no caso das flores iris. Obtenha métricas de desempenho considerando a divisão da base em treinamento (80%) e teste (20%). Você pode seguir os códigos disponíveis em

`https://github.com/cibelerusso/Aprendizado\_de\_Maquina.`

Varie o tamanho das bases de treinamento e teste e verifique o efeito em métricas de classificação.

Exercício 2. Obtenha um classificador por regressão logística para os dados amostra_banco.csv, disponíveis em

`https://github.com/cibelerusso/IntroducaoaInferenciaEstatistica/tree/main/Dados .`

Obtenha métricas de desempenho considerando a divisão da base em treinamento (80%) e teste (20%). Você pode seguir os códigos disponíveis em

`https://github.com/cibelerusso/Modelos-de-Regressao.`

Varie o tamanho das bases de treinamento e teste e verifique o efeito em métricas de classificação.

Exercício 3. Considere um problema de classificação para um conjunto de dados $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$, em que $y_i \in C = \{c_1, c_2, \ldots, c_k\}$, utilizando a perda 0-1, $L(y, \widehat{g}(\mathbf{x})) = \mathbb{I}(Y \neq g(\mathbf{x}))$.

Proponha um breve estudo de simulação para comparar os classificadores obtidos pelos métodos Naive Bayes e Regressão Logística, seguindo os passos:

(1) Gere um conjunto de dados fictício com características ($\mathbf{x}_i$) e rótulos (y_i). O conjunto de dados deve conter duas classes (binário), por exemplo, "spam"e "não spam", ou "inadimplente"e "adimplente". Você pode se inspirar em dados como iris ou amostra_banco, que podem ser obtidos em `https://github.com/cibelerusso/Aprendizado_de_Maquina/tree/main/Dados` ou `https://github.com/cibelerusso/IntroducaoaInferenciaEstatistica/tree/main/Dados` ou criar outros dados de sua preferência.

(2) **Treinamento do Modelo** Divida o conjunto de dados em conjuntos de treinamento e teste (por exemplo, 80% para treinamento e 20% para teste).

(3) Treine um modelo de Naive Bayes e um modelo de Regressão Logística no conjunto de treinamento.

(4) **Avaliação do Desempenho**: Use os modelos treinados para fazer previsões no conjunto de teste.

(5) Calcule a perda 0-1 para cada modelo em relação aos rótulos verdadeiros ($L(y, \widehat{g}(\mathbf{x})) = \mathbb{I}(Y \neq g(\mathbf{x}))$).

(6) Calcule a taxa de erros (taxa de classificação incorreta) para cada modelo. Utilize outras métricas de desempenho.

(7) **Repetição**: Repita as etapas 1 a 6 um número definido de vezes (por exemplo, 100 vezes) para obter médias e desvios padrão das taxas de erro para cada modelo.

(8) **Análise dos Resultados**: Compare as médias das taxas de erro entre o modelo de Naive Bayes e o modelo de Regressão Logística.

Exercício 4. Por que métricas como o erro quadrático médio (EQM) e o erro absoluto médio (EAM) não são adequados para avaliar o desempenho de modelos em problemas de classificação? Explique as razões e forneça exemplos que ilustrem por que essas métricas são inapropriadas.

Exercício 5. Considerando um problema de classificação com duas classes de resposta, podem-se usar várias métricas para avaliar o desempenho de um modelo. Explique as principais diferenças entre as métricas de precisão, revocação (*recall*), F1-score, acurácia e AUC (área sob a curva ROC). Quando você usaria cada uma delas e quais são suas vantagens e limitações?

Exercício 6. Considere um problema de classificação com k classes de resposta. Como definir as métricas de desempenho precisão, revocação (*recall*), F1-score, acurácia e AUC (área sob a curva ROC)? Quais são mais facilmente aplicáveis nesse tipo de problema?

Exercício 7. Pesquise outras possíveis métricas de desempenho em problemas de classificação e apresente referências bibliográficas para elas.

Exercício 8. Considere os dados iris. Obtenha um classificador para a espécie das flores usando o método KNN. Você pode tomar como base os códigos `Aprendizado_Supervisionado_KNN.ipynb` disponíveis em
`https://github.com/cibelerusso/Aprendizado_de_Maquina/blob/main/Codigos%20em%20Python/`
Escolha o número de vizinhos pelo método da validação cruzada.

Exercício 9. Considere os dados iris. Desenvolva códigos para obter um classificador para a espécie das flores usando o método Nadaraya-Watson.

Exercício 10. Considere os dados `spam.txt` disponíveis em
`http://www.rizbicki.ufscar.br/dados/spam.txt` . Obtenha classificadores baseados na regressão logística, Naive Bayes e KNN e compare-os utilizando métricas de desempenho, considerando bases de treinamento (80%) e teste (20%).