Fundamentos de Mineração de Dados e Ciência de Dados

Projeto Final

O objetivo desse projeto final é analisar o desempenho esperado de uma técnica de Aprendizado de Máquina, tanto em dados vistos (conjunto de treino) quanto não vistos (conjunto de teste). Para tal, você deve escolher um método de aprendizado, definir (fixar) seus hiperparâmetros de antemão, treinar o método e testá-lo diversas vezes (via reamostragem), de modo a obter uma distribuição da estimativa de seu desempenho, tanto de treino quanto de teste. Atente para o fato de que nenhuma comparação ou escolha é feita a partir dos dados, ou seja, o experimento de treino e teste será repetido várias vezes apenas para que se obtenha tal distribuição, não havendo, portanto, a necessidade de um conjunto de validação.

Comece então escolhendo um conjunto de dados (disponíveis em https://archive.ics.uci.edu/datasets/), dentre os seguinte, com o qual irá trabalhar:

1. Breast Cancer Wisconsin (Diagnostic)
   https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
   569 instâncias, 30 features. Desbalanceamento: 63/37
2. MAGIC Gamma Telescope
   https://archive.ics.uci.edu/dataset/159/magic+gamma+telescope
   10020 instâncias, 10 features. Desbalanceamento: 65/35
3. Blood Transfusion Service Center
   https://archive.ics.uci.edu/dataset/176/blood+transfusion+service+center
   748 instâncias, 4 features. Desbalanceamento: 76/24
4. Bank Marketing
   https://archive.ics.uci.edu/dataset/222/bank+marketing
   45k instâncias, 16 features. Desbalanceamento: 89/11

Atente para o fato de que alguns desses conjuntos já vêm divididos em conjuntos de treino e teste. Nesses casos, você deve unir esses conjuntos em um único conjunto de dados, antes de prosseguir com seus experimentos.

Na sequência, você deve produzir um notebook python descrevendo seu experimento, desde seu planejamento até sua execução. O notebook deverá necessariamente conter os seguintes passos:

Definição do objetivo

- Qual o objetivo da pesquisa (formule como um problema de classificação)?

- Quais os fatores considerados, e suas variáveis correspondentes?
- Qual a variável-alvo (variável dependente) e quais as independentes?

Caracterização da base

Descreva a base, dando estatísticas básicas como número de variáveis, quantidade de instâncias etc. As seguintes questões devem ser abordadas:

- Quais os tipos de dados de cada variável?
- Qual a distribuição dos valores da variável-alvo?
- Há necessidade de pré-processamento? Qual?
- Que limitações há nessa base, com relação, por exemplo, ao método de amostragem usado para sua criação, aos intervalos de dados etc.?

Metodologia

- Como se dará a manipulação dos dados ao longo do processo? Como será feita a divisão dos dados, em quais conjuntos, e como cada conjunto participará da pesquisa?
- Se necessário pré-processamento, como será feito? Atentem para vazamentos.
- Qual o algoritmo de aprendizado utilizado?
- Como este será utilizado? (Onde será treinado e onde será testado?)
- Como será obtida a distribuição do desempenho em dados vistos (treino) e não vistos (teste)? Que método de reamostragem foi escolhido? (Escolha entre *Repeated Holdout* e *Bootstrapping*)

Como medida de desempenho, você deve usar tanto acurácia quanto macro-F1. Discuta como irá interpretar seus valores (ou seja, faça uma discussão acerca do que cada uma mede e de como elas se complementam).

Descrição dos Resultados

Aqui você deve descrever os resultados do experimento, discutindo-os brevemente.

Para cada medida de desempenho adotada (acurácia e macro-F1), devem constar do notebook:

- Qual o valor esperado do desempenho do método no conjunto de treino e no de teste? (Forneça média e desvio padrão)
- Apresente um boxplot para o desempenho do modelo nos conjuntos de treino e outro para o desempenho nos conjuntos de teste ao longo das repetições feitas (coloque ambos boxes no mesmo gráfico).
- Crie um histograma com a distribuição do desempenho do método nos conjuntos de treino ao longo das repetições feitas.
- Crie um histograma com a distribuição do desempenho do método nos conjuntos de teste ao longo das repetições feitas.
- Compare as distribuições e boxplots no treino e teste. Comente suas observações.

Por fim, compare os resultados obtidos nas duas medidas de desempenho adotadas, discutindo diferenças e semelhanças entre elas.

Apresentação ao aluno: Quarta-feira, 12/11/25, durante os exercícios da aula (11:00)
Entrega pelo aluno: Segunda-feira, 17/11/25, até as 17:00 (17:30 tolerância)
Dúvidas: Sexta-feira, 14/11/25, no horário do exercício prático.