Machine Learning
CPS 863
Terceiro Trimestre de 2024

Professor: Edmundo de Souza e Silva

Lista de Exercícios 2

ATENÇÃO!

- Faça as listas de forma que TODAS AS RESPOSTAS sejam DEVIDAMENTE COMENTADAS (passos para se chegar a resposta).
- A entrega da lista deve ser feita em UM ÚNICO arquivo PDF. Não envie vários pedaços separadamente!
- ATENÇÃO! Faça as listas de forma que TODAS AS RESPOSTAS sejam DEVIDAMENTE COMENTADAS (passos para se chegar a resposta).
  Não procure a solução na Internet ou em livros ou no chatGPT, pois o objetivo é que você mesmo avalie o que sabe. Obviamente, caso você já tenha conhecimento do problema, não leia a resposta (mesmo que já conheça o resultado final) e tente fazer sozinho. Só assim você poderá ter uma ideia melhor dos tópicos que você ainda não domina com desenvoltura.
- Anote as dúvidas encontradas para resolver **sozinho**. Em classe gostaria de saber quais as dúvidas que cada um teve para resolver o problema sem olhar a resposta.
- Qualquer referência a código é MUITO menos importante do que a EXPLICAÇÃO DOS PASSOS que foram realizados. O que mais importa é a explicação de como se chegou na solução.
- Para facilitar escrever a lista de forma clara, é possível traduzir equações a mão para LaTex: https://mathpix.com/, ver também https://www.overleaf.com/learn/latex/Questions/Are_there_any_tools_to_help_transcribe_mathematical_formulae_into_LaTeX%3F

É dado do problema um dataset que contém 3 features e 1000 amostras. Além disso, há uma coluna extra que indica a classe que pertence cada amostra (4 classes). O objetivo é analisar o conjunto de dados, treinar modelos de mistura gaussiana e executar imputação de dados.

Questão 1

Responda as seguintes perguntas usando o dataset fornecido:

1. *Visualize* os dados (em 2 ou 3 dimensões) para entender a estrutura dos dados.
   Explique o que você fez para visualizar as figuras. Descreva sua abordagem para visualizar os dados e quais insights podem ser obtidos do gráfico, se algum.
2. Ajuste uma mistura de gaussianas com 4 componentes ao conjunto de dados. Calcule TODOS os parâmetros necessários para o modelo. Explique todas as etapas. Forneça detalhes de como você determina os parâmetros de melhor ajuste para cada modelo de mistura e descreva o processo de ajuste do modelo.
3. Suponha que as classes 1 2 sejam uma mesma classe. Ajuste uma mistura de gaussianas com 3 componentes ao conjunto de dados.
4. Suponha que as classes 1 2 e 3 sejam uma mesma classe. Ajuste uma mistura de gaussianas com 2 componentes ao conjunto de dados.

5. Mostre como estimar qual dos 3 modelos (2, 3 ou 4 gaussianos) **melhor representa o conjunto de dados original, ignorando as classes do modelo**. Justifique como chegou ao resultado, usando técnicas de avaliação de modelos como AIC, BIC ou outro critério. Estude e explique o que é AIC e BIC.
6. Gere novas amostras com base no melhor modelo pelo seu critério. Plote os resultados. Visualize as amostras geradas e compare-as com o conjunto de dados original.

Questão 2

É dado um novo dataset com vários dados faltantes.

1. Descreva as etapas para executar a imputação de dados para as amostras incompletas.
   Explique sua abordagem. Descreva (mostre a matemática) que você usou para preencher os valores ausentes.
2. Execute os cálculos necessários para imputar os valores ausentes.
   Forneça os detalhes matemáticos e demonstre o processo de imputação de valores ausentes para o conjunto de dados fornecido.