Machine Learning
CPS 863
Terceiro Trimestre de 2024

Professor: Edmundo de Souza e Silva

Lista de Exercícios 3

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

Questão 1

Neste trabalho , você irá ajustar e avaliar três modelos diferentes em um conjunto de dados com três features:

1. **GaussI**: Um modelo de mistura de Gaussianas (GMM) com uma Gaussiana por classe, onde as matrizes de covariância são todas iguais à matriz identidade, i.e., $p(\mathbf{x}|y=c) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \mathbf{I})$.
2. **GaussX**: Um modelo de mistura de Gaussianas (GMM) com uma Gaussiana por classe, sem restrições nas matrizes de covariância, i.e., $p(\mathbf{x}|y=c) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.
3. **LogReg**: Um modelo de regressão logística com características lineares e quadráticas, i.e., função polinomial de grau 2.

A questão inclui dois conjuntos de dados: um conjunto de treino e um conjunto de teste. Cada amostra possui três features. Siga os passos a seguir para cada modelo:

1. Calcule a *log-likelihood*, de forma literal para cada modelo (explique como calcular os parâmetros de cada).
2. Para cada modelo (GaussI, GaussX e LinLog), obtenha os parâmetros usando o conjunto de treino.
3. Para cada modelo (GaussI, GaussX e LinLog), calcule a *log-likelihood* usando do conjunto de teste e os parâmetros obtidos.
4. Avalie o desempenho de cada modelo usando o conjunto de teste e compare os resultados. Discuta qual modelo apresentou o melhor desempenho e tente dar a sua explicação sobre o motivo.

Questão 2

Nesta questão, você usará o classificador Naive Bayes para classificar mensagens SMS como spam ou ham (não spam) usando o conjunto de dados SMS Spam Collection. Esse conjunto de dados contém uma série de mensagens SMS etiquetadas como spam ou ham e será utilizado para treinar e avaliar o desempenho do modelo Naive Bayes.

Objetivos: O objetivo deste exercício é:

- Treinar um classificador Naive Bayes para classificar mensagens de texto.
- Avaliar o desempenho do modelo em um conjunto de teste.
- Discutir o impacto da suposição de independência do Naive Bayes e como ela afeta os resultados.

Dataset O conjunto de dados que será utilizado é o "SMS Spam Collection", disponível no Repositório de Aprendizado de Máquina da UCI. Você pode baixá-lo do link abaixo: https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

O conjunto de dados é composto por:

- **Coluna 1**: A etiqueta ("spam"ou "ham").
- **Coluna 2**: A mensagem SMS em texto.

Siga os passos a seguir para realizar o trabalho.

Preparação dos Dados:

- Carregue o conjunto de dados e converta as etiquetas para formato binário: "ham" = 0 e "spam" = 1.
- Divida o conjunto de dados em um conjunto de treino (70%) e um conjunto de teste (30%).
- Utilize o modelo de bag-of-words para transformar o texto das mensagens em uma representação numérica.

Passo 2: Treinamento do Modelo

- Treine um classificador Naive Bayes multinomial usando o conjunto de treino.

Responda às seguintes perguntas com base nos resultados obtidos:

1. Use o modelo treinado para prever se as mensagens do conjunto de teste são spam ou ham.
2. Calcule a precisão (accuracy), precisão (precision), revocação (recall) e a pontuação F1 (F1-score) para o conjunto de teste.
3. Explique como o modelo Naive Bayes classifica uma mensagem como spam ou ham. Por que o Naive Bayes pode ser eficaz mesmo assumindo independência entre as palavras?
4. Analise as métricas de avaliação (precisão, revocação, F1-score) obtidas. O modelo foi capaz de detectar bem as mensagens spam? Explique com base nas métricas.
5. O Naive Bayes faz uma suposição de independência entre as palavras da mensagem. Discuta como essa suposição pode afetar a classificação de mensagens. Por que, apesar dessa suposição, o modelo ainda pode ter uma boa performance?
6. Discuta um cenário em que a suposição de independência do Naive Bayes pode prejudicar significativamente a precisão do modelo.