Atividade 2 - Classificação com KNN

Prof. Dr. Juliano Henrique Foleis

Descrição da Atividade

Nesta atividade você vai implementar um sistema de classificação usando o classificador KNN. Sua implementação deve ser feita em Python em um caderno no Jupyter.

Nesta atividade vamos trabalhar com um subconjunto da base de dados Water Quality. Esta base de dados contém medidas físico-químicos de 2011 amostras de água. O objetivo é classificar cada amostra como sendo potável ou não-potável. Este subconjunto da base de dados não contém os dados da base original que possuem algum valor faltante.

Documente cada um dos passos indicados a seguir no Jupyter:

1. Explore a base de dados. Conheça a distribuição dos atributos de entrada em relação aos atributos de saída. A partir da exploração é possível ter uma idéia de quais atributos podem conter mais informação útil para realizar a classificação. Algumas idéias: fazer histogramas pra cada atributo, calcular as estatísticas descritivas, visualizar o espaço de características plotando pares de atributos, etc.
2. Visualize o espaço formado pelo conjunto de atributos selecionados no passo **1**. Use PCA para reduzir a dimensionalidade.
3. Avalie o desempenho do classificador KNN usando validação cruzada em dois níveis, conforme discutimos em sala. A validação cruzada no primeiro deve ser em 10 vias, enquanto no segundo nível deve ser em 5 vias. **Dica:** no primeiro nível você deve usar `StratifiedKFold` para gerar os particionamentos, e no segundo nível você deve usar `GridSearchCV`. A validação cruzada no segundo nível deve selecionar o melhor k. Utilize a métrica acurácia para avaliar o desempenho do classificador em ambos níveis. Para avaliar cada particionamento durante a validação cruzada não se esqueça de normalizar os dados de cada particionamento separadamente.

Em vários dos passos acima existem muitas decisões que podem ser tomadas que afetam o desempenho dos classificadores. Justifique suas escolhas. Experimente variações e tente desenvolver um sistema que acerte o máximo possível!

Instruções e Entrega

- A maioria dos passos acima estão prontos nos cadernos das Semanas 3 e 4 disponibilizados no GitHub.
- Capriche no seu *notebook*: coloque textos explicativos, faça gráficos que julgar necessário, etc. Aproveite para aprender como usar as ferramentas!
- A atividade deve ser feita em um Jupyter Notebook. Você pode usar o *Google Colab* se quiser, mas é necessário entregar o arquivo *.ipynb*.
- A entrega deverá ser realizada via Moodle, na *Atividade 2*.
- **Prazo para entrega:** 3/5/2022 às 23:55.
- Esta atividade deve ser realizada individualmente.
- Não é permitido alterar o arquivo que contém a base de dados (**water_potability_nonans.csv**)!