Atividade 3 - Seleção de Modelos com Teste-T

Prof. Dr. Juliano Henrique Foleis

Descrição da Atividade

Nesta atividade você vai implementar um sistema de classificação usando os classificadores KNN e SVM. Sua implementação deve ser feita em Python em um caderno no Jupyter.

Nesta atividade vamos trabalhar com um subconjunto da base de dados MNIST database of handwritten digits. Este subconjunto da base de dados (disponível no Moodle) contém 1500 imagens em escala de cinza de tamanho 28x28 (linearizadas em vetores de 784 pixels). As imagens são de dígitos 0-9 manuscritos. O objetivo do sistema inteligente é classificar cada imagem nas classes que representam os dígitos 0-9. As primeiras 784 colunas da base de dados são os pixels da imagem (valores de 0-255). A ultima coluna representa a variável de saída, ou seja, o dígito que corresponde à imagem. Existem muitas técnicas para gerar descrições de imagens que podem ser usadas para representá-las em sistemas de classificação. Entretanto, neste sistema você deve usar os valores de todos os pixels diretamente para representar as imagens no sistema.

Documente cada um dos passos indicados a seguir no Jupyter:

1. Visualize o espaço formado pelo conjunto de atributos, ou seja, as 784 colunas. Use PCA para reduzir a dimensionalidade.
2. Avalie o desempenho do classificador KNN usando validação cruzada em um nível, conforme discutimos em sala. A validação cruzada deve ser em 10 vias. **Dica:** você deve usar `StratifiedKFold` para gerar os particionamentos. Você deve otimizar o hiperparâmetro `k`, conforme discutimos em sala. Utilize a métrica acurácia para avaliar o desempenho do classificador. Para avaliar cada particionamento durante a validação cruzada não se esqueça de normalizar os dados de cada particionamento separadamente.
3. Avalie o desempenho do classificador SVM usando validação cruzada em um nível, conforme discutimos em sala. A validação cruzada deve ser em 10 vias. **Dica:** você deve usar `StratifiedKFold` para gerar os particionamentos. Você deve otimizar os hiperparâmetros `C` e γ (gamma), conforme discutimos em sala. Utilize a métrica acurácia para avaliar o desempenho do classificador. Para avaliar cada particionamento durante a validação cruzada não se esqueça de normalizar os dados de cada particionamento separadamente.
4. Selecione o classificador que obteve o melhor desempenho usando o teste de hipótese nula baseado no teste-t. Caso a diferença entre os classificadores não seja significativa, indique que a hipótese nula não pode ser refutada.

Instruções e Entrega

- A maioria dos passos acima estão prontos nos cadernos das Semanas 3, 4 e 6 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*. Caso coloque código em arquivos *.py* por favor entregar junto com o *.ipynb* em um arquivo *.zip*.
- A entrega deverá ser realizada via Moodle, na *Atividade 3*.

- **Prazo para entrega:** 24/5/2022 às 23:55.
- Esta atividade deve ser realizada individualmente.
- Não é permitido alterar o arquivo que contém a base de dados (**mini_mnist.csv**)!