Implementação do Algoritmo KNN

Aluno- Ricardo Matsumura

Foram escolhidas três bases de dados.

Spambase, uma base de dados que classificou emails entre ‘Spam’ e ‘Não-Spam’, utilizando 57 atributos (features) em 4601 instâncias. Possui dados faltantes, porém está normalizada. A correção dos dados faltantes foram excluídos no algorítmo através da função dropna().

Qusar Androgen Receptor é uma base de dados que classificou 1687 produtos químicos em duas classes, ‘Andrógeno aglutinante’ e ‘Andrógeno não-algutinante’. Foram utilizados 1024 atributos binários. A base está normalizada.

Wine Quality é uma base de dados originada por testes fisico-químicos em vinhos, para determinar sua qualidade. Possui 11 atributos e não está normalizada. Foram utilizados 4898 instâncias, e não tem valores faltantes. A normalização foi feita utilizando a função scaler.fit\_transform() do MinMaxScaler presente no pacote sklearn.preprocessing. A razão deste último teste ter obtido baixa acurácia, provavelmente se deve à presença de muitas classes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algoritmo |  | Acurácia | | |
|  |  | k = 3 | k=5 | k=7 |
| Knn-spam.py | Base de dados para identificação de e-mails spam | 88,55% | 88,62% | 88,19% |
| Knn-qsar.py | Base de dados de reagentes químicos | 90,32% | 93,68% | 90,12% |
| Knn-iris.py (exemplo) | Base de dados de identificação de flores | 93,33% | 100% | 97,78% |
| Knn-wine | Base de dados de qualidade de vinho | 54,17% | 59,38% | 53.96% |

Repositórios das bases utilizadas.

[QSAR androgen receptor - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/509/qsar+androgen+receptor)

[Spambase - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/94/spambase)

[Wine Quality - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/186/wine+quality)