

Métodos Matemáticos em Engenharia Química

Programa de Pós-Graduação em Engenharia Química - PPGEQ

Trabalho 01: Classificação Binária

Prof. Éliton Fontana

O *sklearn* possui várias funções para gerar conjuntos de dados que podem ser utilizados para testar os algoritmos de classificação. Escolha uma das seguintes funções para gerar um conjunto de dados com dois parâmetros (armazenados na matriz *X*) e com classe 0 ou 1 (armazenado no vetor *y*).

- *make_classification*:

```
from sklearn.datasets import make_classification
```

```
X, y = make_classification(n_features=2, n_redundant=0, n_informative= a,  
n_clusters_per_class= a, n_samples = b, random_state = c)
```

Para o parâmetro *a*, utilize o valor 1 ou 2. Para o parâmetro *b* (número de pontos), utilize um valor entre 200 e 500. Para o parâmetro *c*, utilize algum valor inteiro aleatório com 4 ou 5 dígitos. Isto irá garantir que o conjunto gerado seja sempre o mesmo, permitindo uma comparação dos resultados.

- *make_gaussian_quantiles*

```
from sklearn.datasets import make_gaussian_quantiles
```

```
X, y = make_gaussian_quantiles(n_features=2, n_classes=2,  
n_samples = a, random_state = b)
```

Para o parâmetro *a* (número de pontos), utilize um valor entre 200 e 500. Para o parâmetro *b*, utilize qualquer valor inteiro aleatório com 4 ou 5 dígitos.

- *make_circles*:

```
from sklearn.datasets import make_circles
```

```
X, y = make_circles(n_samples=a, noise=b, random_state= c)
```

Para o parâmetro *a* (número de pontos), utilize um valor entre 200 e 500. Para o parâmetro *b*, utilize um valor real entre 0.1 e 0.2. Para o parâmetro *c*, utilize qualquer valor inteiro aleatório com 4 ou 5 dígitos.

a) Divida o conjunto de dados gerados em um grupo de treinamento e um grupo de teste, utilizando entre 20 - 40% dos dados para o grupo de teste;

b) Aplicando o algoritmo KNN com 3 vizinhos, faça o ajuste do modelo utilizando o grupo de treinamento e avalie a matriz de confusão e a acurácia obtida com o grupo de teste. Avalie diferentes valores para o número de vizinhos, entre 1 e 7, e verifique como isto afeta a acurácia do algoritmo. Escolha o valor que apresente o melhor resultado¹;

c) Altere o número de pontos utilizado na criação do conjunto de dados (parâmetro *n_samples*). Avalie valores contendo 25% - 400% do total de pontos utilizado no conjunto original e verifique como isso afeta a acurácia do algoritmo. Discuta os resultados;

d) Utilize o algoritmo treinado para prever a classe dos seguintes pontos: (0,0); (-1.5, 1); (1, 0.5).

Obs.: A entrega do trabalho deverá ser feita em um arquivo .pdf, contendo os itens c e d, além do arquivo .py com o item b.

¹É normal que diferentes valores apresentem uma acurácia muito próxima.