Aprendizagem Automática 2022/23

Ficha prática 8

Exercício#8.1

1. Carregue o conjunto de dados "breast_cancer" usando a função "load_breast_cancer()".

2. Faça uma divisão treino/teste com os valores por omissão (treino=75%, teste=25%) e confirme o número de exemplos e atributos de cada conjunto.

3. Construa um modelo usando o algoritmo Support Vector Machines (SVM) com um núcleo 'rbf' e 'gamma=1/n_features'. Calcule o desempenho sobre ambos os conjuntos.

4. Construa um gráfico que mostra os valores mínimo e máximo para cada atributo. Confirme a grande variabilidade na magnitude dos valores dos atributos. Utilize uma escala logarítmica, e.g.

```
# gráfico
plt.plot(x_train.min(axis=0), 'o', label="min")
plt.plot(x_train.max(axis=0), '^', label="max")
plt.legend(loc=4)
plt.xlabel("Feature index")
plt.ylabel("Feature magnitude")
plt.yscale("log")
plt.show();
```

5. A função 'MinMaxScaler' transforma os dados de modo a que qualquer atributo varie entre 0 e 1 (por default), ou entre outros valores definidos a e b, e.g. para escalar um conjunto de dados de treino, x_train

```
from sklearn.preprocessing import MinMaxScaler
sc=MinMaxScaler(feature_range=(a,b))
sc.fit(x_train)
x_train_scaled=sc.transform(X_train)
```

Aplique esta normalização sobre os dados (conjunto de **treino**) e confirme a transformação realizada, mostrando os valores mínimo e máximo antes e depois da transformação.

6. Transforme igualmente o conjunto de **teste** (**usando a mesma transformação realizada sobre o conjunto de treino, ou seja usando o objeto "sc" com o fit sobre os dados "x_train"**). Confirme a amplitude dos atributos no conjunto de teste.

7. Construa um novo modelo com o mesmo setup do ponto 3, mas agora utilizando os atributos normalizados. Calcule o desempenho sobre ambos os conjuntos (treino e teste) e compare os resultados.

8. Se considerar que pode existir sub-ajustamento, o que deve fazer? Construa um novo modelo com um valor de C mais adequado e compare os resultados.

Exercício#8.2

Analise a função GridSearchCV para "fine-tuning" de hiper-parâmetros com validação cruzada, na documentação oficial

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

O melhor classificador é definido com a chamada do método fit, e valor dos melhores parâmetros está na variável best_params_

Aplique a procura com grid search ao exemplo anterior fazendo um fine tuning sobre os parâmetros C e gamma.