Universidade Federal do Ceará - UFC
Departamento de Estatística e Matemática Aplicada - DEMA
Disciplina : Introdução à Mineração de Dados
Prof : Tibérius de Oliveira e Bonates

Alunos:
Irlene dos Santos Rabelo - 418147
Thyago Freitas da Silva - 392035

Atividade 5 – Validação cruzada.

1 Scikit-Learn

1.1 Datasets escolhidos

Segue abaixo uma tabela informando quais datasets foram utilizados e algumas informações sobre os mesmos,vale ressaltar que o número de atributos de cada dataset já conta com a classe.

Dataset	Nº de amostras	Nº de atributos	Nº de classes
Wine	178	13	3
Breast Cancer	569	30	2
Iris	150	4	3
Digits	1797	64	10

Tabela 1: Resumo dos resultados obtidos.

1.2 Modelos escolhidos

Os seguintes modelos foram escolhidos :

- K-Nearest Neighbors
- Random Forest
- MultiLayer Perceptron
- Support Vector Machine

1.3 K-Nearest Neighbors

	Parâmetros de teste
n_neighbors	3, 5, 7
weights	uniform, distance

Tabela 2: Conjuntos de parâmetros utilizados na busca em grade.

	Wine	Breast Cancer	Iris	Digits
n_neighbors	3	7	7	3
weights	distance	uniform	uniform	distance

Tabela 3: Melhores parâmetros,por dataset, encontrados na busca em grade.

1.4 Random Forest

	Paramêtros de teste
n_estimators	25, 50
criterion	gini, entropy
max_depth	10, 25

Tabela 4: Conjuntos de parâmetros utilizados na busca em grade.

	Wine	Breast Cancer	Iris	Digits
n_estimators	50	50	25	50
criterion	gini	entropy	entropy	entropy
max_depth	10	15	15	15

Tabela 5: Melhores parâmetros,por dataset, encontrados na busca em grade.

1.5 MultiLayer Perceptron

	Parâmetros de teste
max_iter	500,600
activation	relu,logistic
hidden_layer_sizes	100,120

Tabela 6: Conjuntos de parâmetros utilizados na busca em grade.

	Wine	Breast Cancer	Iris	Digits
max_iter	600	500	500	600
activation	logistic	logistic	logistic	logistic
hidden_layer_sizes	100	120	100	120

Tabela 7: Melhores parâmetros, por dataset , encontrados na busca em grade.

1.6 Support Vector Machine

	Paramêtros de teste
C	0.5, 1, 2
kernel	linear, poly, sigmoid

Tabela 8: Conjuntos de parâmetros utilizados na busca em grade.

	Wine	Breast Cancer	Iris	Digits
C	2	1	0.5	1
kernel	linear	linear	linear	poly

Tabela 9: Melhores parâmetros,por dataset, encontrados na busca em grade.

1.7 Resumo das perfomances.

Segue abaixo uma tabela contendo os valores médios das acurácias de cada modelo,para cada dataset, utilizando cross validation, modelos esses que se utilizam dos melhores parâmetros encontrados através da busca em grade do scikit-learn.

Algoritmo	Wine	Breast Cancer	Iris	Digits
K-Nearest Neighbors	76.0%	93%	97%	99%
Random Forest	98.0%	96%	95%	97%
MultiLayer Perceptron	95.0%	94%	98%	98%
Support Vector Machine	95.0%	94%	98%	98%

Tabela 2: Resumo dos resultados obtidos.