Trabalho #1: Classificação supervisionada

Instruções

Com uma base de dados pública (preferencialmente com duas classes) realizar os procedimentos para a concepção de um sistema de Reconhecimento de Padrões, incluindo pré-processamento, redução de dimensionalidade, classificação e avaliação, conforme visto na aula prática, e considerando as seguintes condições:

1. Pré-processamento: realizar higienização, normalizações, etc necessárias;
2. Redução de dimensionalidade: reduzir a dimensionalidade utilizando a abordagem PCA. Após, selecionar as "n" primeiras Componentes Principais (CPs) que explicam a variabilidade dos dados acumulada até atingir 75%, 90% e 99%, respectivamente; Observe que na prática serão 3 "novas" base de dados, cada uma com o número de CPs (dimensão) correspondente à variância acumulada;
3. Caso a base de dados não forneça os dados particionados em conjuntos de treinamento e teste, separar aleatoriamente o dataset em ⅔ para treinamento e ⅓ para teste;
4. Considerando as bases de dados divididas (⅔ para treinamento e ⅓ para teste), experimentar os algoritmos de classificação lineares e não-lineares abaixo:
    a. Naive Bayes Gaussiano (scikit-learn)
    b. SVM (Support vector machine) utilizando os kernel's linear e RBF. Ver o link http://scikit-learn.org/stable/modules/svm.html
    c. C4.5 (J48 ou CART). Ver link http://scikit-learn.org/stable/modules/tree.html)
5. Fazer uma tabela mostrando o desempenho de classificação de cada classificador (NB, SVM linear e não-linear e C4.5) para o conjunto de teste. Apresentar a matriz de confusão para cada algoritmo http://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
   OBS: observe que o treinamento do modelo deve acontecer **apenas** com os dados de treinamento, enquanto o teste **apenas** com os dados de teste.

Variância acumulada do PCA		Variância 75%	Num CPs 1	Variância 90%	Num CPs 2	Variância 99%	Num CPs 3
Classificadores	Naive Bayes	Acurácia =	0.85	Acurácia =	0.88	Acurácia =	0.86
		TPR =	0.74	TPR =	0.80	TPR =	0.77
		TNR =	0.96	TNR =	0.96	TNR =	0.96
	SVM Linear	Acurácia =	0.86	Acurácia =	0.87	Acurácia =	0.87
		TPR =	0.81	TPR =	0.82	TPR =	0.82
		TNR =	0.91	TNR =	0.93	TNR =	0.93
	SVM RBF	Acurácia =	0.86	Acurácia =	0.87	Acurácia =	0.87
		TPR =	0.81	TPR =	0.82	TPR =	0.82
		TNR =	0.92	TNR =	0.94	TNR =	0.94
	C4.5 (CART)	Acurácia =	0.74	Acurácia =	0.80	Acurácia =	0.83
		TPR =	0.75	TPR =	0.77	TPR =	0.81
		TNR =	0.74	TNR =	0.83	TNR =	0.86