Exercício 3

ATENCAO Houve uma modificacao no enunciado na parte do tratamento de dados faltantes.

Data de entrega: 24/10, as 7:00 (da manha).

Use os dados do dataset SECOM do UCI O arquivo secom.data contem os dados. O arquivo secom_labels.data contem (na 1a coluna) a classe de cada dado.

Usando um 5-fold externo para calcular a accuracia, e um 3-fold interno para a escolha dos hyperparametros, determine qual algoritimo entre kNN, SVM com kernel RBF, redes neurais, Random Forest, e Gradient Boosting Machine tem a maior acuracia.

1. Preprocesse os dados do arquivo: **Substitua os dados faltantes pela media da coluna (imputação pela média)**. Finalmente padronize as colunas para media 0 e desvio padrao 1.
2. Para o kNN, faça um PCA que mantem 80% da variancia. Busque os valores do k entre os valores 1, 5, 11, 15, 21, 25..
3. Para o SVM RBF teste para C=2**(-5), 2**(0), 2**(5), 2**(10) e gamma= 2**(-15) 2**(-10) 2**(-5) 2**(0) 2**(5).
4. Para a rede neural, teste com 10, 20, 30 e 40 neuronios na camada escondida.
5. Para o RF, teste com mtry ou n_featrues = 10, 15, 20, 25 e ntrees = 100, 200, 300 e 400..
6. Para o GBM (ou XGB) teste para numero de arvores = 30, 70, e 100, com learning rate de 0.1 e 0.05, e profundidade da arvore=5.Voce pode tanto usar alguma versao do gbm para R ou SKlearn, ou usar o XGBoost (para ambos).
7. Voce nao precisam fazer os loops da validacao cruzada explicitamente. Pode usar as funcoes como tunegrid (do caret) ou tuneParams (do mlr) ou GridSearchCV do SKlearn..
8. Reporte a acuracia de cada algoritmo calculada pelo 5-fold CV externo..

Detalhes R

Considere usar os pacotes caret ou mlr para fazer os loops de validacao cruzada e usar os diferentes classificadores

---

Last modified: Sun Oct 16 17:46:47 BRST 2016