Exercício 1

Data de entrega: 26/9, as 7:00 (da manha).

Leia os dados do arquivo data1.csv A classe de cada dado é o valor da última coluna (0 ou 1).

1. faca o PCA dos dados (sem a última coluna). Se voce quiser que os dados transformados tenham 80\% da variância original, quantas dimensões do PCA vc precisa manter?

Gere os dados transformados mantendo 80\% da variância. (Atenção este passo não é 100\% correto do ponto de vista de aprendizado de maquina. Não repita este passo em outras atividades).

Considere as primeiras 200 linhas dos dados como o conjunto de treino, e as 276 ultimas como o conjunto de dados

2. Treine uma regressão logística no conjunto de treino dos dados originais e nos dados transformados. Qual a taxa de acerto no conjunto de teste nas 2 condições (sem e com PCA)?

3. Treine o LDA nos conjuntos de treino com e sem PCA e teste nos respectivos conjuntos de testes. Qual a acurácia nas 2 condições?

4. Qual a melhor combinação de classificador e PCA ou não?

Existem vários pacotes em R (caret, e mlr), e em Python (sklearn) que fazem tudo o que eu pedi acima em uma simples chamada de um "workflow de aprendizado de maquina" que passa como parâmetros as funções a serem aplicadas e testadas. Para este e para os próximos exercícios **NAO** usem esses "workflows". Eu quero que voces façam explicitamente os passos pedidos.

Gere um pdf com o código (R ou Python) e as respostas as perguntas. Uma nova versão desta pagina informará como submeter o exercicio

Detalhes R

regressão logística em R é feito pela função glm com parâmetro family="logit". A função predict computa a previsão de um modelo para novos dados. A função predict quando o primeiro argumento é um modelo da classes `gkm` é chamada de `predict.glm`

LDA é feito pela função lda do pacote `MASS`. predict.lda para obter o resultado do modelo em dados novos.

Note que povavelmente a forma de chamada do glm e do lda não é a mesma, e acho que uma delas espera o atributo de saida numérico (0 e 1), e a outra espera um atributo categorico (chamado de `factor` em R.

Detalhes em Python

PCA, regressao logistica e LDA todos do sklearn.

Last modified: Sun Sep 11 18:08:02 BRT 2016