Reconhecimento de Padrões
Exercício prático de implementação de Análise de Componentes Principais e de Máquinas de Vetores de Suporte

Prof. Frederico Coelho

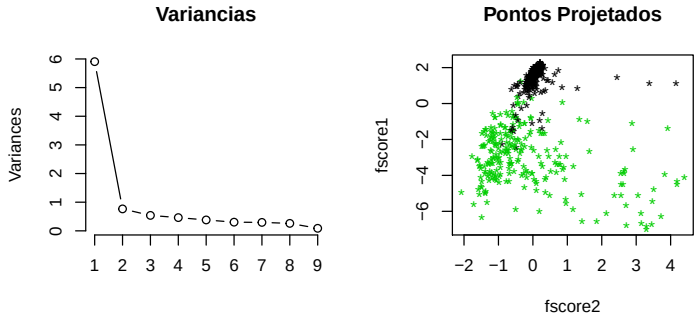
8 de outubro de 2019

1 Introdução

Neste exercício o(a) aluno(a) realizará a implementação do método de Análise de Componentes Principais (PCA) e de uma Máquina de Vetores de Suporte (SVM) para realizar a classificação do conjunto de dados *Breast Cancer*.

2 PCA

Primeiramente o método PCA é empregado para projetar os dados em um novo espaço de caracteristicas, como mostrado na figura a seguir. Para gerar esta figura utilizou-se as 2 primeiras componentes geradas, mas para a solução do problema de classificação o aluno deverá utilizar quantas componentes forem necessárias. Pode-se usar a função nativa do R *prcomp* para extrair as componentes.



3 SVM

Posteriormente, deve-se utilizar as variáveis do novo espaço projetado como a entrada de uma SVM de base radial (com parâmetros: C, h) para realizar a classificação dos dados. A biblioteca *kernlab* possui a função *ksvm* para implementar uma SVM.

O aluno deverá seguir os seguintes passos:

1. Carregar a base e tratar os dados como mostrado abaixo;

```
library(mlbench)
data(BreastCancer) # carrega dados
db <- na.omit(BreastCancer) # elimina dados faltantes
dbLabel[dbClass == "benign"] <- -1 # muda labels
dbLabel[dbClass == "malignant"] <- 1 # muda labels
X <- data.matrix(db[,2:10])
y <- data.matrix(db[,12])
```

2. Aplicar o PCA nos dados;

3. Definir quantas componentes utilizar;

4. Separar os dados em treinamento e teste usando as componentes do PCA e considerando 10 folds para a validação cruzada;

5. Treinar a SVM com os dados no espaço mapeado pelo PCA (atenção para a definição dos parâmetros do kernel escolhido e do parâmetro C de regularização da SVM.);

6. aplicar modelo treinado ao conjunto de testes;

7. Calcular as acurácias para cada classe em cada um dos folds e calcular a acurácia média.

O aluno deverá entregar um relatório PDF contendo a justificativa para a quantidade de componentes escolhida para ser usado na SVM, o gráfico da variância das componentes ordenadas, e a tabela de acurácias mostrando o desempenho em cada classe do problema separadamente em cada fold e na média total.

Material original do Prof. Antônio Braga