Pré-processamento

Redução de Dimensionalidade

Apresenta-se conceitos necessários para análise dos atributos em bases de dados e redução da dimensionalidade usando *Principal Component Analysis* (PCA).

Muitas técnicas de Aprendizado de Máquina (AM) são utilizadas para atratar dados com um número elevado de atributos, por exemplo, em Imagens: pixels, em expressão gênica: gênes, Mineração de textos: palavras. Podemos trabalhar com os atributos mais relevantes que ofereçam ao modelo melhor distinção de padrões, fazendo a seleção e remoção de atributos irrelevantes. Podemos utiliza a abordagem embutida, fazendo a seleção integrada com o algoritmo de aprendizado(decision tree), pode-se utilizar a filtragem de um subconjunto de atributos sem levar em consideração o algoritmo de aprendizado, e também tem o wrapper, que para cada subconjunto de atributos, um algoritmo de aprendizado é consultado.

PCA

Método que utiliza uma tranformação ortogonal para conveter um conjunto de observações correlacionadas em um conjunto de componentes principais(*observações linearmente não-correlacionadas*). Ela traz algumas aplicações, como redução de dimensionalidade, redução de redundância, filtragem de ruído, compressão de dados, preparação para utilização por outras técnicas, e identificação de relacionamento entre variáveis. Antes de aplicar PCA nos dados, é importante realizar uma normalização!

Aplicação do PCA no conjunto de dados Íris

```
# lendo o datset por meio de um arquivo .csv
iris = read.csv('../data/Iris.csv')

# Os dados contêm quatro variáveis contínuas que correspondem às medidas físicas
# das flores e uma variável categórica que descreve as espécies de flores.
head(iris)
```

```
##   Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm     Species
## 1  1           5.1          3.5           1.4          0.2 Iris-setosa
## 2  2           4.9          3.0           1.4          0.2 Iris-setosa
## 3  3           4.7          3.2           1.3          0.2 Iris-setosa
## 4  4           4.6          3.1           1.5          0.2 Iris-setosa
## 5  5           5.0          3.6           1.4          0.2 Iris-setosa
## 6  6           5.4          3.9           1.7          0.4 Iris-setosa
```

Vamos aplicar o PCA às quatro variáveis contínuas e usar a variável categórica para visualizar os PCs posteriormente. Observe que, no código a seguir, aplicamos uma transformação de log às variáveis contínuas, conforme sugerido por [1], e definimos center e scale igual a TRUE na chamada para prcomp padronizar as variáveis antes da aplicação do PCA:

```
# tranformação de log - normalização dos dados
log.iris <- log(iris[,2:5])
iris.species <- iris[,6]

# aplica o PCA - scale = TRUE é aconselhavel
# mas o padrão é false
```

```
iris.pca <- prcomp(log.iris,
                   center = TRUE, scale. = TRUE)
```

Como a assimetria e a magnitude das variáveis influenciam os PCs resultantes, é uma boa prática aplicar a transformação de assimetria, centralizar e dimensionar as variáveis antes da aplicação do PCA. No exemplo acima, aplicamos uma transformação de log às variáveis, mas poderíamos ter sido mais gerais e aplicado uma transformação Box e Cox [2].

Analizando os Resultados

A função prcomp retorna um objeto de classe prcomp, que possui alguns métodos disponíveis. Com aplicação do método podemos o desvio padrão de cada um dos quatro PCs e sua rotação (ou loadings), que são os coeficientes das combinações lineares das variáveis contínuas.

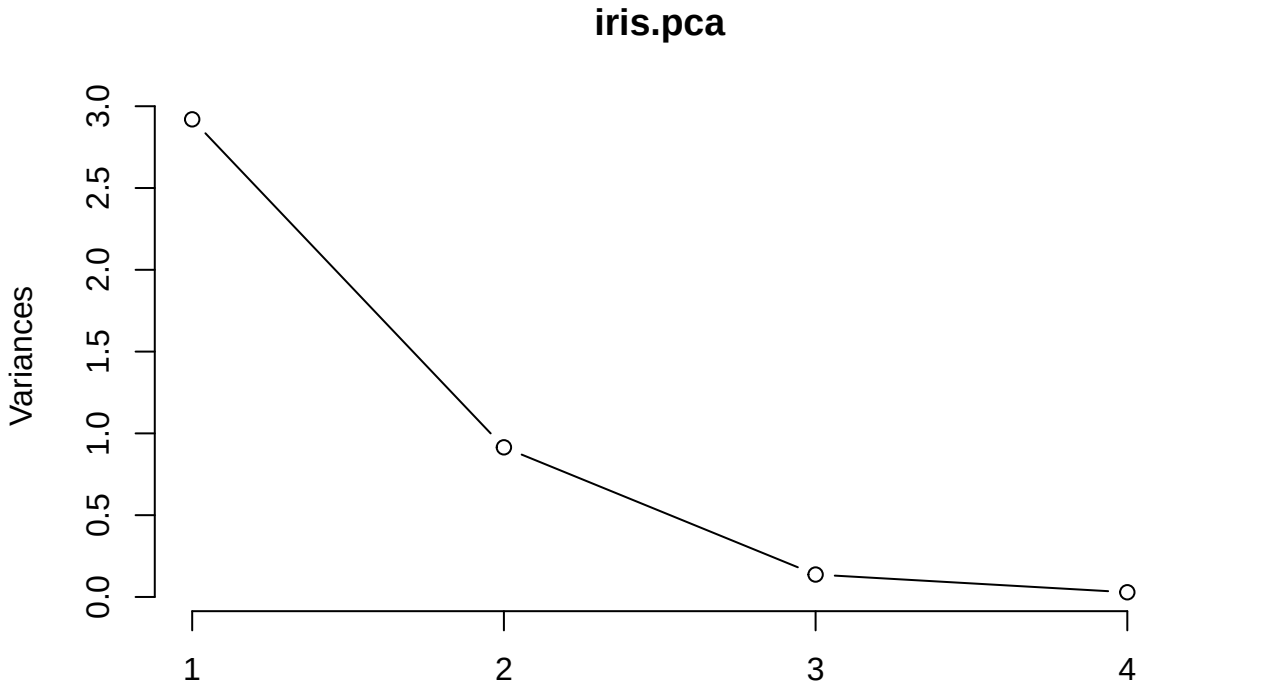
```
iris.pca
```

```
## Standard deviations:
## [1] 1.7087152 0.9563817 0.3700768 0.1693209
##
## Rotation:
##                       PC1         PC2        PC3        PC4
## SepalLengthCm  0.5059018 -0.44704259  0.7084146  0.2058278
## SepalWidthCm  -0.2959102 -0.89323900 -0.3225722 -0.1025106
## PetalLengthCm  0.5781192 -0.02821947 -0.2010441 -0.7902930
## PetalWidthCm   0.5676959 -0.03847958 -0.5947077  0.5679467
```

Encontrando o cotovelo

O método de plotagem retorna um gráfico das variâncias (eixo y) associadas aos PCs (eixo x). A figura abaixo é útil para decidir quantos PCs manter para análise posterior. Neste caso simples, com apenas 4 PCs, isso não é uma tarefa difícil e podemos ver que os dois primeiros PCs explicam a maior parte da variabilidade nos dados.

```
# plot o método
plot(iris.pca, type = 'l')
```



O método **summary** descreve a importância dos PCs. A primeira linha descreve novamente o desvio padrão associado a cada PC. A segunda linha mostra a proporção da variação nos dados explicados por cada componente, enquanto a terceira linha descreve a proporção cumulativa da variação explicada. Podemos ver que os dois primeiros PCs são responsáveis por mais do que pela variação dos dados.

```
summary(iris.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7087 0.9564 0.37008 0.16932
## Proportion of Variance 0.7299 0.2287 0.03424 0.00717
## Cumulative Proportion  0.7299 0.9586 0.99283 1.00000
```

A Figura abaixo é um biplot gerado pelo gráfico **ggbiplot**.

```
library(devtools)
library(ggbiplot)
```

```
## Loading required package: ggplot2
```

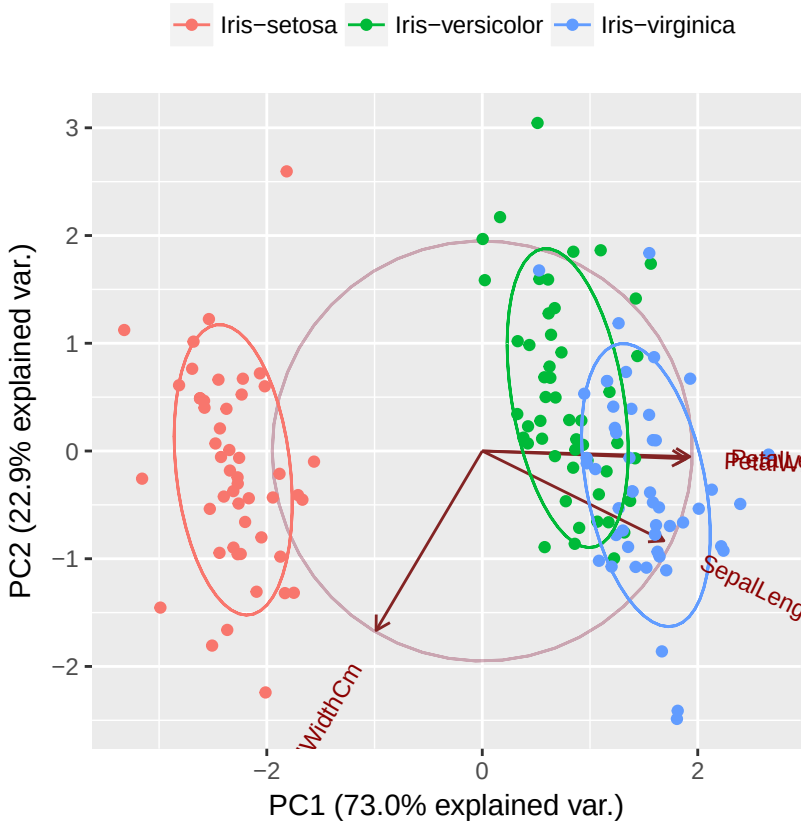
```
## Loading required package: plyr
```

```
## Loading required package: scales
```

```
## Loading required package: grid
```

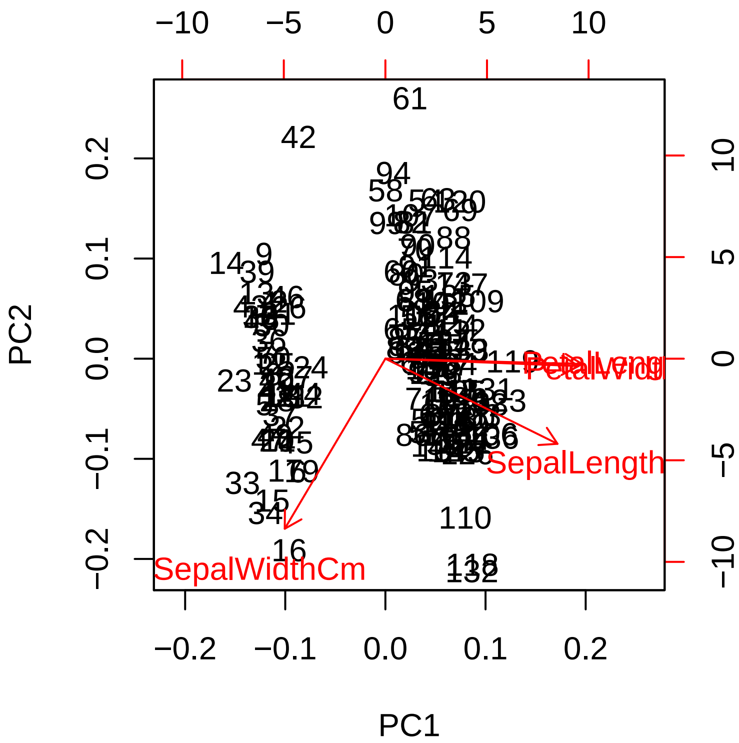
```
g <- ggbiplot(iris.pca, obs.scale = 1, var.scale = 1,
              groups = iris.species, ellipse = TRUE,
              circle = TRUE)

g <- g + scale_color_discrete(name='')

g <- g + theme(legend.direction = 'horizontal',
               legend.position = 'top')

# print g
```

```
g
```



Ele projeta os dados nos dois primeiros PCs. Outros PCs podem ser escolhidos através das escolhas de argumentos da função. Ele colore cada ponto de acordo com as espécies das flores e desenha uma linha de contorno normal com a probabilidade ellipse.prob (padrão) para cada grupo. Mais informações sobre o ggbiplot podem ser obtidas pelo ggbiplot usual. Eu acho que você vai concordar que a trama produzida pelo ggbiplot é muito melhor que a produzida pelo biplot (iris.pca), Figura abaixo.

```
biplot(iris.pca)
```


−10
−5
0
5
10
PC2
0.2
0.1
0.0
−0.1
−0.2
SepalWidthCm
SepalLength
−0.2
−0.1
0.0
0.1
0.2
PC1