

# INF0613 – Aprendizado de Máquina Não Supervisionado

## Trabalho 2 - Redução de Dimensionalidade

Nicole Nogueira Silva

Rodolfo Dalla Costa

O objetivo deste trabalho é exercitar o conhecimento de técnicas de redução de dimensionalidade. Essas técnicas serão usadas tanto para obtenção de características quanto para visualização dos conjuntos de dados. Usaremos a base de dados `speech.csv`, que está disponível na página da disciplina no Moodle. A base contém amostras da pronúncia em inglês das letras do alfabeto.

### Atividade 0 – Configurando o ambiente

Antes de começar a implementação do seu trabalho configure o *workspace* e importe todos os pacotes e execute o pré-processamento da base:

```
# Adicione os demais pacotes usados neste trabalho:
library(magrittr)
library(tidyverse)
library(umap)
library(Rtsne)

# Configure ambiente de trabalho na mesma pasta
# onde colocou a base de dados:
setwd("C:/Users/nicol/Documents/Mineração de dados/inf-0611-0612/inf0613/t2")

# Pré-processamento da base de dados
# Lendo a base de dados
speech <- read.csv("speech.csv", header = TRUE)

# Convertendo a coluna 618 em caracteres
speech$LETRA <- LETTERS[speech$LETRA]
```

### Atividade 1 – Análise de Componentes Principais (3,5 pts)

Durante a redução de dimensionalidade, espera-se que o poder de representação do conjunto de dados seja mantido, para isso é preciso realizar uma análise da variância mantida em cada componente principal obtido. Use função `prcomp`, que foi vista em aula, para criar os autovetores e autovalores da base de dados. Não use a normalização dos atributos, isto é, defina `scale.=FALSE`. Em seguida, use o comando `summary`, analise o resultado e os itens a seguir:

```
# Executando a redução de dimensionalidade com o prcomp

speech_pca <- prcomp(speech[,1:617], scale = FALSE)

# Analisando as componentes com o comando summary
#summary(speech_pca)
importance <- data.frame(summary(speech_pca)$importance)
```

```
importance %>%  
  select(., 1:6)
```

##	PC1	PC2	PC3	PC4	PC5	PC6
## Standard deviation	5.349	3.2000	2.6908	2.2953	2.1593	1.9522
## Proportion of Variance	0.248	0.0889	0.0629	0.0457	0.0405	0.0331
## Cumulative Proportion	0.248	0.3374	0.4003	0.4460	0.4865	0.5196

## Análise

a) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 80% do total?

### Resposta:

O número mínimo de componentes necessárias para que 80% da variabilidade dos dados seja representada é 38.

b) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 90% do total?

### Resposta:

O número mínimo de componentes necessárias para que 90% da variabilidade dos dados seja representada é 91.

c) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 95% do total?

### Resposta:

O número mínimo de componentes necessárias para que 95% da variabilidade dos dados seja representada é 170.

d) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 99% do total?

### Resposta:

O número mínimo de componentes necessárias para que 99% da variabilidade dos dados seja representada é 382.

e) Faça um breve resumo dos resultados dos itens a)-d) destacando o impacto da redução de dimensionalidade.

### Resposta:

A redução de dimensionalidade para bases com grande número de features é um processo muito importante tanto para interpretação quanto para reduzir tempo de processamento e descartar informações redundantes. Utilizando essa técnica no banco de dados `Speech`, a partir de somente 38 componentes, já seríamos capazes de representar 80% da variação total dos nossos dados, já correspondendo a um ótimo resultado. Aumentando para 91, 170 e 382 componentes, teríamos respectivamente 90%, 95%, 99% da variabilidade dos dados representada, enquanto que inicialmente possuíamos 618 atributos a disposição para utilização.

## Atividade 2 – Análise de Componentes Principais e Normalização (3,5 pts)

A normalização de dados em alguns casos, pode trazer benefícios. Nesta questão, iremos analisar o impacto dessa prática na redução da dimensionalidade da base de dados `speech.csv`. Use função `prcomp` para criar os autovetores e autovalores da base de dados usando a normalização dos atributos, isto é, defina `scale.=TRUE`. Em seguida, use o comando `summary`, analise o resultado e os itens a seguir:

```
# Executando a redução de dimensionalidade com o prcomp
# com normalização dos dados
speech_pca_norm <- prcomp(speech[,1:617], scale = TRUE)
# Analisando as componentes com o comando summary
#summary(speech_pca_norm)
summary(speech_pca_norm)$importance %>% data.frame() %>% select(1:4)

##              PC1      PC2      PC3      PC4
## Standard deviation    10.914  7.4033  5.7923  5.2805
## Proportion of Variance  0.193  0.0888  0.0544  0.0452
## Cumulative Proportion  0.193  0.2819  0.3363  0.3815
```

## Análise

a) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 80% do total?

**Resposta:**

O número mínimo de componentes necessárias para que 80% da variabilidade dos dados seja representada é 48.

b) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 90% do total?

**Resposta:**

O número mínimo de componentes necessárias para que 90% da variabilidade dos dados seja representada é 112.

c) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 95% do total?

**Resposta:**

O número mínimo de componentes necessárias para que 95% da variabilidade dos dados seja representada é 200.

d) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 99% do total?

**Resposta:**

O número mínimo de componentes necessárias para que 99% da variabilidade dos dados seja representada é 400.

e) Quais as principais diferenças entre a aplicação do PCA nesse conjunto dados com e sem normalização?

Ao utilizarmos a análise de componentes principais com a normalização dos dados consiste na identificação dos autovalores e autovetores a partir da matriz de correlação, enquanto que sem a normalização identificamos esses fatores através da matriz de covariância.

f) Qual opção parece ser mais adequada para esse conjunto de dados? Justifique sua resposta.

**Resposta:**

Por meio das técnicas apresentadas pelos dois métodos, nota-se que a análise de componentes principais sem realizar a normalização demonstrou-se mais eficaz na redução de dimensionalidade dos nossos dados pois, para todas as variâncias acumuladas avaliadas, o primeiro método necessitou de menos componentes.

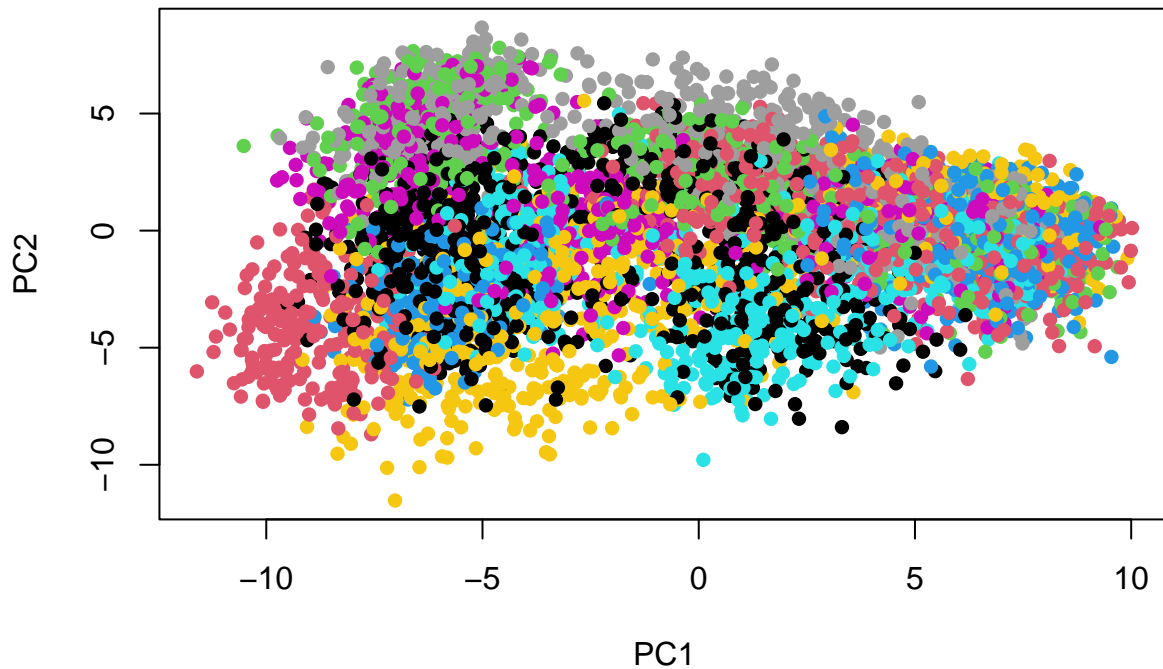
## Atividade 3 – Visualização a partir da Redução (3,0 pts)

Nesta atividade, vamos aplicar diferentes métodos de redução de dimensionalidade e comparar as visualizações dos dados obtidos considerando apenas duas dimensões. Lembre de fixar uma semente antes de executar o T-SNE.

- a) Aplique a redução de dimensionalidade com a técnica PCA e gere um gráfico de dispersão dos dados. Use a coluna 618 para classificar as amostras e definir uma coloração.

```
# Aplicando redução de dimensionalidade com a técnica PCA
speech_pca <- prcomp(speech[,1:617], scale = FALSE)

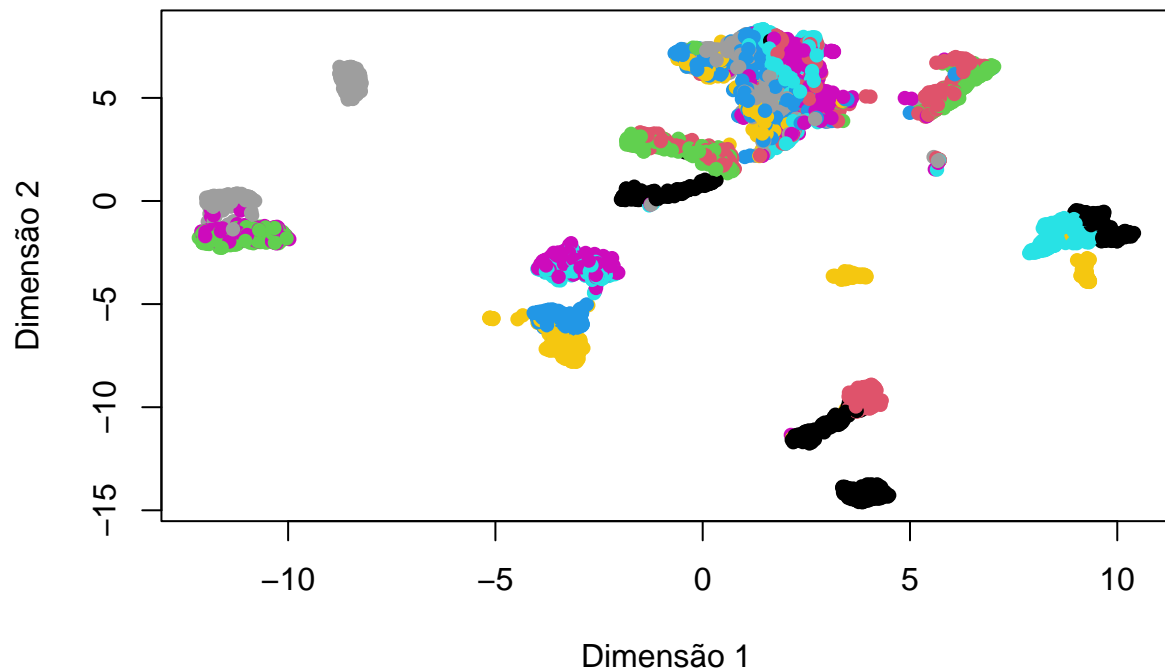
# Gerando o gráfico de dispersão
plot(speech_pca$x[, 1:2], col = as.factor(speech[, 618]), pch = 16)
```



- b) Aplique a redução de dimensionalidade com a técnica UMAP e gere um gráfico de dispersão dos dados. Use a coluna 618 para classificar as amostras e definir uma coloração.

```
# Aplicando redução de dimensionalidade com a técnica UMAP
speech_umap <- umap(as.matrix(speech[,1:617]))

# Gerando o gráfico de dispersão
plot(speech_umap$layout, col = as.factor(speech$LETRA), xlab = "Dimensão 1",
     ylab = "Dimensão 2", pch = 16)
```



- c) Aplique a redução de dimensionalidade com a técnica T-SNE e gere um gráfico de dispersão dos dados. Use a coluna 618 para classificar as amostras e definir uma coloração.

```
# Aplicando redução de dimensionalidade com a técnica T-SNE
```

```
#verificando dados repetidos
```

```
sum(duplicated(speech))
```

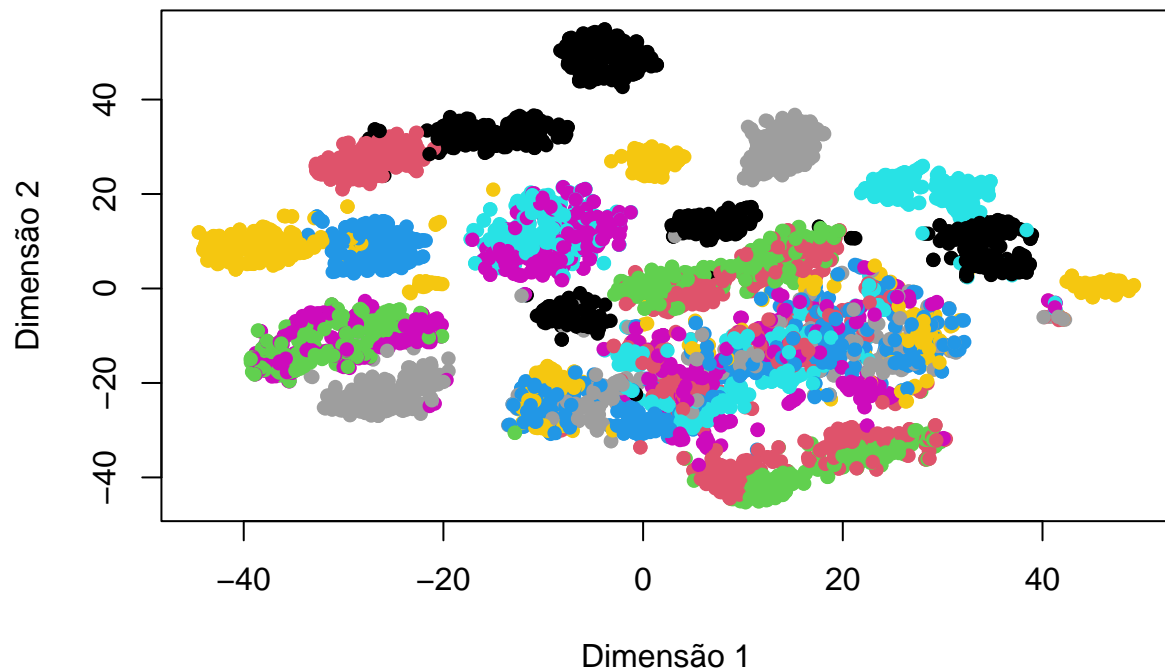
```
## [1] 0
```

```
set.seed(46)
```

```
tsne <- Rtsne(as.matrix(speech[,1:617]), perplexity = 30, dims=2)
```

```
# Gerando o gráfico de dispersão
```

```
plot(tsne$Y, col= as.factor(speech$LETRA) , xlab="Dimensão 1", ylab="Dimensão 2",  
      pch=16)
```



## Análise

d) Qual técnica você acredita que apresentou a melhor projeção? Justifique.

### Resposta:

Analisando os gráficos gerados a partir das 3 técnicas de redução de dimensionalidade, é possível concluir que a técnica de T-SNE mostrou-se mais efetiva para o banco de dados em questão. Na base de dados temos 26 categorias da variável resposta. Ao analisarmos o resultado do PCA, não é possível distinguir nenhum grupo. Utilizando o UMAP, alguns agrupamentos de categorias já ficam mais evidentes, porém alguns registros ficam muito misturados e é com o T-SNE em que conseguimos ter visualizações mais espaçadas e claras de algumas categorias.