**Universidade de Fortaleza - UNIFOR**
**Centro de Ciências Tecnológicas - CCT**
**Engenharia de Computação**

**Sistemas Inteligentes - T951**

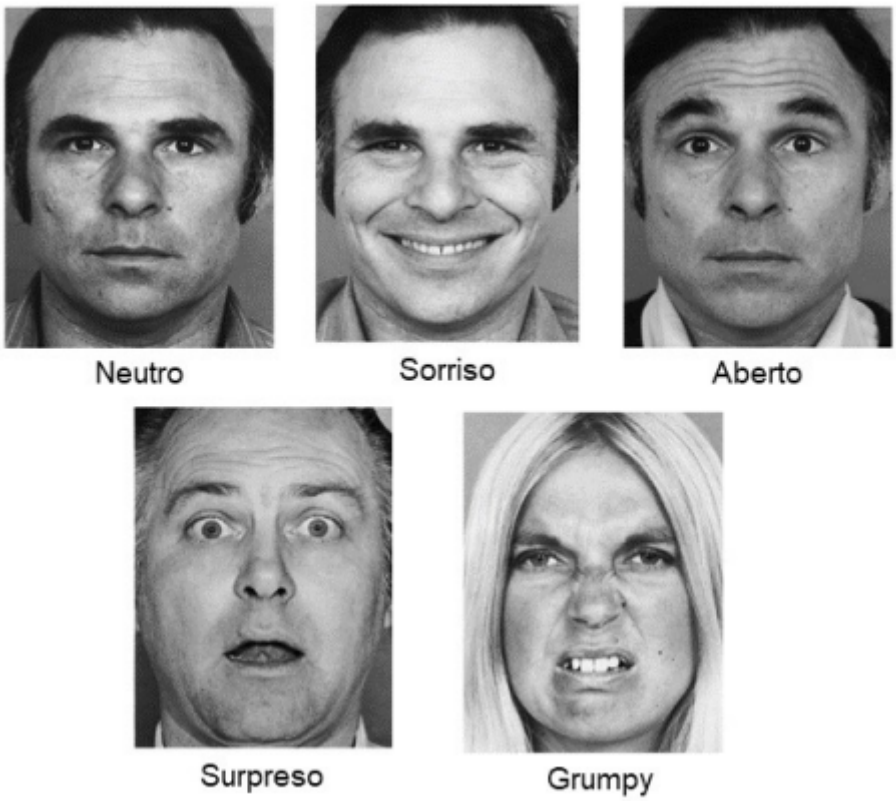
# Trabalho AV1

*Professor: Prof. Msc. Paulo Cirillo Souza Barbosa*

## Discussão Inicial sobre os Dados.

No AVA dois arquivos referente aos dados e rótulos de um conjunto de dados referente aos sinais de eletromiografia, captados nos músculos faciais: Corrugador do Supercílio (Sensor 1); Zigomático Maior (Sensor 2).

No presente conjunto de dados, tem-se 50 000 observações (amostras) para os dois sensores, em classes totalmente balanceadas. No presente conjunto de dados, existem classes referente aos gestos forçados exibidos pela Figura .




No AVA, existem dois arquivos que compõem o presente conjunto de dados. O primeiro, chamado de "EMG.csv"é composto pelas 50000 observações (com a estrutura: $\mathbf{X} \in \mathbb{R}^{N \times p}$), e o segundo arquivo "Rotulos.csv"é composto pelas 50000 classes (com a estrutura: $\mathbf{Y} \in \mathbb{R}^{N \times C}$). Pede-se inicialmente que faça a identificação de $p$ (número de preditores), $N$(Quantidade de amostras) e $C$(Quantidade de classes).

## 1) O que fazer inicialmente?

Após acessar as matrizes $\mathbf{X}$ e $\mathbf{Y}$, faça uma análise inicial dos dados através da construção do gráfico de espalhamento (*scatter plot*). Após essa visualização dos dados, é possível levantar quais hipóteses sobre o problema de classificação?

## 2) Projeto de classificadores.

Pede-se que se desenvolva seguintes algoritmos:

1. Classificador linear baseado no método dos mínimos quadrados ordinários.
2. Classificador linear baseado no método dos mínimos quadrados ordinários e regularizado pelo método de Tikhonov.
3. Classificadores Bayesianos Gaussianos.

## 2) Projeto das rodadas de validação dos modelos.

Antes de projetar cada um dos classificadores propostos, é necessário desenvolver uma rotina que avalia o desempenho dos modelos. Assim, para extrair informação de acurácia mais próxima do real, deve-se realizar 100 rodadas independentes de treino/teste para cada um dos modelos. Em cada rodada, deve-se embaralhar o conjunto de dados original e em seguida segmentar as amostras em dados para treinamento e teste. Tal divisão pode ser realizada utilizando uma das proporções, 80/20, 70/30, 90/10. O processo geral pode ser exemplificado através do algoritmo:

```
import numpy as np
import matplotlib.pyplot as plt

Data = np.loadtxt('EMG.csv',delimiter=',')
Rotulos = np.loadtxt('Rotulos.csv',delimiter=',')
RODADAS = 100

for i in range(RODADAS):
    seed = np.random.permutation(Data.shape[0])
    X = Data[seed,:]
    Y = Rotulos[seed,:]

    #treino / teste:
    Xtreino = X[0:int(X.shape[0]*.8),:]
    Ytreino = Y[0:int(X.shape[0]*.8),:]

    Xteste = X[int(X.shape[0]*.8):,:]
    Yteste = Y[int(X.shape[0]*.8):,:]

    ################################
    ########## MMQ #################
    ################################

    ################################
    ######MMQ Regularizado ########
    ################################

    ################################
    ######## Naive Bayes ##########
    ################################

    ################################
    ########### KNN ################
    ################################

    ################################
    ########### DMC ################
    ################################
```

Após esse processo, faça uma análise comparativa entre os modelos baseado na Acurácia e o custo computacional associado, utilizando a seguinte tabela:

Ao final das 100 rodadas, para cada classificador deve-se ter 100 dados referentes a Acurácia. Com isto posto, deve-se calcular a média, mediana, desvio padrão, menor e maior valor, para cada uma das medidas de desempenho. Faça a mesma análise para o custo computacional associado.

| Classificador | Média | Menor Valor | Maior Valor | Desvio-padrão |
|---|---|---|---|---|
| **OLS** | | | | |
| **OLS Regularizado** | | | | |
| **Naive Bayes** | | | | |
| **Classificador Gaussiano (Pooled)** | | | | |
| **Classificador Gaussiano (Friedman)** | | | | |

# Considerações importantes com relação aos classificadores.

Antes de desempenhar as 100 rodadas, é necessário identificar o valor ideal daqueles classificadores que possuem hiperparâmetro (KNN e OLS Regularizado).

## 1) OLS.

1. Faça a implementação do método dos mínimos quadrados ordinário.
2. Faça a implementação do método dos mínimos quadrados regularizado.
   - Faça o treinamento do modelo utilizando valores de $\lambda = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9, 1\}$
   - **Pergunta:** O que significa $\lambda = 0$??
3. Em sala foi realizada uma discussão sobre a adição de um vetor coluna de **1s** no início da matriz de dados $\mathbf{X}$.
   - O que implicaria adicionar este vetor coluna para o presente trabalho?
   - Qual a interpretação geométrica?
   - O resultado obtido é melhor ou pior?

## 1) Classificadores Bayesianos Gaussianos.

1. Dada as discussões realizadas em sala de aula, faça a estimação dos classificadores bayesianos gaussianos que utilizam-se do critério de máxima verossimilhança para os casos particulares:
2. Considere inicialmente que as covariâncias para todas as classes são iguais e projete o discriminante com essa informação.
3. Faça a abordagem da matriz de covariância agregada (*pooled*) e avalie o desempenho do modelo.
4. Considere uma análise que envolva a estimação do modelo através da regularização por Friedman. Nesta avalie o desempenho do modelo para cada hiperparâmetro $\lambda = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9, 1\}$. Escolha o modelo que tenha o melhor resultado, este será utilizado para as 100 rodadas.
5. Estime o modelo baseado no classificador naive bayes gaussiano.
6. Estime o modelo em que se assume a descorrelação entre os atributos de $\mathbf{x}$.

## 5) Relatório.

Além das implementações, o presente trabalho deve ser entregue em modelo de relatório. Este deve possuir as características descritas nos slides de apresentação do curso. Desta maneira, deve possuir:

1. Título.
2. Resumo.
3. Introdução.
4. Fundamentação Teórica (Revisão Bibliográfica).

5. Metodologia.
6. Resultados.
7. Conclusões.
8. Referências.

O modelo para trabalho pode ser encontrado neste **LINK**