# Relátorio 01 - Modelos de Regressão e Classificação

Bruno Matos de Araújo - 2225673, João Pedro Rego Magalhães - 2517985

***Abstract* — Este relatório apresenta uma atividade prática sobre a aplicação de técnicas de regressão e classificação a partir de dois conjuntos de dados distintos. Na parte de regressão, são analisadas variáveis experimentais —temperatura e pH — para prever o nível de atividade enzimática. Na parte de classificação, sinais de eletromiografia (EMG), obtidos de regiões faciais específicas, são utilizados para identificar expressões emocionais em cinco categorias. Metodologias clássicas e regularizadas, bem como estratégias de validação robusta via simulações de Monte Carlo, são empregadas para fornecer uma comparação detalhada dos desempenhos dos modelos.**

***Palavras-chave — Modelos de regressão, Modelos de Classificação, MQO, Eletromiografia (EMG), Validação de Monte Carlo.***

## I. Introdução

Modelos de inteligência artificial têm se tornado essenciais para replicar processos decisórios humanos em problemas reais. Neste trabalho, exploramos a utilização de métodos preditivos tanto para a regressão—com o objetivo de estimar um nível de atividade enzimática a partir de parâmetros experimentais—quanto para a classificação—na qual sinais provenientes de sensores EMG são processados para identificar emoções faciais. A abordagem segue o paradigma supervisionado, onde dados rotulados são empregados para treinar modelos que minimizam uma função de custo ou maximizam a verossimilhança, possibilitando a extração de padrões complexos em conjuntos de dados de alta dimensão.

## II. Metodologia

### A. *Tarefa de Regressão*

Na tarefa de Regressão, estamos lidando com um problema relacionado à análise de atividade enzimática, um fenômeno biológico que ocorre em condições laboratoriais controladas. O objetivo é prever o nível de atividade enzimática (variável dependente) a partir de dois fatores experimentais cruciais: a temperatura e o pH da solução (variáveis independentes).

A tarefa central é identificar padrões e relações que permitam quantificar e prever como alterações em temperatura e pH impactam a atividade enzimática, utilizando modelos matemáticos baseados em regressão. A correta modelagem desses dados não apenas melhora a compreensão dos mecanismos envolvidos, mas também abre caminho para intervenções práticas e melhorias no controle de tais processos. Isso representa um avanço significativo na aplicação de modelos de IA para resolver problemas de caráter experimental e biológico.

Utilizando o arquivo atividade_enzimatica.csv, realizamos uma análise exploratória por meio de um gráfico de dispersão (figura 1). Nesse gráfico, os eixos representam a temperatura e o pH da solução (variáveis independentes) versus o nível de atividade enzimática (variável dependente).:

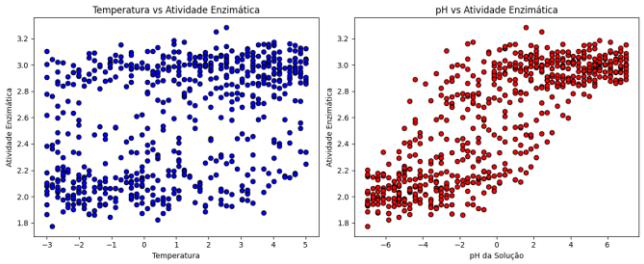


*Figura 1: Gráfico de dispersão*

Para a aplicação dos modelos preditivos, os dados foram estruturados da seguinte forma:

As variáveis regressoras (*X*), que são valores referentes à temperatura e ao pH foram armazenados em uma matriz *X* de dimensão $R^{N\,x\,p}$ (onde *N* é o número total de observações e *p* = 2), de forma a representar cada instância experimental. O vetor da variável dependente (*y*), que é o nível de atividade enzimática foi organizado em um vetor *y* de dimensão $R^{N\,x\,1}$, garantindo a correspondência correta com as entradas de *X*.

Três modelos distintos foram aplicados para prever o nível de atividade enzimática:

**Mínimos Quadrados Ordinários (MQO) Tradicional:** Esse é o método clássico de regressão que estima o vetor de parâmetros β através da equação normal. Para que o modelo tenha maior flexibilidade, é fundamental incluir uma coluna constante em X, o que permite a estimativa do intercepto e evita que o modelo seja forçado a passar pela origem.

A avaliação do desempenho nesse modelo é realizada por meio do Erro Quadrático Médio (EQM). O EQM é definido como a média dos quadrados das diferenças entre os valores previstos pelo modelo e os valores reais observados. Em outras palavras, para cada observação, calcula-se o erro subtraindo o valor real do valor previsto, eleva-se essa diferença ao quadrado e, posteriormente, soma-se todos esses valores quadráticos para dividir pelo número total de observações. Dessa forma, o objetivo central é minimizar o EQM, buscando identificar o modelo que ofereça as previsões mais precisas e que se aproximem ao máximo dos valores reais da atividade enzimática.

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2$$

Em que:

- $n$ representa o total de amostras disponíveis;
- $yi$ corresponde ao valor observado para a $ii$-ésima amostra;
- $y\hat{}i$ indica o valor estimado pelo modelo para a $i$-ésima amostra.

**MQO Regularizado (utilizando Tikhonov/Ridge):** Este método inclui um termo de penalização, ajustado pelo hiperparâmetro λ que controla a magnitude dos coeficientes e ajuda a evitar *overfittin*g. A equação de estimação passa a ser:

$$\beta = (X^T X)^{-1} X^T y$$

onde $I$ é a matriz identidade.

**Média dos valores observáveis:** Um método simples que define a previsão para cada amostra como a média dos valores observados no conjunto de treinamento, funcionando como uma linha de base para comparação com os modelos preditivos mais complexos.

Em todos os métodos, o cálculo e a inclusão do intercepto são realizados para garantir flexibilidade na modelagem dos dados.

No modelo regularizado, o valor de λ é determinante para o equilíbrio entre o ajuste ao conjunto de dados e a penalização dos pesos. Para explorar a sensibilidade do modelo, serão considerados os seguintes valores:

Quando *λ=0*, o método se iguala ao MQO tradicional. Assim, serão geradas 6 estimativas distintas do vetor β (com dimensão *(p+1)×1*), permitindo comparar os efeitos da regularização sobre o desempenho preditivo.

### *B. Tarefa de Classificação*

O segundo bloco de análise trata da classificação de expressões faciais a partir de sinais de eletromiografia. Os dados usados foram recuperados do arquivo EMGsDataset.csv. Esses dados foram coletados através de sensores Myoware Muscle Sensor acoplados a um microcontrolador NODEMCUESP32, com taxa de amostragem de 1 kHz e resolução de 12 bits (0–4095). As medições foram realizadas em dois pontos: o corrugador do supercílio e o zigomático maior, enquanto as expressões foram registradas em cinco categorias (1 – Neutro, 2 – Sorriso, 3 – Sobrancelhas levantadas, 4 – Surpreso e 5 – Rabugento), com repetições que totalizam 50.000 amostras.

Para possibilitar a aplicação dos métodos de classificação, os dados foram organizados em duas configurações:

Para modelos via MQO:

- Matriz X de dimensão $R^{nxp}$ (contendo as leituras dos dois sensores (com N=50000 e p = 2);
- Matriz Y de dimensão $R^{N\,x\,C}$ (em que cada linha codifica uma das cinco categorias).

Para modelos Gaussianos Bayesianos, a organização dos dados é transposta para:

$$X \in R^{P\,X\,N} \; e \; Y \in R^{C\,X\,N}$$

Assim como no modelo de Regressão, uma análise exploratória inicial foi realizada por meio de um gráfico (Figura 02) de dispersão, no qual as amostras são representadas tomando como base as leituras dos sensores. Cada ponto é codificado de acordo com a classe correspondente (neutro, sorriso, sobrancelhas levantadas, surpreso ou rabugento).

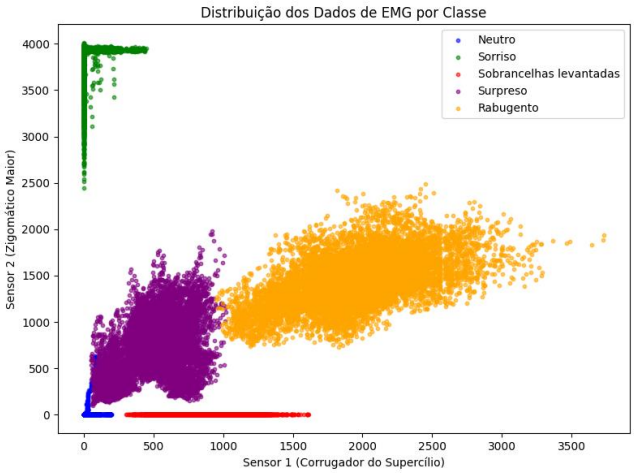


*Figura 2: Distribuição dos dados de EMG por classe*

Seguindo os pseudocódigos fornecidos, serão implementados os seguintes classificadores:

**MQO Tradicional para Classificação:** adapta os métodos dos mínimos quadrados, originalmente empregados em regressão, para problemas de classificação. Os parâmetros são estimados e o intercepto é incorporado da mesma maneira que na regressão, mantendo o mesmo fluxo algorítmico. A única modificação ocorre na etapa de avaliação: em vez de utilizar o Erro Quadrático Médio para medir a diferença entre os valores previstos e os reais, aplicamos uma métrica de acurácia que quantifica a correspondência entre a previsão e o rótulo verdadeiro. Por essa razão, não se utiliza um pseudocódigo distinto, pois os passos matemáticos e a estrutura do algoritmo permanecem inalterados, bastando ajustar a forma de avaliação do desempenho do modelo.

**Classificador Gaussiano por Máxima Verossimilhança:**

*Algorithm 1: Pseudocódigo dado modelo classificador gaussiano (pelo critério de máxima verossimilhança).*

*1: Dividir os dados em treinamento e teste.*
*2: De posse dos dados de treinamento, agrupar os C grupos de dados em ({**X**n1 ,**y**n1},{**X**n2,**y**n2},⋯ ,{**X**nC,**y**nC}).*
*Neste caso, cada **X**i, sendo i=1,2,3,⋯ ,C representa uma matriz de dimensões $R^{p\,x\,ni}$*
*3: Em cada grupo de dados referente a classe, calcular cada vetor médio **μ**i, i=1,2,⋯ ,C*
*4: Em cada grupo de dados referente a classe, calcular cada matriz de covariância **Σ**i, i=1,2,⋯ ,C.*
*5: É de interesse prático, antes de receber uma nova amostra de treinamento, computar $|\boldsymbol{\Sigma i}|$ e $\boldsymbol{\Sigma i^{-1}}$ para i=1,2,⋯ ,C.*
*6: forca da amostra $x_{novo} \in R^{p\,x\,1}$ no conjunto $X_{teste}$ **do***
*7: $\hat{y}i \leftarrow arg\ max\ ln\ [P(\boldsymbol{X}_{novo}|yi)]$ i=1,⋯ ,C*

*8: **end for***

*9: Compute a taxa de acerto e a taxa de erro do modelo*

**Classificador Gaussiano com Covariâncias Iguais:** Parte do mesmo princípio, mas assume-se uma única matriz de covariância Σ para todas as classes. A atribuição de rótulo é baseada na minimização da distância de Mahalanobis em relação aos vetores médios de cada classe.

*Algorithm 2: Pseudocódigo do modelo classificador gaussiano Covariâncias Iguais.*

*1: Dividir os dados em treinamento e teste.*

*2: De posse dos dados de treinamento, estimar a matriz de covariância única Σ, bem como o vetor de ponto médio, $\mu_i$ para cada categoria i = 1, 2, 3, ..., C.*

*3: É de interesse prático, antes de receber uma nova amostra de teste, computar |Σ| e $\Sigma^{-1}$*

*4: for cada amostra $X_{novo} \in R^{p\ x\ 1}$ no conjunto $X_{teste}$ do*

*5: $\hat{y}_i \leftarrow$ arg min $[(X_n - \mu_i)^T \Sigma^{-1} (X_n - \mu_i)]$ i = 1, 2, 3, ..., C*

*6:* ***end for***

*7: Compute a taxa de acerto e a taxa de erro do modelo.*

**Classificador Gaussiano com Covariância Agregada:** Aqui, a matriz de covariância é agregada para todas as classes, permitindo uma abordagem que unifica a variabilidade dos dados.

*Algorithm 3: Pseudocódigo do modelo classificador gaussiano - Covariância Agregada.*

*1: Dividir os dados em treinamento e teste.*

*2: De posse dos dados de treinamento, estimar a matriz de covariância agregada $\Sigma_{agregada}$, bem como o vetor de ponto médio, $\mu_i$ para cada categoria i = 1, 2, 3, ..., C.*

*3: É de interesse prático, antes de receber uma nova amostra de teste, computar $|\Sigma_{agregada}|$ e $\Sigma_{agregada}{}^{-1}$.*

*4: para cada amostra $X_{novo} \in R^{p\ x\ 1}$ no conjunto $\boldsymbol{X}_{teste}$* ***do***

*5: $\hat{y}_i \leftarrow$ arg min $[(X_i - \boldsymbol{\mu_i})^T \Sigma_{agregada}{}^{-1} (X_1 - \mu_i)]$ i = 1, 2, 3, ..., C*

*6:* ***end for***

*7: Compute a taxa de acerto e a taxa de erro do modelo.*

**Classificador Gaussiano Regularizado (Friedman):** As matrizes de covariância são regularizadas por meio de um ajuste com o hiperparâmetro λ. Assim como na regressão, serão testados os valores:

*Algorithm 4: Pseudocódigo do modelo classificador gaussiano (Regularização por Friedman).*

*1: Dividir os dados em treinamento e teste.*

*2: Definir o valor de λ.*

*3: De posse dos dados de treinamento, agrupar os C grupos de dados em ({Xn1, yn1}, {Xn2, yn2}, ... , {XnC, ynC}).*

*Neste caso, cada Xi, sendo i = 1, 2, 3, ..., C representa uma matriz de dimensões $P^{n\ x\ ni}$*

*4: Em cada grupo de dados referente a classe, calcular cada vetor médio $\mu_i$, i = 1, 2, ..., C.*

*5: Em cada grupo de dados referente a classe, calcular cada matriz de covariância $\Sigma_i^{\lambda}$ = 1, 2, ..., C.*

*6: É de interesse prático, antes de receber uma nova amostra de treinamento, computar $|\Sigma_i^{\lambda}|$ e $\Sigma_i^{\lambda\ -1}$ para i = 1, 2, ..., C.*

*7: for cada amostra $X_{novo} \in R^{p\ x\ 1}$ no conjunto $X_{teste}$* ***do***

*8: Faça a atribuição do rótulo $\hat{y}_i$ para a amostra $X_{teste}$ considerando a regra apresentada no slide anterior.*

*9:* ***end for***

*10: Compute a taxa de acerto e a taxa de erro do modelo.*

**Classificador de Bayes Ingênuo:** Adota a hipótese de independência entre as variáveis, simplificando o cálculo da inversa da matriz de covariância (onde os termos fora da diagonal são assumidos como zero). É crucial que essa simplificação permita a inversão dos parâmetros para a correta estimação das probabilidades.

### C. *Validação via Monte Carlo para Regressão e Classificação*

Empregamos a validação por simulações de Monte Carlo em ambas as tarefas, tanto para regressão quanto para classificação, utilizando 500 iterações em que, a cada rodada, os dados foram embaralhados aleatoriamente (por meio de *np.random.permutation*) e divididos em 80% para treinamento e 20% para teste. Essa abordagem consiste em dividir repetidamente o conjunto de dados em subconjuntos de treinamento e teste, o que permite que diferentes combinações de dados sejam utilizadas ao longo de diversas iterações. Dessa forma, cada amostra tem a oportunidade de participar tanto do treinamento quanto da validação em momentos distintos.

No caso da regressão, a cada rodada de Monte Carlo, os modelos foram avaliados com base no Erro Quadrático Médio (EQM), mensurando a diferença entre as previsões e os valores reais de atividade enzimática. Já para a classificação, embora o processo de estimação dos parâmetros (por MQO Tradicional, por exemplo) siga a mesma metodologia, a avaliação final é realizada utilizando a acurácia, que determina a taxa de correspondência entre as previsões e os rótulos verdadeiros.

Os resultados de cada rodada foram armazenados, permitindo o cálculo de estatísticas agregadas (média, desvio padrão, valor máximo e valor mínimo), o que garante um *feedback* confiável do desempenho dos modelos.

## III. RESULTADOS E DISCUSSÕES

Os resultados apresentados na Tabela 1, obtidos em 500 simulações de Monte Carlo, evidenciam diferenças

significativas no desempenho dos modelos avaliados por meio do Residual Sum of Squares (RSS). Em cada iteração, os dados foram divididos aleatoriamente em 80% para treinamento e 20% para teste, e os modelos foram avaliados com base no ajuste obtido.

| Modelo | Média | Std | Maior Valor | Menor Valor |
|---|---|---|---|---|
| **Média de valores observáveis** | 22.9296 | 1.2145 | 26.3616 | 19.2166 |
| **MQO tradicional** | 4.3233 | 0.4499 | 5.6355 | 3.1832 |
| **MQO regularizado (λ = 0.25)** | 4.3237 | 0.4502 | 5.6342 | 3.1858 |
| **MQO regularizado (λ = 0.5)** | 4.3245 | 0.4505 | 5.6332 | 3.1890 |
| **MQO regularizado (λ = 0.75)** | 4.3259 | 0.4509 | 5.6328 | 3.1926 |
| **MQO regularizado (λ = 1)** | 4.3276 | 0.4514 | 5.6394 | 3.1969 |

*Tabela 1: Estatísticas do RSS dos Modelos de Regressão avaliados por Monte Carlo (500 Rodadas)*

O modelo que previu a atividade enzimática utilizando exclusivamente a média dos valores observáveis apresentou um RSS substancialmente mais alto (média ≈ 22,93, desvio ≈ 1,2145; valor máximo ≈ 26,3616 e mínimo ≈ 19,2166), demonstrando que a simples predição pela média não capta a variabilidade dos dados.

Em contraste, o método dos Mínimos Quadrados Ordinários (MQO) tradicional obteve um RSS bem menor (média ≈ 4,3233, desvio ≈ 0,4499; variando de 3,1832 a 5,6355), o que indica um excelente ajuste ao considerar os preditores (temperatura e pH). Nos modelos regularizados (Tikhonov/Ridge), ao testar valores de λ (0,25; 0,5; 0,75; 1), observou-se um padrão consistente: o RSS médio aumenta levemente de aproximadamente 4,3237 para 4,3276, acompanhando um sutil encolhimento dos coeficientes, conforme esperado com a penalização aplicada.

Ademais, as estimativas do vetor β para o MQO tradicional indicam um intercepto de cerca de 2,5024 e coeficientes pequenos para os preditores (0,0724 para a temperatura e 0,0850 para o pH), sugerindo uma relação positiva, embora moderada, entre as variáveis independentes e a atividade enzimática; já o modelo que utiliza a média dos valores observáveis, ao desconsiderar os preditores, fornece um intercepto em torno de 2,5817 e zeros para os demais coeficientes.

Os achados resumidos na Tabela 1, confirmam a superioridade dos modelos que incorporam os preditores, especialmente o MQO tradicional e seus equivalentes regularizados, na obtenção de previsões precisas e na redução do RSS.

Ainda nesse tocante, os resultados da validação por Monte Carlo para a tarefa de classificação, apresentados na Tabela 2, demonstram claramente as diferenças de desempenho entre os modelos avaliados em 500 rodadas.

| Modelo | Média | Desvio | Máximo | Mínimo |
|---|---|---|---|---|
| **MQO** | 72.39% | 0.64% | 74.10% | 70.65% |
| **Gaussiano Tradicional** | 96.70% | 0.17% | 97.28% | 96.20% |
| **Gaussiano Cov. Iguais** | 96.26% | 0.18% | 96.78% | 95.72% |
| **Gaussiano Cov. Agregada** | 94.85% | 0.21% | 95.52% | 94.26% |
| **Naive Bayes** | 79.86% | 0.37% | 80.88% | 78.74% |
| **Friedman (λ=0)** | 96.70% | 0.17% | 97.28% | 96.20% |
| **Friedman (λ=0.25)** | 96.67% | 0.17% | 97.23% | 96.18% |
| **Friedman (λ=0.5)** | 96.64% | 0.17% | 97.18% | 96.16% |
| **Friedman (λ=0.75)** | 96.60% | 0.17% | 97.11% | 96.12% |
| **Friedman (λ=1)** | 96.55% | 0.17% | 97.08% | 96.05% |

*Tabela 2: Estatísticas de Acurácia dos Modelos de Classificação Avaliados por Monte Carlo (500 Rodadas)*

De forma resumida, o modelo MQO adaptado para problemas de classificação obteve uma média de acurácia de 72,39% (com desvio de 0,64%, máximo de 74,10% e mínimo de 70,65%), evidenciando sua limitação em contextos onde as fronteiras entre classes não são bem lineares.

Em contrapartida, os classificadores gaussianos se destacaram: o Gaussiano Tradicional alcançou uma média de 96,70% (desvio de 0,17%), enquanto as versões com covariâncias iguais e com matriz agregada apresentaram médias de 96,26% e 94,85%, respectivamente. O classificador Naive Bayes atingiu uma acurácia moderada de 79,86%, sugerindo que as suposições de independência entre as características limitam sua eficácia. Além disso, os modelos regularizados pelo método Friedman exibiram médias de acurácia muito próximas à do Gaussiano Tradicional – variando de 96,70% para λ = 0 a 96,55% para λ = 1 – indicando que a regularização tem um impacto sutil neste conjunto de dados.

Esses achados ressaltam que, embora todos os modelos compartilhem uma metodologia comum de avaliação (onde o critério final é a acurácia), os classificadores específicos, especialmente os baseados em estatísticas gaussianas, demonstram superioridade ao capturar as nuances dos sinais EMG para a separação das classes.

## IV. Conclusão

O relatório apresentou uma análise do uso de modelos de regressão e classificação aplicados a dados de atividade enzimática e sinais EMG.

Na regressão, os métodos baseados em MQO, especialmente quando regularizados, demonstraram um desempenho preditivo robusto por meio da redução significativa do RSS, evidenciando a importância da inclusão de preditores e da penalização para melhorar a precisão.

Por outro lado, na tarefa de classificação, os classificadores gaussianos se destacaram ao alcançar acurácias superiores a 96%, comprovando sua eficácia em capturar nuances e

separar categorias emocionais, enquanto abordagens como MQO e Naive Bayes, embora consistentes, apresentaram resultados inferiores.

A validação por Monte Carlo, realizada em 500 iterações, assegurou a confiabilidade dos resultados. Em resumo, os achados deste relatório reforçam a relevância de integrar técnicas avançadas de modelagem com validações sólidas, contribuindo significativamente para o avanço dos estudos em predição e classificação de dados experimentais.