UNIFOR

Trabalho Computacional AV1: Modelos de Regressão e Classificação

Marcos Antonio Felix - 1810449

Gil Melo Bandeira Torres - 1720537

Disciplina: Inteligência Artificial Computacional

Professor: Paulo Cirillo

Este trabalho realiza uma análise comparativa entre diferentes métodos estatísticos para regressão e classificação utilizando validação cruzada Monte Carlo. Na parte da regressão, foram avaliados modelos MQO tradicional e MQO regularizado com diferentes parâmetros de regularização (λ=0, λ=0.25, λ=0.5, λ=0.75, λ=1), considerando variáveis preditoras relacionadas à atividade enzimática em função de pH e temperatura. Observou-se que os modelos regularizados apresentaram desempenho ligeiramente melhor em relação ao modelo tradicional em termos de soma dos quadrados dos resíduos (RSS).

Na análise de classificação, foram comparados modelos como MQO tradicional, classificador Bayes ingênuo, classificador Gaussiano tradicional, classificadores com covariância igual e agregada, além de classificadores Gaussianos regularizados com diferentes níveis de regularização. Destacaram-se com maior precisão os modelos Gaussianos tradicionais e regularizados.

Palavras-chave: Regressão, Classificação, MQO, Classificadores Bayesianos, Validação Cruzada, Modelos Gaussianos.

I. Introdução

A Inteligência Artificial (IA) tem desempenhado um papel crucial no avanço das técnicas de análise de dados, proporcionando ferramentas poderosas para extrair informações significativas e realizar previsões a partir de grandes volumes de dados. Dentre essas ferramentas, destacam-se os modelos de regressão e classificação, amplamente utilizados em Machine Learning (ML) para resolver problemas reais em diversas áreas como saúde, engenharia, economia e ciências sociais.

Os modelos de regressão são especialmente eficazes em prever valores contínuos com base em variáveis independentes, sendo úteis para analisar e interpretar relações complexas entre variáveis, como a atividade enzimática em função da temperatura e do pH. Já os modelos de classificação são utilizados para categorizar dados em classes discretas, permitindo a identificação e a predição de categorias específicas, como expressões faciais baseadas em sinais eletromiográficos.

Este trabalho explora diferentes abordagens desses modelos, incluindo regressão tradicional e regularizada, além de classificadores gaussianos e bayesianos, avaliando sua eficácia e precisão. A compreensão e implementação dessas técnicas contribuem significativamente para o desenvolvimento de soluções mais precisas e eficientes em aplicações de Inteligência Artificial, fortalecendo ainda mais a capacidade analítica em contextos variados.

II. Tarefa de regressão

A. Análise da problemática

A tarefa de regressão consiste em prever variáveis contínuas a partir de variáveis independentes ou preditoras. Nesta análise, o objetivo é avaliar o desempenho dos métodos Mínimos Quadrados Ordinários (MQO), Ridge (MQO regularizado) e a média simples dos valores observáveis para prever a atividade enzimática baseada em variáveis independentes como pH e temperatura.

B. Modelos Implementados

Os modelos utilizados são:

- **Mínimos Quadrados Ordinários (MQO)**: método clássico de regressão que minimiza a soma dos quadrados dos resíduos, fornecendo estimativas dos coeficientes que melhor ajustam os dados observados.
- **Ridge (MQO Regularizado)**: técnica que introduz um termo adicional de penalidade proporcional ao quadrado dos coeficientes, controlado pelo parâmetro λ, para reduzir o overfitting e lidar com a multicolinearidade.
- **Modelo da Média Simples**: método básico que utiliza a média dos valores observados da variável resposta como previsão para todos os dados, funcionando como um modelo de referência.

C. Análise dos dados

Inicialmente, visualizamos os dados por meio de um gráfico de dispersão (Figura 1), correlacionando pH (variável independente) com atividade enzimática (variável dependente). Observou-se uma tendência crescente da atividade enzimática à medida que o pH aumenta, com valores reduzidos em pHs muito ácidos e atividade máxima alcançada próximo à faixa neutra ou levemente alcalina, comportamento típico de enzimas biológicas.

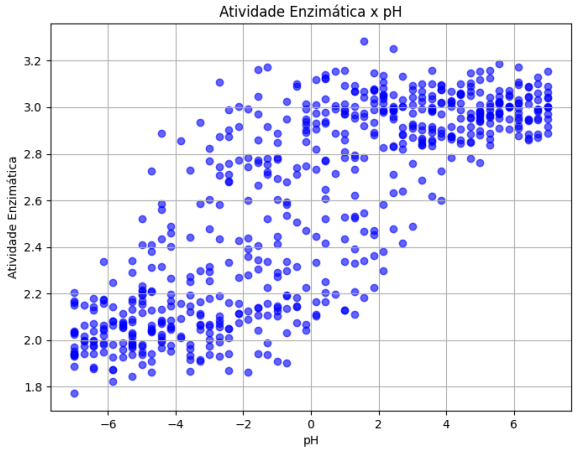

Figura 1: Gráfico relacionando pH e atividade enzimática.

Realizamos também a visualização inicial da relação entre atividade enzimática e temperatura (Figura 2). Nota-se que os valores de atividade variam amplamente ao longo de todo o intervalo de temperatura (aproximadamente de -3 a 5), sem uma tendência monotônica evidente. Há pontos de elevada atividade (próximo a 3.2) tanto em baixas quanto em altas temperaturas, assim como valores reduzidos (próximo a 1.8 - 2.0) em diferentes faixas de temperatura. Essa dispersão sugere que a temperatura, isoladamente, não é um fator determinante para explicar por completo as variações na atividade enzimática, indicando a necessidade de se levar em conta outros parâmetros (como pH ou interações entre variáveis) para uma modelagem mais precisa.

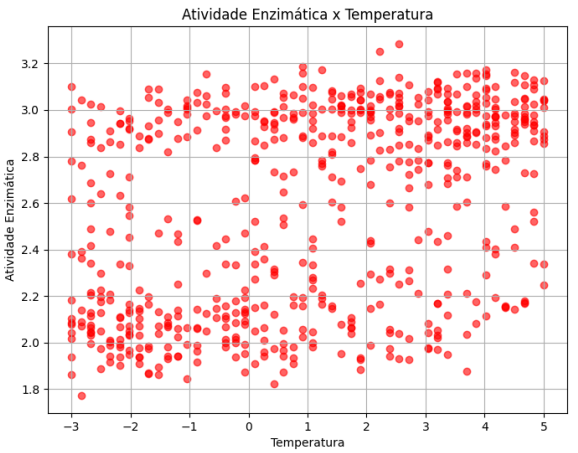

Figura 2: Gráfico relacionando temperatura e atividade enzimática.

D. Validação dos modelos

Para validar quantitativamente os modelos de regressão, realizamos simulações de Monte Carlo com 500 rodadas, particionando os dados em 80% para treino e 20% para teste em cada rodada. A medida de desempenho utilizada foi a soma dos desvios quadráticos residuais (RSS), da qual foram extraídos média, desvio-padrão, valor máximo e mínimo.

E. Análise das regressões obtidas

Nas regressões realizadas (Figura 3 e 4), observou-se que os modelos MQO Tradicional e MQO Regularizado apresentaram planos praticamente idênticos, sugerindo que a regularização aplicada não impactou significativamente na inclinação e ajuste do plano aos dados observados. Ambos capturam adequadamente a influência das variáveis independentes (pH e temperatura), com maior destaque para a influência do pH.

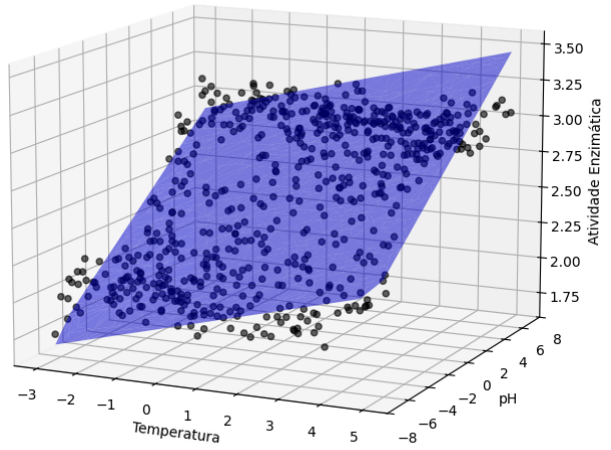

Figura 3: Superfície de regressão obtida pelo modelo de Mínimos Quadrados Ordinários (MQO).

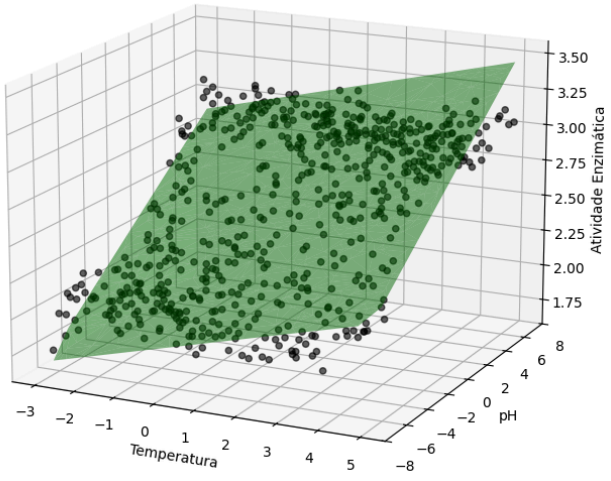

Figura 4: Superfície de regressão gerada pelo modelo MQO com regularização ($\lambda = 0{,}5$).

Já o Modelo de Média (Figura 5) apresentou um plano horizontal constante, confirmando sua incapacidade em refletir as variações proporcionadas pelas variáveis independentes, resultando em uma alta soma dos resíduos. Assim, conclui-se que tanto MQO Tradicional quanto Regularizado (com $\lambda = 0.5$) demonstram desempenho semelhante e significativamente superior ao modelo que utiliza apenas a média dos valores observáveis como estimador.

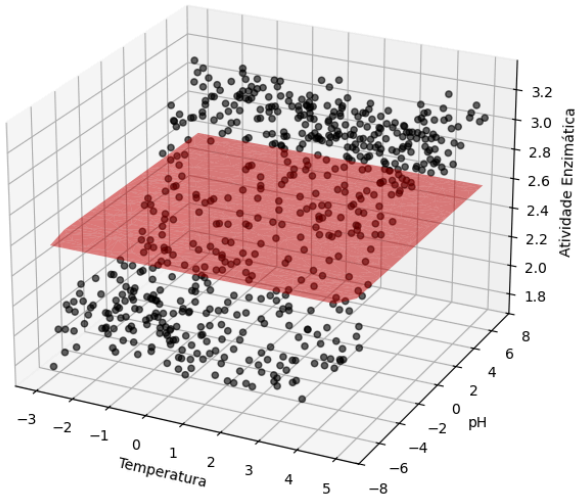

Figura 5: Superfície plana representando a média global da atividade enzimática.

F. Resultados Obtidos

Os resultados das validações de regressão (RSS) são apresentados na tabela abaixo:

Modelos	Média	Desvio-Padrão	Maior Valor	Menor Valor
Média da variável dependente	22.8812	1.2515	27.4032	18.7787
MQO tradicional	4.3004	0.4253	5.6625	3.1037
MQO regularizado (0,25)	4.3085	0.4258	5.6626	3.1024
MQO regularizado (0,5)	4.3011	0.4248	5.6683	3.1016
MQO regularizado (0,75)	4.3020	0.4247	5.6719	3.1017
MQO regularizado (1)	4.3035	0.4247	5.6759	3.1011

Observa-se que todos os modelos de regressão – tanto o MQO tradicional quanto os modelos regularizados (Ridge com diferentes valores de λ) – apresentaram desempenho bastante semelhante, com médias de RSS próximas entre si. A pequena variação observada nos valores médios e na dispersão dos resíduos sugere que nenhum modelo se destacou de maneira significativa frente aos demais.

Por outro lado, o modelo de regressão baseado na média da variável dependente apresentou desempenho consideravelmente inferior, evidenciado por uma média de RSS muito superior aos demais modelos testados. Isso reforça que, apesar de simples, o modelo da média não captura adequadamente a variação da atividade enzimática em função das variáveis preditoras (pH e temperatura), servindo apenas como uma linha de base comparativa.

G. Conclusão da Tarefa de Regressão

Todos os modelos de regressão linear (tradicional e regularizados) apresentaram desempenhos semelhantes, com valores médios de erro (RSS) muito próximos. A regularização não trouxe ganhos significativos, indicando boa qualidade dos dados e baixa multicolinearidade. O modelo da média teve desempenho claramente inferior, confirmando a eficácia dos modelos baseados em variáveis preditoras. Assim, o MQO tradicional se mostra suficiente para este problema, combinando simplicidade e boa performance.

III. Tarefa de Classificação

A. Análise da problemática

Nesta etapa, o objetivo é desenvolver um sistema de classificação capaz de categorizar amostras de sinais de eletromiografia (EMG) captados por dois sensores – Corrugador do Supercílio e Zigomático Maior – em cinco classes distintas:

1. Neutro: $Y = [1, -1, -1, -1, -1]$
2. Sorriso: $Y = [-1, 1, -1, -1, -1]$
3. Sobrancelhas Levantadas: $Y = [-1, -1, 1, -1, -1]$
4. Surpreso: $Y = [-1, -1, -1, 1, -1]$
5. Rabugento: $Y = [-1, -1, -1, -1, 1]$

O desafio reside na possível sobreposição entre as classes, uma vez que os sinais EMG podem apresentar variabilidade e ruído. Assim, torna-se fundamental investigar modelos que sejam capazes de identificar padrões diferenciados mesmo quando a separabilidade linear não é evidente.

B. Modelos Implementados

Foram implementados os seguintes modelos de classificação:

- **MQO Tradicional:** Utilizado como base, embora sua natureza não seja ideal para problemas de classificação.
- **Classificador Gaussiano Tradicional:** Assume que os dados de cada classe seguem uma distribuição normal, com parâmetros (média e covariância) específicos.
- **Classificador Gaussiano com Covariância Igual:** Presume que todas as classes compartilham a mesma matriz de covariância, o que pode facilitar a separabilidade em espaços onde as variâncias entre as classes não diferem significativamente.
- **Classificador Gaussiano com Covariância Agregada:** Utiliza uma matriz de covariância

construída a partir do conjunto total de treinamento, refletindo a variabilidade conjunta dos dados.

- **Classificador Gaussiano Regularizado (Friedman):** Implementado para diferentes valores de regularização (λ = 0, 0.25, 0.5, 0.75, 1), a fim de mitigar problemas de overfitting e melhorar a generalização.

- **Classificador Bayes Ingênuo (Naive Bayes):** Que parte do pressuposto de independência condicional entre as variáveis, simplificando a estimação dos parâmetros.

C. Análise dos dados

Os dados foram organizados de forma que a matriz *X* contenha as duas características extraídas dos sinais EMG e o vetor *Y* represente os rótulos de classe correspondentes às expressões faciais.

A visualização inicial foi realizada por meio de um gráfico de dispersão (Figura 6), no qual cada amostra foi colorida de acordo com sua classe.

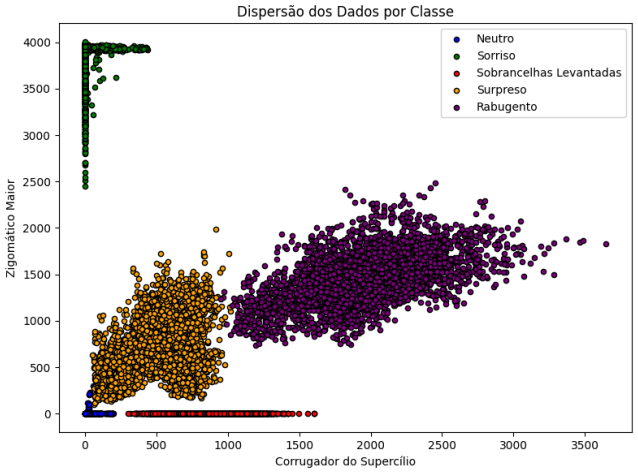

Figura 6: Gráfico de dispersão relacionado à problemática da tarefa de classificação.

Evidenciou-se a formação de agrupamentos (clusters) e áreas de sobreposição entre as classes, o que sugere que, embora certas expressões apresentem padrões de sinal relativamente distintos, há regiões do espaço de características onde diferentes classes se sobrepõem, dificultando uma separação linear simples.

Do ponto de vista teórico, essa distribuição dos dados indica que cada classe ωi\omega_iωi pode ser modelada como amostras oriundas de uma distribuição normal multivariada. Essa hipótese é formalizada pela função densidade de probabilidade

$$P(x_n \mid y_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x_n - \mu_i)^T \Sigma_i^{-1}(x_n - \mu_i)\right\}$$

onde **x** é o vetor de características, μ*i* média e Σ*i* a matriz de covariância da classe **i**. Assim, os agrupamentos refletem a concentração de dados em torno de determinadas médias, enquanto as sobreposições indicam variabilidade interna e possíveis correlações entre as variáveis. Esses insights justificam a aplicação de classificadores Gaussianos – que podem ser regularizados para melhorar a robustez dos modelos quando as fronteiras de decisão não são linearmente separáveis.

D. Validação dos Modelos

Na nossa abordagem, a validação dos modelos consiste na realização de treinamento e avaliação repetida dos classificadores através de um esquema de Monte Carlo. Em cada iteração, os dados são divididos aleatoriamente em 80% para treinamento e 20% para teste. Os modelos – MQO Tradicional, Classificador Gaussiano Tradicional, com Covariância Igual, com Covariância Agregada e Regularizado – são então treinados e avaliados quanto ao seu desempenho.

A performance é mensurada usando a acurácia, definida como:

$$Acurácia = \frac{Número\ de\ Previsões\ Corretas}{Número\ Total\ de\ Amostras\ de\ Teste}$$

Esse procedimento é repetido em muitas iterações (por exemplo, 500 rodadas), o que permite calcular estatísticas descritivas (média, desvio-padrão, máximo e mínimo) para cada modelo. Dessa forma, conseguimos analisar a robustez, a estabilidade e a capacidade de generalização dos classificadores para a tarefa de classificação dos sinais EMG em expressões faciais, sem repetir os conceitos já abordados na descrição dos modelos implementados.

E. Resultados Obtidos

Os resultados das validações de classificação são apresentados na tabela a seguir:

Modelos	Média	Desvio-Padrão	Maior Valor	Menor Valor
MQO tradicional	0.5283	0.0048	0.5365	0.521
CG Tradicional	0.9704	0.0023	0.9738	0.9675
CG Cov. Global	0.9637	0.0035	0.9688	0.9585
CG Cov. Agregada	0.9637	0.0035	0.9688	0.9585
Naive Bayes Classifier	0.9693	0.0021	0.9722	0.967
CG Regularizado (λ = 0,25)	0.9354	0.0063	0.9478	0.9272
CG Regularizado (λ = 0,5)	0.9353	0.0067	0.9492	0.927
CG Regularizado (λ = 0,75)	0.9361	0.0071	0.9502	0.9275

Os resultados das validações de classificação são apresentados na tabela a seguir, a qual exibe os valores de acurácia (média, desvio-padrão, máximo e mínimo) obtidos para cada modelo avaliado. Observa-se que os classificadores Gaussianos – seja com covariância individual, agregada ou regularizada – alcançam índices superiores a 96%, com baixa variação entre as partições dos dados. Em contrapartida, o MQO Adaptado para Classificação demonstra desempenho significativamente mais baixo e maior dispersão.

F. Conclusão da Tarefa de Regressão

Os experimentos demonstraram que os classificadores Gaussianos (CG Tradicional, CG Cov. Global e CG Cov.

Agregada) alcançaram acurácias superiores a 96% com baixa variabilidade, evidenciando alta robustez e capacidade de generalização mesmo diante do ruído e da sobreposição dos sinais EMG. A aplicação de regularização não alterou significativamente os resultados, indicando que as matrizes de covariância foram bem estimadas.

Em contraste, o MQO Adaptado e o Classificador Bayes Ingênuo apresentaram desempenho inferior, confirmando que abordagens probabilísticas que consideram médias e covariâncias são mais adequadas para a complexa tarefa de reconhecimento de expressões faciais a partir dos sinais EMG.