Regressão Logística

1. Introdução

A regressão logística é um dos modelos mais utilizados para tarefas de classificação supervisionada. Embora tenha origem na estatística, seu uso é amplamente difundido em aplicações de aprendizado de máquina, especialmente no processamento de linguagem natural (PLN), sistemas de recomendação e diagnóstico médico. Seu principal objetivo é estimar a probabilidade de um dado pertencer a uma determinada classe com base em características extraídas dos dados de entrada.

2. Como a Regressão Logística Funciona

A base da regressão logística é um modelo matemático que recebe variáveis de entrada numéricas, realiza uma combinação linear com pesos aprendidos durante o treinamento, e em seguida aplica uma função sigmoide para transformar esse valor em uma probabilidade entre 0 e 1.

Etapas do processo:

1. **Extração de características**: os dados de entrada (como um texto ou imagem) são representados numericamente (por exemplo, por meio de vetores de palavras ou atributos estatísticos).
2. **Combinação linear**: cada característica é multiplicada por um peso e somada a um viés.
3. **Função sigmoide**: transforma a soma obtida em um valor entre 0 e 1, que é interpretado como uma **probabilidade da classe positiva**.
4. **Decisão**: se a probabilidade for maior que um determinado **limiar (threshold)**, o exemplo é classificado como pertencente à classe positiva; caso contrário, à negativa.

3. Classificação Binária e Multiclasse

A regressão logística é naturalmente adequada para classificação binária (ex: spam ou não-spam, positivo ou negativo), mas também pode ser estendida para múltiplas classes. Essa versão é chamada de regressão logística multinomial ou softmax regression.

Diferenças principais:

- **Binária**: utiliza a função **sigmoide** para gerar uma probabilidade entre duas classes.
- **Multiclasse**: usa a função **softmax** para calcular a probabilidade de pertencimento a cada uma das n classes possíveis. Essa função garante que todas as probabilidades somem 1.

4. Treinamento do Modelo

Durante o treinamento, o modelo precisa ajustar os pesos e o viés para que as previsões se aproximem o máximo possível dos valores reais. Isso é feito por meio da minimização de uma função de perda, normalmente a perda de entropia cruzada (cross-entropy loss).

Por que a entropia cruzada?

A entropia cruzada mede a diferença entre a distribuição de probabilidades prevista pelo modelo e a real (alvo). Quanto menor essa diferença, melhor o modelo está se ajustando.

Otimização

A minimização da função de perda é um problema convexo, o que significa que não há múltiplos mínimos locais — o que facilita o uso de métodos de otimização como o gradiente descendente para encontrar os melhores pesos.

5. Regularização

Para evitar que o modelo memorize os dados de treino e perca capacidade de generalização (ou seja, para evitar overfitting), utiliza-se regularização.

A regularização penaliza pesos muito altos, incentivando o modelo a encontrar soluções mais simples e robustas. Os dois tipos mais comuns são:

- **L1 (Lasso)**: incentiva a eliminação de características irrelevantes.
- **L2 (Ridge)**: reduz o impacto de pesos muito altos.

6. Interpretação do Modelo

Uma das grandes vantagens da regressão logística é a sua transparência. Diferente de modelos complexos como redes neurais profundas, ela permite uma interpretação direta dos pesos associados a cada característica. Isso é valioso em domínios como saúde, direito ou ciências sociais, onde é importante entender por que o modelo chegou a determinada decisão.

Por exemplo:

- Se uma determinada palavra ou característica tem um peso positivo elevado, isso indica que sua presença aumenta a chance do dado pertencer à classe positiva.
- Um peso negativo forte indica o contrário.

Essa propriedade torna a regressão logística não apenas uma ferramenta preditiva, mas também um instrumento analítico para investigar padrões nos dados.

7. Aplicações Práticas

A regressão logística é amplamente aplicada em diversas áreas:

- **Análise de sentimentos** em textos (positiva/negativa)

- **Diagnóstico médico** (presença ou ausência de uma condição)
- **Detecção de fraudes** (transação legítima ou suspeita)
- **Classificação de clientes** (propenso ou não a comprar um produto)
- **Modelagem de churn** (clientes que vão cancelar um serviço)

8. Conclusão

A regressão logística é um dos modelos mais utilizados e estudados da ciência de dados. Ela oferece uma combinação poderosa de:

- Simplicidade
- Eficiência
- Capacidade preditiva
- Facilidade de interpretação

Embora existam modelos mais complexos e com maior poder de generalização, a regressão logística é muitas vezes o primeiro modelo a ser testado em um problema de classificação, e muitas vezes já fornece resultados competitivos, servindo como baseline confiável.