O que é erro em Aprendizado de Máquina

Nos textos anteriores, formalizamos aprendizado como um problema de minimização de uma função de perda L( ) e discutimos o papel do gradiente na atualização dos parâmetros [1], [2].

2026-02-18

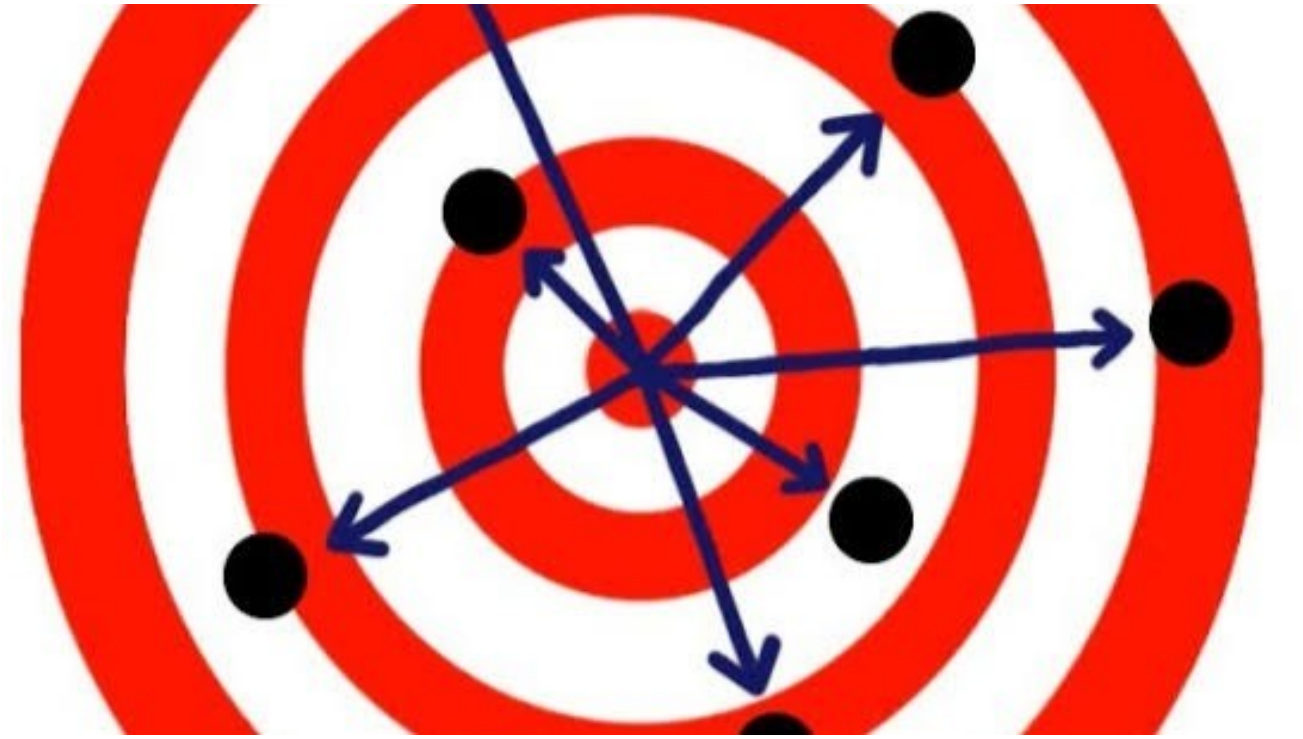


Figure 1: Capa do artigo

Nos textos anteriores, formalizamos aprendizado como um problema de minimização de uma função de perda L( ) e discutimos o papel do gradiente na atualização dos parâmetros [1], [2].

Agora o foco é outro: compreender o erro como objeto matemático, geométrico e estatístico.

Erro como objeto estatístico

Considere a perda:

$$L(f_\theta(x), y)$$

O objetivo teórico do treinamento é minimizar o erro esperado na distribuição D:

$$\mathbb{E}_{(x,y)\sim D}\left[L(f_\theta(x), y)\right]$$

© Artemisia Weyl 2024–2026 — Todos os direitos reservados.

Como D é desconhecida, utilizamos a média empírica:

$$\frac{1}{n}\sum_{i=1}^{n} L(f_\theta(x_i), y_i)$$

A diferença entre essas duas quantidades define o erro de generalização [1].

É nesse ponto que surgem dois regimes clássicos:

1. Underfitting : O modelo apresenta erro elevado tanto na amostra de treino quanto em dados novos.A capacidade do modelo é insuficiente para representar a estrutura dos dados.
2. Overfitting: O erro empírico é reduzido na amostra de treino, mas o erro em dados externos permanece elevado. O modelo ajusta padrões específicos da amostra que não se mantêm na distribuição completa.

Ambos são fenômenos estatísticos ligados à diferença entre média empírica e erro esperado.

Erro como superfície geométrica

A função L( ) define uma superfície no espaço dos parâmetros. Cada vetor corresponde a um ponto nessa superfície. Durante o treinamento buscamos pontos críticos onde:

$$\nabla L(\theta) = 0$$

Para analisar a natureza desses pontos, introduzimos a matriz Hessiana. A Hessiana H( ) é formada pelas segundas derivadas:

$$H_{ij} = \frac{\partial^2 L}{\partial \theta_i \, \partial \theta_j}$$

Ela descreve a curvatura local da superfície de perda [2]. Se a Hessiana for definida positiva, isto é, se todos os autovalores forem positivos, o ponto é um mínimo local.

Se houver autovalores de sinais distintos, o ponto é classificado como ponto de sela.Em espaços de alta dimensão, pontos de sela são estatisticamente mais frequentes que minimos locais.

Convexidade

A complexidade da superfície de erro depende da arquitetura do modelo Essa estrutura geométrica nos leva ao conceito de convexidade.

Funções convexas possuem a propriedade de que qualquer mínimo local é também mínimo global. Modelos lineares com funções de perda adequadas (como MSE) possuem essa propriedades.

Em termos gerais uma função L( ) é convexa quando, para quaisquer 1 e 2 e para todo em [0,1]:

$$L(\lambda\theta_1 + (1-\lambda)\theta_2) \leq \lambda L(\theta_1) + (1-\lambda)L(\theta_2)$$

Se L for duas vezes diferenciável, convexidade equivale à Hessiana ser semidefinida positiva para todo [2]. Em problemas convexos, qualquer mínimo local coincide com o mínimo global.

Redes neurais profundas geram superfícies não convexas devido às composições não lineares entre camadas. Isso resulta em múltiplas regiões de mínimos e exige estratégias de otimização como o Gradiente Descendente Estocástico (SGD).

© Artemisia Weyl 2024–2026 — Todos os direitos reservados.

Regularização e controle da complexidade

Quando o modelo possui muitos parâmetros, ele pode ajustar pequenas flutuações da amostra Esse ajuste excessivo reduz o erro empírico, mas não necessariamente reduz o erro esperado.Além da interpretação estatística, há um efeito geométrico direto.

O controle da complexidade pode ser incorporado diretamente na função objetivo. Para controlar esse comportamento, modificamos a função de perda:

Regularização modifica a função de perda:

$$L_{\text{reg}}(\theta) = L(\theta) + \lambda\Omega(\theta)$$

Onde o termo $\Omega(\ )$ mede complexidade do modelo, o parâmetro controla o peso dessa penalização.

Na regularização L2, adicionamos um termo proporcional ao tamanho dos parâmetros. Isto é fazemos:

$$\Omega(\theta) = \|\theta\|^2$$

Onde:

$$\|\theta\|^2 = \theta_1^2 + \theta_2^2 + \cdots + \theta_k^2$$

Logo:

$$L_{\text{reg}}(\theta) = L(\theta) + \lambda\|\theta\|^2$$

 é um número positivo que controla a intensidade da penalização.

Isso significa que parâmetros com grande magnitude aumentam o valor total da perda. Estatisticamente, reduz a variância do modelo.

Esse termo altera a curvatura da superfície e influencia o condicionamento da Hessiana.Sob a perspectiva estatística, regularização atua como mecanismo de controle de complexidade, reduzindo o risco de overfitting [1].

Sem regularização, modelos complexos tendem a reduzir o erro empírico rapidamente.Com regularização excessiva, o modelo pode perder capacidade expressiva e aumentar o erro tanto em treino quanto em teste.

O parâmetro estabelece um equilíbrio entre:

- Ajuste aos dados observados
- Controle da complexidade do modelo

Essa escolha influencia diretamente o erro de generalização.

Conclusões

Erro em aprendizado de máquina pode ser analisado em três níveis integrados:

- Estatístico — diferença entre erro empírico e erro esperado
- Geométrico — curvatura da superfície definida por L( )
- Estrutural — modificação da perda via regularização

Essa estrutura conecta cálculo diferencial, otimização e teoria estatística

© Artemisia Weyl 2024–2026 — Todos os direitos reservados.

Referências

[1] T. M. Mitchell, *Machine Learning*, 1997.

[2] T. M. Apostol, *Calculus Vol. II*; S. Boyd and L. Vandenberghe, *Convex Optimization*.

Postagem original no LinkedIn: O que é erro em Aprendizado de Máquina

© Artemisia Weyl 2024–2026 — Todos os direitos reservados.