Introdução à otimização aplicada ao aprendizado de máquina supervisionado

1 Resumo

Compõem o que chamamos de *aprendizado de máquina*, modelos matemáticos/algoritmos capazes de predizer a resposta à uma dada situação (por exemplo, se um paciente está doente ou não a partir de exames de imagem; identificar caracteres escritos à mão a partir de padrões preestabelecidos; identificar padrões de rostos, figuras geométricas etc). Assim, o interesse é o estudo de algoritmos e modelos capazes de fornecerem respostas esperadas à dados de entrada desconhecidos, "treinando" o modelo a partir de pares entrada/saída conhecidos (aprendizado *supervisionado*). É neste contexto que esta pesquisa se encaixa. O "treinamento" envolve a otimização dos parâmetros de um modelo matemático e, por conseguinte, algoritmos de otimização são utilizados em sua resolução. O objetivo geral desta pesquisa é estudar conceitos fundamentais do aprendizado de máquina supervisionado, modelos matemáticos e algoritmos de resolução. Especificamente, noções de probabilidade, otimização irrestrita e o método do gradiente estocástico são as principais ferramentas a serem abordadas.
Palavras-chave: Otimização. Probabilidade. Aprendizado de máquina. Gradiente estocástico.

2 Introdução

O termo "inteligência artificial" foi cunhado ainda nos anos 1950. Nos dias atuais, é empregado em incontáveis situações, como forma de transmitir ao público leigo a ideia da construção de máquinas capazes de tomar decisões que imitam a inteligência humana. De fato, são comuns textos jornalísticos, entrevistas, cursos on-line, filmes e produtos com essa temática. Embora muitas vezes o tema seja apresentado de maneira fantasiosa, ou até mesmo falaciosa, a chamada inteligência artificial (IA) possui protocolos bem estabelecidos. Não à toa, o tema hoje está presente nos currículos de cursos de ciência da computação, de engenharias, ou mesmo em cursos de matemática aplicada. A visão moderna de IA baseia-se em técnicas estatísticas, em contraponto à visão clássica, baseada na lógica. Este novo ponto de vista nos permite lidar com uma grande quantidade de dados, hoje largamente disponíveis, e é a chave para o sucesso das inúmeras e crescentes aplicações de IA. Um caso particular é o que chamamos de *aprendizado de máquina supervisionado*: a construção de modelos capazes de predizer a resposta à uma dada situação, ajustados a partir de dados para os quais sabemos as respostas reais. São inúmeras suas aplicações. A grosso modo, devemos ajustar os parâmetros de um modelo/sistema utilizando dados disponíveis no momento (dados de treinamento), para os quais sabemos as respostas (saídas) para cada entrada. A expectativa assim é que o modelo otimizado dê respostas esperadas à dados de entrada ainda desconhecidos, e que não foram utilizados em seu "treinamento". Nesse contexto, a programação matemática/otimização constitui-se como um dos pilares do aprendizado de máquina, ao passo que algoritmos de otimização são utilizados para otimizar o modelo aos dados de treinamento. **Esta pesquisa se propõe ao estudo de modelos e técnicas de otimização utilizados no aprendizado de máquina supervisionado.** Em particular, o método do gradiente estocástico surge como principal ferramenta da área (BOTTOU; CURTIS; NOCEDAL, 2018).

Especificamente, o objetivo é determinar uma função $h: X \to Y$ de um espaço de entrada X a um espaço de saída Y de forma que, dado uma entrada $x \in X$, o valor $h(x)$ forneça uma resposta fidedigna à saída real $y \in Y$. A ideia então é encontrar uma função h que minimiza o risco esperado do erro (BOTTOU; CURTIS; NOCEDAL, 2018)

$$R(h) = P[h(x) \neq y],$$

isto é, a probabilidade de $h(x)$ ser diferente da resposta real y. Evidentemente, minimizar $R(h)$ na prática é inviável, dado que, por exemplo, não conhecemos P. A fim de construir um modelo de otimização tratável, tomamos as seguintes decisões:

1. Fixamos a forma da função h. Usando a notação de Bottou, Curtis e Nocedal (2018), consideramos uma família de funções h parametrizadas por um vetor w:

   $$H = \{h(.,w); w \in \mathbb{R}^d\}.$$

   Dado um par entrada/saída (x, y), definimos uma medida do erro $l(h(x,w), y)$ relativo à saída predita $h(x,w)$ frente a saída real y;

2. O objetivo passa a ser minimizar o erro esperado dentre todas as possíveis entradas, isto é, minimizar o *risco esperado*

   $$R(w) = \int l(h(x,w), y) dP(x,y) = E[l(h(x,w), y)],$$

   onde P é uma distribuição de probabilidade e E denota o valor esperado/esperança. Porém, como já dito, desconhecemos P. A única informação que temos em mãos são pares entrada/saída $(x_1, y_1), \cdots, (x_n, y_n)$ (dados de treinamento). Portanto, nosso objetivo passa a ser minimizar o *risco empírico*

   $$R_n(w) = \frac{1}{n} \sum_{i=1}^{n} l(h(x_i, w), y_i).$$

   Esta função é tratável por métodos de otimização.

Um dos métodos para minimização sem restrições é o clássico *método do gradiente* (RIBEIRO; KARAS, 2013). Basicamente, este procedimento consiste em dar passos na direção contrária à do gradiente da função f a ser minimizada (direção em que a função decresce localmente), resultando na iteração

$$x^{k+1} = x^k - t_k \nabla f(x^k),$$

onde t_k é o tamanho do passo. Este método, em tese, poderia ser utilizado para minimizar R_n. No entanto, isso levaria à escolha de *todos* os pontos de treinamento em *toda* iteração, o que resulta em processo com viés, com perda da aleatoriedade (lembre-se que o objetivo ideal a ser perseguido é minimizar R, R_n é um modelo). Logo a escolha dos dados de treinamento deve ser feita de forma o mais independente possível de uma iteração para outra. Tal tarefa pode ser cumprida com uma variante "estocástica" do método do gradiente: ao invés de considerar a soma completa em R_n, sobre os n termos, tomamos somas parciais em iterações diferentes. A escolha de quais termos considerar numa dada iteração varia: podemos escolher, por exemplo, um único termo ou um número pré-determinado deles. A escolha é feita aleatoriamente, o que justifica o nome "gradiente estocástico". A validação do modelo, isto é, seu teste de eficácia frente a dados desconhecidos, pode ser realizada com parte dos dados disponíveis. Ou seja, separamos uma parte dos dados para treinamento e outra para validação.

O tema escolhido é desafiador pelo seguintes motivos: (i) não é comum ser tratado em cursos de matemática, pelo menos no Brasil. O estudante é aluno do curso de bacharelado em Matemática Industrial (aplicada) da UFES, que visa formar matemáticos capazes de lidar com aplicações industriais. Mesmo entre matemáticos já graduados, incluindo este coordenador, ainda hoje é raro o domínio do tema; (ii) é um tema que, apesar de existir desde os anos 1950, ganhou projeção só muito recentemente, tendo inúmeras novas aplicações; e (iii) para além da conexão entre matemática e ciências da computação (que hoje parece congregar a maioria dos pesquisadores do tema), certamente o conhecimento em matemática contribui para o avanço da pesquisa. Faz sentido portanto a popularização do tema entre matemáticos. Nesse sentido, este subprojeto, além da orientação do estudante em si, representa um primeiro passo deste coordenador ao aprofundamento do tema.

3 Referências

BOTTOU, L.; CURTIS, F. E.; NOCEDAL, J. Optimization Methods for Large-Scale Machine Learning. **SIAM Review**, v. 60, n. 2, p. 223-311, 2018.

GAMBELLA, C.; GHADDAR, B.; NAOUM-SAWAYA, J. Optimization models for machine learning: a survey. **ArXiv:1901.05331**, 2019.

GOODFELLOW, I; BENGIO, Y; COURVILLE, A. **Deep Learning: Adaptive Computation and Machine Learning.** The MIT Press, 2016.

JAMES, B. R. **Probabilidade: um curso em nível intermediário.** 3 ed. Rio de Janeiro: IMPA, 2010.

RIBEIRO, A. A.; KARAS, E. W. **Otimização Contínua.** São Paulo: Cengage, 2013.