A Importância da Qualidade dos Dados em Machine Learning

A importância dos dados para o bom funcionamento dos algoritmos de Machine Learning (ML) é fundamental, pois até os códigos mais sofisticados podem falhar se os dados forem inadequados. Desta forma, é iniciada uma ampla discussão sobre as etapas cruciais na criação de dados de treinamento. Inicialmente, são explorados os métodos de amostragem, tanto probabilísticos quanto não probabilísticos, como ferramentas fundamentais para garantir a qualidade e representatividade dos dados. Além disso, evidência-se a importância da obtenção de rótulos, especialmente em algoritmos supervisionados, e as dificuldades associadas à rotulagem manual. Alternativas, como supervisão fraca e semi-supervisão, surgem como soluções para lidar com esse desafio. Também, é avaliado as implicações do desequilíbrio de classes nos resultados dos algoritmos de ML, assim como as estratégias para mitigar esse problema, incluindo técnicas de reamostragem e ajuste de funções de perda. Por fim, tem-se uma análise das técnicas de aumento de dados, destacando seu papel na melhoria do desempenho e generalização dos modelos, particularmente em tarefas de visão computacional e Processamento de Linguagem Natural (PNL).

Baseado nas ideias do texto[1], pode-se destacar alguns conceitos centrais como:

Amostragem: é parte integrante do fluxo de trabalho de ML, pois aparece em muitas etapas de um projeto, é necessária em casos onde não se tem acesso a todos os dados ou quando processar todos os dados disponíveis é impraticável. A amostragem pode ser não probabilística e aleatória, para garantir a representatividade dos dados de treinamento.

Qualidade dos Dados: destaca-se a importância da qualidade dos dados de treinamento para o desempenho dos algoritmos de ML, enfatizando a necessidade de dados precisos e confiáveis de dados rotulados.

Aquisição de Rótulos: é importante a obtenção de rótulos para os dados de treinamento, especialmente em algoritmos supervisionados, no entanto existem dificuldades associadas à rotulagem manual.

Supervisão Fraca e Semi-Supervisão: aqui é explorada alternativas à rotulagem manual, como supervisão fraca e semi-supervisão, para lidar com a escassez de rótulos e reduzir custos e tempo de rotulagem.

Transferência de Aprendizado: aborda a técnica de transferência de aprendizado como uma forma de utilizar conhecimento prévio de modelos pré-treinados para tarefas semelhantes, reduzindo a necessidade de dados rotulados.

Viés de Dados e Preconceitos: os vieses e preconceitos podem estar presentes nos dados de treinamento e isso pode afetar negativamente o desempenho e a neutralidade dos modelos de ML.

Desequilíbrio de Classe: tem-se os desafios associados ao desequilíbrio de classe nos dados de treinamento e as técnicas para lidar com esse problema, como a reamostragem de dados e o ajuste de funções de perda.

Métricas de Avaliação: Explora a seleção adequada de métricas de avaliação para modelos de ML, levando em consideração o contexto específico do problema e os objetivos de negócio.

Aumento de Dados: a importância das técnicas de aumento de dados para melhorar o desempenho e a generalização dos modelos, especialmente em tarefas de visão computacional e Processamento de Linguagem Natural (PNL).

Validação e Teste de Modelos: Por fim tem-se a necessidade de validação e teste rigorosos dos modelos de ML para garantir sua eficácia e generalização, incluindo a utilização de conjuntos de validação e teste independentes dos dados de treinamento.

Pode-se finalizar com a ideia clara de que, é necessário investir tempo e esforço nos dados de treinamento, para que se tenha um desempenho eficaz dos algoritmos de ML. Este investimento deve estar associado a técnicas específicas, de modo que ao final do processo tenhamos dados com qualidade e representatividade.

---

1 Chip Huyen, “Desegning Machine Learning System”, 2022. Capítulo 4.