Capítulo 4 - Designing Machine Learning Systems - Chip Huyen -
Pontos-chave sobre os dados de treinamento nos sistemas de ML

Os cientistas de dados e os engenheiros de Machine Learning (ML) devem aprender a lidar bem com os dados. Lidar com a modelagem de sistemas é por muitos considerada a parte "divertida" do processo, e por vezes lidar com uma enorme quantidade de dados que muitas vezes nem cabem na memória é frustrante. Para tanto, a criação de dados de treinamento é um processo iterativo, e técnicas são utilizadas para obter ou mesmo criar bons dados de treinamento. Para tanto, são listados os conceitos e percepções mais relevantes sobre os dados de treinamento, a saber:

1. **Amostragem não probabilística**: acontece quando a seleção de dados não baseia-se em nenhum critério de probabilidade, ou seja, elas não são representativas dos dados do mundo real, logo, estão repletas de vieses;
2. **Amostragem aleatória simples:** acontece quando se dá a todas as amostras da população probabilidades iguais de serem selecionadas. No entanto, categorias de dados aqui podem não aparecer;
3. **Amostragem estratificada:** aqui a amostra é dividida em grupos e em cada grupo faz-se uma amostra separadamente. No entanto, quando é impossível dividir todas as amostras em grupo, esse método é inviável. No caso de uma amostra pertencer a vários grupos, como no caso de tarefas com vários rótulos;
4. **Amostragem ponderada:** aqui cada amostra recebe um peso, que determina a probabilidade de ser selecionada. Os dados mais recentes são valiosos, e portanto, pode-se atribuir peso maior para que se tenha uma maior chance de ser selecionada;
5. **Amostragem do reservatório:** aqui é essencial quando precisa-se lidar com dados de fluxo contínuo. Aqui de tal modo todas as amostras têm a mesma chance de serem selecionadas;
6. **Amostragem de importância:** é um dos métodos de amostragem mais importantes. Ela só nos permite fazer uma amostragem de distribuição quando temos acesso a outra distribuição. No ML essa amostragem é utilizada no aprendizado por reforço baseado em políticas;
7. **Rotulagem manual:** os sistemas de ML precisam de dados rotulados para aprender. O desempenho do ML depende muito da qualidade e quantidade dos dados resultados com os quais ele é treinado. Na rotulagem manual é necessário que alguém examine

os dados, e isso nem sempre é possível se os dados tiverem requisito de privacidade. Essa tarefa representa uma ameaça à privacidade dos dados. Também essa técnica é lenta, o que leva a uma velocidade de iteração lenta e o modelo de ML se torna menos adaptável a ambientes e requisitos em constante mudança.

8. **Rotulagem natural:** os rótulos podem ser inferidos no sistema sem a necessidade de anotações humanas. Os rótulos aqui tem previsões do modelo sendo avaliadas automaticamente ou parcialmente pelo sistema. Os rótulos naturais geralmente são atrasados, e o tempo até que uma previsão é fornecida até o momento do *feedback* é o comprimento do *loop* de *feedback.*
9. **Ausência de rótulos - Supervisão fraca**: nessa técnica as pessoas confiram na heurística, que é desenvolvida a partir da experiência no assunto, para rotular os dados. Sendo assim, a supervisão fraca é um paradigma simples, mas poderoso. Porém, não é perfeito. Existem casos em que os dados obtidos podem ser muito ruidosos para serem úteis.
10. **Ausência de rótulos - Aprendizagem por transferência:** aqui um modelo é utilizado como ponto de partida para um novo modelo em uma tarefa subsequente. O modelo básico é treinado, com dados de treinamento baratos e abundantes. Logo, o modelo assim treinado pode ser utilizado para assunto de interesse, como análise de sentimento, detecção de intenção ou respostas a perguntas.

Portanto, compreende-se que os algoritmos de ML funcionam bem em situações em que a distribuição de dados é mais equilibrada. Do contrário, não funciona tão bem quando as classes são desequilibradas. E esse é um problema do mundo real. Logo, para combater a falta de rótulos manuais, técnicas foram desenvolvidas, como a supervisão fraca, aprendizagem ativa, aprendizagem por transferência e a semisupervisão. E também técnicas de aumento de dados que podem ser utilizadas para melhorar o desempenho do modelo por meio de tarefas de visão computacional.