Resumo dos 3 Primeiros Capítulos do Livro: Designing Machine Learning Systems (Chip Huyen 2022)

Autor: José Lindenberg de Andrade

Este resumo está dividido em III partes, correspondentes a cada capítulo lido:

Capítulo I:

O Primeiro capítulo trata de uma compreensão do que é necessário para trazer o Machine Learning (ML) para o mundo real. Foram mostrados uma ampla gama de casos de ML em produção hoje, mostrando que embora a maioria das pessoas esteja familiarizada com ML em aplicativos voltados para o consumidor, a maioria dos casos de uso de ML são para empresas. No capítulo, é apresentado quando as soluções de ML seriam apropriadas.
Embora o aprendizado de máquina (ML) seja eficaz na resolução de muitos problemas, não é capaz de solucionar todas as questões e certamente não é adequado para todos os cenários. No entanto, em situações em que o ML não consegue oferecer uma solução completa, ele ainda pode desempenhar um papel importante como parte da solução.

As diferenças entre ML na pesquisa e ML em Produção também foram destacados no presente capítulo. As diferenças incluem o envolvimento das partes interessadas, prioridade, as propriedades dos dados utilizados, a gravidade das questões de justiça e os requisitos de interpretabilidade. Também foi discutido como os sistemas de ML diferem dos sistemas de software tradicionais.

O capítulo ainda fala que os sistemas de ML são complexos e consistem em muitos componentes diferentes, focar apenas na parte dos algoritmos de ML está longe de ser suficiente. Conhecer outros aspectos do sistema, como a pilha de dados, implantação, monitoramento, manutenção, infraestrutura, além de outros, também é importante no aprendizado de máquina.

Capítulo II:

Este capítulo deu uma introdução ao design de sistemas de ML e fez as considerações que precisamos levar em conta ao projetar um sistema de ML. Todo projeto deve começar explicando por que esse projeto precisa acontecer, e os projetos de ML não são uma exceção. A maioria das empresas não se importam com as métricas de ML, a menos que possam mover as métricas de negócios. Mas, se um sistema de ML for construído para uma empresa, ele deve ser motivado pelos objetivos de negócio, que precisam ser traduzidos em objetivos de ML para orientar o desenvolvimento de modelos de ML.

Antes de construir um sistema de ML, é necessário entender os requisitos que o sistema precisa atender para ser considerado um bom sistema. Os requisitos exatos variam de cada caso de uso, dando relevância aos requisitos gerais: confiabilidade, escalabilidade, capacidade de manutenção e adaptabilidade. Construir um sistema de ML não é uma tarefa única, mas um processo iterativo. Um processo iterativo para desenvolver um sistema de ML que atenda aqueles requisitos gerais.

Capítulo III:

A escolha do formato apropriado para armazenar nossos dados é essencial para facilitar sua utilização futura. Assim, o capítulo 3 aborda diversas formas de dados, destacando as vantagens e desvantagens dos formatos de linha principal em comparação com os formatos de coluna principal, além de discutir as nuances entre formatos de texto e binários. Ele também considera três modelos de dados fundamentais: relacional, de documento e de gráfico.

O modelo relacional é o mais reconhecido devido à sua popularidade, mas é graças ao SQL, que todos os três modelos são amplamente empregados atualmente, sendo cada um adequado para um conjunto específico de tarefas.

Ao compararmos o modelo relacional com o modelo documental, é comum considerar o primeiro como estruturado e o segundo como não estruturado. Porém, a distinção entre dados estruturados e não estruturados é flexível, sendo a questão central sobre quem deve assumir a responsabilidade pela estrutura dos dados. Nos dados estruturados, cabe ao código responsável pela escrita definir a estrutura, enquanto nos dados não estruturados é o código responsável pela leitura que assume essa responsabilidade.

O capítulo 3 continua com mecanismos de armazenamento e processamento de dados. Explorando bancos de dados otimizados para dois tipos diferentes de processamento de dados: transacional e analítico. Analisando os mecanismos de armazenamento e processamento de dados conjuntamente, pois tradicionalmente o armazenamento está intimamente ligado ao processamento: os bancos de dados transacionais são utilizados para processamento transacional, enquanto os bancos de dados analíticos são empregados para processamento analítico.

Contudo, ao longo dos últimos anos, diversos fornecedores têm se empenhado em separar o armazenamento do processamento. Atualmente, existem bancos de dados transacionais capazes de lidar com consultas analíticas e bancos de dados analíticos aptos a processar consultas transacionais.

Quando é discutido sobre formatos de dados, modelos de dados, mecanismos de armazenamento e processamento, é pressuposto que os dados estejam inseridos em um processo. Entretanto, ao operar em um ambiente de produção, é provável que você esteja lidando com diversos processos e, consequentemente, pode ser necessário transferir dados entre eles.

Neste capítulo, foram abordados três métodos de transferência de dados. O método mais direto é a transferência através de bancos de dados. Já o método mais comum de transferência de dados entre processos é realizado por meio de serviços. Nessa abordagem, um processo é exposto como um serviço para que outros processos possam enviar solicitações de dados. Esse método está intimamente ligado às arquiteturas de microsserviços, onde cada componente de um aplicativo é configurado como um serviço independente. Um método de transferência de dados que ganhou popularidade na última década é a transferência em tempo real, utilizando tecnologias como Apache Kafka e RabbitMQ. Esse método se situa entre a transferência por bancos de dados e por serviços, permitindo a transferência assíncrona de dados com baixa latência.

Devido às características distintas dos dados em transportes em tempo real em comparação com os dados em bancos de dados, são necessárias técnicas de processamento diferentes, conforme abordado na seção "Processamento em lote versus processamento em fluxo". Os dados armazenados em bancos de dados geralmente são processados em lotes, resultando em recursos estáticos, enquanto os dados em transportes em tempo real são comumente processados usando técnicas de computação em fluxo, gerando recursos dinâmicos. Alguns argumentam que o processamento em lote é uma forma especializada de processamento em fluxo, e que os mecanismos de computação em fluxo podem ser empregados para unificar ambos os fluxos de processamento.