**Preparação de dados para IA**

Introdução

Você sabia que 80% do tempo de um cientista de dados é gasto na preparação de dados, restando apenas 20% para análise e modelagem? Essa estatística ressalta a importância crucial da preparação de dados em inteligência artificial (IA). Como base de qualquer modelo de aprendizado de máquina (ML) bem-sucedido, a preparação de dados transforma dados brutos em um formato estruturado, garantindo a qualidade e o desempenho dos sistemas de IA.

No mundo atual, impulsionado por dados, empresas e organizações dependem fortemente da IA ​​para tomar decisões informadas, aprimorar a experiência do cliente e impulsionar a inovação. Seja para aprimorar a precisão preditiva na área da saúde, otimizar a logística da cadeia de suprimentos ou personalizar estratégias de marketing, dominar a preparação de dados é essencial. Este treinamento o equipará com as habilidades e técnicas fundamentais necessárias para preparar dados de forma eficaz, tornando-o um recurso inestimável em qualquer equipe de ciência de dados.

Ao longo deste treinamento, você explorará o fluxo de trabalho abrangente da preparação de dados, desde a coleta e o armazenamento de dados até o pré-processamento, a rotulagem e a engenharia de recursos. Ao compreender e aplicar essas etapas, você ganhará confiança para lidar com diversos tipos e formatos de dados, garantindo que seus modelos de ML sejam construídos sobre uma base sólida.

Ao final deste treinamento, você terá uma sólida compreensão do processo de preparação de dados, o que lhe permitirá contribuir significativamente para o desenvolvimento de modelos de IA de alta qualidade. Você também estará preparado para enfrentar desafios do mundo real, tomando decisões baseadas em dados que podem levar a melhores resultados de negócios e maior satisfação do cliente. O treinamento o equipará com as habilidades essenciais para sua jornada como cientista de dados proficiente, pronto para causar um impacto significativo na área de IA.

Para aproveitar ao máximo este treinamento, sugerimos que você leia o texto a seguir, visualize os gráficos que o acompanham e assista aos vídeos do treinamento. Em seguida, serão feitas perguntas de revisão para que você possa se autoavaliar na preparação de dados para IA antes de passar para a próxima seção.

**Visão geral do fluxo de trabalho de processamento de dados**

Neste tópico, você revisará o fluxo de trabalho de preparação de dados e sua importância. Um modelo de ML aprende padrões a partir de dados históricos para fazer previsões ou decisões informadas sobre dados novos e desconhecidos. A eficácia do modelo de ML depende da disponibilidade de dados confiáveis; portanto, um processo eficaz de preparação de dados é essencial para gerar previsões confiáveis ​​e precisas. A preparação de dados consiste em várias etapas, como coleta, pré-processamento, armazenamento e recuperação de dados. Dados imprecisos, incompletos, inconsistentes ou tendenciosos levarão a previsões defeituosas do modelo, resultando potencialmente em conclusões incorretas, decisões de negócios ruins e impactos negativos na experiência do cliente. Técnicas aprimoradas de preparação de dados, com foco na rotulagem, ampliação e gerenciamento eficientes de dados, mantendo uma arquitetura de modelo fixa, melhoraram significativamente os resultados do modelo.

A preparação de dados transforma dados brutos em um conjunto de dados mais estruturado e utilizável, que revela padrões de dados incorporados. Essas técnicas removem o ruído irrelevante e identificam padrões e insights significativos para abordar o seu problema.

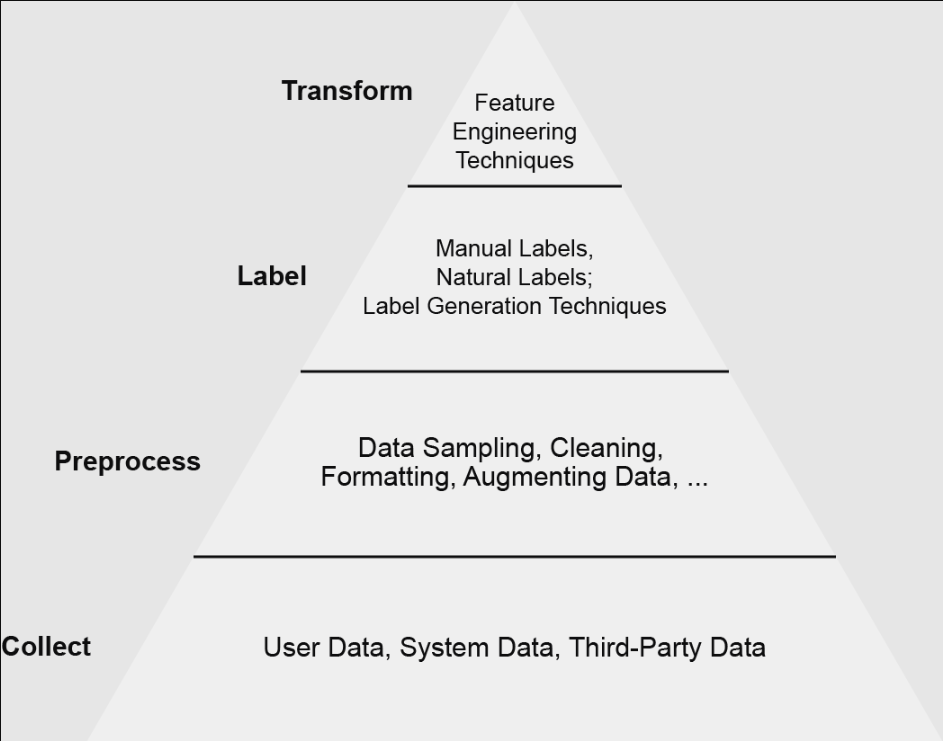
A equipe de anotação de dados deve garantir que conjuntos de dados de alta qualidade, cuidadosamente limpos, rotulados e selecionados estejam disponíveis para que os cientistas de ML treinem seus modelos. A aquisição e o pré-processamento de dados são processos iterativos e contínuos. Automatizá-los o máximo possível é uma boa prática. Também é importante incentivar a colaboração entre engenheiros de dados, cientistas de dados e especialistas de domínio para entender melhor as necessidades e os desafios dos dados.

Embora cada fluxo de trabalho de preparação de dados de IA seja único, as etapas gerais podem ser resumidas da seguinte forma:

1. Coletar e armazenar dados
2. Pré-processar dados
3. Rotule os dados
4. Transformar os dados (engenharia de recursos)

**Observação**

*O processo de transformação de dados, também conhecido como engenharia de recursos, envolve a seleção, a transformação e a criação de variáveis ​​(recursos) a partir de dados brutos para permitir o aprendizado eficiente do modelo. Recursos de ML são propriedades ou atributos mensuráveis ​​dos objetos observados, representados como amostras de dados de entrada. Os recursos podem ser representados como um número ou categoria.*



A complexidade do modelo de IA depende do tipo de problema que ele aborda. Estruturas de programação definem as bibliotecas e técnicas disponíveis para cientistas de dados, mas não afetam a precisão do modelo. Os recursos e capacidades da infraestrutura (em termos de recursos como unidades de processamento gráfico (GPUs), CPUs, armazenamento e rede) influenciam a velocidade do treinamento do modelo e a tomada de decisões, mas não a precisão.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**Fases do fluxo de trabalho de processamento de dados**

Existem diversas técnicas comuns usadas em diferentes fases do fluxo de trabalho de preparação de dados. Essas fases incluem a coleta de dados brutos, o pré-processamento, a rotulagem e a transformação dos dados para torná-los utilizáveis ​​no treinamento dos seus modelos.

**Coletar e armazenar dados**

Você pode adquirir conjuntos de dados externos, como estatísticas de comportamento de usuários online, ou até mesmo gerá-los com técnicas de IA. Dados gerados por IA fornecem vastos conjuntos de dados de forma econômica, mas carecem de realismo e podem resultar em previsões imprecisas.

Três tipos de dados são usados ​​para treinar modelos de IA:

* Dados estruturados baseados em texto (armazenados em bancos de dados relacionais e data warehouses)
* Dados semiestruturados, como texto ou documentos (armazenados não apenas em bancos de dados de Linguagem de Consulta Estruturada [NoSQL] e data lakes)
* Texto não estruturado, imagens, gravações de áudio e vídeo (armazenados em data lakes)

Os dados vêm em vários formatos, que podem ser classificados da seguinte forma:

* **Baseado em texto (legível por humanos)** : esses dados são ideais para configurações, pequenos conjuntos de dados e troca de dados entre sistemas onde a legibilidade humana é importante.
* **Binário (legível por máquina)** : Esses dados são adequados para conjuntos de dados maiores, estruturas de dados complexas e cenários onde eficiência e desempenho são críticos. São frequentemente usados ​​em bancos de dados, computação científica e aplicações de processamento de dados de alto desempenho.
* **Observação**
* O tipo e formato de dados ideais dependem dos objetivos que você deseja atingir e da estrutura dos seus dados de entrada.

| Formatar | Binário ou texto | Formatar | Uso |
| --- | --- | --- | --- |
| Notação de Objeto JavaScript (JSON) | Texto | Semiestruturado | Arquivos de configuração |
| XML | Texto | Semiestruturado | Serviços web, armazenamento de documentos e troca |
| Valores separados por vírgula (CSV) | Texto | Estruturado | Ferramentas e bibliotecas de processamento de dados |
| Parquet | Binário | Orientado a colunas | Apache Spark, Amazon Redshift |
| Avro | Binário | Orientado para linha | Hadoop |
| Protobuf | Binário | Mecanismo extensível para serializar dados estruturados. | TensorFlow |
| Salmoura | Binário | Formato de serialização binária específico do Python para serializar e desserializar objetos Python. Útil para salvar modelos de ML, resultados intermediários e estruturas de dados personalizadas. | Scikit-Learn |
| Troca de Rede Neural Aberta (ONNX) | Binário | ONNX é um formato de código aberto para representar modelos de ML. O ONNX foi projetado para facilitar a transferência de modelos entre diferentes estruturas. | PyTorch, TensorFlow |

**Observação**

*A serialização de dados é o processo de conversão de estruturas ou objetos de dados em um formato que possa ser facilmente armazenado, transmitido e reconstruído posteriormente. A serialização é essencial para a troca de dados entre vários sistemas, especialmente quando são escritos em diferentes linguagens de programação ou executados em diferentes plataformas.*

A estrutura de dados pode ser orientada por linha ou coluna. O formato orientado por linha é otimizado para acessar a amostra de dados específica (com todos os seus recursos), enquanto o formato orientado por coluna é otimizado para acessar um único recurso (em todas as amostras).

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Os dados podem ser gerados pelo usuário (como texto e imagens) ou gerados pelo sistema (como logs de infraestrutura e aplicativos). Existem dois tipos distintos de dados:

* **Dados históricos:** Esses dados são armazenados em sistemas de armazenamento, como bancos de dados e data warehouses. Geralmente, são processados ​​periodicamente em tarefas em lote usando ferramentas como MapReduce ou Spark. O processamento em lote é usado para calcular recursos estáticos, que raramente mudam.
* **Dados em streaming:** sistemas de transporte em tempo real, como Apache Kafka ou RabbitMQ, processam esse tipo de dados. O processamento em fluxo é usado para calcular recursos dinâmicos, que mudam com mais frequência.
* **Observação**
* É útil registrar a origem de cada amostra de dados e seus rótulos, o que é conhecido como linhagem de dados. A linhagem de dados ajuda a identificar possíveis vieses nos seus dados e a solucionar problemas nos seus modelos.

**Pré-processamento de dados**

O pré-processamento de dados inclui amostragem, limpeza, formatação e transformação de dados. O pré-processamento é realizado para tornar os dados utilizáveis ​​para a etapa de aprendizado (treinamento do modelo). O pré-processamento normalmente envolve algumas das seguintes transformações de dados:

* **Amostragem de dados** : Amostragem é o processo de seleção de dados do conjunto de dados que será usado para treinamento. Normalmente, são criadas diversas divisões de dados para treinamento, validação e teste. Existem duas famílias de amostragem: amostragem não probabilística (a seleção não se baseia em nenhum critério probabilístico) e amostragem aleatória.
* **Limpeza de dados** : a limpeza envolve a exclusão de dados não utilizados ou inválidos.
* **Transformando o formato** : Este processo altera a estrutura de dados para atender às necessidades do seu aplicativo.
* **Tratamento de dados ausentes** : campos vazios geralmente são substituídos por um valor médio, interpolados ou excluídos.
* **Lidando com valores discrepantes** : pontos de dados raros que se desviam significativamente do conjunto de dados devem ser identificados e excluídos.
* **Tratamento de erros de dados** : Dados inválidos devem ser transformados ou excluídos.
* **Lidando com dados desbalanceados** : Este processo lida com o desequilíbrio de classes usando alguma forma de técnica de reamostragem ou dados gerados sinteticamente. Se houver um desequilíbrio de dados, você terá significativamente mais pontos de dados para uma classe específica, o que leva a modelos tendenciosos. Como alternativa, você pode tornar seu algoritmo mais robusto a dados desbalanceados.
* **Aumento de dados** : se você não tiver dados suficientes, poderá usar algumas técnicas de aumento de dados (perturbação, transformação de dados com preservação de rótulos) em dados existentes para aumentar a quantidade de dados de treinamento.

Uma imagem contendo Diagrama

O conteúdo gerado por IA pode estar incorreto.

**Rotulagem de dados**

A maioria das técnicas de IA/ML utilizadas na prática pertence a uma classe de técnicas de aprendizado supervisionado. O aprendizado supervisionado depende de rótulos de dados, que precisam aprender a reconhecer padrões corretamente nos dados. A rotulagem de dados atribui tags ou anotações significativas aos dados brutos. Os rótulos normalmente definem a saída alvo que o modelo deve prever. Os rótulos fornecem contexto aos dados para auxiliar no processo de aprendizado do modelo de ML. Uma equipe de anotadores dedicados geralmente rotula esses dados manualmente.

Algumas tarefas podem evitar a rotulagem manual usando rótulos naturais já presentes nos dados. Rótulos naturais ocorrem em tarefas em que o aplicativo avalia automaticamente as previsões do modelo. Exemplos são um sistema de recomendação (o sistema sabe se você aceitou a recomendação) ou o Google Maps (o modelo estima o tempo de chegada para uma determinada rota no Google Maps e, quando você chega, ele mede a precisão da previsão).

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

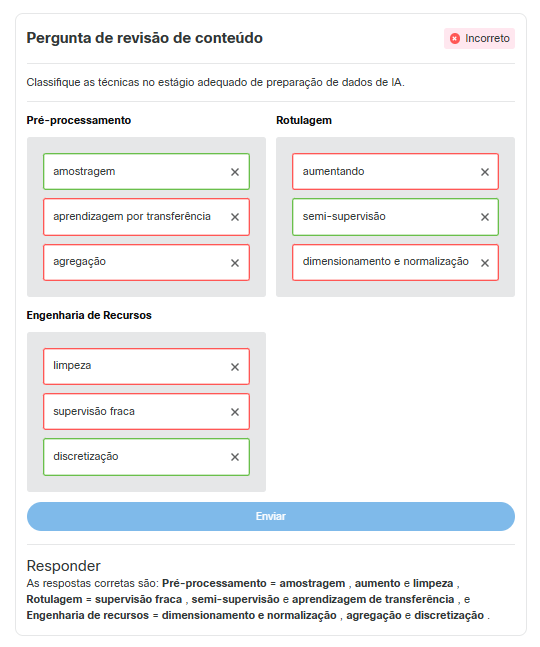
Como não é fácil obter rótulos de alta qualidade, existem várias técnicas para gerar ou aprimorar conjuntos existentes de rótulos para melhorar a qualidade do modelo:

* **Supervisão fraca:** Esta técnica usa regras simples e eficientes para gerar rótulos.
* **Semi-supervisão:** Esta técnica usa suposições estruturais para gerar rótulos.
* **Aprendizagem de transferência:** o modelo já está treinado para outras tarefas e reutilizado como ponto de partida.
* **Aprendizado ativo:** essa técnica aumenta a eficiência do aprendizado ao escolher os rótulos de amostra de dados que são úteis para seu modelo.

**Engenharia de Recursos (Transformação de Dados)**

A engenharia de recursos requer expertise na área específica para extrair recursos relevantes de dados brutos. Cientistas de dados transformam dados brutos em recursos específicos necessários para realizar tarefas específicas. Os recursos de dados extraídos alimentam o processo de treinamento do modelo e permitem que você tenha modelos precisos. A engenharia de recursos envolve as seguintes tarefas:

* **Dimensionamento** : esta tarefa normaliza pontos de dados para intervalos semelhantes.
* **Discretização** : Esta tarefa transforma um valor contínuo em um valor discreto.
* **Codificação** : esta tarefa transforma dados contínuos em categorias.
* **Decomposição** : esta tarefa divide dados complexos em componentes mais simples e gerenciáveis.
* **Agregação** : esta tarefa combina vários recursos de dados em um.
* **Redução de dimensionalidade** : esta tarefa transforma um conjunto de dados usando menos recursos (dimensões) sem reduzir o valor dos dados.



**Resumo**

Neste treinamento, você aprendeu sobre as etapas e técnicas essenciais envolvidas na preparação de dados para aplicações de IA. A eficácia de um modelo de ML depende fortemente da qualidade dos dados com os quais ele é treinado, o que torna a preparação de dados um processo crítico. O fluxo de trabalho inclui coleta e armazenamento de dados, pré-processamento, rotulagem e engenharia de atributos. Cada etapa envolve tarefas específicas, como amostragem, limpeza, transformação e ampliação de dados para garantir que sejam adequados para o treinamento do modelo. A rotulagem de dados, frequentemente feita manualmente, fornece o contexto necessário para o aprendizado supervisionado. A engenharia de atributos transforma dados brutos em atributos significativos que aumentam a precisão do modelo. Entender esses processos e seus desafios é crucial para o desenvolvimento de modelos de IA confiáveis ​​e precisos.

Agora que você concluiu o treinamento, reflita sobre as seguintes perguntas:

* Coleta e armazenamento de dados
  1. Quais tipos de dados (estruturados, semiestruturados, não estruturados) são mais prevalentes no seu ambiente de trabalho e como eles são armazenados atualmente?
  2. Como você pode melhorar o processo de coleta de dados para garantir dados de maior qualidade e mais relevantes para seus modelos de IA?
* Pré-processamento de dados
  1. Quais problemas comuns de qualidade de dados (por exemplo, valores ausentes, valores discrepantes) você encontra e quais estratégias pode implementar para lidar com eles de forma eficaz?
  2. Como você pode automatizar partes do fluxo de trabalho de pré-processamento de dados para economizar tempo e reduzir erros?
* Engenharia de recursos
  1. Quais recursos específicos de domínio você pode desenvolver a partir de seus dados brutos para melhorar o desempenho de seus modelos de IA?
  2. Como você pode usar técnicas como dimensionamento, discretização e redução de dimensionalidade para aumentar a eficiência do seu processo de treinamento de modelo?