Introdução à Machine Learning

Clube do Código - Análise Macro

Vítor Wilher

Cientista de Dados

Plano de Voo

Introdução

Nessa primeira edição dos Exercícios de Machine Learning do Clube do Código, vamos fazer uma introdução à área, de modo a apresentá-la a nossos membros e alunos. Esperamos que essa introdução os ajude a compreender os conceitos fundamentais que definem e diferenciam as abordagens mais utilizadas de machine learning.[1]

[1] Essa introdução é baseada principalmente em Lantz [2013].

Introdução

Você aprenderá nessa introdução:

- As origens e aplicações práticas de machine learning;
- Como transformar dados e conhecimento em ação;
- Como adequar algoritmos de machine learning a dados reais.

Origens

Diversos aspectos da nossa vida são registrados. Governos, administradores e indivíduos estão todo o tempo registrando e reportando informação. Esse dilúvio de dados tem levado muitos a constatar que vivemos uma era de **Big Data**, mas isso pode ser um termo impróprio. Isto porque, temos estado desde sempre cercados de um amontado de dados. O que diferencia a era atual é que temos um vasto conjunto de *dados registrados*. Essa riqueza de informações tem o potencial de gerar ação, dada uma sistematização que transforme conjuntos de informação aparentemente confusos que façam sentido.

Origens

O campo de estudo interessado no desenvolvimento de algoritmos que transformam dados em ações inteligentes é conhecido como **machine learning**. Esse campo se originou em um ambiente onde dados disponíveis, métodos estatísticos e poder computacional rápido e simultaneamente se desenvolveram. O crescimento dos dados gera uma necessidade de poder computacional, o que por sua vez transborda para o desenvolvimento de métodos estatísticos para analisar grandes conjuntos de dados. Isso criou um ciclo de progresso, permitindo que conjuntos ainda maiores e mais interessantes de dados pudessem ser coletados.

Aplicações

Machine learning é mais bem sucedido quando ela aumenta ao invés de substituir um conhecimento especializado de um assunto específico. Ela funciona com médicos em busca da cura do câncer, programadores comprometidos em construir casas e automovéis inteligentes ou ajudando cientistas sociais a construir conhecimento sobre como as sociedades funcionam. Algumas outras aplicações podem ser listadas abaixo:

Aplicações

- Identificação de mensagens indesejáveis;
- Segmentação do comportamento de consumidores para propaganda direcionada;
- Previsão do tempo e de mudanças climáticas de longo-termo;
- Redução de fraudes em transações de cartão de crédito;
- Estimativas atuariais de danos financeiros associados a tempestades e desastres naturais;
- Previsão de eleições;
- Desenvolvimento de algoritmos para drones e carros sem motoristas;
- Otimização do consumo de energia em casas e escritórios;
- Projeção de áreas onde é mais provável ocorrer crimes;
- Descobrimento de sequências genéticas associadas a doenças.

Como as máquinas aprendem?

Uma definição formal de machine learning foi proposta por Tom M. Mitchell: `Maquinas aprendem sempre que são capazes de utilizar suas experiências de modo que essa performance melhora uma experiência similar no futuro`. Embora essa definição seja intuitiva, ela ignora por completo o processo exato sobre como essa experiência é transformada em ação futura.

Enquanto o cérebro humano é naturalmente capaz de aprender desde o nascimento, as condições necessárias para que computadores aprendam precisam ser especificadas. Por essa razão, embora não seja estritamente necessário entender as bases teóricas do processo de aprendizado, essa fundação ajuda a entender, distinguir e implementar algoritmos de machine learning.

Como as máquinas aprendem?

Em termos gerais, o processo de aprendizado pode ser dividido em quatro componentes:

- **Armazenamento de dados**: utiliza observação, memória e recordações para prover uma base factual para aumentar o raciocínio;
- **Abstração**: envolve a tradução de dados armazenados em representações e conceitos mais amplos;
- **Generalização**: usa dados abstraídos para criar conhecimento e inferências que direcionam ações em novos contextos;
- **Avaliação**: provê um mecanismo de feedback para medir a utilidade do conhecimento aprendido e informar possíveis melhorias.

Machine Learning na prática

Até aqui, temos focado como machine learning trabalha em teoria. De modo a aplicar o processo de aprendizado a tarefas do mundo real, nós podemos utilizar um processo de cinco etapas:

- **Coleta de dados**: envolve reunir o material de aprendizado que o algoritmo irá usar para gerar conhecimento tangível;
- **Exploração de dados e preparação**: A qualidade de qualquer projeção de machine learning está de longe baseada na qualidade dos dados imputados. Por isso, é importante aprender mais sobre os dados e seus nuances durante o processo de *exploração*. Um trabalho adicional é necessário para preparar os dados para o processo de aprendizagem. Isso envolve corrigir ou limpar dados desestruturados, eliminando dados desnecessários e guardando os dados conforme as expectativas de aprendizado;

Machine Learning na prática

- **Treinamento do modelo**: Desde que os dados estão preparados para a análise, você está pronto para ter alguma dimensão sobre o que pode ser aprendido a partir dos dados. A tarefa específica de machine learning irá informar o algoritmo apropriado e este irá representar os dados na forma de um modelo;
- **Avaliação do modelo**: Dado que cada modelo de machine learning resulta em uma solução viesada para o problema de aprendizado, é importante avaliar quão bem o algoritmo aprende a partir dessa experiência. A depender do tipo de modelo usado, você pode ser capaz de avaliar a acurácia do modelo usando um conjunto de dados de teste (*dataset test*) ou criar medidas de performance específicas para uma dada aplicação;

Machine Learning na prática

- **Aperfeiçoamento do modelo**: Se uma melhor performance for necessário, pode ser importante utilizar estratégias mais avançadas de modo a aumentar a acurácia do modelo. Em alguns casos, pode ser necessário mudar o tipo de modelo, adicionar outras variáveis ou mesmo refazer o trabalho de preparação dos dados.

Tipos de Algoritmos

Algoritmos de machine learning são divididos em categorias de acordo com os seus propósitos. Entender as categorias dos algoritmos de aprendizado é um primeiro passo essencial na direção de usar os dados para direcionar uma ação desejada.

Um **modelo de previsão** é utilizado para tarefas que como o nome diz exigem a previsão de um valor utilizando outros valores do conjunto de dados. O algoritmo de aprendizado busca descobrir e modelar a relação entre o *objetivo*, a variável a ser prevista, e outras variáveis. Dado que em modelos de previsão está muito claro sobre o que e como eles precisam aprender, o processo de treinar um modelo de previsão é conhecido como **aprendizado supervisionado**. Dado um conjunto de dados, um algoritmo de aprendizado supervisionado busca otimizar um função - o modelo - de modo a encontrar uma combinação de valores que resultam no *output* esperado.

Tipos de Algoritmos

A tarefa de machine learning supervisionada frequentemente utilizada para prever a qual categoria pertence um exemplo é conhecida como **classificação**. Potenciais usos:

- Um e-mail é um spam;
- Uma pessoa tem câncer;
- Um time de futebol irá perder ou ganhar;
- Um tomador não irá pagar um empréstimo.

Tipos de Algoritmos

Em algoritmos de classificação, o objetivo a ser previsto é um aspecto categórico conhecido como **classe** e é dividida em categorias chamadas de **níveis**. Algoritmos supervisionados podem ser utilizados também para **previsões numéricas**.

Um **modelo descritivo**, por outro lado, é utilizado para tarefas que se beneficiam dos insights gerados pela sumarização dos dados em um novo e interessante modo. Como oposição aos modelos preditivos que prevem um objetivo de interesse, em um modelo descritivo, uma variável não é mais importante do que outra. Dado que não um objetivo implícito a ser aprendido, o processo de treinamento de um modelo descritivo é conhecimento como **aprendizado não supervisionado**.

Tipos de Algoritmos

Uma tarefa de modelagem descritiva conhecida como **detecção de padrões** é utilizada, por exemplo, para identificar associações úteis nos dados. Já a tarefa de dividir um conjunto de dados em grupos homogêneos é chamada de **clustering**.

Por fim, uma classe de algoritmos de machine learning conhecida como **meta-aprendizagem** não está associada a uma tarefa de aprendizado específica, mas por outro lado está focada em como aprender mais efetivamente.

Dados de entrada e algoritmos

Abaixo, listamos os tipos gerais de algoritmos de machine learning de acordo com as tarefas de aprendizado. Primeiro os algoritmos de aprendizado supervisionado com os nomes em inglês:

Nearest Neighbor	Classificação
Naive Bayes	Classificação
Decision Trees	Classificação
Classification Rule Learners	Classificação
Linear Regression	Previsão numérica
Regression Trees	Previsão numérica
Model Trees	Previsão numérica
Neural Networks	Uso dual
Support Vector Machine	Uso dual

Dados de entrada e algoritmos

Abaixo, algoritmos de aprendizado não supervisionado.

Association Rules	Detecção de padrões
k-means clustering	Clustering

Dados de entrada e algoritmos

Por fim, os algoritmos de meta-aprendizagem.

Bagging	Uso dual
Boosting	Uso dual
Random Forests	Uso dual

Conclusão

Machine learning se origina da intersecção entre estatística, bases de dados e ciência da computação. É uma poderosa ferramenta, capaz de encontrar insights interessantes em grandes conjuntos de dados.

Conceitualmente, o aprendizado envolve a abstração dos dados em uma representação estruturada e a generalização dessa estrutura em uma ação em que sua utilidade pode ser avaliada. Em termos práticos, utiliza-se dados contendo exemplos e amostras de onde pode ser aprendido algo útil. Sumarizamos esses dados na forma de um modelo, que pode ser utilizado para previsão ou propósitos descritivos. Esses propósitos podem ser agrupar em tarefas, incluindo classificação, previsão numérica, detecção de padrões e *clustering*. Algoritmos de machine learning são escolhidos de acordo com os dados de entrada e a tarefa de aprendizagem.

Conheça o Clube do Código

Essa apresentação faz parte dos *Exercícios de Machine Learning* do Clube do Código.

Conheça o Clube em **Clube do Código**.

Brett Lantz. *Machine Learning with R*. Packt Publishing, 2013.