

Aprendizagem de máquina básico: uma abordagem teórico-prática com Python



Giseldo Neo (versao alpha 0.1)

Sumário

1	Introdução	5
1.1	Inteligência artificial	5
1.2	Aprendizado de Máquina	6
1.2.1	Classificação	8
1.2.2	Exemplo de AM	9
2	Estatística Básica	11
2.1	Tipo de variável	12
2.1.1	Variável quantitativo	12
2.1.2	Dado qualitativo	13

Capítulo 1

Introdução

1.1 Inteligência artificial

O termo “inteligência” tem várias definições que dependem do contexto. Isso pode trazer certa confusão no entendimento e delimitação do tema. Menos abrangente, porém mais confuso ainda, é o termo “inteligência artificial”. Portanto, dado as diversas definições de inteligência artificial (IA), ou *artificial intelligence* em inglês) vamos delimitar um pouco o significado destas palavras.

Nós humanos somos da espécie Homo-Sapiens. Espero que o leitor ainda o seja, pois esse texto pode estar sendo processado para treinar o mais novo modelo de IA como a do filme “2001 uma odisseia no espaço”, clássico de Kubric, ou como a do mais recente filme “Ela”, com o ator Joaquim Phenix, espero que com sorte, por um garoto interessado em aprender.

Homo-Sapiens vem do latim e significa homem sábio [4]. A importância da sapiência (que é um sinônimo de inteligência) é tamanha que define a nossa própria espécie. Porém, dentro do nosso contexto consideramos que um gato e um cachorro são seres dotados de muita inteligência, uma abelha, então nem se fala, praticamente uma cientista. Portanto, seremos mais contidos e reservados quanto ao termo inteligência.

No entanto, várias questões relacionadas a inteligência também guiam inúmeras pesquisas científicas, por exemplo: como funciona nossa inteligência? Nossa percepção do ambiente é próxima da realidade objetiva? Ainda não é dessa inteligência que estamos falando. Essas perguntas estão mais próxima da neurociência e da filosofia.

Além disso, “inteligência” e “artificial” são palavras que têm significado implícito para pessoas que não são da área de computação, naturalmente surge o desejo de médicos, advogados, engenheiros (só para citar alguns) de verificar como a “inteligência artificial” pode ser inserida na sua rotina diária. Por exemplo, o meu dentista já quis saber como a IA iria afetar seus procedimentos odontológicos. Porém, ninguém nunca me perguntou em como a “Transformada de Fourier” poderia melhorar o seu dia-a-

dia, mesmo sabendo que ela já é utilizada em vários domínios do conhecimento e com entusiasmo [5].

A nossa “inteligência artificial” está mais relacionada com a capacidade de realizar coisas que seres inteligentes (um gato, um bebê, ou um cientista) realizam, como por exemplo puxar a mão (ou pata) instantaneamente ao tocar em uma superfície quente (inteligência reativa), ou realizar uma prova de anatomia (inteligência cognitiva). Se conseguimos que programas realizem ações realizadas por entidades dotadas de inteligencia, e realizamos isso de forma computacional, estamos próximos da nossa definição desejada de “inteligência artificial”.

Russel e Norvig (2020) em um dos livros mais lidos em todas as universidades do mundo tem uma boa definição sobre esse tema: “O campo da inteligência artificial [...] tenta não apenas compreender, mas também construir entidades inteligentes” (tradução nossa) [2]. Logo, dado o ambicioso desejo de compreender a inteligência, temos o também audacioso objetivo de construir agentes inteligentes dotados dessa inteligência.

A origem do termo “inteligência artificial”, nesse contexto, é atribuída a John McCarthy, professor de Matemática da Universidade Dartmouth College[1], ele organizou uma conferência com duração de oito semanas com mais alguns colegas em 1956, alguns anos após a segunda guerra, e desde então o termo vem sendo utilizado para designar parte de conteúdos estudados em ciência da computação. Porém, um pouco antes, o artigo seminal de Alan Turing já demonstrava um bom ensaio sobre as possibilidades de uma máquina inteligente [3].

Foi na década de 1970 que o uso da IA começou a ser mais difundido. Esses primeiros sistemas de IA foram chamados de Sistemas Especialistas (veja um exemplo na Figura 1.1)e dependiam muito dos Homo-Sapiens para transformar o conhecimento tácito (baseado em sua experiência) em explícito (formalizado, documentado), que era então codificado na forma de um software contendo regras em lógica formal. O processo de aquisição do conhecimento por esses especialistas acabou sendo um grande obstáculo na adoção em massa dessa abordagem.

Nestas últimas décadas houve um crescimento exponencial das tecnologias que estão ao redor da IA, tais como, o aumento capacidade de processamento e armazenamento dos computadores e a geração de grandes volumes de dados, além dos avanços científicos e tecnológicos em outras áreas, tais como chips supercondutores e eficiência energética.

1.2 Aprendizado de Máquina

O Aprendizado de Máquina (AM) é uma subárea da IA (veja Figura 1.2). que foi motivada pelo desenvolvimento de softwares mais independentes da intervenção

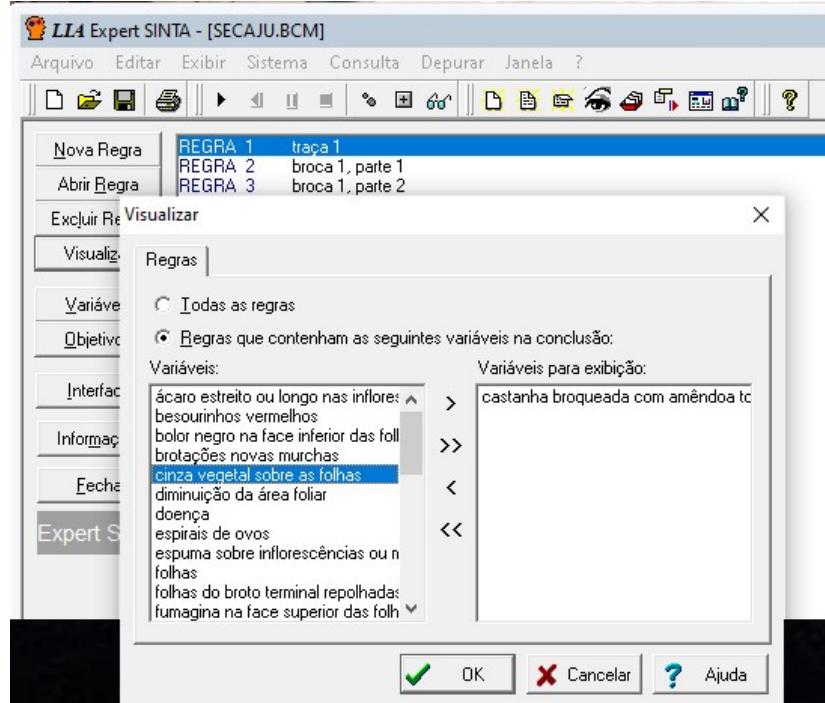


Figura 1.1: ExpertSinta. Uma interface de um Sistema Especialista

humana para extração do conhecimento, o que era uma dificuldade nos Sistemas Especialistas. Geralmente aplicações de AM utilizam **heurísticas** (regra do dedão) que buscam por modelos capazes de representar o conhecimento existente nos dados.

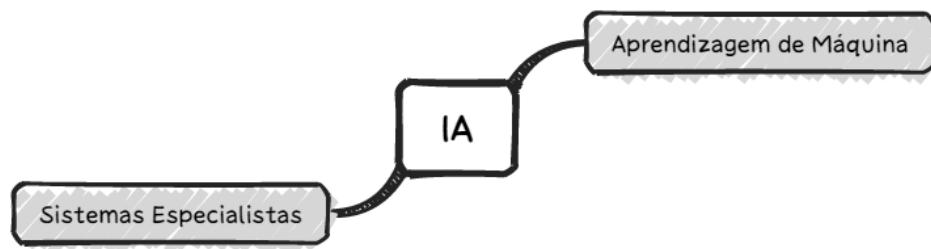


Figura 1.2: AM é uma parte da IA

Na Figura 1.3, é possível identificar alguns usos de AM integrado em diversas atividades cotidianas. São elas, (a) Um smartphone com um assistente de voz fornecendo atualizações meteorológicas. (b) Um sistema de casa inteligente ajustando o termostato com base nas preferências do usuário. (c) Um carro autônomo dirigindo em uma rua movimentada da cidade. (d) Uma plataforma de compras online recomendando produtos a um usuário com base em suas compras anteriores. Essa figura foi criada inclusive com inteligência artificial.



Figura 1.3: Exemplos AM

1.2.1 Classificação

As tarefas de aprendizado de máquina podem ser divididas entre tarefas preditivas, que visam inferir o atributo alvo de uma nova entrada a partir da exposição prévia aos dados rotulados durante o treinamento do modelo, e descritivas, que buscam extrair padrões dos atributos preditivos. Por conseguinte, uma vez que pertencem a este paradigma, as tarefas de aprendizado descritivas não possuem atributos alvo. Noutras palavras, tarefas preditivas analisarão os atributos preditivos, comparando-os com os atributos alvo (rótulos), ao passo que tarefas descritivas utilizaram os atributos preditivos entre si para buscar por padrões e correlações.

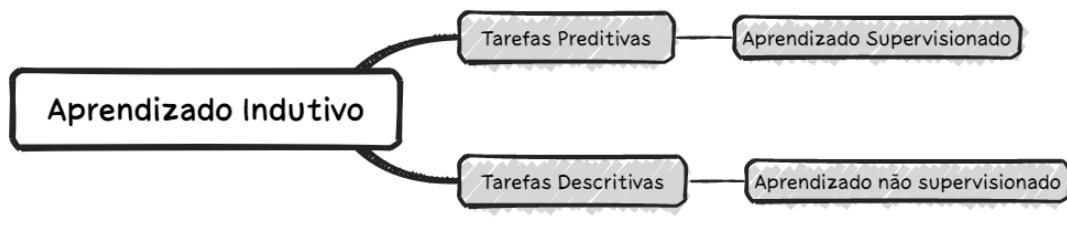


Figura 1.4: Classificação AM

Ambas as tarefas podem ser categorizadas sob o conceito de aprendizado indutivo, que é a capacidade de generalizar a partir de exemplos específicos, isto é, do conjunto de dados de treinamento. Em se tratando de tarefas preditivas, os algoritmos

poderão implementar tarefas de classificação, nas quais o atributo alvo (rótulo) é discreto (enumerável ou finito), ou de regressão, em que o atributo alvo (rótulo) é contínuo (não enumerável ou infinito). Já as descritivas distinguem-se entre agrupamento, que busca por similaridades, associação, que busca por padrões frequentes, e summarização, que resulta em um resumo do conjunto de dados.

1.2.2 Exemplo de AM

A seguir um exemplo de modelo preditivo em Python. O modelo utiliza o algoritmo SVM e o conjunto de dados iris, que é um conjunto de dados conhecido e bastante utilizado como em demonstrações em outros livros e sites.

```
1  from sklearn import svm
2  from sklearn.datasets import iris
3  iris = load_iris()
4  X = iris.data
5  y = iris.target
6  model = svm.SVC()
7  model.fit(X, y)
8  model.predict([[2., 2., 2., 2.]])
```

Listing 1.1: Exemplo de código que usa AM

Capítulo 2

Estatística Básica

Um conjunto de dados geralmente é uma estrutura tabular com linhas e colunas, o nome da coluna é o identificador do dado (também chamado de variável) disposto naquela coluna. Cada linha da coluna é chamada de observação (ou registro), e representa uma instância daquele elemento. Por exemplo, uma tabela com dados do cliente é apresentado na Tabela 2.1, a primeira linha (em negrito) é o nome da coluna, são elas: nome, endereço, telefone; cada linha abaixo do nome da coluna representa um cliente. Portanto a linha 1, teria o nome das 3 colunas; já a linha 2, teria um determinado cliente, e a linha 3, outro cliente.

Tabela 2.1: Tabela cliente.

nome	endereço	telefone
Giseldo Neo	Rua das alamedas, n 27, Corumbá MS	222 66666
Alex Neo	Avenida Fernandes 325, Macieó, AL	333 6589

Em resumo, a Tabela 2.1 apresenta 2 observações (ou registros) de clientes. A primeira coluna é uma descrição das informações que existirão naquela coluna, por exemplo “nome” significa que provavelmente todos os dados dessa coluna são referentes ao nome de determinado cliente, já cada linha abaixo da primeira linha são os dados de um cliente em específico.

Quando vamos realizar um resumo estatístico (tal como média, mediana ou moda), ou algum gráfico, ou alguma inferência estatística mesmo vamos executar nossos modelos preditivos, temos que conhecer o tipo teórico, daquele dado (ou variável, ou coluna) para podermos selecionar qual a técnica mais adequada que devemos aplicar. Por exemplo, em nenhuma das colunas da tabela cliente devemos calcular a média.

Utilizando uma classificação para esse tipo de variável, podemos definir qual é o tipo possível de operação. Portanto, vamos acrescentar mais uma informação (teórica) a coluna da tabela para que possamos claramente definir o que deve, ou pode ser feito com ela, chamamos essa informação de tipo de dado, ou tipo de variável, que é o termo mais usado.

2.1 Tipo de variável

Uma variável, em conceitos estatísticos, é uma característica do que foi observado naquele universo (amostra ou população). No nosso exemplo do cliente (Tabela 2.1) é o nome, endereço e telefone. Ela é registrada em uma tabela após um processo de medição ou contagem do elemento observado.

É útil definir de qual tipo é determinada variável, pois, existem técnicas adequadas a cada tipo para a análise estatística, para a elaboração de gráficos, para as técnicas de pré-processamento e para as técnicas de aprendizagem de máquina.

O tipo da variável, neste contexto, pode ser **quantitativo** ou **qualitativo** (veja a Figura 2.1).

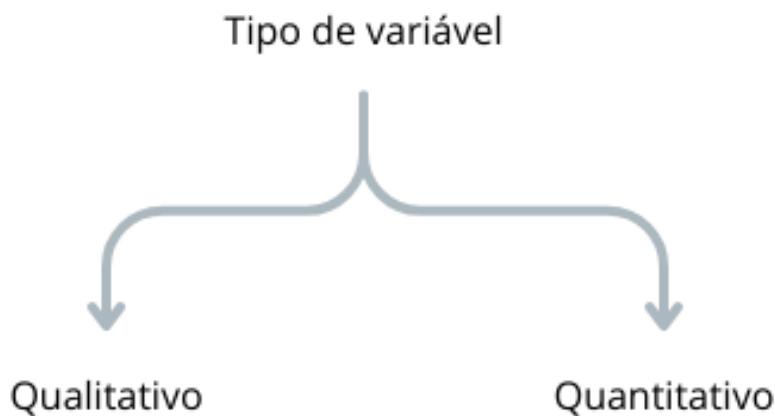


Figura 2.1: Tipo de dado

A variável do tipo **quantitativa** é expressa geralmente como um número inteiro ou real. Porém, existem casos em que números inteiros também expressam dados do tipo qualitativo, cada caso deve ser analisado. Já a variável do tipo **quantitativa** está relacionado a [...].

2.1.1 Variável quantitativo

Já sabemos que um dado do tipo **numérico** é expresso geralmente como um número inteiro ou real. Além disso, esse tipo de dado ainda pode ser subclassificado em **contínuo** ou **discreto** (veja na Figura 2.2).

Um dado **numérico contínuo** é quando o dado pode ser qualquer número em um intervalo de números reais - lembrando que o conjunto de números reais engloba os números inteiros. Geralmente é o resultado de uma medida, por exemplo, a altura dos estudantes é um dado do tipo numérico contínuo.

O dado numérico discreto geralmente é resultado de uma contagem - um número inteiro. Por exemplo, a idade é uma contagem de anos do estudante, logo é um dado do tipo **numérico discreto**.

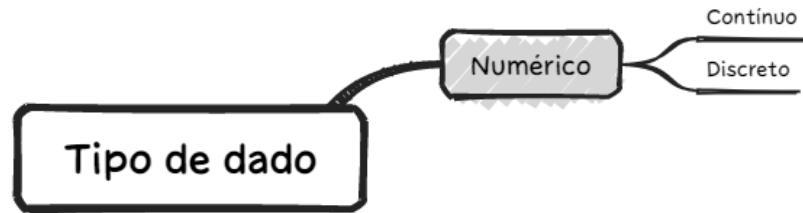


Figura 2.2: Tipo dado numérico

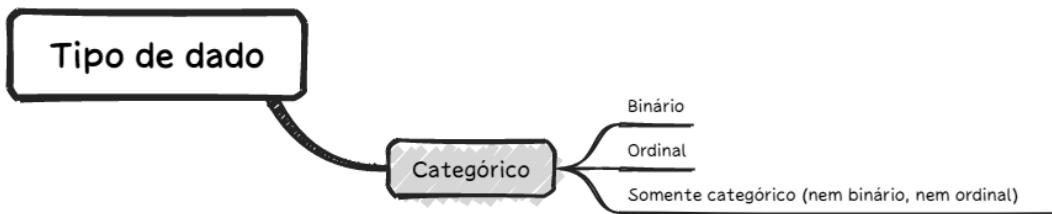


Figura 2.3: Subtipos Categórico

Por exemplo, a idade de um estudante é um dado do tipo **numérico discreto**.

Cada linguagem de programação tem tipos de dados específicos para suas variáveis, que estão relacionados às vezes com essa classificação teórica (numérico contínuo ou numérico discreto). Por exemplo no python temos o tipo da variável *int* (inteiro), que é equivalente ao numérico discreto, já o tipo *float* (flutuante) é equivalente ao numérico contínuo.

Aqui está um exemplo de código em Python:

```

1  >>> idade = 25
2  >>> type(idade)
3  <class 'int'>
  
```

Listing 2.1: Exemplo de declaração de variável do tipo inteiro, equivalente ao tipo de dado numérico contínuo.

2.1.2 Dado qualitativo

Um dado é do tipo categórico quando ele faz parte de um conjunto, de uma classe ou de uma categoria.

O dado categórico pode ser binário ou ordinal, ou nenhuma das duas subcategorias (Figura 2.3).

Um exemplo de dado categórico, é uma lista com as cores preferidas dos estudantes, ou o estado civil de uma pessoa.

O dado do tipo categórico binário é um tipo especial quando ele somente pode assumir dois valores no universo de valores possíveis. Por exemplo 0 ou 1, existente ou ausente, true ou false, sim e não.

O dado do tipo categórico ordinal também é um tipo especial, é quando ele faz parte de um conjunto com determinada ordem, por exemplo, imagine a classificação de altura de estudantes somente com os valores alto, médio e baixo. Nesse exemplo existe uma ordem, o aluno com altura classificado como baixo tem uma altura menor do que o aluno com altura média.

Referências Bibliográficas

- [1] blipblog. <https://www.take.net/blog/tecnologia/inteligenciaartificial/>.
- [2] P. R. Norvig and S. A. Intelligence. A modern approach. *Prentice Hall Upper Saddle River, NJ, USA: Rani, M., Nayak, R., & Vyas, OP (2015). An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage. Knowledge-Based Systems*, 90:33–48, 2002.
- [3] A. M. Turing. Computing Machinery and intelligence. *Mind*, 49(8):433–460, 1950.
- [4] Wikipedia. <https://pt.wikipedia.org/wiki/Humano>.
- [5] wikipedia. <https://pt.wikipedia.org/wiki/TransformadadeFourier>.