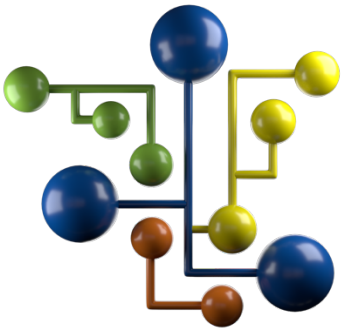

Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Reconhecimento de Voz

O reconhecimento de voz é a tarefa de identificar uma sequência de palavras proferidas por um falante, dado um sinal acústico. Tornou-se uma das aplicações principais de IA — milhões de pessoas interagem com os sistemas de reconhecimento de voz a cada dia para navegar por sistemas de correio de voz, pesquisar na Web a partir de telefones móveis e outras aplicações. A voz é uma opção atraente quando há necessidade de operação com as mãos livres e ao operar máquinas. O reconhecimento de voz é difícil porque os sons feitos por um falante são ambíguos e… ruidosos. Como exemplo bem conhecido, o sintagma "recognize speech" soa quase o mesmo que "wreck a nice beach", quando falado rapidamente. Esse exemplo curto mostra vários dos problemas que tornam a voz problemática. Primeiro, a segmentação: as palavras escritas em inglês têm espaços entre elas, mas na fala rápida não há pausas em "wreck a nice" que a distinguisse como uma frase de várias palavras em oposição à palavra "recognize". Segundo, a coarticulação: quando se fala rapidamente, o som do "s" ao final de "nice" se funde com o som de "b" do início de "beach", produzindo algo como "sp". Outro problema que não aparece nesse exemplo é o de homófonos — palavras como "to", "too" e "two", que têm o mesmo som, mas diferem em significado.

Podemos ver o reconhecimento de voz como um problema na explicação da sequência mais provável. Esse é o problema de calcular a sequência mais provável das variáveis de estado, **x**1:*t*, dada uma sequência de observações **e**1:*t*. Nesse caso, as variáveis de estado são as palavras, e as observações são os sons. Mais precisamente, uma observação é um vetor de características extraído do sinal de áudio. Como de costume, a sequência mais provável pode ser calculada com a ajuda da regra de Bayes:

$$\underset{palavra_{1:t}}{argmax}\, P(palavra_{1:t} \mid som_{1:t}) = \underset{palavra_{1:t}}{argmax}\, P(som_{1:t} \mid palavra_{1:t})\, P(palavra_{1:t})$$

Aqui *P*(*som*1:*t* | *palavra*1:t) é o **modelo acústico**. Descreve os sons das palavras — que "**ceiling**" (teto) começa com um "c" suave e soa como "*sealing*" (vedação). P(*palavra*1:*t*) é conhecido como **modelo de linguagem**. Especifica a probabilidade antes de cada emissão — por exemplo, que "*ceiling fan*" (ventilador de teto) tem cerca de 500 vezes mais probabilidade como sequência de palavras que "*sealing fan*" (ventilador de vedação).

Essa abordagem foi nomeada por Claude Shannon (1948) de **modelo de canal ruidoso**. Ele descreveu uma situação em que uma mensagem original (as *palavras* em nosso exemplo) é transmitida através de um canal ruidoso (como uma linha de telefone) de tal forma que uma mensagem alterada (os *sons* no nosso exemplo) é recebida na outra ponta. Shannon mostrou que não importa o quão ruidoso seja o canal, é possível recuperar a mensagem original com erro arbitrariamente pequeno se codificarmos a mensagem original de forma bastante redundante. A abordagem de canal ruidoso tem sido aplicada para reconhecimento de voz, tradução automática, correção ortográfica e outras tarefas.

Uma vez que definimos os modelos acústicos e de linguagem, podemos resolver, para a sequência de palavras mais provável, utilizando o algoritmo de Viterbi. A maioria dos sistemas

de reconhecimento de voz usa um modelo de linguagem que realiza a suposição de Markov — que o estado atual *Palavrat* depende apenas de um número fixo *n* de estados anteriores — e representa *Palavrat* como uma única variável aleatória que assume um conjunto finito de valores, que a torna um Modelo Oculto de Markov (MOM). **Assim, o reconhecimento de voz torna-se uma aplicação simples da metodologia do MOM, uma vez definidos os modelos acústico e de linguagem.** Falaremos sobre ele em seguida.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig