MAP2010 - Trabalho de Formatura
Projeto de Pesquisa
Perceptrons

Eduardo Galvani Massino
Orientador: José Coelho de Pina Junior

1 Introdução

De tempos pra cá, ler e ouvir falar de **ciência de dados** tornou-se muito comum, tanto nos meios profissionais e científicos quanto na mídia. Existem atualmente aplicações em praticamente todas as áreas do conhecimento humano, da agricultura à indústria e ao entretenimento.

Uma busca rápida na *Wikipedia* (*Ciência de Dados* 2020) define ciência de dados como um conjunto de ferramentas que extrai informações ou previsões a partir de um grande volume de dados, que podem ser números, textos, áudio, vídeo, entre outros, para ajudar na tomada de decisões de negócios.

Apesar de não ser a única definição para o termo, Pedro A. Morettin e Julio M. Singer (Pedro A. Morettin, 2020) nos lembram que essa também é uma definição da estatística. Eles comparam o uso dos termos e apontam que o trabalho dos *cientistas de dados* diferem dos *estatísticos* apenas quando eles usam dados de natureza multimídia como áudio e vídeo, por exemplo. Mas que, uma vez que esses dados são processados e tornam-se números, as técnicas e conceitos utilizados pelos primeiros passam a ser basicamente os mesmos utilizados pelos segundos.

Na verdade, Morettin & Singer (Pedro A. Morettin, 2020) citam que na década de 80 houve uma primeira tentativa de aplicar o rótulo *ciência de dados*, (*Data Science*), ao trabalho feito pelos estatísticos aplicados da época, como uma forma de dar-lhes mais visibilidade. Curiosamente, fato mencionado pelos autores, existem atualmente

cursos específicos de ciência de dados em universidades ao redor do mundo, mas a maioria deles situada em institutos de áreas aplicadas como Engenharia e Economia, e raramente nos institutos de Estatística propriamente ditos.

Para entender um pouco mais de seu escopo, David M. Blei e Padhraic Smyth (David M. Blei, 2017) discutem ciência de dados sob as visões estatística, computacional e humana. Eles argumentam que é a combinação desses três componentes que formam a essência do que ela é e, assim como, do conhecimento que ela é capaz de produzir.

Em resumo, a estatística guia a coleta e análise dos dados. A computação cria algoritmos, técnicas de processamento em paralelo e gerenciamento de memória eficazes para que sua execução seja efetiva. E o papel humano é o de avaliar quais tipos de dados, técnicas de análises, algoritmos e modelos são apropriados para responder ao problema em questão. Este é o papel do *cientista de dados*.

2 Aprendizado de máquina

Enquanto isso, algoritmos de **aprendizado de máquina** vem sendo utilizados em grande parte dos modelos de ciência de dados. Mas o que é aprendizado de máquina? Ou então, o que significa dizer que o computador, neste caso a "máquina", está *aprendendo*?

Antes da definição formal, Aurélien Géron (Géron, 2019) nos dá uma ideia geral lembrando que uma das primeiras aplicações de sucesso de aprendizado de máquina foi o filtro de *spam*, criado na década de 90. Uma das fases de seu desenvolvimento foi aquela em que os usuários assinalavam que certos e-mails eram *spams* e outros não eram. Hoje em dia, raramente temos que marcar ou desmarcar e-mails, pois a maioria dos filtros já "aprenderam" a fazer seu trabalho de forma muito eficiente.

Dentre as muitas definições de aprendizado de máquina, de modo geral uma área da ciência da computação, no contexto de ciência de dados, Joel Grus (Grus, 2016) define-o como a "criação e o uso de modelos que são ajustados a partir dos dados". Seu objetivo é usar dados existentes para desenvolver modelos que possamos usar para *prever* possíveis saídas para dados novos. Exemplos, além do filtro de *spams* podem ser: prever transações de crédito fraudulentas, prever a chance de um cliente clicar em uma propaganda ou então prever qual time de futebol irá vencer o Campeonato Brasileiro.

Pode-se classificar as técnicas de aprendizado de várias formas, de acordo com alguma de suas características. Por exemplo, Géron (Géron, 2019) utiliza o grau de supervisão humana durante o seu funcionamento para classificá-los em aprendizado supervisionado ou não-supervisionado.

Aprendizado supervisionado

Um algoritmo de **aprendizado supervisionado** é usado quando conhecemos os rótulos dos dados que estamos utilizando para o treinamento, ou seja, temos a resposta correta da aprendizagem. Por exemplo, se estamos classificando fotos de animais, possuímos um conjunto de fotos em que já sabemos de antemão quais são de gatos, cachorros, etc.

O ato de rotular ou classificar os dados que usamos no aprendizado é o que designamos de supervisão humana. Uma vez *treinado*, o algoritmo recebe uma foto nova e então a classifica como sendo uma foto de um gato, ou cachorro, ou qualquer outra resposta daquelas que foram dadas como exemplos durante o treinamento.

Dentro do aprendizado supervisionado temos duas técnicas principais. A regressão é usada para prever valores, e a classificação é usada para prever os rótulos dos dados, que também são chamados de classes. Neste texto os termos "algoritmo" e "técnica" serão usados livremente como sinônimos, pois uma técnica de aprendizado de máquina, no contexto atual, é obviamente um algoritmo executado no computador.

Aprendizado não-supervisionado

Nesse tipo de aprendizado de máquina, não sabemos os rótulos dos dados que estamos lidando, assim o algoritmo poderá agrupar os dados de forma automática, por exemplo, se estiver sendo usado um algoritmo classificador.

Alguns métodos não-supervisionadas de aprendizado foram enumeradas por Géron (Géron, 2019). O **agrupamento** de dados similares, sendo essa similaridade podendo ser uma distância no espaço dos dados (inspiração geométrica), e utiliza-se algoritmos como k-Vizinhos, k-Means, k-Medians, etc. Exemplos de aplicações são agrupamento de produtos em supermercados, interesses comuns de clientes em sites de conteúdo digital, etc.

Outra técnida é a **detecção de anomalias**, cujo objetivo é ter uma descrição de como os dados considerados "normais" se parecem, e usa esse agrupamento para detectar se novos dados estariam "fora" desse padrão. Um exemplo é a detecção de fraudes.

Também pode-se citar sobre a técnica de **estimação de densidades**, que tem como objetivo a estimação da função densidade de probabilidade de um conjunto de dados gerados por algum processo aleatório.

3 Técnicas de Classificação

Sendo uma das duas técnicas principais do aprendizado supervisionado, problemas desse tipo buscam aprender com um conjunto de dados previamente rotulados, como "se parecem" os dados que pertencem às classes que queremos classificar, para que quando processarmos novos dados, o algoritmo usado possa identificar, o mais corretamente possível, as classes às quais pertecem esses dados, dentro do conjunto de classes que já definimos ao rotular os dados iniciais.

Existem vários tipos de algoritmos de classificação, dentre eles podemos mencionar: Máquina de Vetor Suporte, em inglês Support Vector Machine (SVM), Árvores de decisão, Florestas aleatórias (podendo ser entendidas como um conjunto de centenas de árvores de decisão aleatoriamente definidas) e as Redes neurais artificiais.

Todos essas técnicas podem ser usadas para classificação linear ou não-linear, no sentido em que valores eles estão classificando, assim como na forma que está sendo feita essa classificação. Se imaginamos um espaço bidimensional, um algoritmo de classificação linear irá separar as classes de dados por retas, enquanto que um classificador não-linear poderá usar outra curva qualquer para a separação. Abstraindo o espaço bidimensional para os espaços multidimensionais dos dados que são comumente analisados, podemos pensar em hiperplanos (estruturas (n–1-dimensionais de espaços n-dimensionais) para o caso dos classificadores lineares, ou subespaços quaisquer para os não-lineares.

4 Redes Neurais

Uma **rede neural** é um exemplo de modelo preditivo de aprendizado de máquina. Tal modelo foi criado com inspiração no funcionamento do cérebro biológico, e que, apesar de terem sido as primeiras a serem criadas, conforme descrito por David Kopec (Kopec, 2019), vem ganhando nova importância na última década, graças ao avanço computacional, uma vez que exigem muito processamento, e também porque podem ser usadas para resolver problemas de aprendizagem dos mais variados tipos.

Uma rede neural artificial é um dentre vários métodos de classificação, ou seja, de aprendizado supervisionado. De acordo com Kopec (Kopec, 2019), ele é utilizado como um classificador não-linear, e por isso pode ser utilizado para prever tipos de dados genéricos, que podem ou não ser lineares.

5 Perceptron

Atualmente existem vários tipos de redes neurais, porém este trabalho lida principalmente com aquele tipo que foi originalmente criado sob a inspiração do funcionamento do cérebro, chamado de **perceptron**, e que portanto tenta imitar o comportamento dos neurônios e suas conexões, aprendendo padrões a partir de dados existentes e tentando prever o comportamento de dados novos a partir do padrão aprendido.

Mais recentemente surgiu a rede perceptron de várias camadas (*multi-layer perceptron*), àquela mais simples e antiga dá-se o nome de perceptron de única camada (*single-layer perceptron*), uma ilustração dela está na Figura 1.

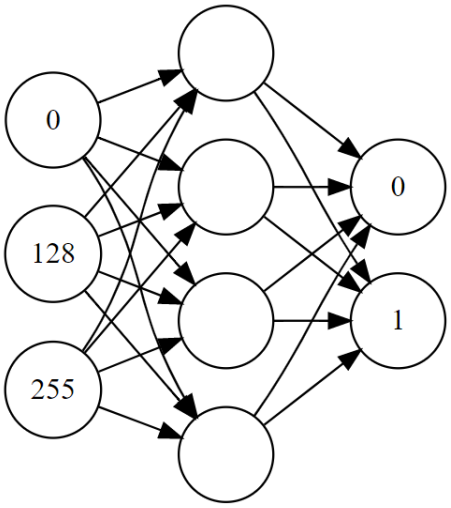

Figura 1: *Rede neural simples, o perceptron de única camada.*

O perceptron de camada única consiste de uma camada de neurônios de entrada, uma camada oculta de neurônios usados na otimização, e uma camada de saída, que irá conter os dados previstos, ou ainda as probabilidades do dado pertencer a alguma das classes que a rede poderá classificá-lo.

Os neurônios são representados por círculos, cada coluna de neurônios representa uma camada, nesse caso, da esquerda para a direita temos a camada de entrada, a camada oculta e a camada de saída. As linhas representam as ligações entre os neurônios, sendo que cada neurônio de uma camada está ligado a todos da camada anterior.

Referências

[*Ciência de Dados* 2020] *Ciência de Dados*. https://pt.wikipedia.org/wiki/Ci%C3%AAncia_de_dados. Mar. de 2020 (citado na pg. 1).

[David M. Blei 2017] Padhraic Smyth David M. Blei. "Science and data science". Em: *PNAS* 114.33 (ago. de 2017), pgs. 8689–8692 (citado na pg. 2).

[Géron 2019] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2º. O'Reilly, 2019 (citado nas pgs. 2, 3).

[Grus 2016] Joel Grus. *Data Science do Zero*. 1º. O'Reilly, 2016 (citado na pg. 2).

[Kopec 2019] David Kopec. *Problemas Clássicos de Ciência da Computação com Python*. 1º. Novatec, 2019 (citado na pg. 4).

[Pedro A. Morettin 2020] Julio M. Stinger Pedro A. Morettin. *Introdução à Ciência de Dados - Fundamentos e Aplicações*. Departamento de Estatística. Universidade de São Paulo, 2020 (citado na pg. 1).