XXIX Congresso de Iniciação Científica UNICAMP 2021

Introdução à computação científica: aprendizagem de máquina e algoritmos de classificação

Palavras-chave: apredizagem de máquina, algoritmos de classificação, classificador bayesiano, k-NN, classificador linear.

Beatriz Belucci [Unicamp]
Profº Dr. Ricardo Biloti (orientador) [Unicamp]

1 Introdução

Este projeto foi planejado com o objetivo de realizar o estudo da Computação Científica, através do estudo de modelagens matemáticas e ferramentas teóricas que auxiliassem na resolução de um problema selecionado. Para tal, foi escolhido o tema Inteligência Artificial como área de estudo, focando particularmente nos algoritmos de Aprendizagem de Máquina. Dentre esses algoritmos, os estudados foram os algoritmos de classificação.

Para desenvolver o projeto, foi utilizado como material principal de estudo o livro de Kubat (2017), *An Introduction to Machine Learning*. O material apresenta os fundamentos da aprendizagem de máquina, enfatizando os métodos de classificação, exibindo a teoria relacionada a diferentes métodos classificadores, assim como formas de implementar seus algoritmos e aplicações. Foram estudados o Classificador Bayesiano, o método dos k-Vizinhos Mais Próximos e os Classificadores Lineares.

Quando necessário, foram realizadas leituras complementares de outras referências, a fim de aprimorar o aprendizado. Para o estudo de probabilidades foi utilizado o texto de Ross (2009). Já para auxiliar na implementação dos algoritmos em Octave, foi utilizado o livro Quarteroni e Saleri (2007); e as vídeo aulas Biloti (2020a) e Biloti (2020b). Esse projeto foi executado com uma bolsa financiada pelo SAE/Unicamp.

2 Metodologia

O estudo teórico foi feito seguindo os capítulos de Kubat (2017), ou seja, o livro foi utilizado como um roteiro de estudos para o projeto. A cada semana o conteúdo era avançado com a leitura da teoria apresentada no livro, bem como com a elaboração de exemplos de aplicações dos métodos aprendidos. Ademais, semanalmente foram realizadas reuniões com o orientador e com colegas de iniciação científica, onde cada aluno teve a oportunidade de expor o conteúdo estudado ao longo da semana através de uma apresentação de slides. Com as reuniões, abriram-se discussões a cerca do tema e, além disso, era possível, com o professor, solucionar as dúvidas a respeito do conteúdo, enriquecendo o estudo.

3 Teorias estudadas

Nos algoritmos de classificação, queremos que o algoritmo receba um objeto de classe desconhecida e que ele seja capaz de nos retornar a classe a qual esse objeto pertence. Para isso, é necessário, primeiramente, que o algoritmo saiba como realizar uma classificação. Vejamos então, como o algoritmo aprende a classificar.

Para induzir um classificador, é preciso, antes de mais nada, treinar o algoritmo. Isso é feito através do *conjunto de treinamento*. Os elementos do conjunto de treinamento, por estarem já classificados, são explorados na determinação de padrões. Esses padrões são, numa etapa de classificação, empregados na inferência da classe sempre que um novo elemento for analisado. Isso permite que o algoritmo consiga, quando receber um objeto de classe desconhecida, analisar com quais elementos do conjunto de treinamento ele melhor se assemelha e, consequentemente, a qual classe ele tem maiores chances de pertencer.

Para que o algoritmo seja capaz de transformar as informações que recebeu em conhecimento, é necessário que essas informações sejam passadas a ele de maneira adequada. Podemos transmitir essas informações utilizando vetores, os vetores de atributos. No conjunto de treinamento podemos extrair muitas informações a respeito de seus elementos. Entretanto os vetores de atributos armazenam apenas as características que julgamos serem suficientes para a decisão de pertencimento a uma classe ou outra. Essas características podem ser tanto discretas como contínuas.

Uma maneira de testar se o classificador induzido está correto é dividir o conjunto de objetos conhecidos em dois subconjuntos: conjunto de treinamento e conjunto de teste. Fazendo isso, podemos utilizar o conjunto de treinamento para induzir o classificador e o conjunto de teste para verificar se as classificações estão corretas, umas vez que essas classes eram previamente conhecidas.

Alguns métodos que induzem classificadores foram estudados. Neste resumo serão apresentados três deles. O primeiro deles é o classificador Baysiano, um método que utiliza de probabilidades e da fórmula de Bayes para realizar as classificações. Esse método pode ser apli-

cado de duas formas distintas, levando em consideração dois casos: quando os dados utilizados na classificação são discretos e quando são contínuos. Nesse classificador o conjunto de treinamento é utilizado para o cálculo das probabilidades necessárias ao método. Esse cálculo é feito através da fórmula de Bayes. Para o caso discreto, a fórmula de Bayes é empregada utilizando-se os conceitos de probabilidade condicional e frequência relativa. Já para o caso contínuo, utiliza-se os conceitos de probabilidade condicional e função densidade de probabilidade. Por fim, o algoritmo atribui ao objeto a classe com maior probabilidade.

Em seguida, foi estudado o método dos k-Vizinhos Mais Próximos (k-NN). Nele são utilizadas as distâncias geométricas entre k pontos no espaço para realizar as classificações. Nesse método os vetores de atributos $\mathbf{x} \in \mathbb{R}^n$ do conjunto de treinamento são representados como pontos no espaço n-dimensional e as métricas são calculadas entre tais pontos e o ponto que representa o objeto que queremos classificar. Consideramos que pontos mais próximos tendem a ser mais semelhantes e então, a classificação é feita escolhendo k pontos, os k-vizinhos, mais próximos do objeto que está sendo classificado. Dentre esses k pontos, a classificação é feita atribuindo ao objeto a classe com maior representatividade.

Por fim, temos o classificador linear. Assim como no método k-NN, este método parte da representação dos vetores de atributos no espaço n-dimensional. Ele é empregado apenas nos casos em que as classes existentes no conjunto de treinamento são linearmente separáveis, ou seja, os pontos representantes de cada classe estão aglomerados em regiões que podem ser separadas por uma função linear. O objetivo desse classificador é determinar a função linear que faz essa separação. Para isso podem ser utilizados dois algoritmos: *Perceptron Learning* e *WINNOW*. Tais algoritmos têm como saída os coeficientes que definem a função linear desejada. Para uma melhor aplicação desses algoritmos, os atributos devem ser variáveis booleanas representadas pelos inteiros 1 ou 0 ou valores reais normalizados no intervalo [0,1]. A função linear que separa as classes é denominada *superfície de decisão*. Essa superfície pode ser uma reta, um plano ou um hiperplano. Ela é uma reta no caso em que temos os vetores de atributos $\mathbf{x} \in \mathbb{R}^2$, um plano quando $\mathbf{x} \in \mathbb{R}^3$ e um hiperplano quando $\mathbf{x} \in \mathbb{R}^n$, com $n \geq 4$. A partir dessa superfície, a classificação é feita utilizando-se a equação linear encontrada.

4 Resultados e discussão

Para cada um dos métodos de classificação citados, foram criados algoritmos para implementá-los. Nesta seção, são apresentados alguns dos exemplos criados para ilustrar os diferentes classificadores, bem como os resultados das classificações realizadas em cada algoritmo.

Para ilustrar o classificador Bayesiano para atributos discretos, empregamos o algoritmo para classificar animais em duas classes, os "selvagens" e os "domésticos". Para tal, foi criado um conjunto contendo 26 animais de classes conhecidas. Cada animal foi descrito por um vetor com quatro características: pele, tamanho, alimentação e tipo. Foi criado ao todo 13 tipos de atributos. Para o atributo "pele" temos: pelo, escama e pena. Para o atributo "tamanho":

pequeno e grande. Para "alimentação": carnívoro, herbívoro, onívoro e insetívoro. E, por fim, para "tipo", temos: ave, mamífero, réptil e peixe. A título de exemplo, temos o animal "Cachorro", que é descrito pelo vetor $\mathbf{x} =$ (pelo, pequeno, carnívoro, mamífero) e pertence à classe "domésticos".

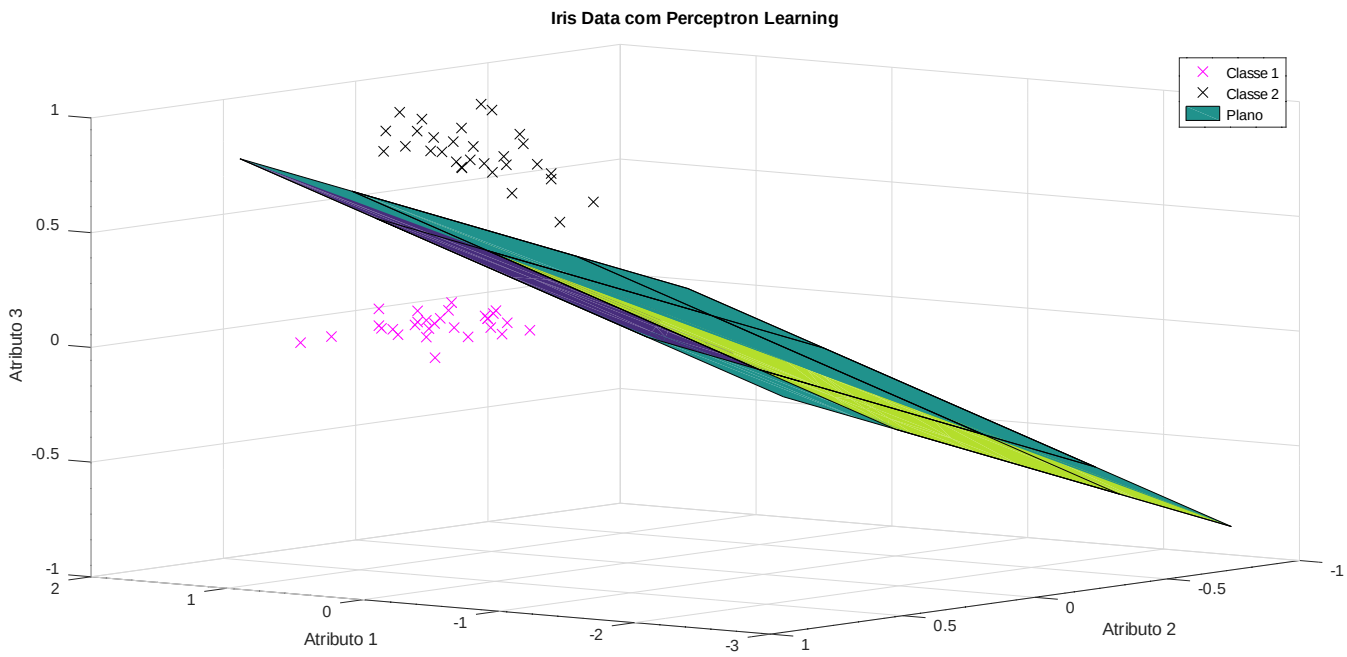
Para calcular as probabilidades necessárias para a implementação do método, foram escolhidos aleatoriamente 14 animais. Em seguida, foi utilizada a fórmula de Bayes para classificar os 12 animais restantes. Para esse método, 12 objetos foram classificados. O algoritmo classificou corretamente 7 deles, resultando em aproximadamente 58% de acerto.

Já para o classificador Bayesiano para atributos contínuos, para criar o conjunto de treinamento e o conjunto de teste, foi utilizado Fisher (1936). Nesse conjunto de dados temos 150 flores *Iris* de 3 espécies distintas: *Iris Setosa*, *Iris Versicolor* e *Iris Virginica*. Cada flor é descrita por um vetor com 4 atributos: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. Dentre esses 150 exemplos, 90 deles foram utilizados para treino e os 60 restantes para teste.

Para esse caso, o classificador Bayesiano apresentou bons resultados. Dentre 60 classificações, 58 delas foram bem sucedidas, resultando em aproximadamente 96% de acerto. Essa melhoria, quando comparado ao caso de vetores discretos, pode ser resultado do uso de um conjunto de dados melhor, ou seja, um conjunto de dados maior e com atributos mais representativos. Como uma grande quantidade de elementos foi utilizada para treino, as probabilidades puderam ser estimadas com maior acurácia. Todos esses fatores influenciam na qualidade da classificação.

Para o método k-NN foi empregado o mesmo conjunto de dados utilizado no método do classificador Bayesiano para criar o conjunto de treinamento e o conjunto de teste. Esses dados foram preparados para a implementação do algoritmo da mesma forma como feito anteriormente. Foi utilizado $k = 6$. Esse método também classificou corretamente 58 dos 60 objetos, obtendo aproximadamente 96% de acerto. O fato de ter sido utilizado um bom conjunto de dados também pode ser mencionado aqui.

Finalmente, para ilustrar o classificador linear, é apresentado no gráfico a seguir o resultado da implementação do algoritmo *Perceptron Learning* utilizando o conjunto de dados *Iris*. Para essa implementação, a fim de exibir o gráfico, o quarto elemento dos vetores de atributos foi omitido. Além disso, estão representados graficamente apenas as classes 1 e 2, pois são linearmente separáveis. Os pontos apresentados no gráfico são os representantes do conjunto de treinamento utilizado, já o plano exibido é a superfície de decisão determinada pelo algoritmo. A equação dessa superfície é usada na etapa de classificação para a decisão de pertencimento à uma classe ou outra.



5 Considerações finais

Em conclusão, podemos notar, a partir do resultados obtidos para os diferentes métodos, que o pré processamento dos dados é fundamental para uma boa classificação. Um conjunto de dados com um número significativo de elementos e um conjunto de treino representativo e com atributos relevantes, permite que as estimativas dos dados necessários a cada método sejam feitas de maneira menos perturbada, gerando resultados melhores.

Referências

Biloti, R. (2020a). Introdução ao Octave – Parte I. https://www.ime.unicamp.br/~biloti/an/211/oct-01.html (acessado em 29/11/2020).

Biloti, R. (2020b). Introdução ao Octave – Parte II. https://www.ime.unicamp.br/~biloti/an/211/oct-02.html (acessado em 29/11/2020).

Fisher, R. (1936). Iris data set. https://archive.ics.uci.edu/ml/datasets/Iris (acessado em 02/03/2021).

Kubat, M. (2017). *An Introduction To Machine Learning*. Springer.

Quarteroni, A. M. e Saleri, F. E. (2007). *Cálculo Científico com MATLAB e Octave*. Springer.

Ross, S. (2009). *Probabilidade: um curso moderno com aplicações*. Bookman Editora.