Relatório de Inteligência Artificial
Aprendizado de Máquina

Hugo Carneiro, Mariana Werneck

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)

{hugocarneiro, marianawerneck}@id.uff.br

Resumo. *O objetivo do nosso trabalho é construir um modelo de aprendizado de máquina, através da implementação de 3 algoritmos de aprendizado supervisionado, e então avaliar os resultados da nossa implementação em função de métodos de avaliação apropriados*

1. Introdução

Para a implementação do nosso trabalho, começamos escolhendo entre uma das bases de dados fornecidas para o trabalho. Nosso grupo iniciou com a base de revisões de diferentes tipos de vinhos, mas voltamos e preferimos mudar para a base relativa ao comportamento do tráfego urbano na cidade de São Paulo. A escolha se deu principalmente para evitar o uso de uma base textual. Em seguida determinamos quais métodos utilizaríamos para aproveitar os atributos da base, escolhendo como target a porcentagem de atraso no tráfego causada pelos acidentes.

2. Metodologia

Nós escolhemos como ferramenta de resolução do trabalho o SciKit, devido principalmente à familiaridade do grupo com a linguagem Python. Devido à natureza discreta dos dados fornecidos pela base de dados, escolhemos adotar dois métodos classificadores para aprendizado: KNN e regressão logística.

2.1. KNN

O algoritmo conhecido como KNN (K-Nearest Neighbor) envolve determinar um ou mais pontos em um conjunto de dados que sejam os mais próximos de um ponto de consulta dado[Beyer et al. 1999]. Ele é melhor explicado pensando-se nos dados dos atributos como pontos ao longo de um plano cartesiano: assim como podemos utilizar técnicas de interpolação ao longo de pontos para descrever uma função em um plano, podemos utilizar o KNN para determinar o valor de um ponto aleatório no nosso plano, avaliando os K vizinhos mais próximos para chegar a uma conclusão adequada.

Uma vantagem do algoritmo de KNN é que a base de dados não precisa ser treinada antes da utilização desse algoritmo. Porém, é importante levar em consideração a busca pelos vizinhos mais próximos, que é essencial para a determinação da acurácia do algoritmo, assim como potencialmente custoso para a máquina em casos de um conjunto de dados muito grande, devido ao alto gasto de memória. Como um dos dois algoritmos cuja implementação pôde ser feita diretamente pela ferramenta, e dado a natureza da base de dados, escolhemos a classe *KNeighborsClassifier* do SciKit.

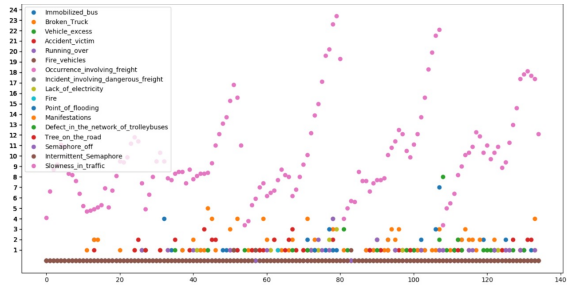

Figure 1. Dados representados com o auxílio do Matplotlib

Devido à influência no número de vizinhos K escolhido na determinação da acurácia do algoritmos, fizemos 3 (três) testes diferentes, utilizando $K = 2$, $K = 7$ e $K = 10$, respectivamente.

2.2. Regressão Logística

Métodos de regressão são um componente importante de qualquer análise de dados preocupada com a relação entre uma variável de resposta e duas ou mais variáveis utilizadas para chegar à resposta. Com frequência ocorrer da variável resultante ser discreta, com dois ou mais resultados possíveis. Nesses casos, regressão logísticas tem se tornado o método de escolha para resolução de problemas. O que distingue um modelo de regressão logística da regressão linear comum, que não discutiremos nesse relatório, é que a saída do algoritmo é discreta, e não contínua.

Como um dos métodos que são implementados já utilizando as funcionalidades fornecidas pela ferramenta, nós não nos ateremos a complexa matemática das regressões lineares, responsáveis pelo método de treinamento. Porém, é interessante apontar que para se avaliar adequadamente o método é importante ser capaz de dar um significado adequado à diferença entre duas unidades lógicas, ou *logits*, especialmente em pontos extremos do espaço amostral[Pregibon et al. 1981]. Assim como no método anterior, a implementação foi feita usando as técnicas já fornecidas pela ferramenta SciKit, no caso a classe *LogisticRegression*.

2.3. Bagging

Nosso método de Ensemble escolhido foi Bagging (uma abreviação de *Bootstrap aggregating*), proposto por Leo Braiman em 1994 para aumentar a acurácia de métodos de classificação através da combinação de um conjunto de métodos de classificação gerados aleatoriamente[Breiman 1996]. A ideia é que cada um dos métodos gerados aleatoriamente recebe uma amostra dos dados de treinamento, e através de um método de votação o algoritmo original pode determinar o resultado do ponto de consulta melhor do que qualquer um dos conjuntos individualmente[Bauer and Kohavi 1999].

Na nossa implementação, nosso algoritmo de votação é relativamente simples: nenhum peso é associado a nenhuma amostra em especial, e simplesmente escolhemos o resultado discreto que for mais prevalente dentro do espaço amostral fornecido. O Bagging utiliza, portanto, exclusivamente de métodos classificadores, embora existam outras formas de algoritmos de Ensemble que façam resoluções de regressão, tal como tirar a média ou mediana de resultados contínuos. Para a resolução do resultado, empregamos

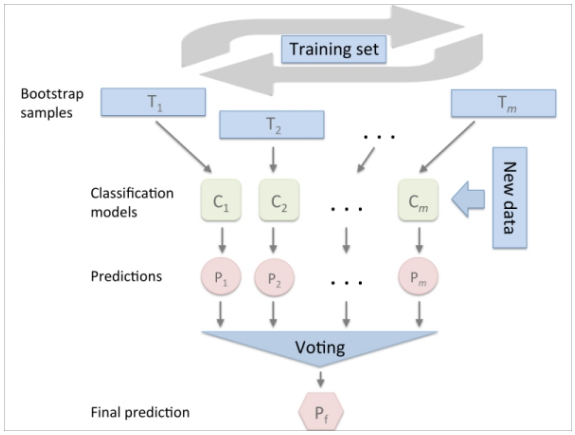

Figure 2. Ilustração do funcionamento do algoritmo Bagging

mais uma vez o KNN, utilizando como amostras metade do espaço amostral disponível para eles. Numa tentativa de diminuir a probabilidade de empates no espaço amostral, o número escolhido de KNNs gerados foi um número primo, 53.

3. Conclusão

O target do nosso trabalho é determinar, dados alguns atributos correspondentes a vários tipos diferentes de acidentes de trânsito, calcular como esses atributos afetam o trânsito paulista. Para determinar a eficácia do algoritmo, utilizamos calculo de acurácia do algoritmo, precisão e f1.

References

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 36(1-2):105–139.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Pregibon, D. et al. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724.