Atividade 2: Exploração de dados criminais

Rodney Rick
RA 181.814
Disciplina MO444
Prof. Dr. Anderson Rocha

***Resumo*—O objetivo deste trabalho é estudar a representação dos dados, exploração e tranformação dos mesmo para posterior análises com modelos de Regressão Logística e Rede Neural Multi-Camada (*MLP*), a fim de prever a categoria do Crime com base em um determinado conjunto de dados entrada. Discussão dos dados de saída e problemas enfrentados para as execuções.**

Palavras-chave: Transformação dos dados, Regressão Logística e *MLP*.

I. Introdução

Nas seções seguintes o conjunto de dados será apresentado, seguido por análise sobre as alguns pontos de vista das características dos atributos e dos modelos abordados, como o número e a distribuição de amostras, a natureza do atributos e funções de custo de aprendizagem (computacionalmente).

Para o escopo deste trabalho, foi necessário a extração de alguns dados dos atributos a fim de gerar novos campos interessantes para uso, essa é a etapa de engenharia dos dados. Transformar para obter mais recursos para explorar.

Utilizará duas técnicas classificativas, uma de Rede Neural, conhecida como *MLP* (*Multi-Layer Perceptron*) e outra mais simples para comparação, a Regressão Logística.

Dados as técnicas, avalia-se atributos descritivos e a combinação dos dados sugeridas para análise (Dia da Semana, Distrito Policial, Latitude e Longitude). Assim, caso haja uma entrada criminal, é possível sua classificação conforme esses dados? Outras discussões são levantadas (seção II) e conforme os gráficos existentes no Apêndice A, verifica-se, de modo visual, qual seriam os possíveis dados mais destacados a serem trabalhados. Esse tipo de visualização, normalmente, é consideração a etapa inicial a fase exploratória dos dados.

Ao final do relatório (seção V) faz-se o levantamento dos resultados e discussão das técnicas aplicadas conforme a resolução.

Ainda no Apêndice A, os gráficos apresentados exibem, por exemplo, dados de Distritos Policiais que mais recebem/tratam dos crimes e quais crimes. Na figura 1, uns levantamento em formato de distribuição dos crimes dentro da base de dados.

II. A amostragem dos dados

O conjunto de dados referente a base criminal contém 878.049 amostras. Cada amostra está descrita pelo conjunto de informações descritas na tabela I.

As amostras são completas, porém nas 39 Categorias de crimes que a quantidade de dados mostra-se desproporcional conforme a figura 1.

Tabela I: Tabela de descrição dos campos da base de entrada

Atributo	Descrição dos campos
Data	Data e hora do acontecimento do crime
Categoria	categoria do Incidente que o crime é classificado
Descrição	descrição detalhada do incidente crime
Dia da semana	-
Distrito Policial	nome do Departamento de Polícia do Distrito que tratou o crime
Resolução do Crime	Descrição de como foi resolvido o crime
Endereço	Local do crime
Longitude	-
Latitude	-

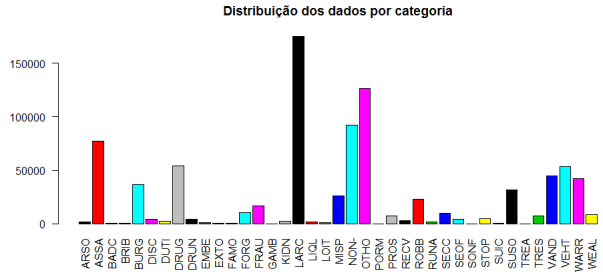

Figura 1: Distribuição dos crimes (os nomes das categorias foram substituídas por abreviações).

Outros levantamentos (citados abaixo) foram feitos e as figuras para demonstração do conceito estão presentes no Apêndice A.

(i) Distrito Policial → Distribuição crimes por local de atendimento (figura 5);
(ii) Resolução do Crime → Distribuição dos dados conforme a resolução atribuída ao registro (figura 6). Esta parece pouco promissor, por existirem muitos crimes "Sem Resolução"como dados da amostragem.
(iii) Anos → Distribuição dos dados está ao longo de 2003 até início de 2015 e dos crimes realizados (figura 7);
(iv) Mês → Distribuição dos dados está ao longo dos meses do ano, mas que apresenta uma homogeneidade dos crimes (figura 8);
(v) Dias da semana → Distribuição dos dados está ao longo dos dias da semana, mas que também apresenta uma homogeneidade dos crimes (figura 9);
(vi) Horas → Distribuição dos dados está ao longo das horas dos dias e que demonstra-se relevantes para certos crimes menos cometidos (figura 10);
(vii) Dia da Semana por hora → Mapa de calor da distribuição

da quantidade de ocorrências ao longo das horas e dias da semana (figura 11).

III. Exploração dos dados

Esta seção dedica-se a exploração dos dados, extração e transformação para enriquecimento da base de entrada. Alguns desses campos podem ser trabalhados e extraído a fim de obter dados extras para posterior análises, campos como:

(a) Data → Pode contribuir para a construção de dados para análise de distribuição. Extrair dados como dia, mês, ano e horário do crime para análises com dados contínuos;

(b) Descrição e Resolução → Torna-se interessante extrair palavras chaves de campos descritivos para aplição de técnicas de reconhecimento de padrões. Para isso, dado um texto trabalha-se com a maior valorização das palavras contidas, logo torna-se necessário a remoção de palavras específicas como artigos definidos e indefinidos, pronomes pessoais, possessivos, demonstrativos, relativos e outros subgrupos de palavras tratadas comumente. Um próximo para, é garantir que, para aumentar o reconhecimento de padrões, torna-se interessantes trabalhar palavras na mais pura forma, ou seja, sua raiz. Assim são trabalhadas palavras iguais desconsiderando as conjugações verbais, as derivações entre masculino ou feminino, entre outros. Este tipo de tratamento de remoção em inglês denomina-se como *stopwords*.

Dados descritivos podem resultar em grandes ganhos, porém para que sejam utilizados, devem ser transformados nas entradas específicas para os algoritmos deste trabalho. Dado um conjunto de treino, capta-se todas as palavras de textos, trabalha-se a remoção da palavras conforme citado anteriormente e coloca-se os dados em formato de bolsão de palavras (em inglês isso é conhecido como *Bag-of-words model*[1]) conforme figura 2. Esse bolsão de palavras torna-se um grande vetor binário, onde cada posição representa uma palavra do bolsão.

Figura 2: Técnica *Bag-of-words* para visualização dos dados, onde o tamanho da palavra representa maior aparecimento durante os textos análisados.

Após a criação desse bolsão, para cada amostra de dados, trabalha-se toda a descrição, varrendo e sempre que existir uma palavra presente na descrição e presente no vetor binário, muda-se o *bit* de 0 para 1. Ao final do processamento, os campos descritivos estão representados em formato adequado para posterior uso.

Outro ponto a se levantar, para trabalhar com horas e minutos, ou seja, valores temporais, torna-se interessante o uso da transformação dos dados para o círculo trigonométrico. Por exemplo, um crime, que normalmente acontece entre 1:50 da madrugada e outro às 2:00 da madrugada do dia seguinte, em locais próximos, pode ter a transformação para os eixos de seno e cosseno a fim de obter algumas informações extras os quais podem relacioná-los. Assim dados com localização diferente, podem ter influências temporais importantes e talvez categorizar crimes associados.

Outros meio de tratamento de dados é a Normalização. Todos os dados que são contínuos, podem ser transformados. Assim, dentro dos atributos, todas as amostras de um determinado atributo passam a variar dentro de um intervalo de escopo menor, por exemplo, através da utilização de mínimo e máximo e garantir a variação entre 0 e 1 (dependendo até trabalhar com o intervalo -1 a 1). Esta etapa é interessante (e algumas vezes necessárias) para permitir o uso de campos como Latitude e Longitude com técnicas como Regressão Logística e Rede Neural para convergir corretamente, oferecendo assim, aos atributos a mesma influência sobre o resultado final.

Feito os devidos ajustes na amostragem de dados, o próximo passo é o balanceamento desigualdade da base (conforme figura 1) para os dados de treino, a fim de compor uma grande quantidade das categorias da base toda e garantir a predição, com o menor erro possível. No entanto, possivelmente ocorrerá uma precisão pior para categorias menos abastecidas de amostras de treino, ou ainda inexistentes.

IV. Os Algoritmos de Classificação

Finalmente, a quantidade de dados a serem estudados torna-se adequados após os tratamento e balanceamentos. Todavia ainda quando comparado com o tempo computacional e capacidade de dispositivos de retenção, algumas das análises efetuadas sobre o conjunto de dados de treino é pequena com relação a base como um todo.

O objetivo das técnicas a serem trabalhadas é obter, dado uma entrada criminal operar com: (a) descrição; (b) a combinação entre localização (uso de Latitude e Longitude, talvez o uso de endereço), dia da semana (segunda-feira, terça-feira...), Distrito Policial; (c) trabalhar com horário, dia da semana e e Distrito Policial; ou outras combinações entre os atributos as quais podem ser exploradas. Assim, dada uma entrada, classificar, por predição, o tipo de crime realizado.

Para o escopo desse projeto, foram escolhidos os itens (a) e (b) e as técnicas de Regressão Logística (subseção IV-A) e Rede Neural em Multi-Camadas (*MLP*) (subseção IV-B).

A. Regressão Logística

Para começar com um algoritmo mais simples que a Rede Neural, nesta etapa trabalha-se com a Regressão Logística por ser mais rápida para execução, convergindo para um resultado adequado (ou ótimo) em menos tempo de execução.

[1]Técnica de *Bag-words* para visualização dos dados https://en.wikipedia.org/wiki/Bag-of-words_model_in_computer_vision

A figura 3 (exemplificando somente com 3 classes) demonstra como é o uso da técnica. Necessita-se verificação de uma classe contra todas as outras (em inglês *One-vs-All*), onde torna-se binário o problema para depois os cálculos e cortes no espaço amostral.

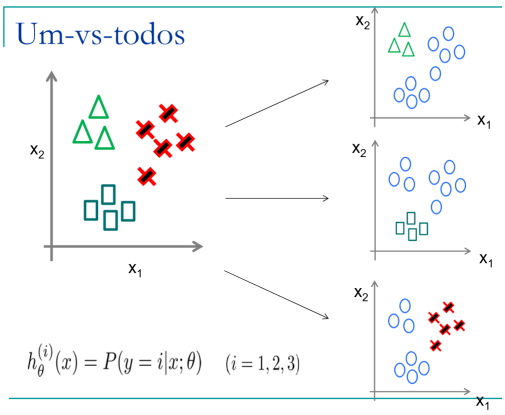

Figura 3: Classificação de Um-vs-Todos(slide da retirado da apresentação da Universidade de Algarve (Portugal) sobre *Machine Learning*)

As equações 1 e 2 representa, respectivamente, as funções *sigmoid* e de custo aplicado para esse processo, esta última estimativa da probabilidade de uma entrada X ser de uma categoria de crime específica. Para o parâmetro θ, inicialmente foram utilizados valores aleatórios, mas que ao decorrer das iterações recebeu forma durante o cálculo da função de gradiente descente (equação 3).

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

$$h(x) = \frac{1}{1 + e^{\theta^T x}} \tag{2}$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2 \tag{3}$$

Para o entendimento de como seria elaborado o montagem desta estrutura foram consultados os livros Bishop (capítulo 4) e Wasserman (capítulo 14), Seltman (capítulo 9) e notas das aulas.

B. MLP

Por outro lado, no caso da *MLP*, torna-se necessário trabalhar com um escopo um pouco diferente. Após a binarização dos vetores, tanto para dados descritivos quanto para a combinação (Dia da Semana, Distrito Policial, Latitude, Longitude). Também é feito uma vetor binário no qual cada posição representa uma Categoria do crime, atribuindo 1 somente uma vez no vetor para identificação do crime e 0 para todas as outras posições.

A figura 4 representa como entrada a combinação (Dia da Semana, Distrito Policial, Latitude, Longitude) e após a transformação em um vetor binário. Cada neurônio de entrada $I_x, x = [1, 18]$ são os neurônios de entrada. Uma aplicação de uma camada oculta ($H_x, x = [1, 10]$) composta por 10 neurônios. E na camada de saída, 38 categorias ($O_x, x = [1, 38]$) que foram detectadas durante o treino. Lembrando, que mesmo efetuado o balanceamento dos dados, mantendo a aleatoriedade dos dados também para a base de treino, perde-se 1 categoria de crime.

Tabela II: Comparação dos resultados obtidos para a classificação

	Regressão Logística	*MLP*
Descrição	92%	65%
Combinação[2]	12%	23%

Para o entendimento de como seria feita o montagem desta estrutura foram consultados os livros Bishop (capítulo 5), Carvalho, A. (capítulo 7) e notas das aulas.

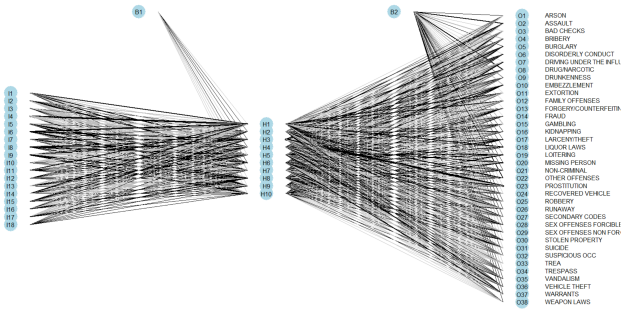

Figura 4: Rede Neural para trabalhar na combinação (Dia da Semana, Distrito Policial, Latitude, Longitude).

V. Resultados

Os resultados obtidos são demonstrados na tabela II. Na qual representa a porcentagem de acerto dado uma entrada X_{ij} onde i representa a amostra dos dados de validação ou teste e j a posição no vetor da amostra (sendo esse dado 0 ou 1).

Os resultados na tabela estimam valores condizentes ao escopo da execução de 30.000 amostras para treino (mas por questão de tempo e hardware, a MLP necessitou uma amostragem menor para processamento), 20.000 amostras de validação e 40.000 para teste. A amostragem inicialmente sugerida, de 700.000 exemplares para treino, e 178.000 para teste, não foi possível ser feita devido os recursos computacionais.

O tempo computacional usado para a execução dos algoritmos, trabalhando com 1000 iterações, MLP (6 horas, lembrando com um escopo menor de dados) foi extraordinariamente maior que da Regressão Logística (40 minutos). Porém para uso de dados formados pela combinação dos atributos, a *MLP* obteve um valor melhor (23%) que a Regressão Logística (12%).

Foram feitos alguns testes para a utilização da hora e minutos (10), combinado com dia da semana (9) e também

combinados junto os atributos de Latitude e Longitude. Mas não foi obtido valores acima entre 5% e 10%. Independente de trabalhar com uma amostragem maior dos dados, ou no caso da MLP, incluir mais camadas ocultas de neurônios (no caso de 8, 10 e 6 neurônios). Como os resultados não foram promissores, o modelo este modelo foi abandonado.

No geral, pode-se concluir (através da comparação da base de treino, e posterior testes e validações) que o pré-processamento, transformação ajudou para otimização e exploração das técnicas usadas e classificação dos próximos dados de entrada, diminuindo o erro médio e aumento a porcentagem de probabilidade da classificação de saída. Trabalhar com o aumento do número de iterações ou de uma amostragem não afetou de forma substancial os resultados (e algumas vezes variou na margem de 1.5% para erro da classificação, ou para menos ou para mais). Para melhorar ainda mais o resultado, seria necessário para tanto adicionar mais atributos (através da combinação dos já existentes ou obtê-los a partir de outras fontes) ou aplicar modelos mais sofisticados, com funções de custo mais relevantes, combinação de modelos e os modelos não lineares.

Resumindo, não há nenhuma única para melhor solução, no entanto, dentre as propostas neste trabalho, seleciona-se as mais viáveis em resultados promissores, e continua-se lapidando os resultados trabalhando as técnicas e/ou explorar mais formas de extrair conteúdos dos atributos disponibilizados .

Referências

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[2] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2010.

[3] H. J. Seltman, "Experimental Design and Analysis," 2010. [Online]. Available: http://www.stat.cmu.edu/\~{}hseltman/309/Book/

[4] Carvalho, A., *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. LTC. [Online]. Available: https://books.google.com.br/books?id=4DwelAEACAAJ

Apêndice

Abaixo seguem figuras 5, 6, 7, 8, 9, 11 comentadas anteriormente e separadas por alguns critérios conforme a classificação dos crimes.

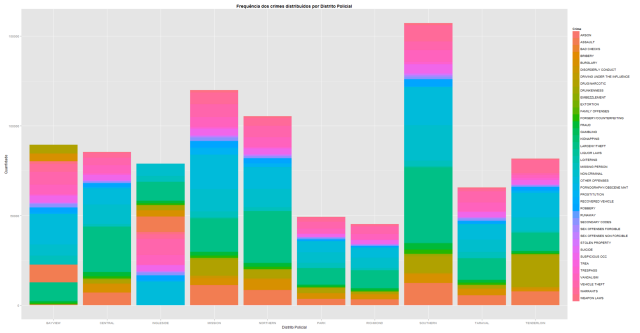

Figura 5: Distribuição dos crimes pelos distritos policiais.

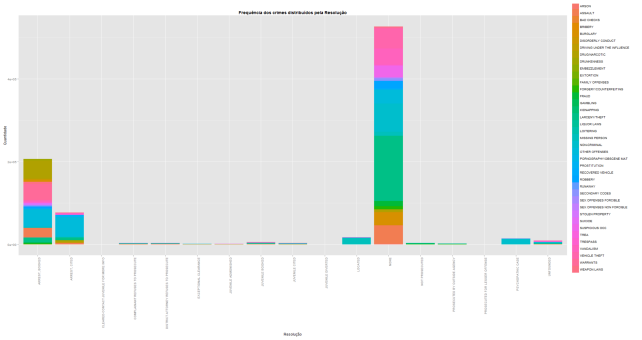

Figura 6: Distribuição dos crimes conforme a resolução.

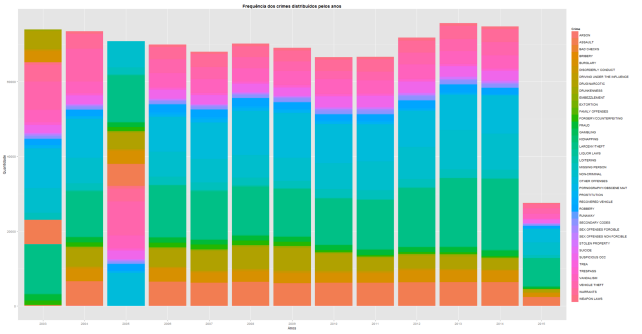

Figura 7: Distribuição dos crimes baseado pelos anos.

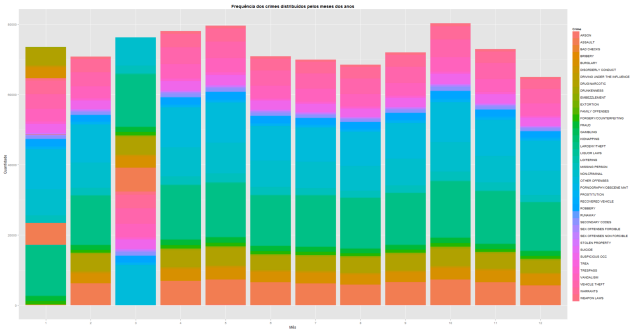

Figura 8: Distribuição dos crimes baseado no mês.

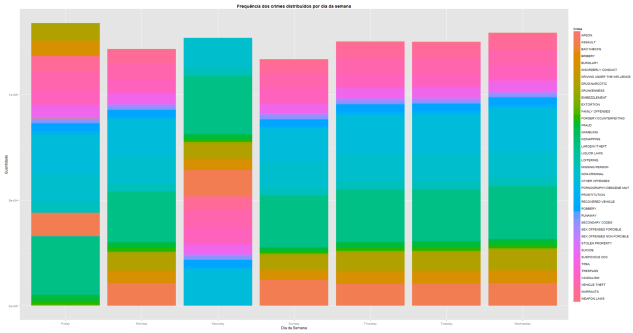

Figura 9: Distribuição dos crimes baseado no dia da semana.

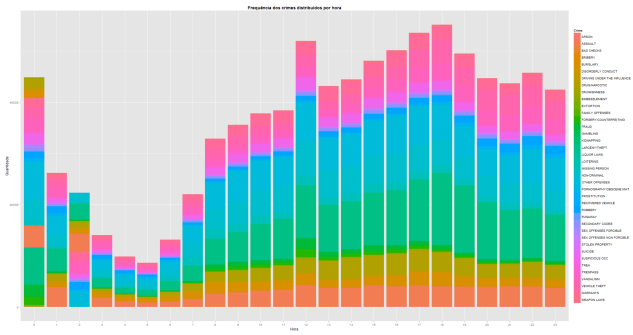

Figura 10: Distribuição dos crimes baseado no horário.

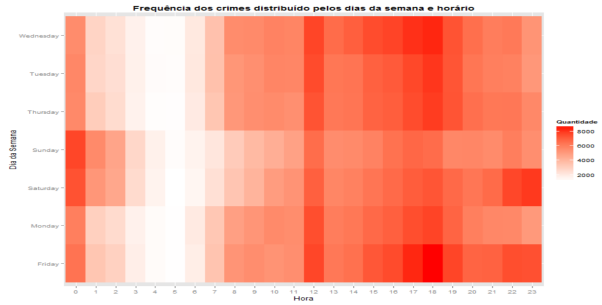

Figura 11: *Heatmap* da distribuição dos dados pela semana e hora.