Machine Learning Engineer Nanodegree

Trabalho de Conclusão de Curso

Análise de Rota de Veículos

Jessé Berto de Andrade

23 de Outubro de 2017

I. Definição

Visão Geral do Projeto

Análise de rotas de veículos que transitam em uma cidade através de dados de passagens de veículos por pontos de monitoramento. A partir de um conjunto de dados não rotulados, e com o auxílio de um segundo conjunto de dados rotulados, tentar-se-á rotular os dados não rotulados.

Problema

Dadas as passagens de veículos por pontos de monitoramento na cidade, a proposta deste trabalho é identificar padrões de rotas dos veículos, considerando a quantidade de vezes em que cada veículo passou em cada local. A partir dos padrões definidos, um segundo conjunto de dados será utilizado. Este segundo conjunto contém veículos com restrições de envolvimento em crimes, como roubo/furto e veículos clonados. A proposta é verificar se em algum dos padrões encontrados há um predomínio de veículos com restrição, e a partir de então, encontrar veículos que, mesmo sem restrição, se comportam de forma parecida com os veículos rotulados com restrição.

Métricas

Para definição de qual é a rota de um veículo, será utilizada a quantidade de vezes em que este veículo passou por cada ponto de monitoramento, não será levado em consideração a ordem pela qual o veículo transita entre os pontos. Desta forma, as características de cada veículo terão a dimensão do total de pontos de monitoramento, sendo o valor de cada atributo a contagem de passagem do veículo por aquele ponto.

Para avaliação do projeto, será verificado se o sistema encontra um padrão de rotas no qual há uma maior concentração de veículos alvo (veículos com restrição). Para isso, a métrica a ser utilizada será o percentual de veículos alvo em

uma determinada rota, ou seja, será calculada a razão entre a frequência de veículos alvo e a frequência de veículos em geral para determinada rota.

$$r = \frac{f_{alvo}}{f_{geral}}$$

Ao encontrar um padrão que maximize esta razão, o sistema estará identificando um grupo no qual o comportamento mais se assemelha com o comportamento dos veículos alvo.

II. Análise

Exploração dos Dados[1]

O conjunto principal a ser utilizado possui as características:

- loc_placa – placa do veículo
- loc_latitude – latitude do ponto de monitoramento
- loc_longitude – longitude do ponto de monitoramento
- count – contagem da quantidade de vezes que determinado veículo passou pelo ponto

Tabela 01 – Amostra das 6 primeiras linhas do conjunto de dados

	loc_placa	loc_latitude	loc_longitude	count
2	NLQ2918	1010	9237	1
3	NYU1477	4944	9278	1
4	OTS5855	1412	9134	1
5	APS9101	5825	9794	1
6	HRN6599	9222	5020	2

O conjunto de dados dos veículos em geral possui no total 3.196.578 passagens de 618.757 veículos em 135 pontos de monitoramento. Esses dados são referentes aos dias 14 e 15 de setembro de 2017, quinta e sexta-feira.

O conjunto de dados dos veículos com restrição possui as características:

- placa – placa do veículo
- loc_latitude – latitude do ponto de monitoramento
- loc_longitude – longitude do ponto de monitoramento

[1] **Política de privacidade e confidencialidade das informações**: para que não seja infligida a privacidade dos donos dos veículos e a confidencialidade dos dados, tanto as placas quanto os dados dos pontos de controle foram cifrados para que se tornassem ininteligíveis.

- count – contagem da quantidade de vezes que determinado veículo passou pelo ponto

Tabela 02 – Amostra do conjunto de dados de veículos com restrição

	placa	loc_latitude	loc_longitude	count
1	NQO7323	312	1961	3
2	PSN6821	802	2054	1
3	PDM5733	2833	1278	2
4	OHT0583	7262	8878	1
5	ORF0008	2088	6750	1
6	OLS2179	4750	5075	4

Este conjunto de dados de veículos com restrição possui no total 4.234 passagens de 2.377 veículos em 198 pontos de monitoramento. Esses dados são referentes aos dias 21 e 22 de setembro, quinta e sexta-feira da semana subsequente à semana do primeiro conjunto.

Um terceiro conjunto de dados, também com veículos com restrição, porém dos dias 12 e 13 de outubro de 2017, uma quinta e sexta-feira do mês subsequente, será utilizado para validação dos padrões encontrados

Tabela 03 – Amostra do conjunto de dados de veículos com restrição que será utilizado para validação

	placa	loc_latitude	loc_longitude	count
1	FHN8871	9666	833	1
2	HBY5919	6072	9488	1
3	OLE8486	132	5022	1
4	OHT0583	7262	8878	1
5	PFM6271	8525	2161	1

Este conjunto de dados de veículos com restrição possui no total 4.394 passagens de 2.458 veículos em 196 pontos de monitoramento.

Visualização exploratória

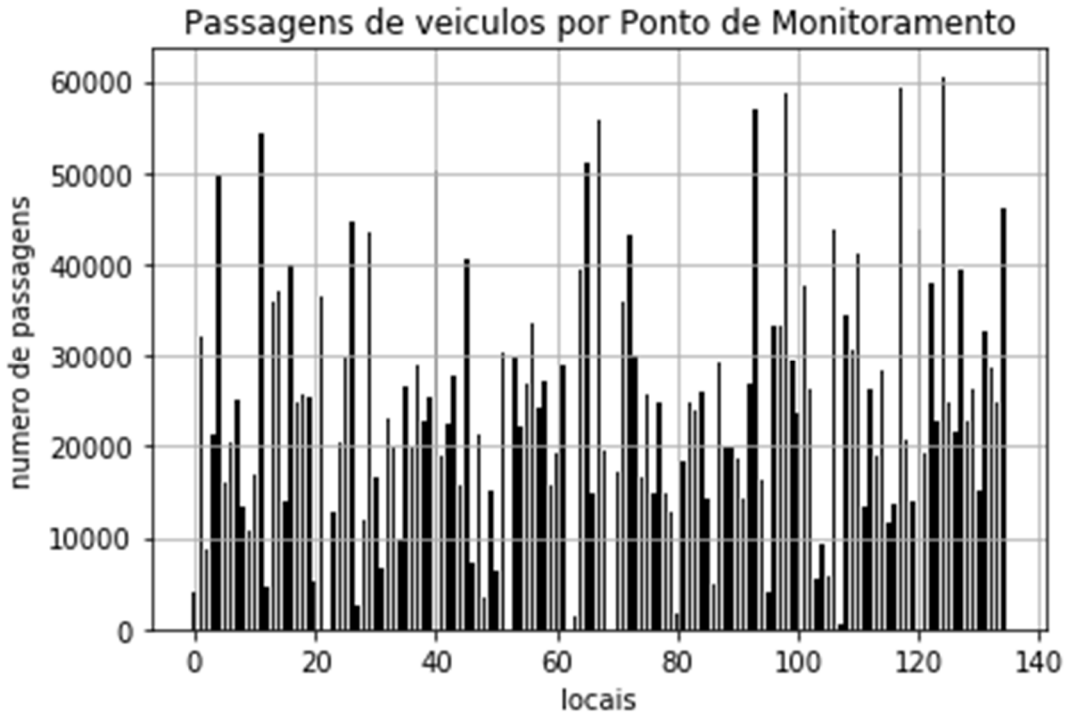

Figura 1 – Distribuição de frequência dos veículos nos pontos de monitoramento

A Figura 1 mostra a distribuição de frequência dos veículos nos pontos de monitoramento. Para este projeto, estes pontos de monitoramento serão utilizados como dimensões, e cada veículo possuirá como características a quantidade de passagens por cada ponto.

Tabela 04 – Estatísticas de oito locais de monitoramento

Local	5613 8278	4738 6727	4332 2860	8053 4258	3444 967	1412 9134	5880 8554	2833 1278
count	618757	618757	618757	618757	618757	618757	618757	618757
mean	0.0063	0.0520	0.0141	0.0342	0.0802	0.0257	0.0331	0.0406
std	0.0956	0.4741	0.1485	0.2859	0.8146	0.1849	0.2314	0.2447
min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
25%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
50%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
75%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
max	11	295	21	39	578	22	77	12

Pela impossibilidade visual de apresentar as estatísticas dos 135 pontos, a Tabela 01 mostra, a título de exemplificação, estatísticas de oito pontos de monitoramento. É interessante notar que até o terceiro quartil tem valor zero, por mais que possa parecer errado, a informação está correta, pois a grande maioria

dos veículos não passa por cada local específico, ou seja, cada local possui zero passagens para grande maioria dos veículos. Portanto, dos 618.757 veículos, o número de passagens de veículos em cada ponto é tão baixo que se aproxima de zero e pelo fator de arredondamento é apresentado o valor zero. Isso dá uma idéia de o que se espera dos dados, cada veículo possui valor zero para a maioria dos pontos de monitoramento com alguns valores diferentes de zero em poucos pontos de monitoramento, como exemplificado na tabela 02.

Tabela 05 – Passagens de veículos por locais

Local	5613 8278	4738 6727	4332 2860	...	4853 8470	9430 4736	5303 7658	7723 5667	9666 833
HXX6311	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0
OES0945	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
HWC4722	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
HXX6312	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	1.0
NMQ8671	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
NTM2610	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
ODI1168	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
OES0947	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
NTQ9976	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
ODI1161	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0

Algorítimos e Técnicas

PCA – Principal Component Analisys

A Análise de Componentes Principais (ACP) ou Principal Component Analysis (PCA) é um procedimento matemático que utiliza uma transformação ortogonal (ortogonalização de vetores) para converter um conjunto de observações de variáveis possivelmente correlacionadas num conjunto de valores de variáveis linearmente não correlacionadas chamadas de componentes principais. O número de componentes principais é menor ou igual ao número de variáveis originais. Esta transformação é definida de forma que o primeiro componente principal tem a maior variância possível (ou seja, é responsável pelo máximo de variabilidade nos dados), e cada componente seguinte, por sua vez, tem a máxima variância sob a restrição de ser ortogonal a (i.e., não correlacionado com) os componentes anteriores. Os componentes principais são garantidamente independentes apenas se os dados forem normalmente distribuídos (conjuntamente). O PCA é sensível à escala relativa das variáveis originais. Dependendo da área de aplicação, o PCA é também conhecido como transformada de Karhunen-Loève (KLT) discreta, transformada de Hotelling ou decomposição ortogonal própria (POD). [1]

O PCA será utilizado para fornecer uma dimensão reduzida dos dados, pois suas componentes possuem a perspectiva mais informativa dos dados. Para um

problema que se inicia com 135 dimensões, a utilização de uma técnica de redução de dimensionalidade como o PCA é essencial.
Para implementação do PCA será utilizada a biblioteca do Scikit-learn sklearn.decomposition.PCA. Será a princípio calculado o PCA com 50 componentes, e após a análise da variança explicada acumulada até a quinquagésima componente, será verificado se o PCA pode ser recalculado com menos componentes ou se necessitará de um número maior de componentes.

K-means

O algoritmo KMeans agrupa dados tentando separar amostras em n grupos de variância igual, minimizando um critério conhecido como inércia ou soma de quadrados dentro do cluster. Este algoritmo exige que o número de clusters seja especificado. Escala bem para um grande número de amostras e foi usado em uma ampla gama de áreas de aplicação em muitos campos diferentes.[2]

O algoritmo k-means divide um conjunto de N amostras X em K clusters disjuntos C, cada um descrito pela média μ_j das amostras no cluster. Os meios são comumente chamados de "centroides" do cluster; note que eles não são, em geral, pontos de X, embora vivam no mesmo espaço. O algoritmo de K-means visa escolher centroides que minimizem a inércia ou o critério de soma dos quadrados no cluster:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (\| x_j - \mu_i \|^2)$$

O k-means será o método de clusterização utilizado, pois é de simples implementação, fácil interpretação do resultado, rápido e eficiente em termos de custo computacional. Para implementação do k-means será utilizada a biblioteca do Scikit-learn sklean.cluster.KMeans.

O grande desafio do K-means é determinar o número de clusters a ser utilizado, pois a técnica em si não determina qual é o melhor ajuste, então deve ser dado como entrada no algorítimo de clusterização a quantidade de clusters desejada. Inicialmente os dados serão segmentados em 30 clusters, posteriormente esse número de clusters será otimizado.

Métrica para avaliação

Lembrando o objetivo de encontrar um seguimento dos veículos que circulam na cidade no qual possui um comportamento mais parecido com o dos veículos alvo, passamos a analisar os seguintes fatores:

- Seria desejável encontrar um seguimento dos veículos que transitam na cidade, no qual todos os veículos alvo se encaixam. Definimos então a taxa de veículos alvo que se encaixam em cada seguimento como:

$$t_{va} = \frac{n_{as}}{NA}$$, onde:

t_{va} é a taxa de veículos alvo no seguimento,
n_{as} é o número de veículos alvo no seguimento, e
NA é o número total de veículos alvo.

Porém, se considerarmos a existência de um único seguimento para todos os veículos, essa taxa será máxima, pois todos os veículos alvo também estarão neste seguimento. Tal situação inviabiliza a utilização de t_{va} como o único fator de pontuação.

Definiremos então f_s como sendo a frequência de veículos gerais em cada seguimento e a taxa de ocupação de veículos alvo no seguimento como:

$$t_{oc} = \frac{n_{as}}{f_s}$$, onde:

t_{oc} é a taxa de ocupação de veículos alvo no seguimento,
n_{as} é o número de veículos alvo no seguimento, e
f_s é a frequência de veículos gerais no seguimento.

Quanto maior a taxa de ocupação de veículos alvo no seguimento, mais podemos considerar que os veículos deste seguimento se comportam como os veículos alvo, pois uma taxa igual a 1 significaria total identidade entre os veículos do seguimento e os veículos alvo. Porém, se for utilizada essa taxa como o único critério de pontuação, correremos o risco de um seguimento com pouca representatividade numérica possuir a maior pontuação. Por exemplo, um seguimento que contenha apenas um veículo, e um veículo alvo fosse ajustado neste seguimento teria t_{oc} igual a 1, que seria a pontuação máxima, e ainda assim esse seguimento não teria representatividade no universo de amostras.

Na busca de um único valor que pontue o desempenho do sistema, a solução aqui proposta será unificar essas duas taxas em uma única pontuação, para isso definiremos a pontuação, s, a ser utilizada nesse trabalho como:

$$s = t_{va} \times t_{oc} = \frac{n_{as}}{NA} \times \frac{n_{as}}{f_s}$$

Considerando que NA (número de veículos alvo) é uma constante, para fins de comparação entre pontuações pode ser desconsiderado, o que ajuda também na

redução do custo computacional de score, portanto, podemos simplificar a expressão de s para:

$$s = n_{as} \times \frac{n_{as}}{f_s} \text{ ou } s = \frac{n_{as}^{\,2}}{f_s}$$

Será calculada a pontuação s para cada seguimento de veículos, definiremos agora a pontuação *score* como sendo o máximo s para uma determinada segmentação.

$$score = \max(s)$$

ou

$$score = \max\left(\frac{n_{as}^{\,2}}{f_s}\right)$$

A partir de agora, poderemos otimizar o sistema, buscado uma segmentação na qual se maximize a pontuação score, portanto, definiremos o melhor número de segmentos, n_{best}, como:

$$n_{best} = \arg\max\left(\frac{n_{as}^{\,2}}{f_s}\right)$$

III. Metodologia

Pré-processamento de dados

As tabelas de dados carregadas no projeto estão arquivos .csv, estes arquivos são carregados em data frames, como exemplificado anteriormente nas Tabelas 01, 02 e 03.

Os data frames possuem em suas linhas a contagem das passagens de cada veículo por cada latitude e longitude;

Para cada data frame são seguidos os seguintes passos:

- O data frame é reindexado sendo coluna de placas o novo index;
- É criada uma nova coluna no data frame denominada “local” que concatena a latitude e a longitude, as colunas loc_latitude e loc_longitude são removidas;

- É utilizada uma função denominada create_df_placaporpontos, que recebe o data frame e retorna um novo data frame;
- O novo data frame é indexado pelas placas dos veículos e possui uma coluna para cada "local", assim cada placa possui como características suas passagens em cada ponto. Este novo data frame, exemplificado na Tabela 05, é que será utilizado para o desenvolvimento do projeto.

Não foram utilizadas técnicas para excluir *outliers*, pois esse projeto trata justamente de encontrar padrões que divergem do comportamento comum, portanto os *outliers* podem estar em os alvos do projeto.

Implementação

PCA

O data frame que contém os dados dos veículos em geral é denominado "data". Como "data" possui 135 características, será utilizado o PCA para descobrir qual dimensão dos dados melhor maximizam a variância dos atributos envolvidos. Além de descobrir essas dimensões, o PCA também irá reportar a razão da variância explicada de cada dimensão, ou seja, quanto da variância dentro dos dados é explicada pela dimensão sozinha.

Inicialmente é aplicado aos dados um PCA com 50 componentes principais, após a análise da variância explicada acumulada, vê se que até a 30ª dimensão obtém-se 80% da variância explicada. Este valor é considerado aceitável e os dados então são reduzidos para 30 dimensões através de um PCA com 30 componentes principais. A Tabela 06 exemplifica os dados reduzidos.

Tabela 06 – Exemplo de amostras reduzidas pelo PCA

	Dimension 0	Dimension 1	Dimension 2	...	Dimension 28	Dimension 29
0	-0.337746	0.134597	-0.092640	...	0.089238	0.030261
1	0.046368	0.786888	-0.231215	...	0.025529	-0.028116
2	-0.322102	0.156673	-0.120398	...	-0.034227	-0.050598
3	-0.338625	0.148377	-0.077651	...	0.079058	-0.002922
4	0.028331	0.476897	-0.257953	...	-0.187332	0.049972
5	-0.334127	0.142917	-0.038640	...	0.005598	-0.041838
6	-0.117623	0.377637	-0.194715	...	-0.009458	-0.153634
7	-0.290270	0.184811	-0.132350	...	0.048488	0.034825
8	-0.337188	0.125138	-0.093539	...	0.027748	-0.027319
9	-0.339927	0.158331	-0.108718	...	-0.033818	-0.007216

Clustering

Para segmentar os dados, é utilizado o KMeans, importado de sklearn.cluster. Foi criada uma função denominada make_clusterer() que recebe os dados reduzidos da tabela geral de veículos e o número de clusters. Esta função retorna um objeto clusterer ajustado a estes dados.

Inicialmente é criado um clusterer para 30 segmentos. Os dados reduzidos da tabela geral são preditos neste clusterer e é calculado um array com a distribuição de frequências de cada segmento.

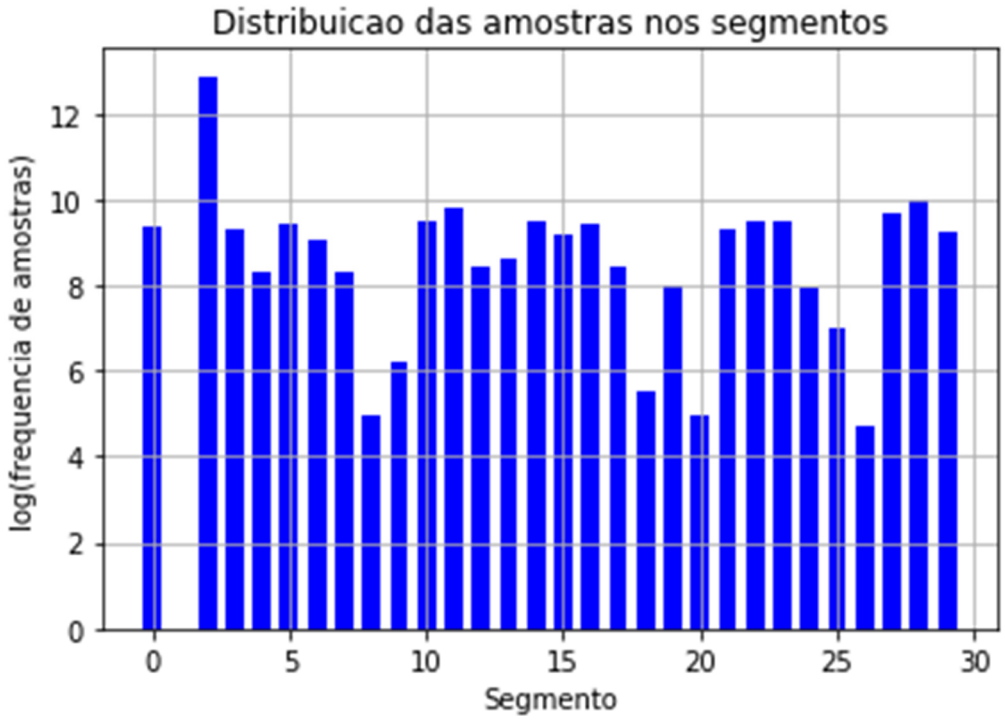

Figura 2 – Distribuição de frequências das amostras nos segmentos

Agora que os padrões de rotas de veículos na cidade foram definidos, os dados de veículos com restrição são utilizados para buscar padrões dos veículos em geral que se assemelhem com padrões de veículos com restrição. Portanto, os veículos com restrição serão encaixados nessa clusterização.

Primeiramente os dados dos veículos com restrição são ajustados as dimensões reduzidas através do PCA e depois são preditos pelo clusterer já treinado. É então calculada a frequência de veículos com restrição nos segmentos.

De posse tanto da frequência geral de veículos, quanto da frequência dos veículos com restrição, é feita uma comparação entre essas duas frequências como mostrado na Figura 3.

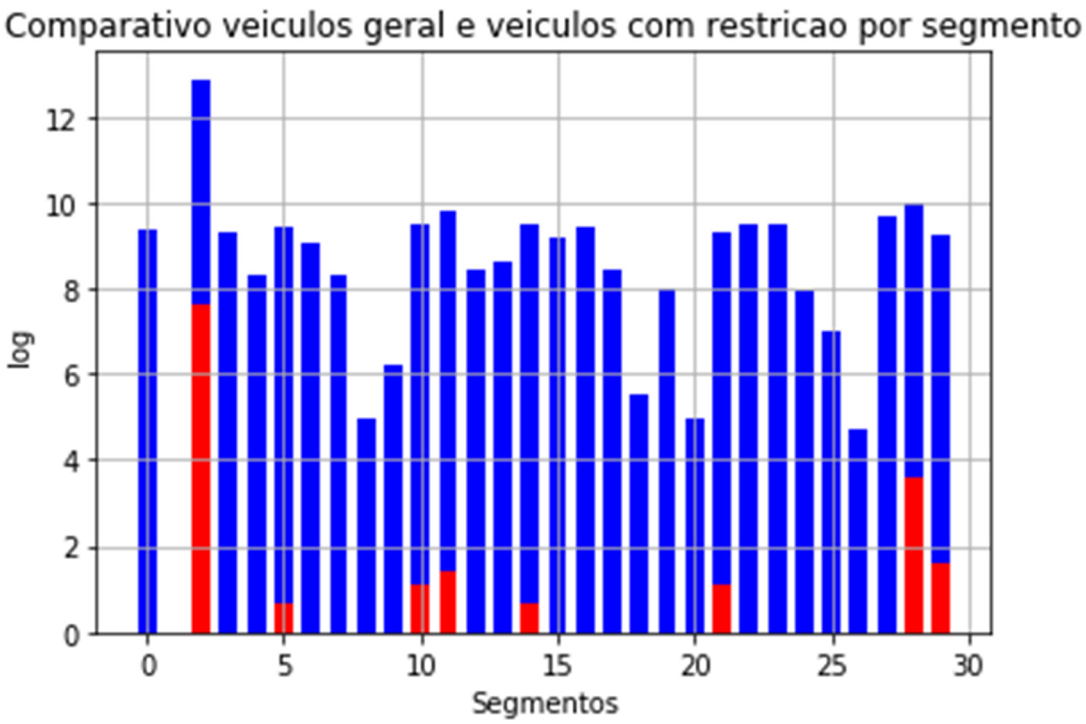

Figura 3 – Comparativo de veículos geral e veículos com restrição por segmento

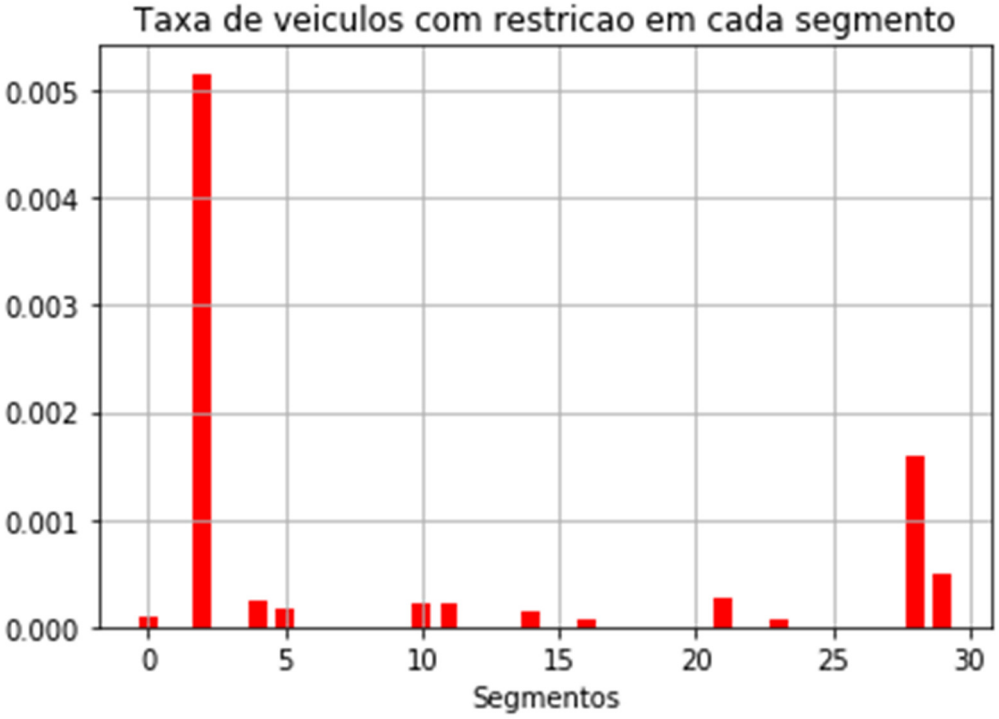

Figura 4 – Taxa de veículos com restrição por segmento

Otimização

Definidos a forma de segmentação e comparação entre os veículos, é necessário otimizar a segmentação de forma a encontrar qual o número de seguimentos no qual um padrão de rotas de veículos não rotulados mais se aproxime de um padrão de rotas de veículos rotulados com restrição.

Para encontrar o número de segmentos ótimo, é utilizada a equação de n_{best} utilizada anteriormente.

$$n_{best} = \arg\max\left(\frac{{n_{as}}^2}{f_s}\right)$$

Foi criada a função make_score() que recebe a frequência geral de veículos, a frequência de veículos com restrição, e retorna a pontuação para esse conjunto.

Foi criada também a função find_better_clustering() que, a partir de um número n de clusters inicial, procura um número melhor com base na pontuação score.

Ainda uma função test_range_of_clusters(), que recebe os dados gerais, os dados alvo e um array com os números de clusters a serem testado. Esta função retorna a pontuação para cada clusterização testada e um objeto com o melhor clusterer encontrado.

A busca pelo melhor número de clusteres é feita pela função test_range_clusters(). Esta função recebe a tabela principal reduzida, a tabela alvo reduzida, um vetor com as segmentações a serem testadas. Retorna o score para cada segmentação testada e o objeto clusterer da segmentação de maior pontuação. A figura 4 mostra um gráfico com o score para diferentes segmentações.

Figura 5 – Score alcançado para diferentes segmentações

O melhor ajuste encontrado foi com 90 clusteres, que atingiu um score de 12.62. Definido o melhor número de segmentos, os dados são ajustados para essa segmentação como mostrado pela Figura 5.

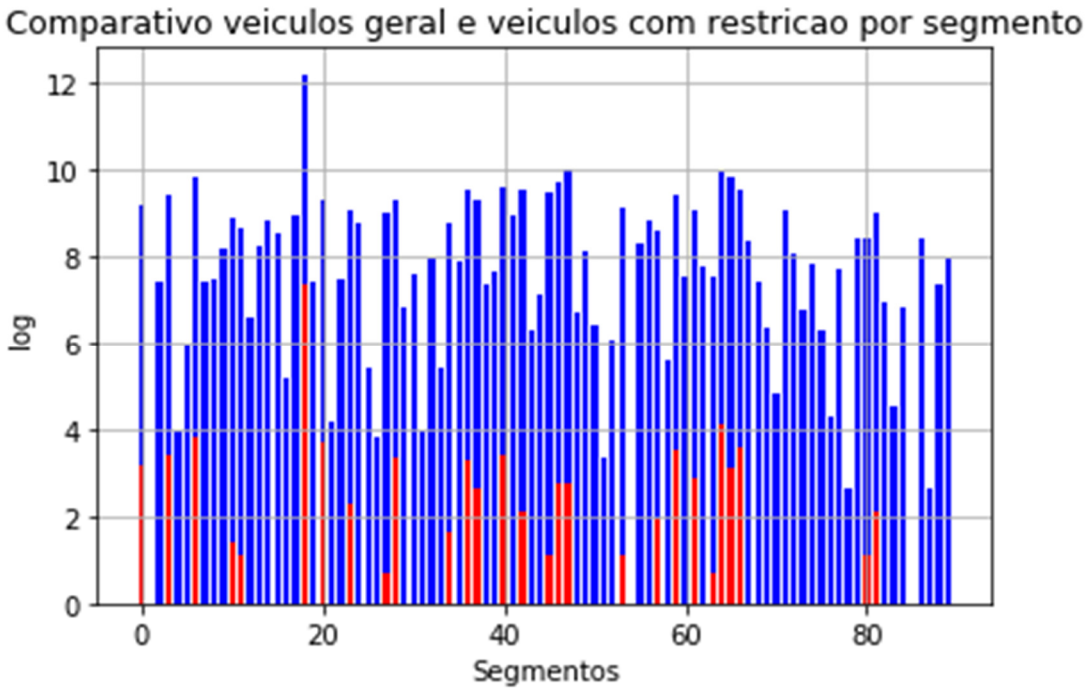

Figura 6 – Comparativo de veículos geral e veículos com restrição para a segmentação otimizada.

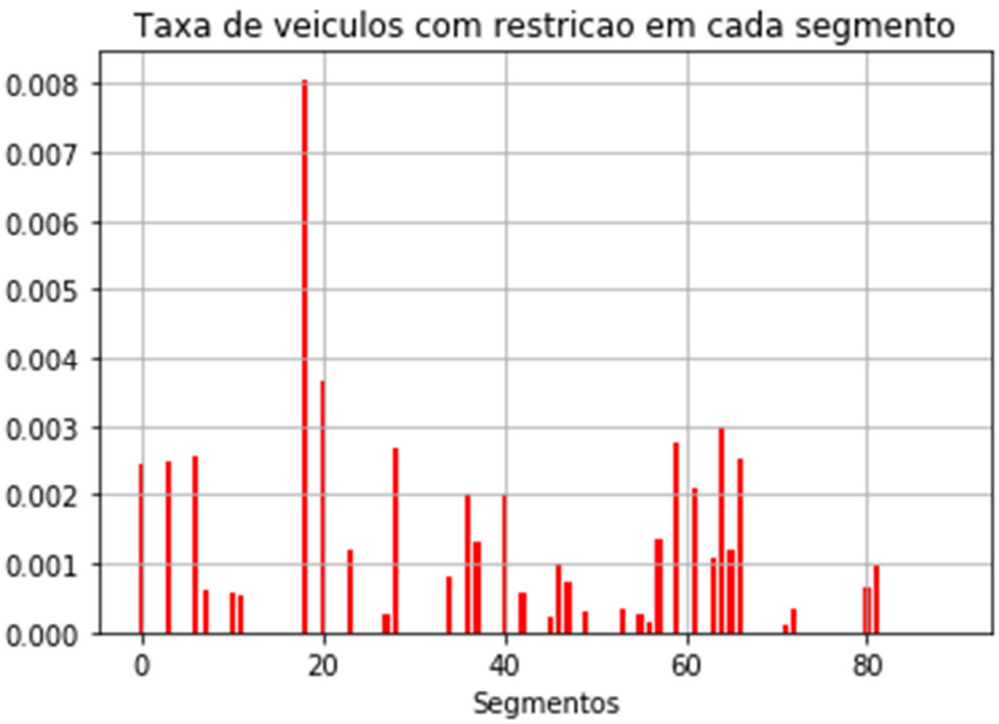

Figura 7 – Taxa de veículos com restrição para a segmentação otimizada

O Segmento 18 tem maior taxa de veículos com restrição, 0.0081.

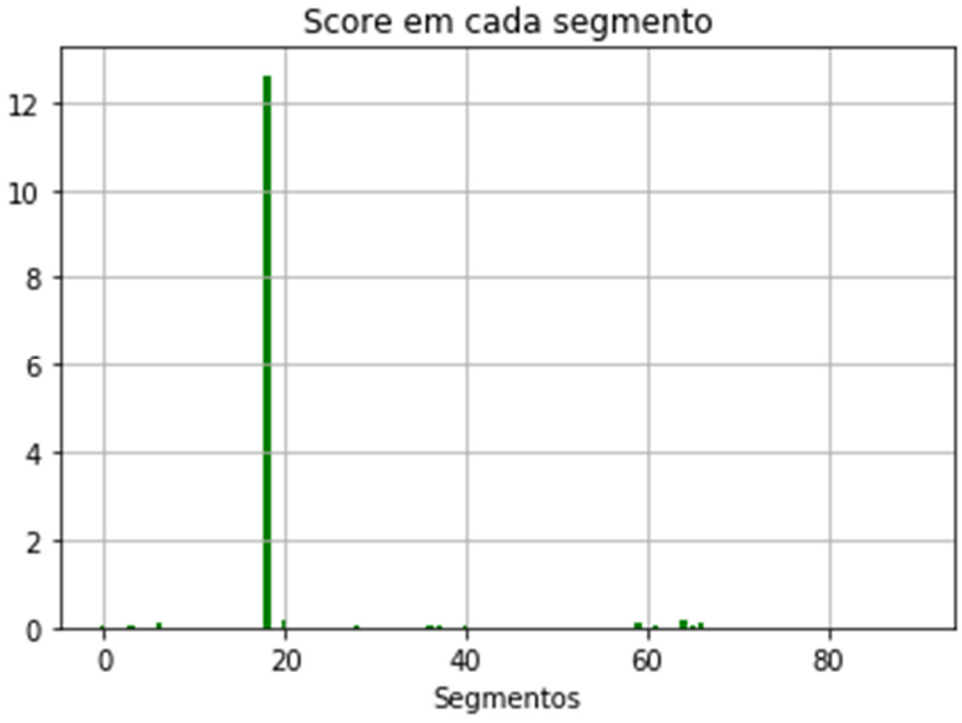

Figura 8 – Score para a segmentação otimizada.

O Segmento 18 possui maior score, 12.6172. Este segmento representa 75.63% dos veículos com restrição.

IV. Resultados

Avaliação e validação de modelos

Para avaliar e validar o modelo treinado é utilizado o conjunto de dados de veículos com restrição dos dias 12 e 13 de outubro. Os dados passam pela mesma redução de dimensionalidade PCA e são preditos pelo clusterizador otimizado. A Figura 9 apresenta o comparativo entre os veículos geral e novo conjunto de veículos com restrição. Em seguida são apresentadas a taxa de veículos com restrição em cada segmento, figura 10, e o score para cada segmento, figura 11.

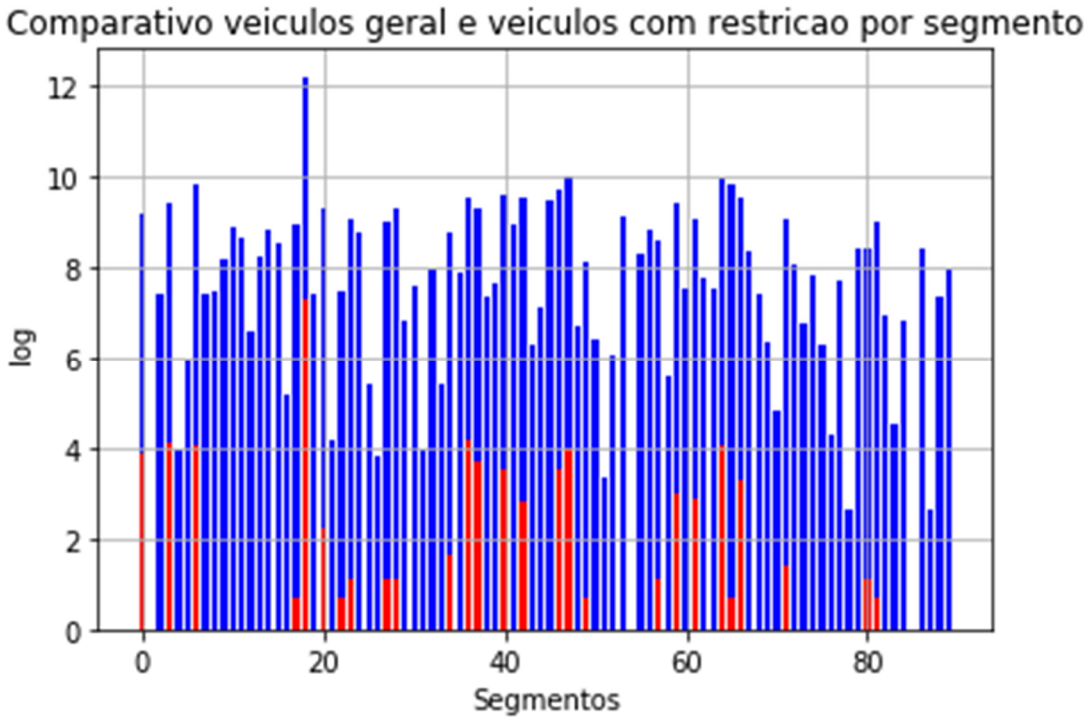

Figura 9 – Comparativo de veículos geral e veículos com restrição de validação

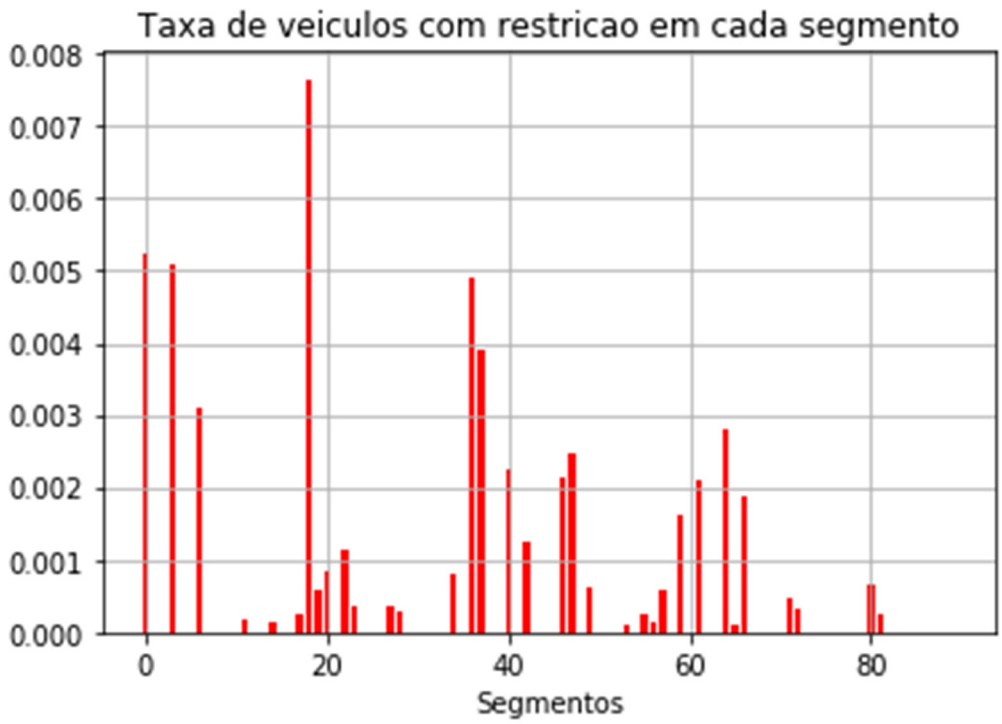

Figura 11 – Taxa de veículos com restrição para os dados de validação

O segmento 18 é o segmento com maior taxa de veículos com restrição, 0,0076.

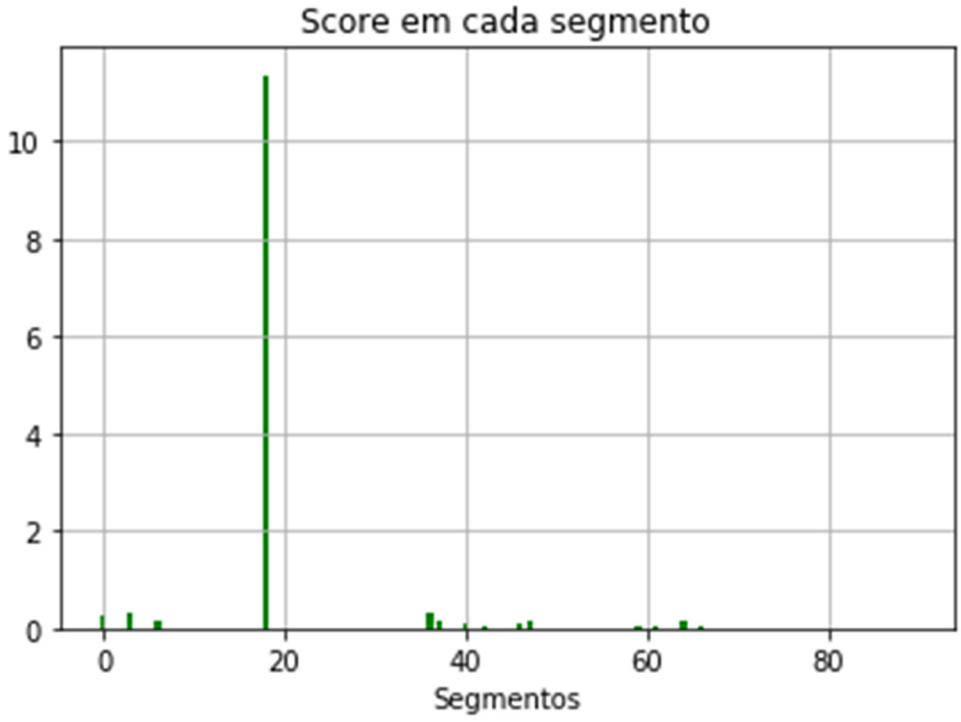

Figura 8 – Score para os dados de validação.

Justificativa

O segmento 18 alcançou um score de 11,3441. Este segmento possui 71,75% dos veículos com restrição. Analisando a segmentação do conjunto principal, o segmento 18 representa 31,33% do total.

O sistema foi capaz de encontrar um segmento que contém 31,33% do total de veículos, porém ao analisar os veículos com restrição, este segmento possui 71,75% do total de veículos com restrição.

Ainda há de considerar que o conjunto de veículos com restrição de validação possui 2.068 veículos, o conjunto principal de dados possui 618.757, portanto, os veículos com restrição representam 0,33% do total. No segmento destacado, a taxa de veículos com restrição foi de 0,76%, ou seja, ao selecionar um veículo deste segmento, a chance de encontrar um veículo com restrição é 2,3 vezes maior do que se fosse selecionado um veículo da população geral.

V. Conclusão

Reflecção

A proposta deste trabalho era encontrar veículos, que mesmo não rotulados, tivesse comportamento parecido com veículos rotulados como “com restrição”, pois tais veículos podem não ser roubados ou clonados, mas por ter um comportamento semelhante, podem estar cometendo outros ilícitos. Portanto, estes novos veículos não seriam rotulados como veículos “com restrição”, mas apenas como “veículos de interesse”.

Para solução do problema a população geral de veículos foi segmentada com técnicas de aprendizagem não supervisionada. Através dessa segmentação foi possível identificar grupos de veículos que se comportam de forma parecida. A partir dos dados segmentados, os veículos com rótulo de “com restrição” foram encaixados nos segmentos na tentativa de se encontrar segmentos nos quais os veículos com restrição tivessem uma maior concentração.

O sistema foi otimizado, a partir de um critério de pontuação que levava em conta a representatividade dos veículos alvo em cada segmento e a concentração destes veículos em cada segmento.

A estratégia seguida neste trabalho obteve sucesso ao encontrar um segmento da população geral que concentrava grande parte dos dados alvo. Porém, há de se fazer aqui uma ressalva, o objetivo desse trabalho era encontrar rotas nas quais havia grande incidência de veículos com restrição, porém ao analisar o comportamento dos veículos que foram classificados no segmento de maior pontuação, verificou-se que esses veículos possuem apenas uma passagem em um dos pontos, e essa foi provavelmente a característica em comum que os uniu. O

problema aqui apontado se deve a uma limitação do conjunto de dados dos veículos com restrição, no qual a grande maioria dos veículos possuem apenas uma passagem, e não da técnica empregada no trabalho.

Melhorias

O projeto aqui apresentado, apesar de não ter alcançado totalmente seu objetivo inicial, se mostra muito promissor. Primeiramente, deve-se buscar um conjunto de dados alvo que contenha mais passagens de veículos pelos pontos de monitoramento.

Os dados alvo não precisam se limitar a veículos com restrição, quaisquer que sejam as características a serem buscadas na população geral, a mesma técnica pode ser aplicada, como por exemplo veículos que vieram a sofrer acidentes, veículos com alta incidência de multas de trânsito, veículos que foram apreendidos com contrabando/descaminho, etc.

A base de dados extraída para este trabalho foi de dois dias consecutivos, porém há de se fazer também uma otimização para verificar qual o melhor período a ser analisado.

O sistema se limitou a um período de dois dias na semana, porém essa mesma análise pode ser feita para períodos ininterruptos que cubram toda a semana.

VI. Referências

[1] https://pt.wikipedia.org/wiki/An%C3%A1lise_de_componentes_principais#Algoritmos_iterativos
[2] http://scikit-learn.org/stable/modules/clustering.html#k-means