# **Machine Learning Engineer Nanodegree**

## **Análise de Rota de Veículos**

Jessé Berto de Andrade

15 de Outubro de 2017

## **I. Definição**

**Visão Geral do Projeto**

Análise de rotas de veículos que transitam em uma cidade através de um conjunto de dados de passagens de veículos por pontos de monitoramento. A base de dados contém a placa do veículo e a latitude e longitude do ponto de monitoramento. Esses dados se referem aos dias 14 e 15 de setembro de 2017.

### **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.



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

## **II. Análise**

*(approx. 2-4 pages)*

### **Exploração dos Dados**

O conjunto principal a ser utilizado possui as características:

1. loc\_placa – placa do veículo
2. loc\_latitude – latitude do ponto de monitoramento
3. loc\_longitude – longitude do ponto de monitoramento
4. count – contagem da quantidade de vezes que determinado veículo passou pelo ponto

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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 |

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

O conjunto de dados dos veículos em geral possui 618.757 veículos passando por 135 pontos.

Este conjunto de dados será carregado em um dataframe, data\_csv, e então será reindexado pela placa e uma nova coluna, que concatenará a latitude e longitude, será criada com o nome “local”.

Para preparar a tabela que será utilizada no projeto, foi criada uma função “create\_df\_placaporpontos(data, placa\_column, local\_column)” que ce

In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input files), or even an environment. The type of data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics about the input or environment). Any abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of outliers). Questions to ask yourself when writing this section:

* *If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?*
* *If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?*
* *If a dataset is not present for this problem, has discussion been made about the input space or input data for your problem?*
* *Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)*

### **Exploratory Visualization**

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

* *Have you visualized a relevant characteristic or feature about the dataset or input data?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

### **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.

**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 µj 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:



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

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

* *Are the algorithms you will use, including any default variables/parameters in the project clearly defined?*
* *Are the techniques to be used thoroughly discussed and justified?*
* *Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?*

### **Benchmark**

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

* *Has some result or value been provided that acts as a benchmark for measuring performance?*
* *Is it clear how this result or value was obtained (whether by data or by hypothesis)?*

## **III. Methodology**

*(approx. 3-5 pages)*

### **Data Preprocessing**

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

* *If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?*
* *Based on the Data Exploration section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?*
* *If no preprocessing is needed, has it been made clear why?*

### **Implementation**

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

* *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
* *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
* *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

### **Refinement**

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

* *Has an initial solution been found and clearly reported?*
* *Is the process of improvement clearly documented, such as what techniques were used?*
* *Are intermediate and final solutions clearly reported as the process is improved?*

## **IV. Results**

*(approx. 2-3 pages)*

### **Model Evaluation and Validation**

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

* *Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?*
* *Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?*
* *Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?*
* *Can results found from the model be trusted?*

### **Justification**

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

* *Are the final results found stronger than the benchmark result reported earlier?*
* *Have you thoroughly analyzed and discussed the final solution?*
* *Is the final solution significant enough to have solved the problem?*

## **V. Conclusion**

*(approx. 1-2 pages)*

### **Free-Form Visualization**

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

* *Have you visualized a relevant or important quality about the problem, dataset, input data, or results?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

### **Reflection**

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

* *Have you thoroughly summarized the entire process you used for this project?*
* *Were there any interesting aspects of the project?*
* *Were there any difficult aspects of the project?*
* *Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

### **Improvement**

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

* *Are there further improvements that could be made on the algorithms or techniques you used in this project?*
* *Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?*
* *If you used your final solution as the new benchmark, do you think an even better solution exists?*

Before submitting, ask yourself. . .

* Does the project report you’ve written follow a well-organized structure similar to that of the project template?
* Is each section (particularly Analysis and Methodology) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
* Would the intended audience of your project be able to understand your analysis, methods, and results?
* Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
* Are all the resources used for this project correctly cited and referenced?
* Is the code that implements your solution easily readable and properly commented?
* Does the code execute without error and produce results similar to those reported?

Referências

[1] [https://pt.wikipedia.org/wiki/An%C3%A1lise\_de\_componentes\_principais#Algoritmos\_iterativos](https://pt.wikipedia.org/wiki/Análise_de_componentes_principais#Algoritmos_iterativos)

[2] <http://scikit-learn.org/stable/modules/clustering.html#k-means>