Análise descritiva dos dados

Descrição dos dados

Os dados coletados podem ser vistos como um conjunto de itens, cada item com um ou mais atributos, que podemos também chamar variáveis. Por exemplo, um item pode ser um usuário do foursquare que vai ter os atributos nome, idade, número de checkins na última semana, etc.

Alguns conjuntos de dados têm mais de um tipo de item. Por exemplo, no foursquqre, um outro item pode ser um local, que terá como atributos latitude, longitude, número de checkins, etc. Geralmente se os dados têm múltiplos tipos de itens, temos também relacionamentos entre itens, como por exemplo, um usuário pode fazer checkin em um lugar. Nestes casos, os relacionamentos geralmente também têm atributos: por exemplo, o checkin foi feito em uma determinada data, então a data é um atributo do relacionamento entre usuários e local. Quando o conjunto de dados tem relacionamentos entre itens, o conjunto de dados é um grafo. Caso haja apenas um tipo de item, ele é uma tabela.

Nesta etapa do projeto, você deve preencher a seção de descrição dos seus dados no documento de seu projeto. Para isso: (i) identifique os itens em seus dados e os atributos de cada item; (ii) descreva cada atributo explicando o seu significado e identifique o tipo de variável associada a cada atributo (ex. escala de razão, nominal, etc.) e (iii) se houver, identifique os relacionamentos existentes entre os itens, e os atributos destes relacionamentos.

Lembre também de especificar, quando cabível, a que período de tempo se referem os dados (por exemplo, foram trending topics coletados de 1 de maio de 2011 a 23 de junho de 2011), a periodicidade em que eles foram coletados (a cada hora? a cada dia?) e quaisquer outras meta-informações sobre os seus dados que sejam importantes serem identificadas aqui.

Tratamento dos dados

1. É necessário limpar os dados de alguma forma? Existe lixo nos seus dados que precisa ser removido? Explique que "limpeza foi realizada" e por quê.

2. É preciso realizar algum tratamento em seus dados, por exemplo, conversões, normalizações, contagens? Explique nesta seção os tratamentos que você irá adotar nos seus dados. Repita o estudo de normalidade com os dados após o tratamento e use os dados tratados nas seções que seguem.

Distribuição das variáveis (atributos)

Como são as distribuições de onde suas variáveis vêm? Existem outliers? Quais os tamanhos de suas amostras?

1. Crie (e coloque no documento) visualizações e analise as distribuções de cada uma das suas variáveis já tratadas. Existe simetria? Caudas longas?

2. Há outliers? Explique como você os identificou.

3. Analise de forma crítica se os outliers existentes devem fazer parte da análise, por serem dados importantes que você quer considerar. Ou será que eles representam eventos extraordinários que você não quer considerar em suas análises? De forma muito consciente e com justificativas claras, você pode remover *outliers*. Deixe clara sua posição em retirar ou manter outliers, com justificativa.

4. Qual o tamanho das amostras que você tem de cada variável. Você tem amostras de tamanhos diferentes? Você tem amostras muito pequenas e inexpressivas (menor que 30, pelo menos). Você vai poder usar todas as variáveis que planejava?

Índices de tendência central e dispersão

Para cada uma de suas variáveis/atributos decida qual o melhor índice de tendência central a usar (média, mediana, média geométrica, etc.) e o melhor índice de dispersão a usar (desvio padrão, IQR, etc.). Sumarize os seus dados usando estes índices escolhidos.

Correlações entre atributos de um mesmo item

Podem existir correlações entre atributos de um item, por exemplo, ao analisar os dados pode-se identificar que quanto mais jovens são os usuários do foursquare, mais checkins eles fazem. Analise se existe correlação entre os atributos dos itens em seus dados. Isso pode ser feito com matrizes de dispersão (splom) para enxergar visualmente as correlações possíveis, além de gerar as matrizes de correlação para ver a intensidade e direção das correlações entre seus atributos[1]. Escolha um método de correlação adequado para este estudo. Para dados categóricos talvez sejam necessários outros testes diferentes dos coeficientes mais comumente usados para analisar correlações, como por exemplo testes de Chi-quadrado (lembrem de Fundamentos de Analytics).

Análise de viabilidade

1. Revisite as suas *use cases* e o seu planejamento inicial. Para cada use case identifique quais atributos serão usados, o tipo de análise que será realizado e os atributos que serão geradas após a análise.

2. Existem atributos coletados desnecessariamente?

3. Todas as use cases definidas são viáveis?

---

[1] Isso pode ser feito quando seus atributos forem de escala de razão ou ordinais.