Capítulo 4

Análise de Dados

Não tem faltado empenho na automação e na visualização dos dados com a finalidade de aprimorar a análise destes para tomadas de decisão ou solução de problemas mal-condicionados. No Capítulo 1 foi mencionada uma lista de áreas de pesquisa ativas nesta busca.

Com o acelerado aumento da capacidade de processamento das máquinas, técnicas relacionadas com a mineração de dados conseguem gerar hoje em dia, em tempo aceitável, um resultado a partir de um volume gigantesco de dados. Diante do nosso total desconhecimento acerca de um plausível resultado, o grande problema que enfretamos é sabermos o grau de credibilidade do resultado gerado, o quanto ele está próximo do resultado correto. Por outro lado, no contexto das técnicas de visualização interativa/exploratória tem-se desenvolvido interfaces que permitem especialistas acompanharem uma análise feita pelos computadores e guiarem esta análise com base nos seus conhecimentos. Neste contexto, temos no entanto o problema da exibição de um grande volume de dados de forma inteligível para os humanos.

Um sistema de analítica visual nasceu da ideia de combinar a capacidade (limitada) de mineração de dados e descoberta de conhecimentos (*Knowledge discovery and data mining*, KDD) das máquinas com o processamento cognitivo dos humanos na análise dos dados complexos. Na Seção 4.1. do livro-texto[1] são listadas algumas áreas de aplicação que se beneficiaram desta combinação de análise automática da máquina e da análise visual humana: bioinformática, análise da mudança climática, identificação de padrões extraídos automaticamente, mineração de dados espaço-temporais,

[1]Daniel Keim e colegas. Mastering the information age: solving problems with visual analytics

e indústrias.

Mostramos no Capítulo 3 que a estatísitca descritiva nos fornece ferramentas para organizar e sintetizar um grande volume de dados em medidas escalares e que os estatísticos desenvolveram diversas técnica de visualização destes dados com a expectativa de que os gráficos e diagramas ajudem a revelar algum padrão ou anormalidade nos dados. Neste capítulo apresentamos alguns conceitos da estatística inferencial que nos permitem fazer algumas conclusões a partir das medidas sintetizadas. Como já escreveu Friedman em 1997[2], as técnicas de amostragem e de inferência de uma característica específica a partir de uma amostra de N observações desenvolvidas na estatística inferencial podem ser úteis no processamento de uma quantidade volumosa de dados:

> *Most DM (Data Mining) applications routinely require data sets that are considerably larger than those that have been addressed by our traditional statistical procedures (kilobytes). However, it is often the case that the questions being asked of the data can be answered to sufficient accuracy with less than the entire (giga- or terabytes) data base. Sampling methodology, which has a long tradition in Statistics, can profitably be used to improve accuracy while mitigating computational requirements. Also, a powerful computationally intense procedure operating on a subsampling of the data may in fact provide superior accuracy than a less sophisticated one using the entire data base.*

Adicionalmente, selecionamos duas técnicas clássicas, amplamente aplicadas em Aprendizado de Máquina para mineração de dados, a fim de ilustrar como se pode construir alguns modelos computacionais equivalente aos modelos visuais mostrados na Fig. 4.1 do nosso livro-texto[3].

4.1 Estatística Inferencial

Essencialmente, a pergunta que a estatística inferencial procura responder é se podemos inferir a partir de uma amostra aleatória extraída de uma população a característica média e a variância desta característica da população. Veremos nesta seção que os conceitos desenvolvidos são aplicáveis nos testes (estatísticos) de hipóteses e na avaliação da credibilidade destes

[2] `http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf`

[3] Daniel Keim e colegas. Mastering the information age: solving problems with visual analytics

testes. Finalmente, mostramos uma técnica simples de estimativa de uma relação (linear) entre duas populações a partir das suas amostras.

4.1.1 Estimativa Intervalar

A média μ e a variância σ^2 de uma população, que vimos na Seção 3.2, são constantes, porém geralmente desconhecidas. Por outro lado, a média amostral $\overline{x}$ e a variância amostral $\sigma^2_{\overline{x}}$, mais fáceis de serem coletadas, variam usualmente de uma amostra para outra em função das observações feitas em N amostras de uma população. A **estatística inferencial** nos provê ferramentas para inferir os parâmetros fixos, μ e σ, de uma população através das variáveis amostrais, $\overline{x}$ e $\sigma^2_{\overline{x}}$, também denominadas **variáveis aleatórias** ou **estimadores aleatórios**. Ao invés de uma amostra (**estimativa pontual**), procura-se estimar os parâmetros fixos levando em conta um conjunto de amostras que tenham uma boa probabilidade para cobrir μ.

Para isso, o **Teorema de Limite Central**[4] tem um papel fundamental. Ele nos diz que a distribuição das médias $\overline{x}$ de um conjunto de amostras de tamanho N, extraídas praticamente de qualquer população, tende para uma **distribuição normal** com média μ e desvio padrão $\frac{\sigma}{\sqrt{N}}$. O fato da média amostral $\overline{x}$ ser um bom estimador da média μ de uma população não implica em $\overline{x} = \mu$. Na verdade, temos um **intervalo** centrado em μ, dentro do qual se encontra a média amostral $\overline{x}$ com um nível **de confiança** $p \in [0, 1]$. Isso equivale a dizer que tem um intervalo centrado em $\overline{x}$ no qual μ pode ser encontrado com uma probabilidade de p. Esta estimativa de μ com base na média amostral dentro de um intervalo de p de confiança é conhecida por **estimativa intervalar**.

Quando se conhece o desvio-padrão σ de uma população, o intervalo de $(1 - 2 \times p)$[5] de confiança, em que se encontra μ, pode ser computado com uso da variável z_p (escore-z) extraído da **tabela da distribuição normal padronizada** a partir de uma amostra de N observações[6]

$$z_p = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{N}}}, \tag{4.1}$$

isto é,

$$\mu \in [\overline{x} - z_p \times \frac{\sigma}{\sqrt{N}}, \overline{x} + z_p \times \frac{\sigma}{\sqrt{N}}]. \tag{4.2}$$

[4] `https://pt.wikipedia.org/wiki/Teorema_central_do_limite`

[5] São removidas as probabilidades de ocorrência das duas caudas, a inferior $\leq p$ e a superior $\geq p$.

[6] Uma tabela que relaciona a probabilidade acumulada $1-p$ de uma distribuição normal do menos infinito até z_p

Veja na nota de aula [7] como se obtém um extremo z_p do intervalo de confiança que acumula até z_p uma probabilidade acumulada, 1-p, de ocorrências.

Infelizmente, o desvio-padrão σ de uma população não é conhecido em muitas aplicações. Usualmente, ele precisa ser estimado e o candidato natural para esta estimativa é o desvio-padrão amostral s que vimos na Seção 3.2. Substituindo σ na Eq. 4.1 por s da Eq. 3.4, temos **escore-t** t_p no lugar do escore-z z_p

$$t_p = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{N}}}. \tag{4.3}$$

e a nova distribuição de observações, conhecida por **distribuição t de Student**[8], tem uma dispersão um pouco maior do que a da distribuição normal pela incerteza adicional que o estimador s embute. Diferente do escore-z, o escore-t depende não só da probabilidade de confiança como também do **número de graus de liberdade** (*degree of freedom*, df) [9]. O número de graus de liberdade é o divisor de s^2 no cômputo de variância amostral (Eq.3.4). Quanto maior o número de graus de liberdade, mais a distribuição t se aproxima da distribuição normal. Na prática, usa-se somente a distribuição t quando σ é desconhecido e o tamanho de amostras muito pequeno.

Com base nas ferramentas até agora apresentadas, podemos **comparar duas populações independentes (não correlacionadas)**, com variâncias populacionais conhecidas σ_1 e σ_2. De forma análoga ao procedimento desenvolvido para uma população, podemos usar, no lugar de uma estimativa pontual da diferença entre as duas médias da população

$$\mu_1 - \mu_2, \tag{4.4}$$

uma estimativa intervalar da diferença com $(1 - 2 \times p)$ de confiança

$$\begin{aligned} (\mu_1 - \mu_2) &= (\overline{x}_1 - \overline{x}_2) \pm z_p var(\overline{x}_1 - \overline{x}_2) \\ & (\overline{x}_1 - \overline{x}_2) \pm z_p \sqrt{\frac{{\sigma_1}^2}{N_1} + \frac{{\sigma_2}^2}{N_2}}, \end{aligned} \tag{4.5}$$

onde $\overline{x}_1$ e $\overline{x}_2$ são as médias amostrais para as respectivas N_1 e N_2 observações.

[7] `https://social.stoa.usp.br/articles/0027/6301/memoria-da-aula-pratica-4.pdf`

[8] `http://www.portalaction.com.br/probabilidades/64-distribuicao-t-de-student`

[9] `https://www.dummies.com/education/math/statistics/how-to-use-the-t-table-to-solve-statistics-problems/`

Se as variâncias populacionais não forem conhecidas, lançamos mão nas variâncias amostrais s_1 e s_2 e na distribuição de t_p com o número de graus de liberdade igual a $\min\{N_1-1,N_2-1\}$, ou seja, as diferenças entre as duas médias amostrais, $\overline{x}_1$ e $\overline{x}_2$, flutuam no intervalo

$$(\overline{x}_1 - \overline{x}_2) \pm t_p \sqrt{\frac{{s_1}^2}{N_1} + \frac{{s_2}^2}{N_2}} \tag{4.6}$$

Exemplos para cômputo de intervalos de confidência em R podem ser encontrado na *internet*[10].

Vale observar que, se as **duas populações não forem independentes**, é mais natural fundir as duas populações numa única população cujos elementos/observações são as diferenças entre pares correlacionados, e aplicar as estimativas pontual e intervalar sobre a população destas diferenças.

4.1.2 Teste de Hipóteses

O teste (estatístico) de hipósteses consiste em testar a existência de uma relação entre duas populações. Modelando a hipótese de comparação entre as duas populações como diferença Δ entre elas, podemos aplicar Eq.4.5 e Eq.4.6 para determinar o intervalo de flutuações das diferenças entre as médias amostrais das duas populações, respectivamente, com variâncias conhecidas e desconhecidas.

Se o intervalo de flutuações de diferenças contiver a relação pressuposta, dizemos que a hipótese é **plausível** ao nível de $1 - 2 \times p$[11] de confiança. E se ela não estiver no intervalo, a hipótese é então rejeitada e dizemos que a diferença entre as duas médias amostrais é **estatisticamente significante, ou discernível, ao nível de significância** $2 \times p$.

Uma **hipótese nula**, H_0, é aquela que assume que a média de uma população seja igual à pressuposta, ou seja, assume que seja verdadeiro o que se considerou. Se conseguirmos mostrar que tal hipótese não seja corroborada pelos dados disponíveis, a **hipótese alternativa**, H_1, passa a ser verdadeira.

Para "provar" que uma hipótese nula seja verdadeira, usamos o **nível do teste**, α, que costuma ser pré-fixado em 0.05, 0.01, 0,005 ou 0.001, e o valor de prova, **p-valor**, da média amostral $\overline{x}$ dos N dados observados. Este valor de prova não é nada mais do que a probabilidade acumulada da

[10] `https://www.cyclismo.org/tutorial/R/confidence.html`

[11] Lembra-se de que as probabilidades relacionadas a ambas as caudas da distribuição devem ser subtraídas.

distribuição normal a partir do ponto em que situa $\overline{x}$[12]. Ele pode ser obtido a partir das tabelas de distribuição com uso dos escores calculados (Eqs. 4.1 e 4.3). Rejeitamos a hipótese H_0 se, e somente se, o valor de prova da média amostral $\overline{x}$ estiver menor ou igual ao nível de teste α.

Vale destacar que, caso uma hipótese verdadeira ser rejeitada, dizemos que ocorreu um erro do tipo I. A probabilidade da ocorrência de um erro do tipo I é naturalmente o nível de teste α.

No ambiente R pode-se computar facilmente os p-valores para testes de hipóteses H_0[13].

4.1.3 Regressão

A **regressão linear** consiste, de fato, uma aproximação de um conjunto de pares ordenados por uma função linear. A técnica numérica mais utilizada é o ajuste de mínimos quadrados (*least-square fitting*). Dada a equação de uma reta:

$$y = a + bx. \tag{4.7}$$

O ajuste dos m pares ordenados (x_i, y_i) a esta reta pelos mínimos quadrados corresponde a minimizar a soma q dos quadrados das distâncias d_i para cada x_i

$$q = \sum_{i=1} m d_i^2 = \sum_{i=1} m(y_i - a - bx_i)^2. \tag{4.8}$$

Após algumas manipulações algébricas, chega-se ao seguinte sistema de equações, denominado **equações normais** do problema:

$$\begin{cases} ma + (\sum_{i=1} mx_i)b = \sum_{i=1}^{m} y_i \\ (\sum_{i=1} mx_i)a + (\sum_{i=1} mx_i^2)b = \sum_{i=1}^{m} x_i y_i \end{cases} \tag{4.9}$$

As duas icógnitas a e b correspondem exatamente aos coeficientes da reta procurada.

Se considerarmos x_i e y_i observações de duas populações, a resolução do sistema de equações 4.9 nos dá a relação linear que melhor se ajusta às observações.

É possível usar o ambiente R para computar regressões lineares[14].

[12] Valor de prova unilateral: de $-\infty$ até -(p-valor) ou de (p-valor) até ∞; Valor de prova bilateral: soma dos dois valores unilaterais.

[13] `http://www.r-tutor.com/elementary-statistics/hypothesis-testing`

[14] `https://www.datacamp.com/community/tutorials/linear-regression-R`

4.2 Mineração de Dados

Como já comentamos anteriormente, o interesse pela mineração de dados nasceu da evolução de *Data Base Management Systems* (DBMS) para *Decision Support Systems* (DBB) com suporte a *online analytic processing* (OLAP). Originalmente, os usuários formulavam as potencialmente relevantes perguntas (*queries*) para o sistema interativamente até chegarem a informações apropriadas ou até se cansarem como escreveu Friedman em 1997[15]. Literalmente, usuários guiavam o processo de busca por informações relevantes.

Sob forte pressão de demanda por sistemas mais "inteligentes" capazes de minerarem padrões, anomalias e previsões automaticamente a partir de perguntas vagas dos usuários, o interesse pela Mineração de Dados cresceu vertiginosamente. Em vista da área emergente de Aprendizado de Máquina, acreditava-se que uma possível solução seria "instruir" a máquina com os procedimentos manuais que os especialistas usavam. Procurou-se então focar no aprimoramento das técnicas de uso consolidado em Aprendizado de Máquina, como escreveu Friedman em 1997:

> *From the perspective of statistical data analysis one can ask whether DM (Data Mining) methodology is an intellectual discipline. So far, the answer is – not yet. DM packages implement well known procedures from the fields of machine learning, pattern recognition, neural networks, and data visualization. They emphasize "look and feel" (GUI) and the existence of functionality . There seems to be no real regard for performance (wht's under the hood). The goal is to get to market quickly. Most academic research in this area so far has focused on incremental modifications to current machine learning methods, and the speed-up of existing algorithms.*

Só para ilustrar a relação entre técnicas de mineração de dados e descoberta automática de "informações", apresentamos nesta seção duas técnicas bem populares na área de Mineração de Dados: análise de componentes principais e clusterização por *k-means*.

4.2.1 Análise de Componentes Principais

Embora seja considerada por alguns uma técnica (não-supervisionada) de Aprendizado de Máquina, muitos autores a classificam como uma técnica

[15] http://statweb.stanford.edu/ jhf/ftp/dm-stat.pdf

estatística de pré-processamento dos dados de treinamento. Pois, ela permite organizar os dados pela variância (estatística) dos seus atributos e reduzir o volume de dados, como por exemplo antes de uma técnica de Aprendizado de Máquina como clusterização (Seção 4.2.2). É também muito utilizada em técnicas supervisionadas de Aprendizado de Máquina quando usamos os dados reduzidos para treinamento e confrontar os resultados com os obtidos a partir de dados originais.

Um conjunto de dados com M atributos pode ser representado por M populações e ao ordenarmos as suas variâncias podemos identificar os atributos/componentes estatisticamente mais significativos/relevantes, ou seja, os componentes que apresentam maiores variâncias. Na prática, os dois primeiros componentes de maiores variâncias são considerados os mais importantes.

Caso os atributos sejam valores numéricos, podemos representar o conjunto de dados de M atributos como um espaço de vetores de M dimensões e reduzir o problema a um problema algébrico de cômputo de autovalores e autovetores deste espaço e ordenar a importância dos atributos pelos seus autovalores. Quanto maior o autovalor, maior a variância e maior a importância sob o ponto de vista estatístico.

A técnica numérica mais popular para decompor um conjunto de vetores de dimensão M em M valores singulares é a Decomposição em Valores Singulares (*Singula Value Decomposition*, SVD). Ela consiste essencialmente na fatorização de uma matriz A constituída por todos os dados de interesse em

$$A = U \sum V^T, \tag{4.10}$$

onde U é uma matriz de autovestores AA^T, $\sum$ á uma matriz diagonal formada pelos autovalores, e V uma matriz de autovetores $A^T A$.

Dispõem-se em R 5 funções relacionadas com a análise de componentes principais (*Principal Components Analysis*, PCA). A função **prcomp**, baseado no algoritmo de SVD, faz parte do pacote básico de R[16].

4.2.2 Clusterização

Clusterização é uma das técnicas muito utilizadas em Mineração de Dados para revelar possíveis padrões de similaridade. Um dos métodos mais conhecidos de clusterização é a clusterização **k-means**, considerada como uma técnica não-supervisionada de Aprendizado de Máquina. Pois, somente

[16] https://www.datacamp.com/community/tutorials/pca-analysis-r

a partir dos dados de entrada, a máquina consegue agrupá-los iterativamente em k grupos em torno dos centróides estimados.

Quando os dados são descritíveis pelos vetores e a dissimilaridade entre eles representável pela distância euclidiana, podemos aplicar o algoritmo de Lloyd proposto em 1957 para agrupá-los de forma que a distância entre um dado x_{ij} em relação ao centróide c_i do grupo com n_i elementos, ao qual ele pertence, seja a menor[17]:

$$c_i = \frac{\sum x_{ij}}{n_i} \tag{4.11}$$

Clusterização por k-means é suportada pelo R[18].

4.3 Visualização

Para suportar esta etapa de análise de dados, Keim e colegas levantaram três características importantes para um ambiente interativo de suporte a mineração de dados[19]:

1. suficientemente rápido para interações efetivas. Vimos na Seção 2.4 que, de acordo com Jakob Nielsen[20], para uma resposta em até 1s a um evento gerado pelo usuário, o usuário ainda consegue manter o seu fluxo de raciocínio sem se distrair.

2. os parâmetros dos modelos são representados pelas formas gráficas e atributos gráficos de forma intuitiva. Além disso, é fundamental que haja uma correspondência biunívoca entre os parâmetros dos modelos e as formas e atributos gráficos associados.

3. os parâmetros dos modelos podem ser ajustados interativamente.

De fato, estas propriedades não são específicas para mineração de dados. Elas devem ser consideradas no projeto de qualquer interface interativa e responsiva.

Ressaltamos que a aplicação de técnicas de mineração de dados e de estatísticas para selecionar, agrupar, reduzir as dimensões dos dados, e predizer a significância dos dados, pode reduzir o domínio dos dados relevantes

[17] `https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm`

[18] `https://www.datacamp.com/community/tutorials/k-means-clustering-r`

[19] Daniel Keim e colegas. Mastering the information age: solving problems with visual analytics

[20] `https://www.nngroup.com/articles/response-times-3-important-limits/`

que devem ser renderizados. Isso pode aumentar a legibilidade dos dados exibidos e contribuir para uma melhor análise visual dos dados. Porém, dados pré-processados podem não ter mais referências espaciais. É, então, necessário elaborar uma forma de mapear estes dados abstratos em formas espaciais e atributos gráficos antes de renderizá-los numa tela de computador.

4.4 Exercícios

1. Qual é a diferença entre Estatística Descritiva e Estatística Inferencial?

2. Qual é a diferença entre uma distribuição normal e uma distribuição t de Student? Qual(is) critério(s) utilizado(s) na escolha de uma delas para estimar um intervalo de confiança em que se encontra a média populacional?

3. Supondo que *Fisher's Iris dataset* contenha dados de quatro populações (largura e comprimento de sépalas, largura e comprimento de pétalas). Construa aleatoriamente 12 conjuntos de 5 observações para cada uma destas populações. Plote a distribuição das 12 médias amostrais. Determine o intervalo de confiança para cada média amostral, considerando (a) as variâncias conhecidas (as que você computou na questão 4 do Capítulo 3) (b) as variâncias desconhecidas.

4. Há alguma diferença estatisticamente significativa entre as larguras de sépalas e os comprimentos de pétalas no conjunto *Fisher's Iris dataset*? Qual é o intervalo de confiança para a sua afirmação? Justifique. Use R na sua solução, mas faça uma descrição sucinta do procedimento que pode ser utilizado.

5. Prove estatisticamente a credibilidade das seguintes hipóteses nulas H_0 em relação a *Fisher's Iris dataset* ao nível do teste 0.05 (5%): média dos comprimentos de sépalas é 5.84; média das larguras de sépalas é 3.054; média dos comprimentos de pétalas é 3.7586663; média das larguras de pétalas é 1.1986667. Use R na solução mas faça uma descrição sucinta do algoritmo. Repita as mesmas provas ao o nível de teste 0.1%. Houve alguma diferença? Explique.

6. Aproxime todas as possíveis relações entre as quatro variáveis na questão 4 do Capítulo 3 por funções lineares em R. Qual é a função de aproximação para cada par? Plote a reta e verifique visualmente

7. Quais três características da Íris são estatisticamente mais relevantes? Você consegue extrair as 3 variedades de Íris em *Fisher's Iris dataset* com estas três características?

8. Aplique a técnica de clusterização *k-means* para agrupar as Íris em 3 grupos com uso (a) das quatro características; (b) das 3 características mais significativas estatisticamente; (c) das 2 características mais significativas estatisticamente; (d) da caractarística mais significativas estatisticamente. Compare os resultados.

9. Qual é a representação gráfica mais apropriada para visualizar *outliers* nos dados coletados para as três variedades de íris na questão 4 e os quatro conjuntos de dados na questão 5? Visualize os gráficos com uso de R[21].

10. Plote no ambiente R as observações das Íris com uso das 2 características mais significativas estatisticamente. Qual é a diferença desta forma de visualização em relação à *scatterplots*? (Dica: Lembre-se que a técnica de redução de dimensões é muito utilizada para exibir dados de alta dimensão numa tela do compurador.)

11. O ambiente R suporta interações coordenadas entre os gráficos[22]. Crie um ambiente de *brushing and linking* na identificação de um ponto nos gráficos *scatterplots* e *parallel coordinates plot* dos quatro conjuntos de dados da questão 5.

12. Numa visualização coordenada é possível ter mais de uma janela associada a um mesmo conjunto de dados, cada janela é usualmente associada a uma representação. Nesta arquitetura uma das fontes de problema são as interações inconsistentes num mesmo conjunto de dados através de diferentes janelas. Como você evitaria estas inconsistências? Discuta brevemente a sua solução.

[21] http://r-statistics.co/Linear-Regression.html
[22] https://www.statmethods.net/advgraphs/interactive.html