Análise Multivariada e Aprendizado Não Supervisionado

Vicente G. Cancho
garibay@icmc.usp.br

Departamento de Matemática Aplicada e Estatística
Universidade de São Paulo

Objetivos

- Introduzir os fundamentos básicos de análise multivariado e aprendizado não supervisionado.
- Apresentar os principais métodos estatísticos para a modelagem e análise de dados multivariados.

Programa

1. Introdução;
2. Visualização de dados multivariados e análise descritiva;
3. Vetores aleatórios e álgebra de matrizes;
4. Média e matriz de covariâncias amostrais;
5. Distribuições multivariadas;
6. Inferências sobre um vetor de médias;
7. Análise de variância multivariada;
8. Regressão linear multivariada;
9. Análise de componentes principais;
10. Análise fatorial;
11. Análise de correlação canônica;
12. Análise de agrupamentos;
13. Análise discriminante;
14. Análise de correspondência;

Bibliografia

1. Johnson, R. A. and Wichern, D. W. (2007) Applied Multivariate Statistical Analysis. 5th edition. Prentice-Hall.
2. Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). Multivariate Analysis. Academic Press.
3. Friedman, J., Hastie, T., and; Tibshirani, R. (2001). The elements of statistical learning. New York, NY,USA: Springer Series in Statistics.
4. James, G., Witten, D., Hastie, T., and; Tibshirani, R. (2013). An introduction to statistical learning (with applications in R). New York: Springer.

Apoio computacional

1. R (www.R-project.org).
2. Minitab, S-PLUS, SAS, SPSS, Statistica, etc.

Avaliação

A avaliação consta de duas provas, um serie de lista de atividades.
O conteúdo das provas serão aquelas que foram ministrados até uma semana antes do dia da prova.
Os pesos da nota final da disciplina são:

Prova	Peso	Data
1	40%	04/10/2022
2	30%	06/12/2022

A nota média de atividades tem peso de 30% .

Introdução

Análise Multivariada é utilizado em situações nas quais várias características (variáveis) são medidas simultaneamente, em cada elemento amostral. (As variáveis são aleatórias e correlacionadas).

As técnicas multivariados são comumente utilizados:

- Redução de Dimensão;
- Geração de agrupamentos homogêneos (clusters);
- Investigação de dependência entre variáveis;
- Previsão;
- construção de hipóteses e testes;

Redução de dimensão

- O objetivo é encontrar um meio de condensar a informação em um número de variáveis originais en um conjunto de menor de variáveis ;
- Esses métodos geram novas variáveis compostas pelas as variáveis originais.
- O fenômeno que esta sendo estudado pode ser representado de forma mais simples, a fim de propiciar uma interpretação mais fácil.

Exemplo-1

Um analista financeiro está interessado em estudar a saúde financeira de empresas. Para isso, identificou 18 indicadores (liquidez corrente, giro do ativo, receita operacional líquida, lucro líquido, endividamento geral, endividamento corrente, ...). Entretanto, a tarefa do analista seria simplificada se os 18 indicadores pudessem ser reduzidos para poucos índices, independentes (fatores que impactam na saúde financeira).

Geração de agrupamentos homogêneos

Grupos de observações ou variáveis similares podem ser estabelecidos maximizando a similaridade dentro do agrupamento e a dissimilaridade entre agrupamentos.

Exemplo-2

1. O analista financeiro gostaria de segmentar as empresas analisadas de acordo com os fatores (que impactam na saúde financeira) identificados.
2. O fabricante de bens de consumo, após mapear a estrutura de mercado e determinar os fatores que diferenciam os produtos/marcas, gostaria de segmentar os produtos/marcas.

Investigação da dependência entre variáveis

A natureza das relações entre as variáveis é de interesse.

Exemplo-3

Os dados sobre diversas variáveis foram usadas para identificar os fatores que foram responsáveis para sucesso do cliente quanto da prestação do serviço de consultores externos.

Previsão

As relações entre as variáveis devem ser determinadas com o objetivo de prever os valores de um ou mais variáveis com base em observações nas outras variáveis.

Exemplo-4

Medidas de variáveis contábeis e financeiros foram usados para desenvolver um modelo, para identificar clientes potencialmente inadimplentes.

Construção de testes de hipóteses

hipóteses estatísticas específicas, formuladas em termos dos parâmetros de populações multivariadas, são testadas a fim de que seja possível validar as considerações iniciais do pesquisador.

Exemplo-5

Variáveis relacionadas a poluição foram medidos para determinar se os níveis de poluição de uma grande área metropolitana foram mais ou menos constantes ao longo da semana, ou se houve uma diferença notável entre dias úteis e fins de semana.

Classificação de técnicas multivariadas

Sintetização da estrutura de variabilidade dos dados

- Componentes principais,
- Análise fatorial,
- Análise de agrupamento,
- Análises discriminante,
- Análise de correlações canônicas,
- Análise de correspondência

Classificação de técnicas multivariadas

1. técnicas exploratórias de simplificação da estrutura de variabilidade dos dados
   - Componentes principais,
   - Análise fatorial,
   - Análise de agrupamento,
   - Análises discriminante,
   - Análise de correlações canônicas,
   - Análise de correspondência
2. Técnicas de inferência estatística
   - Estimação de parâmetros,
   - Testes de hipóteses,
   - Análise de variância e de covariância,
   - Regressão multivariada.

O que é aprendizado Estatístico ?

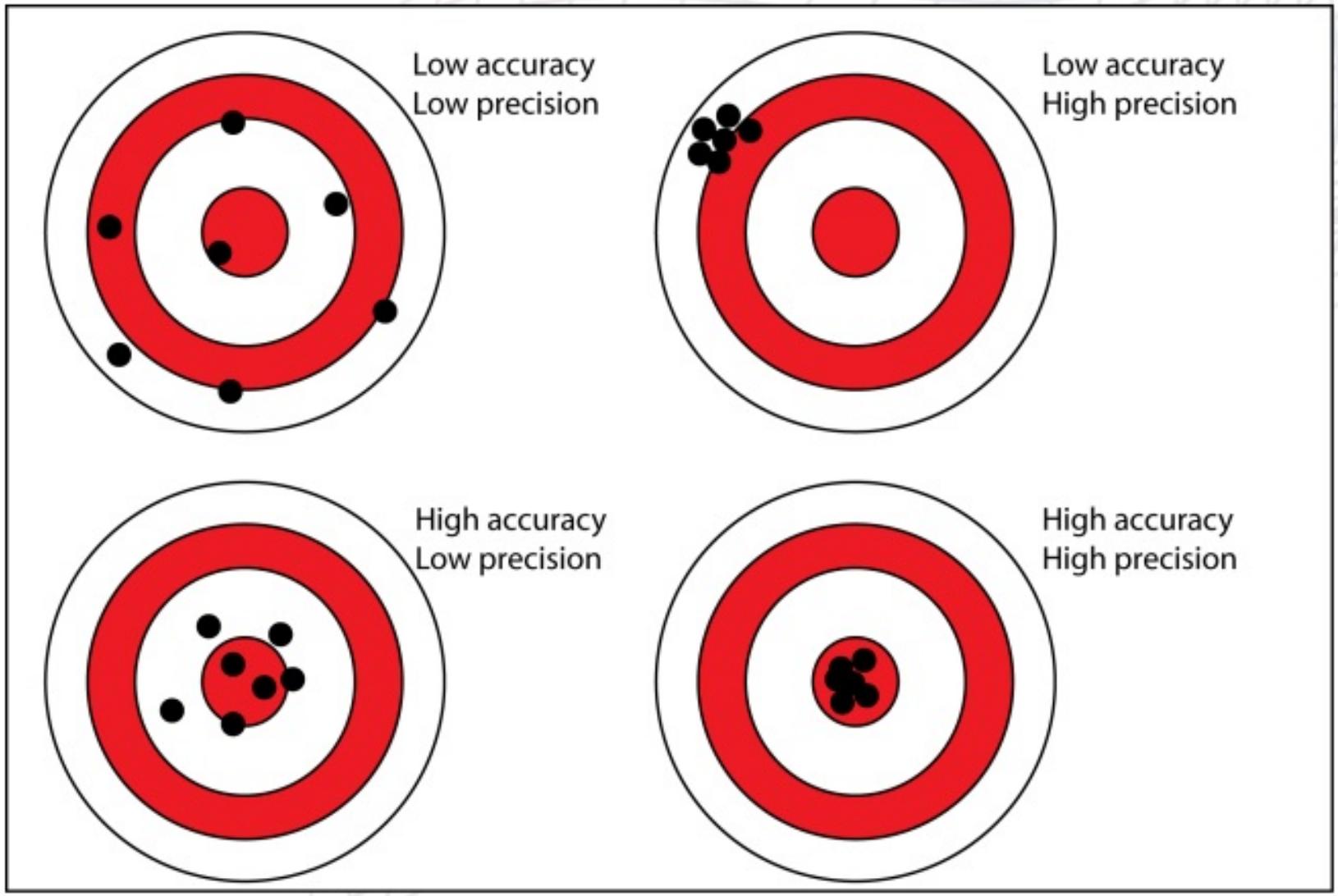
Aprendizagem Estatística (Statistical Learning (SL)) é um conjunto de ferramentas cujo objetivo principal é obter maior conhecimento possível de um conjunto de dados.

- De maneira geral, as técnicas de aprendizagem em SL pode-se classificar en dois grandes grupos: Supervisionado y No Supervisionado.
- Um dos objetivos do SL é a previsão.
- É uma área fértil para pesquisa e desenvolvimento de soluções e aplicações que vem crescendo especialmente durante os últimos 30 anos

Statistical Learning (SL) vs Machine Learning (ML)

- ML surgiu (nos anos 80) como parte da Inteligência Artificial;
- SL surgiu como um ramo da Estatística;
- Há muito em comum entre as duas técnicas, tanto o objetivo principal (previsão), como classificação de problemas (supervisionado vs. não supervisionado) ;
- ML: aborda problemas de grande escala e / ou ritmo reduzido. Focado em soluções com precisão ;
- SL: enfatiza nos modelos e na interpretação, Focado em soluções de problemas com equilíbrio entre Precisão e Acurácia,

Statistical Learning (SL) vs Machine Learning (ML)



Aprendizado supervisionado

- Y : Variável de resposta (também chamada de variável dependente, alvo).
- $\boldsymbol{X} = (X_1, ..., X_p)$: Conjunto de variáveis ??preditivas (também chamadas de variáveis independentes, regressores, entrada, covariáveis);
- Em problemas de regressão: Y é contínuo (exemplo: salário, altura);
- Em problemas de classificação: Y é binário, ordinal ou nominal (exemplo: gênero, classe social, cor dos olhos);
- Temos um conjunto de observações relacionadas $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$.

Aprendizado supervisionado

Neste tipo de problema, gostaríamos de:

- Prever, com um nível adequado de precisão e exatidão, a valor de Y_{new} nos casos em que $\boldsymbol{X}_{new}$ não foi observado
- Compreender e interpretar os fatores que afetam o resultado de esta predição
- Quantificar a certeza das nossas previsões

Algumas técnicas de aprendizado supervisionado:

- Modelos de regressão linear,
- Regressão logística,
- Análise fatorial,
- Análise discriminante linear (LDA).

Aprendizado Não supervisionado

- No existe una variable respuesta Y;
- $\boldsymbol{X} = (X_1, ..., X_p)$: Conjunto de características, ja seja por cada individuo, tempo, etc;
- O objetivo desses problemas está relacionado à identificação grupos ou padrões de comportamento implícitos nos dados; item Difícil de medir a qualidade do nosso modelo .

Algumas técnicas de aprendizado não-supervisionado:

- Análise de componentes principais;
- Análise de agrupamentos;
- Análise de correlações canônicas;
- Análise de correspondência.