MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 1

13 de março de 2023

Sumário

Paradigma

1. Modelo, padrão a ser seguido.
2. Um pressuposto filosófico, uma teoria, um conhecimento, que origina o estudo de um campo científico.
3. Aquilo que os membros de uma comunidade científica partilham.

Exemplos:

- Inferência Frequentista, Inferência Bayesiana
- Data Mining, Neural Networks, Data Science
- Statistical Learning, Machine Learning

Aprendizado com Estatística

- Aprendizado com Estatística (AE)/Statistical Learning (SL): nomenclatura nova, mas a maioria dos conceitos foram desenvolvidos desde o Século 19. Métodos estatísticos para previsão, classificação, análise de agrupamentos etc. Inferência é o objetivo e interpretação é importante.
- Aprendizado com (ou de) Máquina (AM)/Machine Learning (ML): métodos para "aprender" padrões ocultos em dados. Usados para previsão, classificação, reconhecimento de padrões, análise de agrupamentos etc. Pouca atenção à inferência (do ponto de vista computacional) e à interpretabilidade.
- When ML methods are statistically sound they are called Statistical Learning (SL) methods.
- Nosso foco: Métodos de AE.

Probabilidade

1. Início em 1654 com Fermat (1601-1665), Pascal (1623-1662): jogos de dados
2. Huygens (1629-1695): primeiro livro de probabilidade em 1657.
3. Bayes (1702-1761): primeira versão do Teorema de Bayes, publicado em 1763.

Gauss e Legendre

1. Gauss (1777-1856) inventou o método de mínimos quadrados (MQ) na última década do século 18 (1795) e o usou regularmente depois de 1801 em cálculos astronômicos.
2. Legendre (1752-1833): publicou no apêndice de "Nouvelles Methodes pour la Détermination des Orbites des Còmetes". Nenhuma justificação.
3. Gauss (1809): deu justificativa probabilística do método. Em "The Theory of the Motion of Heavenly Bodies".
4. Implementaram o que é hoje chamado de **regressão linear**.

Bayes ou Laplace?

1. Laplace (1749-1761): desenvolveu o Teorema de Bayes independentemente, publicado em 1774.
2. 1812: Théorie Analytique dés Probabilités: aplicações científicas e práticas.
3. 1814: Essais Philosophiques sur les Probabilités: interpretação Bayesiana das probabilidades
4. Inferência Bayesiana $\leftrightarrow$ Inferência Laplaciana. Usada a partir de 1800.
5. Fisher e Neyman: início do século 20.
6. Jeffrey (1939). Theory of Probability. Considerado como o re-início da Inferência Bayesiana
7. de Finetti, Savage, Lindley etc.

Fisher e Neyman

1. Inferência Frequentista (testes de hipóteses, estimação, planejamento de experimentos e amostragem) foi iniciada por R. Fisher(1890-1962) e J. Neyman (1894-1981).
2. Fisher (1925): Statistical Methods for Research Workers. (14ª Edição: 1970)
3. Fisher (1935): The Design of Experiments (8ª Edição: 1966)
4. Fisher (1936): propõe a **análise discriminante linear**.

Gosset/Student

1. W. Gosset (1876-1937): Em 1908 publicou sob o pseudônimo de Student um artigo que iniciou um novo paradigma em "Pequenas Amostras". Resultado provado por Fisher em 1912-1915.
2. Student (1908). The probable error of a mean. *Biometrika*, 6, 1-25.
3. Fisher (1922). On the mathematical foundation of theoretical statistics. *Phil. Trans. Royal Society*, A, 222, 308-368.
4. Stigler: "The most influencial article on Theoretical Statistics in the 20th Century"
5. Hald: "For the first time in the history of Statistics a framework for frequency-based general theory of parametric statistical inference was clearly formulated."

Neyman e Pearson

1. Neyman and Pearson (1933a) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Royal Society*, A, 231, 289-331.
2. Neyman and Pearson (1933b). On the testing of statistical hypothesis in relation to probabilities a priori. *Proc. Cambridge Philos. Society*, 24, 429–510.
3. Livros de E. Lehmann sobre estimação e testes de hipóteses. 1967, 1983.
4. Era do "Small Data".

1940 - 2000

1. 1940: propostas de abordagens alternativas: regressão logística.
2. 1970: Nelder e Wedderburn: generalized linear models para uma classe de métodos que incluem regressão linear e logística como casos especiais.
3. 1970: Efrom: bootstrap; Hoerl e Kennard: ridge regression.
4. até o final de 1970: métodos lineares.
5. 1980: tecnologia computacional possibilita a aplicação de métodos não lineares.
6. 1984: Friedman et al. introduzem CART (Classification and regression trees) e propõem uma implementação prática de método para seleção de modelos, incluindo CV (cross validation)
7. 1986: Hastie e Tibshirani: estendem os MLG pra os modelos GAM (generalized additive models).
8. 1996: Tibshirani: introduz o LASSO; extensões para outros métodos de regularização.

O que é Estatística?

[1] Coleta de dados: amostras, planejamento de experimentos, estudos observacionais.

[2] Modelagem e análise de dados.

[3] Tomada de decisões.

Eras da Estatística

A história da Estatística pode ser dividida em três eras:

(1) A era de Quetelet (astrônomo, matemático, estatístico belga, 1796-1874) e seus sucessores, na qual o obejtivo era obter grandes conjuntos de dados (censos) em ciências sociais.

(2) O período clássico de Pearson (1857-1936), Fisher (1890-1962), Neyman (1894-1981), Hotelling (1895-1973) a seus sucessores, que desenvolveram a teoria de inferência ótima; métodos apropriados para *small data sets*.

(3) A era da produção de dados em massa (*big data sets)*, com novas tecnologias como *microarrays*, dados de alta frequência em finanças, dados astronômicos etc.

Algoritmo e Inferência

Análise Estatística:

(a) algoritmica
(b) inferencial

Exemplo: Considere estimar a média $\mu = E(X)$ de uma v.a. X, definida sobre uma população.
Para uma AAS $X_1, \ldots, X_n$, considere o estimador

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Este é o **algoritmo**.
Quão acurado e preciso é $\overline{X}$. Esta é a parte da **inferência**.

Algoritmo e Inferência

Algoritmos: é o que os estatísticos fazem.
Inferência: porque os estatísticos usam os algoritmos. (Efrom, 2016).

Conjuntos de dados enormes (*Big Data*) requerem novas metodologias. Esta demanda está sendo atendida por algoritmos estatísticos baseados em computação intensiva.

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- $\rightarrow$ 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960$\rightarrow$ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980$\rightarrow$: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C_+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- $\rightarrow$ 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960$\rightarrow$ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980$\rightarrow$: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C_{+}, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- $\rightarrow$ 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960$\rightarrow$ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980$\rightarrow$: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C_+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data"e da "Data Science".

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- $\rightarrow$ 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960$\rightarrow$ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980$\rightarrow$: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C_{+}, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- $\rightarrow$ 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960$\rightarrow$ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980$\rightarrow$: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C_+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data"e da "Data Science".

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- $\rightarrow$ 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960$\rightarrow$ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980$\rightarrow$: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C_+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

Referências

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Berlin: Springer.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning*. Second Edition. Springer

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Morettin, P.A. and Singer, J.M. (202). *Estatística e Ciência de Dados*. Rio de Janeiro:LTC.

MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 2

23 de março de 2023

Sumário

Ciência de Dados

- Atualmente, os termos *Data Science* (Ciência de Dados) e *Big Data* (Megadados) são utilizados em profusão, como se fossem conceitos novos, distintos daqueles com que os estatísticos lidam há cerca de dois séculos.
- Na década de 1980, numa palestra na Universidade de Michigan, EUA, C.F. Jeff Wu já sugeria que se adotassem os rótulos *Statistical Data Science*, ou simplesmente, *Data Science*, em lugar de *Statistics*, para dar maior visibilidade ao trabalho dos estatísticos.
- Talvez seja Tukey (1962, 1977), sob a denominação *Exploratory Data Analysis* (Análise Exploratória de Dados), o primeiro a dar importância ao que hoje se chama Ciência de Dados, sugerindo que se desse mais ênfase ao uso de tabelas, gráficos e outros dispositivos para uma análise preliminar de dados, antes que se passasse a uma **análise confirmatória**, que seria a **inferência estatística**.

Ciência de Dados

- Outros autores, como Chambers (1993), Breiman (2001) e Cleveland (1985, 1993, 2001), também enfatizaram a preparação, apresentação e descrição dos dados como atividades preparatórias para inferência ou modelagem.
- Basta uma procura simples na Internet para identificar novos centros de Ciências de Dados (CD) em várias universidades ao redor do mundo, com programas de mestrado, doutorado e mesmo graduação.
- O interessante é que muitos desses programas estão alojados em escolas de Engenharia, Bioestatística, Ciência da Computação, Administração, Economia etc., e não em departamentos de Estatística.
- Paradoxalmente, há estatísticos que acham que Estatística é a parte menos importante de CD! Certamente isso é um equívoco. Como ressalta Donoho (2017), se uma das principais características de CD é analisar grandes conjuntos de dados (Megadados), há mais de 200 anos os estatísticos têm se preocupado com a análise de vastos conjuntos de dados provenientes de censos, coleta de informações meteorológicas, observação de séries de índices financeiros etc., que têm essa característica.

Ciência de Dados

- Outro equívoco consiste em imaginar que a Estatística Clássica (frequentista, bayesiana etc.) trata somente de pequenos volumes de dados, conhecidos como *Small Data*.
- Essa interpretação errônea vem do fato de que muitos livros didáticos apresentam conjuntos de dados, em geral de pequeno ou médio porte, para que as metodologias apresentadas possam ser aplicadas pelos leitores, mesmo utilizando calculadoras ou aplicativos estatísticos (pacotes). Nada impede que essas metodologias sejam aplicadas a grandes volumes de dados a não ser pelas dificuldades computacionais inerentes.
- Talvez seja este aspecto computacional, aquele que mascara os demais componentes daquilo que se entende por CD, pois em muitos casos, o interesse é dirigido apenas para o desenvolvimento de algoritmos cuja finalidade é aprender a partir dos dados, omitindo-se características estatísticas.

Ciência de Dados

- Ciência de Dados(CD) é "filha" da Estatística e da Ciência da Computação.
- Perspectiva não é nova: Tukey (1962): The future of Data Analysis, AMS.
- Cientistas de diversas disciplinas estão sendo confrontados com conjuntos enormes de dados: sequenciamento genético, grandes arquivos de textos, dados astronômicos, dados financeiros de alta frequência.
- Perspectiva da Estatística, da Computação e Humana.
- Ciência de Dados: redes neuronais, support vector machines, machine learning, deep learning, classification and regression trees (CART), random forests etc.

Ciência de Dados

- Ciência de Dados(CD) é "filha" da Estatística e da Ciência da Computação.
- Perspectiva não é nova: Tukey (1962): The future of Data Analysis, AMS.
- Cientistas de diversas disciplinas estão sendo confrontados com conjuntos enormes de dados: sequenciamento genético, grandes arquivos de textos, dados astronômicos, dados financeiros de alta frequência.
- Perspectiva da Estatística, da Computação e Humana.
- Ciência de Dados: redes neuronais, support vector machines, machine learning, deep learning, classification and regression trees (CART), random forests etc.

Ciência de Dados

- Ciência de Dados(CD) é "filha" da Estatística e da Ciência da Computação.
- Perspectiva não é nova: Tukey (1962): The future of Data Analysis, AMS.
- Cientistas de diversas disciplinas estão sendo confrontados com conjuntos enormes de dados: sequenciamento genético, grandes arquivos de textos, dados astronômicos, dados financeiros de alta frequência.
- Perspectiva da Estatística, da Computação e Humana.
- Ciência de Dados: redes neuronais, support vector machines, machine learning, deep learning, classification and regression trees (CART), random forests etc.

Ciência de Dados

- Ciência de Dados(CD) é "filha" da Estatística e da Ciência da Computação.
- Perspectiva não é nova: Tukey (1962): The future of Data Analysis, AMS.
- Cientistas de diversas disciplinas estão sendo confrontados com conjuntos enormes de dados: sequenciamento genético, grandes arquivos de textos, dados astronômicos, dados financeiros de alta frequência.
- Perspectiva da Estatística, da Computação e Humana.
- Ciência de Dados: redes neuronais, support vector machines, machine learning, deep learning, classification and regression trees (CART), random forests etc.

Ciência de Dados

- Ciência de Dados(CD) é "filha" da Estatística e da Ciência da Computação.
- Perspectiva não é nova: Tukey (1962): The future of Data Analysis, AMS.
- Cientistas de diversas disciplinas estão sendo confrontados com conjuntos enormes de dados: sequenciamento genético, grandes arquivos de textos, dados astronômicos, dados financeiros de alta frequência.
- Perspectiva da Estatística, da Computação e Humana.
- Ciência de Dados: redes neuronais, support vector machines, machine learning, deep learning, classification and regression trees (CART), random forests etc.

CD: Perspectiva da Estatística

- Estatística "serve" a Ciência guiando na coleta e análise de dados.
- Dados envolvem incertezas: como foram coletados, medidos ou como foram gerados. A modelagem estatística ajuda a quantificar e racionalizar incertezas de maneira sistemática.
- Conjuntos de dados são complexos: tipos diferentes de dependência (ao longo do tempo, sobre escalas espaciais, entre variáveis diferentes)
- Dados de alta dimensão: medimos milhares de variáveis para cada unidade amostral.

CD: Perspectiva da Estatística

- Estatística "serve" a Ciência guiando na coleta e análise de dados.
- Dados envolvem incertezas: como foram coletados, medidos ou como foram gerados. A modelagem estatística ajuda a quantificar e racionalizar incertezas de maneira sistemática.
- Conjuntos de dados são complexos: tipos diferentes de dependência (ao longo do tempo, sobre escalas espaciais, entre variáveis diferentes)
- Dados de alta dimensão: medimos milhares de variáveis para cada unidade amostral.

CD: Perspectiva da Estatística

- Estatística "serve" a Ciência guiando na coleta e análise de dados.
- Dados envolvem incertezas: como foram coletados, medidos ou como foram gerados. A modelagem estatística ajuda a quantificar e racionalizar incertezas de maneira sistemática.
- Conjuntos de dados são complexos: tipos diferentes de dependência (ao longo do tempo, sobre escalas espaciais, entre variáveis diferentes)
- Dados de alta dimensão: medimos milhares de variáveis para cada unidade amostral.

CD: Perspectiva da Estatística

- Estatística "serve" a Ciência guiando na coleta e análise de dados.
- Dados envolvem incertezas: como foram coletados, medidos ou como foram gerados. A modelagem estatística ajuda a quantificar e racionalizar incertezas de maneira sistemática.
- Conjuntos de dados são complexos: tipos diferentes de dependência (ao longo do tempo, sobre escalas espaciais, entre variáveis diferentes)
- Dados de alta dimensão: medimos milhares de variáveis para cada unidade amostral.

CD: Perspectiva da Computação

- Particularmente importante na análise de dados contemporâneos, onde frequentemente nos deparamos com a dicotomia entre acurácia e precisão estatística e recursos computacionais (tempo e memória).
- Exemplos: otimização, bootstrap, MCMC.
- Distribuição de conjuntos de dados enormes por múltiplos processadores (velocidade) e múltiplos equipamentos de armazenamento (memória).

CD: Perspectiva da Computação

- Particularmente importante na análise de dados contemporâneos, onde frequentemente nos deparamos com a dicotomia entre acurácia e precisão estatística e recursos computacionais (tempo e memória).
- Exemplos: otimização, bootstrap, MCMC.
- Distribuição de conjuntos de dados enormes por múltiplos processadores (velocidade) e múltiplos equipamentos de armazenamento (memória).

CD: Perspectiva da Computação

- Particularmente importante na análise de dados contemporâneos, onde frequentemente nos deparamos com a dicotomia entre acurácia e precisão estatística e recursos computacionais (tempo e memória).
- Exemplos: otimização, bootstrap, MCMC.
- Distribuição de conjuntos de dados enormes por múltiplos processadores (velocidade) e múltiplos equipamentos de armazenamento (memória).

CD: Perspectiva Humana

- CD liga modelos estatísticos e métodos computacionais para resolver problemas específicos de outras disciplinas.
- Entender o domínio de um problema, decidir quais dados obter, como processá-los, explorar e visualizar os dados, selecionar um modelo estatístico e métodos computacionais apropriados, comunicar os resultados da análise.
- Estas habilidades não são usualmente ensinadas em disciplinas tradicionais de Estatística ou Computação, mas são adquiridas por meio da experiência e colaboração com outros pesquisadores.

CD: Perspectiva Humana

- CD liga modelos estatísticos e métodos computacionais para resolver problemas específicos de outras disciplinas.
- Entender o domínio de um problema, decidir quais dados obter, como processá-los, explorar e visualizar os dados, selecionar um modelo estatístico e métodos computacionais apropriados, comunicar os resultados da análise.
- Estas habilidades não são usualmente ensinadas em disciplinas tradicionais de Estatística ou Computação, mas são adquiridas por meio da experiência e colaboração com outros pesquisadores.

CD: Perspectiva Humana

- CD liga modelos estatísticos e métodos computacionais para resolver problemas específicos de outras disciplinas.
- Entender o domínio de um problema, decidir quais dados obter, como processá-los, explorar e visualizar os dados, selecionar um modelo estatístico e métodos computacionais apropriados, comunicar os resultados da análise.
- Estas habilidades não são usualmente ensinadas em disciplinas tradicionais de Estatística ou Computação, mas são adquiridas por meio da experiência e colaboração com outros pesquisadores.

Aprendizado com Estatística

- O Aprendizado com Estatística (AE) pode ser supervisionado ou não supervisionado.
- No AE supervisionado, o objetivo é prever o valor de uma variável resposta (*output*) a partir de variáveis preditoras (*inputs*).
- A variável reposta pode ser quantitativa ou qualitativa. No caso de variáveis respostas quantitativas, um dos modelos estatísticos mais utilizados é o de **regressão**; quando a variável resposta é qualitativa, utilizam-se geralmente modelos de **regressão logística** para a análise.
- Adicionalmente, para variáveis qualitativas (categóricas), com valores em um conjunto finito, os modelos mais comuns são os de classificação, em que a partir de um conjunto $(x_i, y_i), i = 1 \ldots, N$ de dados, chamado de conjunto de treinamento, obtemos, por exemplo, obtemos uma regra de classificação.

Aprendizado com Estatística

- No caso de AE não supervisionado, temos apenas um conjunto de variáveis (*inputs*) e o objetivo é descrever associações e padrões entre essas variáveis. Nesse caso, não há uma variável resposta.
- Um algoritmo de AE não supervisionado pode ter por objetivo aprender a distribuição de probabilidades que gerou os dados, para efeito de estimação de densidades, por exemplo.
- Dentre as técnicas mais utilizadas nesta situação temos a **análise de agrupamentos**, a **análise de componentes principais** e a **análise de componentes independentes** (ambas proporcionando a redução da dimensionalidade dos dados).

Inteligência Artificial

- Inteligência Artificial (IA) é um tópico de extremo interesse e que aparece frequentemente nas mídias escritas e faladas. Normalmente o termo suscita questões do tipo: computadores no futuro vão se tornar inteligentes e a raça humana será substituída por eles? Ou que todos perderemos nossos empregos, por que seremos substituídos por robôs inteligentes? Pelo menos até o presente esses receios são infundados.
- Acredita-se que o artigo de Turing (1950) seja o primeiro a tratar do tema. A primeira frase do artigo diz:

  I propose to consider the question, "Can machines think?
- De modo informal, a IA é um esforço para automatizar tarefas intelectuais usualmente realizadas por seres humanos.
- Jordan (2019). Segundo esse autor, o que é rotulado hoje como IA, nada mais é do que aquilo que chamamos de Aprendizado de Máquina (ML).

Inteligência Artificial

Jordan (2019): Harvard Data Science Review, Issue 1.

- The problem had to do not just with data analysis, but with what database researchers call provenance–broadly, where did data arise, what inferences were drawn from the data, and how relevant are those inferences to the present situation?
- I'm also a computer scientist, and it occurred to me that the principles needed to build planetary-scale inference-and-decision-making systems of this kind, blending computer science with statistics, and considering human utilities, were nowhere to be found in my education.
- It occurred to me that the development of such principle–which will be needed not only in the medical domain but also in domains such as commerce, transportation, and education–were at least as important as those of building AI systems that can dazzle us with their game-playing or sensorimotor skills.
- This new engineering discipline will build on ideas that the preceding century gave substance to, such as information, algorithm, data, uncertainty, computing, inference, and optimization. Moreover, since much of the focus of the new discipline will be on data from and about humans, its development will require perspectives from the social sciences and humanities.

Inteligência Artificial

- While the building blocks are in place, the principles for putting these blocks together are not, and so the blocks are currently being put together in ad-hoc ways.
- Humans are proceeding with the building of societal-scale, inference-and-decision-making systems that involve machines, humans, and the environment.
- Just as early buildings and bridges sometimes fell to the ground—in unforeseen ways and with tragic consequences—many of our early societal-scale inference-and-decision-making systems are already exposing serious conceptual flaws.
- Unfortunately, we are not very good at anticipating what the next emerging serious flaw will be. What we're missing is an engineering discipline with principles of analysis and design.

Inteligência Artificial

- Most of what is labeled AI today, particularly in the public sphere, is actually machine learning (ML), a term in use for the past several decades.
- ML is an algorithmic field that blends ideas from statistics, computer science and many other disciplines to design algorithms that process data, make predictions, and help make decisions.
- The phrase data science emerged to refer to this phenomenon, reflecting both the need of ML algorithms experts to partner with database and distributed-systems experts to build scalable, robust ML systems, as well as reflecting the larger social and environmental scope of the resulting systems.
- This confluence of ideas and technology trends has been rebranded as AI over the past few years.

Inteligência Artificial

- Three types of AI:
  - (i) Human-imitative AI : the artificially-intelligent entity should be one of us, if not physically then at lest mentally;
  - (ii) Intelligence Augmentation (IA): computation and data are used to create services that augment human intelligence and creativity, eg, natural language translation, which augments the ability of a human to communicate;
  - Intelligent Infrastructure (II): a web of computation, data and physical entities exists tha makes human environments more supportive, interesting and safe.
- We are very far from realizing human-imitative AI aspirations, that gives rise to levels of over-exuberance and media attention that is not present in other areas of engineering.
- Sucess in these domains is neither sufficient nor necessary so solve important IA and II problems.

Aprendizado de Máquina=ML

- A IA está intimamente ligada ao desenvolvimento da computação (ou programação de computadores) e até a década de 1980, a IA era entendida como na programação clássica: temos um sistema computacional (SC) (um computador ou um *cluster* de computadores ou nuvem etc.) no qual se alimentam dados e uma regra de cálculo e se obtém uma reposta.
- Exemplo: regressão, usando-se MQ para se obter os EMQ. A regra de cálculo é um algoritmo que resolve o problema e pode ser programado em alguma linguagem (Fortran, C, S etc). A maioria dos pacotes computacionais existentes funciona dessa maneira.
- A partir da década de 1990, o aprendizado de máquina (AM–ML) criou um novo paradigma. A programação clássica não resolve problemas mais complicados, como reconhecimento de imagens, voz, escrita etc.

Aprendizado de Máquiana=ML

- Então a ideia é treinar um SC no lugar de programá-lo. Isso significa que se apresentam muitos exemplos relevantes a determinada tarefa (**dados de treinamento**) ao SC, de modo que esse encontre uma estrutura estatística nesses exemplos, produzindo uma regra automatizada. Ou seja, no AM, a entrada é constituída de dados e respostas, e a saída é uma regra de cálculo. Com um novo conjunto de observações (**dados de teste**) procura-se obter a eficácia do método segundo algum critério.
- Existem atualmente, muitos procedimentos que são usados em AM (ou em AE): SVM (*support vector machines*), métodos baseados em árvores de decisão (árvores, florestas, *bagging*, *boosting*), redes neuronais etc. O objetivo é obter algoritmos que tenham um alto valor preditivo em problemas de regressão, agrupamento, classificação e previsão.
- AM está fortemente relacionado com Estatística Computacional, que também trata de fazer previsões com o auxílio de computador. Tem relação forte com otimização, que fornece métodos, teoria e aplicações a este campo.

Redes Neuronais

- As contribuições pioneiras para a área de Redes Neuronais (RN) (também denominadas redes neurais) foram as de McCulloch e Pitts (1943), que introduziram a ideia de RN como máquinas computacionais, de Hebb (1949), por postular a primeira regra para aprendizado organizado e Rosenblatt (1958), que introduziu o *perceptron*, como o primeiro modelo de aprendizado supervisionado.
- O algoritmo do perceptron (programado para o IBM 704) foi implementado por uma máquina, chamada Mark I, planejada para reconhecimento de imagens. O modelo consiste de uma combinação linear das entradas, incorporando um viés externo. A soma resultante é aplicada a um limitador, na forma de uma função degrau (ou uma sigmóide).
- Se $\mathbf{x} = (+1, x_1, x_2, \ldots, x_p)^\top$ contém as entradas, $\mathbf{w} = (b, w_1, w_2, \ldots, w_p)^\top$ são os pesos, a saída é dada por

$$v = \sum_{i=0}^{p} w_i x_i = \mathbf{w}^\top \mathbf{x}.$$

Redes Neuronais

- Atualmente, a RN mais simples consiste de entradas, de uma camada intermediária escondida e das saídas.
- Sejam $\mathbf{X} = (X_1, \ldots, X_p)^\top, \mathbf{Y} = (Y_1, \ldots, Y_M)^\top$ e $\mathbf{W} = (W_1, \ldots, W_K)^\top$ e sejam, os vetores de pesos $\boldsymbol{\alpha}_j$, $j = 1, \ldots, M$, $\boldsymbol{\beta}_k$, $k = 1, \ldots, K$, de ordens $p \times 1$ e $M \times 1$, respectivamente.
  Essa rede neural simples pode ser representada pelas equações:

$$
\begin{aligned}
Y_j &= f(\alpha_{0j} + \boldsymbol{\alpha}_j^\top \mathbf{X}),\ j = 1, \ldots, M, && (1)\\
W_k &= \beta_{0k} + \boldsymbol{\beta}_k^\top \mathbf{Y},\ k = 1, \ldots, K, && (2)\\
f_k(\mathbf{X}) &= g_k(\mathbf{W}),\ k = 1, \ldots, K. && (3)
\end{aligned}
$$

- A função f é chamada **função de ativação** e geralmente é usada a sigmóide $f(x) = 1/(1 + e^{-x})$.
  Os pesos α_{0j} e β_{0k} têm o mesmo papel de b no perceptron e representam vieses. A saída final é $g_k(\mathbf{W})$. Em problemas de regressãao, $g_k(\mathbf{W}) = W_k$ e em problemas de classificação, $g_k(\mathbf{W}) = e^{W_k} / \sum_i e^{W_i}$, que corresponde a uma logística multidimensional.
  Os Y_j constituem a camada escondida e não são observáveis.

Redes Neuronais

- O ajuste de modelos de RN é feito minimizando a soma dos quadrados dos resíduos, no caso de regressão, onde a minimização é sobre os pesos. No caso de classificação, usamos a taxa de erros de classificaçao. Nos dois casos é usado um algoritmo chamado de backpropagation. É necessário escolher valores iniciais e regularização (usando uma função penalizadora), porque o algoritmo de otimização é não convexo e instável.
- No caso de termos várias camadas intermediárias obtém-se o que é chamado aprendizado profundo (deep learning). A complexidade do algoritmo é proporcional ao número de observações, número de preditores, número de camadas e número de épocas de treinamento. Para detalhes sobre esses tópicos, veja Hastie et al. (2017) e Cholet (2018).
- Leo Breiman (2001) distingue dois paradigmas em modelagem estatística: data model e algorithmic model, onde o segundo engloba os algoritmos usados em ML. Segundo ele, a maioria dos métodos importantes estão na categoria 2.

Redes Neuronais: Perceptron

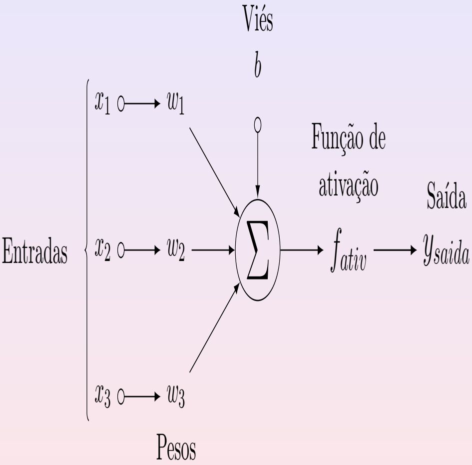
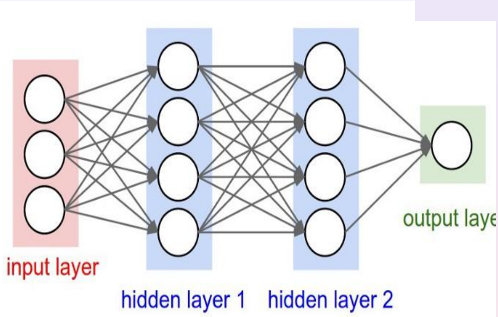


Figura: Perceptron de Rosenblatt

Redes Neuronais



Referências

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science*, **16**, 199–231.

Chambers, J. M. (1993). Greater or lesser Statistics: A choice for future research. *Statistics and Computing*, **3**, 182–184.

Chollet, F. (2018). *Deep Learning with R*. Manning.

Cleveland, W. M. (1985). *The Elements of Graphing Data*. Monterey: Wadsworth.

Cleveland, W. M. (1993). *Visualizing Data*. Summit, New Jersey: Hobart Press.

Cleveland, W. M. (2001). Data Science: An action plan for expanding the technical areas of the field of Statistics. *International Statistical Revue*, **69**, 21–26.

Donoho, D. (2017). 50 years of Data Science. *Journal of Computational and Graphical Statistics*, **26**, 745–766.

Referências

Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning*, 2nd Edition, Springer.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Jordan, M. I. (2019). Artificial inteligence – The revolution hasn't heppened yet. *Harvard Data Science Review*, Issue 1.1.

McCulloch, W. S. and Pitts, W. A. (1943). Logical calculus of the ideas immanent in nervous activity. *Butt. math. Biophysics*, S, 115–133.

Rosenblatt, F. (1958). The perceptron: A theory of statistical separability in cognitive systems. Buffalo: Cornell Aeronautical Laboratory, Inc. Rep. No. VG-1196-G-1.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, **33**, 1–67.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

Turing, A. (1950). Computing machinery and intelligence". *Mind*, LIX (236).

MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 3

20 de março de 2023

Sumário

Notação

1. Matriz de dados $\mathbf{X}$, de ordem $n \times p$; n amostras (indivíduos), p variáveis.

$$\mathbf{X} = [x_{ij}] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ldots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

2. linhas de $\mathbf{X}$: $x_1, \ldots, x_n$; cada x_i é um vetor $p \times 1$;

   colunas de $\mathbf{X}$: $\mathbf{x}_1, \ldots, \mathbf{x}_p$; cada $\mathbf{x}_j$ é um vetor $n \times 1$.
3. Podemos escrever

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p] = \begin{bmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{bmatrix}.$$

4. y_i=i-ésima observação, $\mathbf{y} = (y_1, \ldots, y_n)^\top$. No caso de AE supervisionado, y_i é resposta aos preditores x_i, num problema de regressão, e é o rótulo da i-ésima classe, num problema de classificação.
   Dados: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, cada x_i um vetor $p \times 1$.

Tipos de dados

- grande número de amostras, n, e pequeno número de variáveis, p;
- pequeno número de amostras, n, e grande número de variáveis, p;
- grande número de amostras, n, e grande número de variáveis, p.
- Dados de Alta Dimensão: $n < p$.
- SAS (www.sas.comenus/insightsbig–datawhat–is–big–data.html): "Big data is a term that describes a large volume of data - both structured and unstructured - that inundates a business on a day-to-day basis. But it is not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves."

Dados de alta dimensão

- Alta dimensão relativa: modelos com muita variáveis (p) comparado com o número de amostras (n), mas usualmente com $p < n$;
- Alta dimensão moderada: modelos com número de variáveis proporcional ao número de amostras, usualmente $p > n$;
- Alta dimensão: modelos com mais variáveis do que amostras, e o número de variáveis cresce polinomialmente ou exponencialmente com n.

Tipos de dados

- Dados estruturados: informação organizada que se ajusta a estruturas usuais de bases de dados, relativamente fáceis de armazenar e analisar. Exemplos usuais de dados numéricos ou não, que podem ser dispostos em uma matriz de dados.
- Dados não estruturados: tudo que não se encaixa no item anterior, como arquivos de textos, páginas da *web*, email, midias sociais etc.
- Os quatro V's dos Big Data: VOLUME (escala dos dados); VARIEDADE (formas diferentes de dados); Velocidade (análise de *streaming data*); VERACIDADE (incerteza sobre os dados). (www.ibmbigdatahub.com)

Tipos de dados

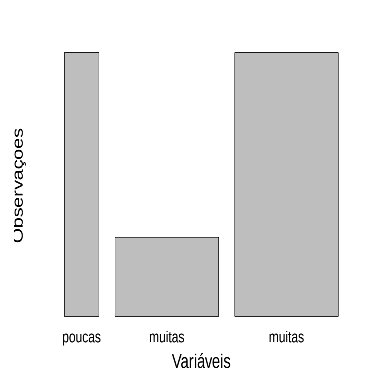


Figura 1: Tipos de dados

AE supervisionado

- Big Data implica em Big Models
- Big model: número grande de parâmetros (p) a serem estimados por algum método estatístico, comparado com o número de observações (n).
- Exemplo: regressão linear

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

- Problema se $p >> n$ (dados de alta dimensão)
- Estimação em modelos lineares com dados de alta dimensão pode ser tratada:
  (a) usando técnicas de redução da dimensão, como por exemplo ACP, AF e ACI;
  (b) usando estimação penalizada (regularização);
  (c) métodos bayesianos;
- No caso de modelos não lineares: árvores de decisão, redes neurais, bagging, boosting.

AE supervisionado: Exemplo 1

- Se quisermos saber se há relação entre o consumo privado (variável C) e renda disponível (variável Y) de indivíduos de uma população, podemos escolher uma amostra de n indivíduos dessa população e medir essas duas variáveis nesses indivíduos, obtendo-se o conjunto de dados $\{(Y_1, C_1), \ldots, (Y_n, C_n)\}$.
- Em Economia, sabe-se, desde Keynes, que o gasto com o consumo de pessoas (C) é uma função da renda pessoal disponível (Y), ou seja
$$C = f(Y),$$
para alguma função f.
- Para se ter uma ideia de como é a função f para essa comunidade, podemos construir um gráfico de dispersão entre Y e C. Com base em um conjunto de dados hipotéticos com $n = 20$, esse gráfico está apresentado na Figura 1 e é razoável postular o modelo
$$C_i = \alpha + \beta Y_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$
em que (Y_i, C_i), $i = 1, \ldots, n$ são os valores de Y e C efetivamente observados e ε_i, $i = 1, \ldots, n$ são variáveis não observadas, chamadas **erros**. Aqui, $n = 20$ e $p = 1$.

AE supervisionado: Exemplo 1

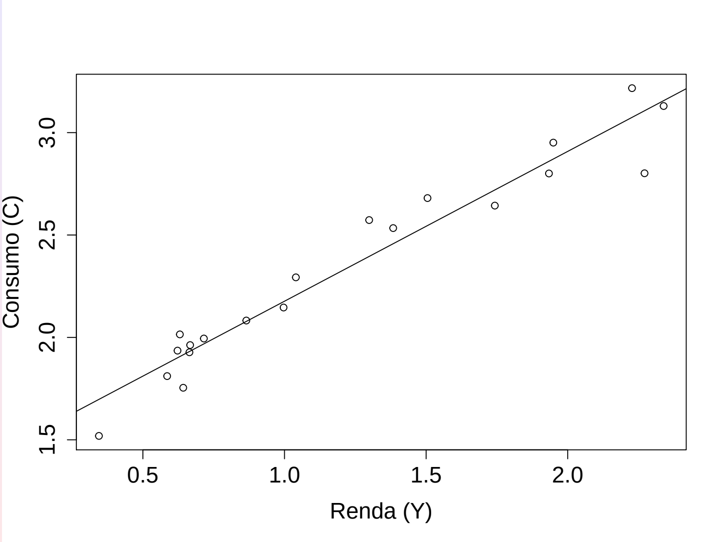


Figura 1: Relação entre renda e consumo de 20 indivíduos.

AE supervisionado: Exemplo 2

- Os dados apresentados na planilha **Esforço** são provenientes de um estudo sobre teste de esforço cardiopulmonar em pacientes com insuficiência cardíaca realizado no InCor da Faculdade de Medicina da USP. Um dos objetivos do estudo é comparar os grupos formados pelas diferentes etiologias quanto às respostas respiratórias e metabólicas obtidas do teste de esforço cardiopulmonar. Outro objetivo do estudo é saber se alguma das características observadas (ou combinação delas) pode ser utilizada como fator prognóstico de óbito.
- Nosso objetivo poderia ser desenvolver um modelo que possa ser usado para prever o consumo de oxigênio **VO2** (variável resposta, y) com as informações sobre **Carga** na esteira ergométrica (x_1) e **IMC** (x_2) (preditores). Um modelo de regressão linear seria

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, \ldots, 127. \tag{3}$$

Aqui, $p = 2$ e teremos ques estimar $\beta_0, \beta_1, \beta_2$ mais os parâmetros associados à variável ε_i. A Figura 2 mostra que o modelo (plano) não parece ser adequado, há vários pontos bastante afastados do plano.

AE supervisionado: Exemplo 2

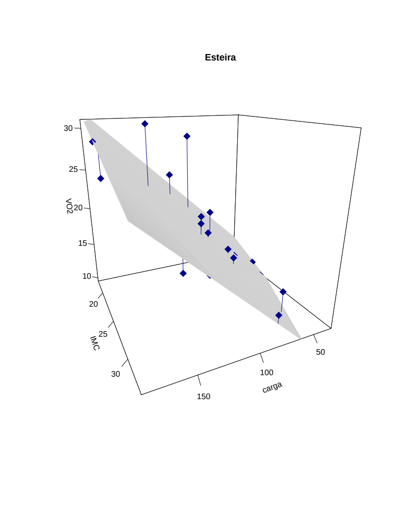


Figura 2: Consumo de oxigênio como função de Carga e IMC.

AE supervisonado: regressão

- Dados o vetor $\mathbf{y}$ de variáveis respostas e os preditores $\mathbf{x}_i$, o modelo geral é da forma

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4}$$

  com $E(\varepsilon_i) = 0$, ε_i ortogonal a $\mathbf{x}_i$ e f desconhecida, chamada de informação sistemática.
- O objetivo do AE é encontrar métodos para estimar f.
- Dois motivos para estimar f: Previsão e Inferência

Previsão

- Obtido o estimador $\hat{f}$, obtemos o previsor de $\mathbf{y}$
$$\hat{\mathbf{y}} = \hat{f}(\mathbf{X}).$$
- A acurácia de $\hat{\mathbf{y}}$ como previsor de $\mathbf{y}$ depende de dois erros:

  erro redutível: introduzido pelo estimador de f; assim chamado porque podemos melhorar a acurácia de $\hat{f}$ usando uma técnica de AE mais apropriada;

  erro irredutível: mesmo usando o melhor estimador de f, esse erro depende de ε, que não pode ser previsto usando $\mathbf{X}$.
- Supondo $\hat{f}$ e $\mathbf{X}$ fixos, pode-se ver que

$$\begin{aligned} E(\mathbf{y} - \hat{\mathbf{y}})^2 &= E[f(\mathbf{X}) + \varepsilon - \hat{f}(\mathbf{X})]^2 \\ &= E[f(\mathbf{X}) - \hat{f}(\mathbf{X})]^2 + \mathrm{Var}(\varepsilon). \qquad (5) \end{aligned}$$

  O primeiro termo do segundo membro representa o erro redutível e, o segundo termo, o erro irredutível. O objetivo é minimizar o erro redutível.

Inferência

O interesse pode não ser fazer previsões, mas entender como a resposta é afetada pela variação dos preditores.
Interesse nos seguintes tópicos:

- Identificar alguns preditores importantes, dentre todos os preditores;
- Relação entre a resposta e cada um dos preditores; no Exemplo 2, y cresce com a carga, mas decresce com o IMC;
- A relação entre a resposta e cada preditor é linear, ou mais complicada? Modelos lineares fornecem interpretações mais simples, mas em geral previsões menos acuradas. Modelos não lineares fornecem previsões mais acuradas, mas o modelo perde em interpretabilidade.

Métodos paramétricos

- Fazemos alguma suposição sobre a forma de f, por exemplo, o modelo de regressão dado em (1). O problema simplifica-se, pois temos que estimar $p+1$ parâmetros.
- Selecionado o modelo, temos que ajustá-lo aos dados de treinamento (*treinar* o modelo). No caso do modelo (1), o método mais usado é Mínimos Quadrados (MQ). Mas há outros métodos, como SVM.
- O ajuste do modelo (1) por MQ pode ser pobre, como no Exemplo 2 (ver Figura 2).
- Nesse caso, pode-se tentar modelos mais flexíveis, escolhendo-se outras formas funcionais para f, incluindo-se modelos não lineares.
- Modelos mais flexíveis envolvem estimar um número muito grande de parâmetros, aparecendo o problema do super-ajustamento (*overfitting*).

Métodos não paramétricos

- Nesse caso, não fazemos nenhuma hipótese sobre a forma funcional de f.
- Como o problema não se reduz a estimar um número pequeno de parâmetros, necessitaremos de um número grande de observações para obter estimadores acurados de f.
- Vários métodos podem ser usados:

  ◇ usando kernels
  ◇ usando polinômios locais (e.g Lowess)
  ◇ usando splines
  ◇ usando polinômios ortogonais (e.g. Chebyshev)
  ◇ usando outras bases ortogonais (e.g. Fourier, ondaletas)

Acurácia e interpretação

- métodos menos flexíveis (e.g regressão linear) (ou mais restritivos) em geral são menos acurados e mais fáceis de interpretar.
- métodos mais flexíveis (e.g splines) são mais acurados e mais difíceis de interpretar.
- Para cada conjunto de dados, um método pode ser preferível a outros.
- Escolha do método é a parte mais difícil do AE/ML.

Qualidade do ajuste

- No caso de regressão, a medida de ajuste mais usada é o Erro Quadrático Médio (EQM), dado por

$$\text{EQM} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2, \tag{6}$$

  onde $\hat{f}(x_i)$ é o preditor de y para a i-ésima observação.
- O EQM acima é calculado no conjunto de treinamento que produz $\hat{f}$. É chamado *EQM de treinamento*.
- Todavia, estamos mais interessados na acurácia do ajuste para os dados de teste.
- Se tivermos um grande número de dados de teste, poderemos calcular o *EQM de teste*,

$$\text{Média}(y_0 - \hat{f}(\mathbf{x}_0))^2, \tag{7}$$

  que é o erro de previsão quadrático médio para as observações teste $(\mathbf{x}_0, y_0)$.
- Se não tivermos observações teste use (6). Na maioria dos casos o EQM de treinamento é menor que o EQM de teste.
- Para calcular o EQM de treinamento, usa-se CV.

Viés/variância

- Para um dado $(\mathbf{x}_0, y_0)$,

$$E[y_0 - \hat{f}(\mathbf{x}_0)]^2 = \mathrm{Var}[\hat{f}(\mathbf{x}_0)] + [\mathrm{Vies}(\hat{f}(\mathbf{x}_0))]^2 + \mathrm{Var}(\varepsilon). \tag{8}$$

- Para minimizar (8), selecionamos um método que simultaneamente tiver baixo viés e baixa variância. O EQM de teste, em geral, apresenta uma forma de U, resultante da competição entre viés e variância.
- Métodos de AE mais flexíveis têm viés baixo e variância grande.
- Na prática, f não é conhecido e não é possível calcular o EQM de teste, viés e variância para um método de AE.

Classificação

- No caso de respostas $y_1, \ldots, y_n$ qualitativas temos um problema de **classificação**.
- Formalmente, seja $(\mathbf{x}, y)$, de modo que $\mathbf{x} \in \mathbb{R}^p$ e $y \in \{-1, 1\}$. Então, um **classificador** é uma função $g : \mathbb{R}^p \to \{-1, 1\}$ e a **função erro** ou **risco** é a probabilidade de erro, $L(g) = P\{g(X) \neq Y\}$.
- Obtendo-se um estimador de g, digamos $\hat{g}$, sua acurácia pode ser medida pelo estimador de $L(g)$, chamado de **taxa de erro de treinamento**, que é a proporção de erros gerados pela aplicação de $\hat{g}$ às observações de treinamento, ou seja,

$$\hat{L}(\hat{g}) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i), \tag{9}$$

sendo que $\hat{y}_i = \hat{g}(x_i)$ é o rótulo (-1 ou 1) da classe prevista usando $\hat{g}$.

Classificador de Bayes

- O interesse está na **taxa de erro de teste**

$$\text{Média}(I(y_0 \neq \hat{y}_0), \tag{10}$$

  para observações de teste $(\mathbf{x}_0, y_0)$. Um bom classificador tem (10) pequeno.

- Pode-se provar que (10) é minimizado, em média, por um classificador que associa cada observação à classe mais provável, dados os preditores; ou seja, temos que maximizar

$$P(y = j|\mathbf{x} = \mathbf{x}_0). \tag{11}$$

  Tal classificador é chamado de Bayes.

- No caso de duas classes, classificar na classe -1 se $P(y = -1|\mathbf{x} = \mathbf{x}_0) > 0,5$ e na classe 1 c.c. O classificador de Bayes produz a menor taxa de erro e será dada por $1 - \max_j P(y = j|\mathbf{x} = \mathbf{x}_0)$. A taxa de erro de Bayes global é dada por $1 - E(\max_j P(y = j|\mathbf{x} = \mathbf{x}_0))$, onde $E(\cdot)$ é caculada sobre todos os valores de $\mathbf{x}$.

K-ésimo vizinho mais próximo

- O classificador de Bayes não pode ser calculado na prática, pois não conhecemos a distribuição condicional de y dado $\mathbf{x}$.
- Uma possibilidade é estimar a distribuição condicional e, então, estimar (11).
- O classificador K-ésimo vizinho mais próximo (K-nearest neighbors, KNN) estima tal distribuição da seguinte maneira:
  - (i) Escolha $K > 0$ inteiro e uma observação teste $\mathbf{x}_0$.
  - (ii) O classificador KNN primeiro identifica os K pontos do conjunto de treinamento mais próximos de $\mathbf{x}_0$; chame-os de $\mathcal{N}$.
  - (iii) Estime a probabilidade condicional da classe j por

$$P(y = j|\mathbf{x} = \mathbf{x}_0) = \frac{1}{K} \sum_{i \in \mathcal{N}} I(y_i = j). \tag{12}$$

  - (iv) Classifique $\mathbf{x}_0$ na classe com a maior probabilidade condicional.
- Escolha de K crucial; resultado depende dessa escolha.

Modelos para séries temporais

(1) Consideremos uma série temporal multivariada $\mathbf{z}_t, t = 1, \ldots, T$ em que $\mathbf{z}_t$ contém valores de d variáveis e uma série temporal univariada Y_t, $t = 1, \ldots, T$. O objetivo é fazer previsões de Y_t, para horizontes $h = 1, \ldots, H$, com base nos valores passados de Y_t e de $\mathbf{z}_t$. Uma suposição básica é que o processo $\{Y_t, \mathbf{z}_t\}$, $t \geq 1$ seja estacionário fraco (ou de segunda ordem), veja Morettin (2017), com valores em $\mathbb{R}^{d+1}$. Para $p \geq 1$, consideramos o processo vetorial n-dimensional

$$\mathbf{x}_t = (Y_{t-1}, \ldots, Y_{t-p}, \mathbf{z}_t^\top, \ldots, \mathbf{z}_{t-r}^\top)^\top, \quad \text{com } n = p + d(r+1).$$

(2) O modelo a considerar é uma extensão do modelo (4), ou seja,

$$Y_t = f(\mathbf{x}_t) + e_t, \quad t = 1, \ldots, T. \tag{13}$$

O modelo geral

(3) Como em (4), f é uma função desconhecida e e_t tem média zero e variância finita. O objetivo é estimar f e usar o modelo para fazer previsões para um horizonte h por meio de

$$Y_{t+h} = f_h(\mathbf{x}_t) + e_{t+h}, \quad h = 1, \ldots, H, \; t = 1, \ldots, T.$$

Para uma avaliação do método de previsão, a acurácia é

$$\Delta_h(\mathbf{x}_t) = |\widehat{f}_h(\mathbf{x}_t) - f_h(\mathbf{x}_t)|.$$

(4) Em geral, a norma L_q é $E\{|\Delta_h(\mathbf{x}_t)|\}^q$, sendo que as mais comumente usadas consideram $q = 1$ (correspondente ao erro de previsão absoluto médio) ou $q = 2$ (correspondente ao errode previsão quadrático médio). A raiz quadrada dessas medidas também é utilizada.
Escolhendo-se uma função perda, o objetivo é selecionar f_h a partir de um conjunto de modelos que minimize o risco, ou seja, o valor esperado da norma L_q.

Modelos lineares

(1) Modelos frequentemente usados têm a forma (13) em que $f(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$, com $\boldsymbol{\beta}$ denotando um vetor de $\mathbb{R}^n$. Estimadores de mínimos quadrados não são únicos se $n > T$. A ideia é considerar modelos lineares com alguma função de penalização, ou seja que minimizem

$$Q(\boldsymbol{\beta}) = \sum_{t=1}^{T-h} (Y_{t+h} - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + p(\boldsymbol{\beta}),$$

em que $p(\boldsymbol{\beta})$ depende, além de $\boldsymbol{\beta}$, de $\mathbf{Z}_t$, de um parâmetro de suavização λ e de eventuais hiperparâmetros.

(2) Este processo denomina-se **regularização**. Há várias formas de regularização como aquelas conhecidas por *Ridge*, *Lasso*, *Elastic Net* e generalizações. Detalhes sobre esses processos de regularização são apresentados no Capítulo 8.

Modelos não lineares

(1) Num contexto mais geral, o objetivo é minimizar

$$S(f) = \sum_{t=1}^{T-h} [Y_{t+h} - f(\mathbf{x}_t)]^2,$$

para f pertencendo a algum espaço de funções $\mathcal{H}$.

(2) Por exemplo, podemos tomar f como uma função contínua, com derivadas contínuas, ou f apresentando alguma forma de descontinuidade etc. Esses espaços, são, em geral, de dimensão infinita e a solução pode ser complicada. Para contornar esse problema, pode-se considerar uma coleção de espaços de dimensão finita $\mathcal{H}_d$, para $d = 1, 2, \ldots$, de tal sorte que $\mathcal{H}_d$ convirja para $\mathcal{H}$ segundo alguma norma. Esses espaços são denominados **espaços peneira** (*sieve spaces*).

(3) Para cada d, consideramos a aproximação

$$h_d(\mathbf{x}_t) = \sum_{j=1}^{J} \beta_j h_j(\mathbf{x}_t),$$

em que $h_j(\cdot)$ são **funções base** para $\mathcal{H}_d$ e tanto J como d são funções de T. Podemos usar *splines*, polinômios, funções trigonométricas, ondaletas etc. como funções base.

Modelos não lineares

(4) Se as funções base são conhecidas, elas são chamadas *linear sieves* e se elas dependem de parâmetros a estimar, são chamadas *nonlinear sieves*.

(5) Exemplos de non linear sieves são as árvores de decisão e as redes neurais.

(6) Métodos em AE ou ML são dedicados a observações de variáveis independentes e identicamente distribuídas. O caso de séries temporais é mais complicado e foge ao escopo deste curso. Apresentaremos apenas algumas ideias relacionadas a esse tópico.

Referências

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *An Introduction to Statistical Learning*. Springer.

Morettin, P. A. e Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC.

MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 4

27 de março de 2023

Sumário

Modelos de regressão linear

1. Um dos modelos estatísticos mais usados na prática: **modelo de regressão**.
2. Exemplo mais simples: dados $(x_1, y_1), \ldots, (x_n, y_n)$ de duas variáveis contínuas X e Y num contexto em que sabemos a priori que a distribuição de probabilidades de Y pode depender de X, ou seja, X é a variável explicativa (ou preditora) e Y é a variável resposta.
3. Trata-se de um **modelo linear**, ou seja, os parâmetros aparecem no modelo de forma linear.
4. Trata-se, também, de um **modelo paramétrico**.

Regressão linear simples

- **Exemplo**: objetivo é avaliar como a distância com que indivíduos conseguem distinguir um determinado objeto (doravante indicada simplesmente como distância) varia com a idade.
- Figura 1: gráfico de dispersão: tendência decrescente da distância com idade.
- modelo: $y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n$
- α, β **parâmetros**
- modelo: regressão linear simples – RLS
- mais adequado: $y_i = \alpha + \beta(x_i - 18) + e_i, \quad i = 1, \ldots, n.$

Regressão linear simples

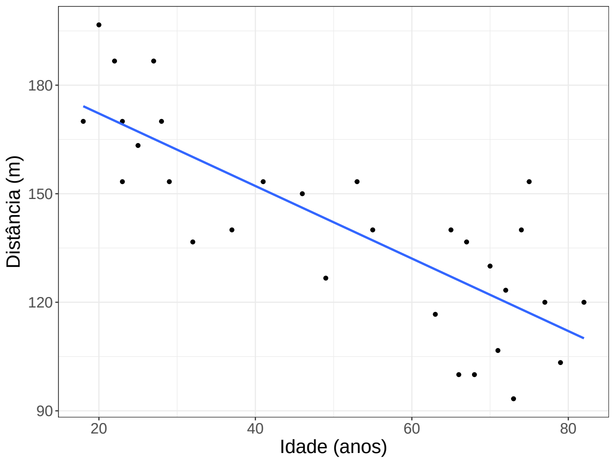


Figura 1: Gráfico de dispersão para os dados **distância**

Estimação do modelo RLS

- SQE: $Q(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$.
- Estimadores de mínimos quadrados (EMQ): minimizam a SQE.
- $$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{1}$$

  e

  $$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \tag{2}$$

  em que $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$ e $\overline{y} = n^{-1} \sum_{i=1}^{n} y_i$
- Um estimador de σ^2 é

  $$S^2 = \frac{1}{n-2} Q(\widehat{\alpha}, \widehat{\beta}) = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{e}_i^2 = \sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2, \tag{3}$$

  em que onde $Q(\widehat{\alpha}, \widehat{\beta})$ é a **soma dos quadrados dos resíduos**, abreviadamente, *SQRes*.

Uso do R para o ajuste

- função lm() do pacote MASS
- modelo: $\text{distancia}_i = \alpha + \beta(\text{idade}_i - 18) + e_i, \quad i = 1, \ldots, n$
  $\hat{y}_i = 174.23 - 1.004(x_i - 18).$
- modelo: $\text{distancia}_i = \alpha + \beta\text{idade}_i + e_i, \quad i = 1, \ldots, n$
  $\hat{y}_i = 192.3 - 1.004x_i,$
- Residual standard error: 16.6 on 28 degrees of freedom
- Multiple R-squared: 0.6424, Adjusted R-squared: 0.6296

RLS-Avaliação do ajuste

- Uma vez ajustado o modelo, convém avaliar a qualidade do ajuste e um dos indicadores mais utilizados para essa finalidade é o **coeficiente de determinação** definido como

$$R^2 = \frac{SQTot - SQRes}{SQTot} = \frac{SQReg}{SQTot} = 1 - \frac{SQRes}{SQTot}$$

  em que a soma de quadrados total é $SQTot = \sum_{i=1}^{n}(y_i - \overline{y})^2$, a soma de quadrados dodos resíduos é $SQRes = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ e a soma de quadrados da regressão é $SQReg = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$.
- Esse coeficiente mede a porcentagem da variação total dos dados (em relação à sua média) explicada pelo modelo de regressão. O coeficiente de determinação deve ser acompanhado de outras ferramentas para a avaliação do ajuste, pois não está direcionado para identificar se todas as suposições do modelo são compatíveis com os dados sob investigação. Em particular, mencionamos os gráficos de resíduos, gráficos de Cook e gráficos de influência local.

R^2 ajustado

- Toda a vez que incluimos um preditor ao modelo, o R^2 aumenta; nunca decresce. Consequentemente, um modelo com mais parâmetros pode parecer que tenha o melhor ajuste (sobreajuste).
- O R^2 ajustado cresce somente quando o preditor incluído realmente aumenta o poder preditivo do modelo de regressão.
- O R^2 ajustado é calculado por

$$\overline{R}^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right],$$

em que n é o tamanho da amostra e k o número de preditores do modelo.

RLS-Gráfico de resíduos

O gráfico de resíduos correspondente ao modelo ajustado aos dados **distancia** está apresentado na Figura 2.

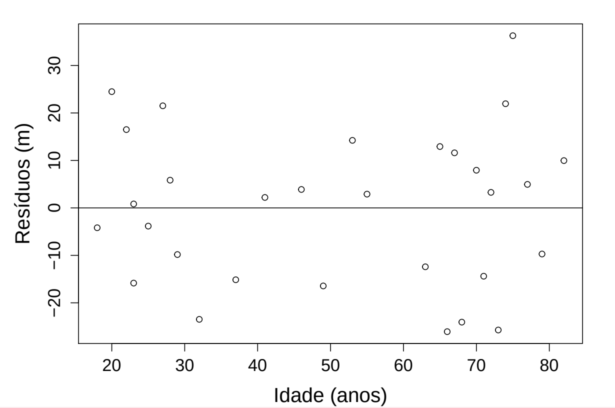


Figura 2: Gráfico de resíduos para o ajuste do modelo de regressão linear simples aos dados **distância**.

RLS-resíduos padronizados

- Para facilitar a visualização em relação à dispersão dos resíduos e para efeito de comparação entre ajustes de modelos em que as variáveis resposta têm unidades de medida diferentes, convém padronizá-los, *i.e.*, dividi-los pelo respectivo desvio padrão para que tenham variância igual a 1.
- Como os resíduos (ao contrário dos erros) são correlacionados, pode-se mostrar que

$$DP(\widehat{e}_i) = \sigma\sqrt{1 - h_{ii}} \quad \text{com} \quad h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2},$$

de forma que os **resíduos padronizados**, também chamados de **resíduos studentizados** são definidos por

$$\widehat{e}_i^* = \widehat{e}_i / (S\sqrt{1 - h_{ii}}) \tag{4}$$

- Os resíduos padronizados são adimensionais e têm variância igual a 1, independentemente da variância da variável resposta. Além disso, para erros com distribuição Normal, cerca de 99% dos resíduos padronizados têm valor entre -3 e +3.
- h_{ii} : leverage (alavancagem) do i-ésimo preditor.

RLS- Resíduos padronizados

O gráfico de resíduos padronizados correspondente àquele da Figura 2 está apresentado na Figura 3.

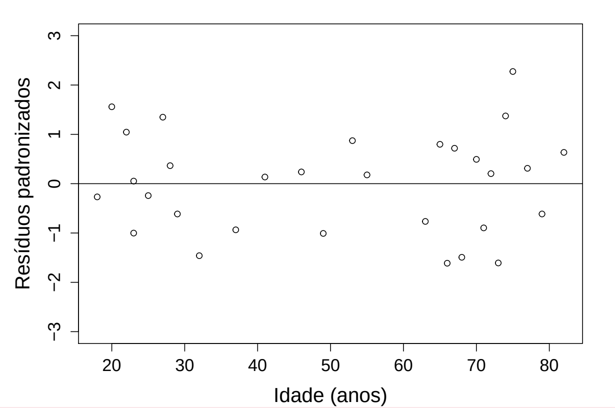


Figura 3: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear simples aos dados **distancia**.

RLS-Distância de Cook

Exemplo: Consideremos agora os dados (hipotéticos) dispostos na Tabela 1, aos quais ajustamos um modelo de regressão linear simples.

Tabela 1: Dados hipotéticos

X	10	8	13	9	11	14	6	4	12	7	5	18
Y	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68	6,31

RLS-Distância de Cook

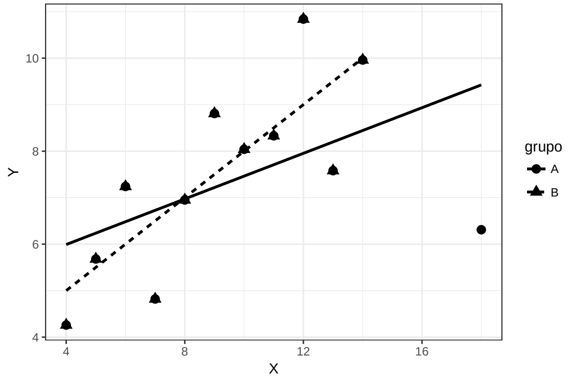


Figura 4: Gráfico de dispersão (com retas de regressão sobrepostas) para os dados da Tabela 1; curva sólida para dados completos e curva interrompida para dados com ponto influente eliminado.

RLS-Distância de Cook

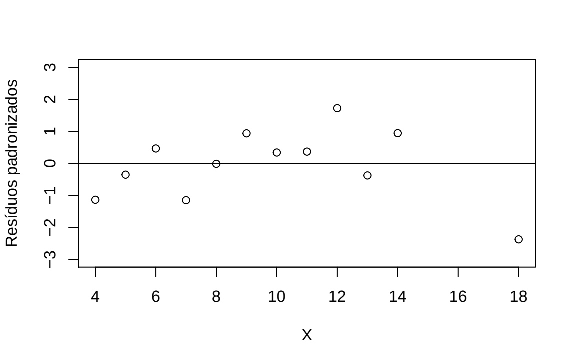


Figura 5: Gráfico de resíduos padronizados para o ajuste do modelo de regressão linear aos dados da Tabela 1.

RLS-Distância de Cook

- A distância de Cook é uma maneira de identificar pontos influentes (outliers) em um conjunto de preditores, que afetam o modelo. É uma combinação da alavancagem de cada observação e dos resíduos. Quanto maior a alavancagem, maior é a distância de Cook.
- Denotando por $\widehat{\mathbf{y}}$ o vetor (de dimensão n) com os valores preditos obtidos do ajuste do modelo baseado nas n observações e por $\widehat{\mathbf{y}}^{(-i)}$ o correspondente vetor com valores preditos (de dimensão n) obtido do ajuste do modelo baseado nas $n-1$ observações restantes após a eliminação da i-ésima, a **distância de Cook** é definida como

$$D_i = \frac{(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(-i)})^\top (\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(-i)})}{(p+1)S}$$

  em que p é o número de coeficientes de regressão e S é uma estimativa do desvio padrão.
- Pode-se mostrar que a distância de Cook (D_i) pode ser calculada sem a necessidade de ajustar o modelo com a omissão da i-ésima observação por meio da expressão

$$D_i = \frac{1}{p+1} \widehat{e}_i^2 \frac{h_{ii}}{(1-h_{ii})^2},$$

  lembrando que h_{ii} é a leverage.

RLS-Distância de Cook

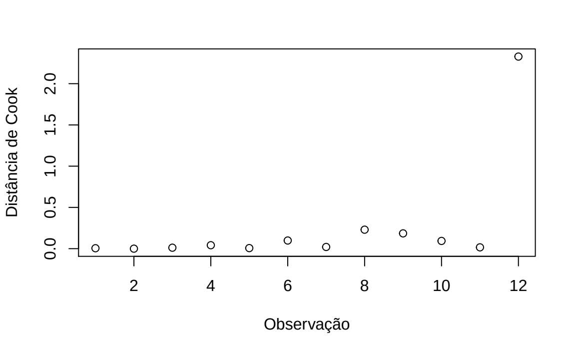


Figura 6: Gráfico de Cook correspondente ao ajuste do modelo de regressão linear aos dados da Tabela 1.

RLS-Gráficos QQ

- Nos casos em que se supõe que os erros têm distribuição Normal, pode-se utilizar gráficos QQ (quantis–quantis)com o objetivo de avaliar se os dados são compatíveis com essa suposição. É importante lembrar que esses gráficos QQ devem ser construídos com os quantis amostrais baseados nos resíduos e não com as observações da variável resposta, pois apesar de suas distribuições também serem normais, suas médias variam com os valores associados da variável explicativa, ou seja, a média da variável resposta correspondente a y_i é $\alpha + \beta x_i$.
- Convém observar que sob normalidade dos erros, os resíduos padronizados seguem uma distribuição t com $n - 2$ graus de liberdade e é dessa distribuição que se devem obter os quantis teóricos para a construção do gráfico QQ. Também deve-se lembrar que para valores de n maiores que 20 ou 30, os quantis da distribuição t se aproximam daqueles da distribuição Normal, tornando-as intercambiáveis para a construção do correspondente gráfico QQ.

RLS-Gráficos QQ

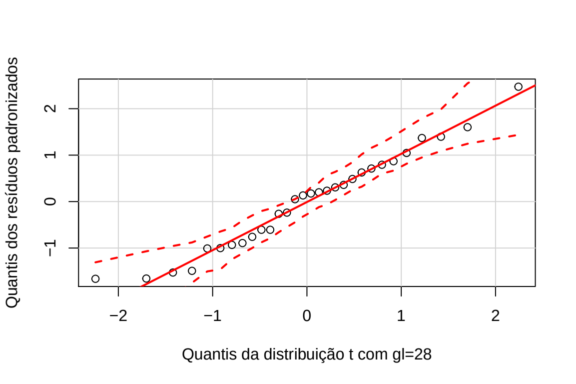


Figura 7: Gráfico QQ correspondente ajuste do modelo de regressão linear aos dados **distancia**.

RLS-Gráficos QQ

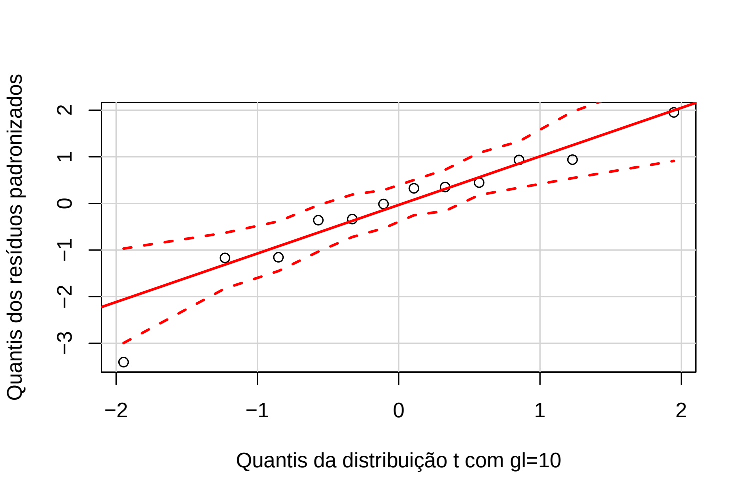


Figura 8: Gráfico QQ correspondente ajuste do modelo de regressão linear aos dados da Tabela 1 (com todas as observações).

RLS-Gráficos QQ

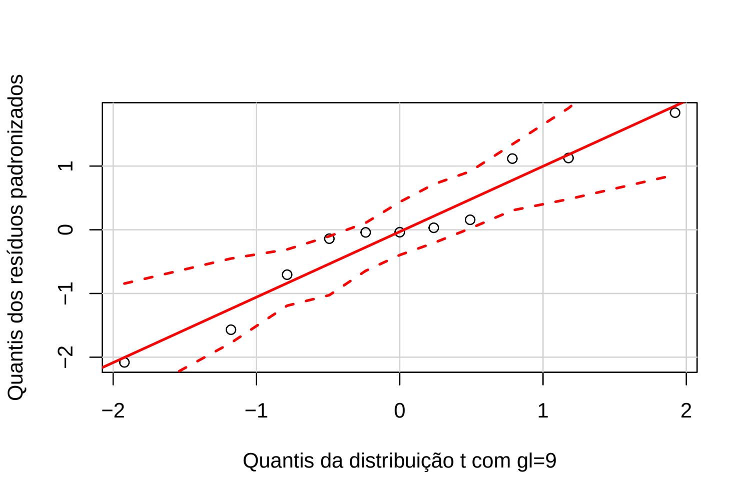


Figura 9: Gráfico QQ correspondente ajuste do modelo de regressão linear aos dados da Tabela 1 (sem a observação influente).

RLS-Dados correlacionados

Exemplo: Na Tabela 2 apresentamos valores do peso de um bezerro observado a cada duas semanas após o nascimento com o objetivo de avaliar seu crescimento nesse período. O gráfico de dispersão correspondente está disposto na Figura 10.

Tabela 2: Peso (kg) de um bezerro nas primeiras 26 semanas após o nascimento

Semana	Peso
0	32,0
2	35,5
4	39,2
6	43,7
8	51,8
10	63,4
12	76,1
14	81,1
16	84,6
18	89,8
20	97,4
22	111,0
24	120,2
26	134,2

RLS-Dados correlacionados

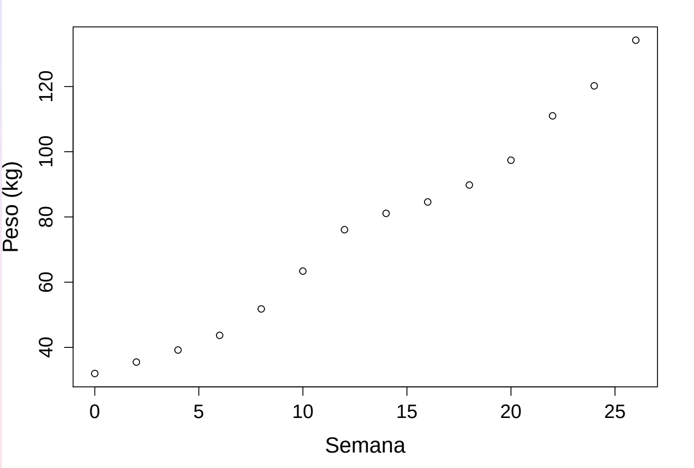


Figura 10: Gráfico de dispersão para os dados da Tabela 2

Dados correlacionados

- Tendo em vista o gráfico de dispersão, um possível modelo seria

$$y_t = \alpha + \beta t + \gamma t^2 + e_t, \tag{5}$$

  $i = 1, \ldots, 14$ em que y_t representa o peso do bezerro no instante t, α denota o valor esperado de seu peso ao nascer, β e γ representam os componentes linear e quadrático da curva que rege a variação temporal do peso no intervalo de tempo estudado e e_t denota um erro aleatório. Utilizamos t como índice para salientar que as observações são colhidas sequencialmente ao longo do tempo.
- O coeficiente de determinação ajustado, $R^2_{aj} = 0,987$ indica que o ajuste (por mínimos quadrados) do modelo com $\widehat{\alpha} = 29,9$ $(2,6)$, $\widehat{\beta} = 2,7$ $(2,5)$ e $\widehat{\gamma} = 0,05$ $(0,02)$ é excelente (sob essa ótica, obviamente).
- Por outro lado, o gráfico de resíduos apresentado na Figura 11 mostra sequências de resíduos positivos seguidas de sequências de resíduos negativos, sugerindo uma possível correlação positiva entre eles (autocorrelação).

RLS-Dados correlacionados

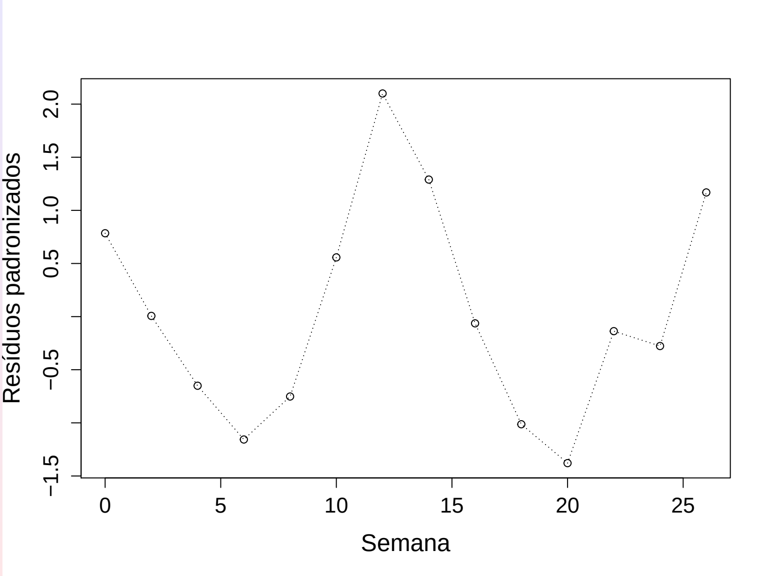


Figura 11: Resíduos studentizados obtidos do ajuste do modelo (5)

RLS-Dados correlacionados

- Uma maneira de contornar esse problema, é modificar os componentes aleatórios do modelo para incorporar essa possível autocorrelação nos erros. Nesse contexto, podemos considerar o modelo (5) com

$$e_t = \rho e_{t-1} + u_t, \quad t = 1, \ldots, n \tag{6}$$

em que $u_t \sim N(0, \sigma^2)$, $t = 1, \ldots, n$, independentes e e_0 é uma constante (geralmente igual a zero). Essas suposições implicam que $\text{Var}(e_t) = \sigma^2/(1-\rho^2)$ e que $\text{Cov}(e_t, e_{t-s}) = \rho^s[\sigma^2/(1-\rho^2)]$.

- Para testar a hipótese de que os erros são não correlacionados pode-se utilizar a **estatística de Durbin-Watson**:

$$D = \sum_{t=2}^{n} (\widehat{e}_t - \widehat{e}_{t-1})^2 / \sum_{t=1}^{n} \widehat{e}_t^2, \tag{7}$$

em que $\widehat{e}_t$, $t = 1, \ldots, n$ são os resíduos obtidos do ajuste do modelo (5) por mínimos quadrados.

RLS-Dados correlacionados

- Expandindo (7) podemos verificar que

$$D \approx 2 - 2\frac{\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1}}{\sum_{t=1}^{n} \widehat{e}_t^2}, \tag{8}$$

- Se os resíduos não forem correlacionados, então $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx 0$ e consequentemente, $D \approx 2$; se, por outro lado, os resíduos forem altamente correlacionados, esperamos que $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx \sum_{t=2}^{n} \widehat{e}_t^2$ e então $D \approx 0$; finalmente, se os resíduos tiverem uma grande correlação negativa, esperamos que $\sum_{t=2}^{n} \widehat{e}_t \widehat{e}_{t-1} \approx -\sum_{t=2}^{n} \widehat{e}_t^2$ e nesse caso, $D \approx 4$.
- Durbin and Watson (1950), Durbin and Watson (1951) e Durbin and Watson (1971) produziram tabelas da distribuição da estatística D que podem ser utilizados para avaliar a suposição de que os erros são não correlacionados.
- O valor da estatística de Durbin-Watson para os dados do Exemplo sob o modelo (5) é $D = 0,91$ ($p < 0,0001$), sugerindo um alto grau de autocorrelação dos resíduos. Uma estimativa do coeficiente de autocorrelação ρ é 0,50. Nesse caso, o modelo (5) - (6) poderá ser ajustado pelo **método de mínimos quadrados generalizados** ou por métodos de **Séries Temporais**.

RLS - Inferência

a) $E(\widehat{\alpha}) = \alpha$ e $E(\widehat{\beta}) = \beta$, ou seja, os EMQ são não enviesados.

b) $\mathrm{var}(\widehat{\alpha}) = \sigma^2 \sum_{i=1}^{n} x_i^2 / [n \sum_{i=1}^{n} (x_i - \overline{x})]^2$.

c) $\mathrm{var}(\widehat{\beta}) = \sigma^2 / \sum_{i=1}^{n} (x_i - \overline{x})^2$.

d) $\mathrm{cov}(\widehat{\alpha}, \widehat{\beta}) = -\sigma^2 \overline{x} / \sum_{i=1}^{n} (x_i - \overline{x})^2$.

RLS - Inferência

Com a suposição adicional de normalidade, pode-se mostrar que

e) $y_i \sim N(\alpha + \beta x_i, \sigma^2)$

f) as estatísticas

$$t_{\widehat{\alpha}} = \frac{\widehat{\alpha} - \alpha}{S}\sqrt{\frac{n\sum(x_i - \overline{x})^2}{\sum x_i^2}}$$

e

$$t_{\widehat{\beta}} = \frac{\widehat{\beta} - \beta}{S}\sqrt{\sum(x_i - \overline{x})^2}$$

têm distribuição t de Student com $(n-2)$ graus de liberdade. Nesse contexto, os resíduos padronizados também seguem uma distribuição t de Student com $(n-2)$ graus de liberdade. Daí a denominação alternativa de resíduos studentizados

g) Com esses resultados é possível testar as hipóteses $H_0 : \alpha = 0$ e $H_0 : \beta = 0$, bem como construir intervalos de confiança para esses parâmetros.

RLS - Inferência

- **Teorema de Gauss-Markov**: EMQ têm variância mínima na classe dos estimadores não enviesados que sejam funções lineares das observações y_i (que não depende da suposição de normalidade dos erros).
- Quando os erros não seguem uma distribuição Normal, mas o tamanho da amostra é suficientemente grande, pode-se mostrar com o auxílio do **Teorema Limite Central** que sob certas condições de regularidade (usualmente satisfeitas na prática), os estimadores $\widehat{\alpha}$ e $\widehat{\beta}$ têm distribuições aproximadamente normais com variâncias que podem ser estimadas pelas expressões indicadas nos itens b) e c).

RLS - Previsão

- Um dos objetivos da análise de regressão é fazer previsões sobre a variável resposta com base em valores das variáveis explicativas.
- Uma estimativa para o valor esperado $E(Y|X = x_0)$ da variável resposta Y dado um valor x_0 da variável explicativa é $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x_0$ e com base nos resultados anteriores pode-se mostrar que a variância de $\widehat{y}$ é

$$\text{var}(\widehat{y}) = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right].$$

- Então os limites superior e inferior para um intervalo de confiança aproximado com coeficiente de confiança de 95% para o valor esperado de Y dado $X = x_0$ são

$$\widehat{y} \pm 1,96S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

com S^2 denotando uma estimativa de σ^2. Podemos dizer que esse intervalo deve conter o verdadeiro valor esperado de $E(Y|X = x)$, *i.e.*, a média de Y para todas as observações em que $X = x_0$, com coeficiente de confiança de 95%.

RLS - Previsão

Isso não significa que esperamos que o intervalo contenha o verdadeiro valor de Y, digamos Y_0 para uma unidade de investigação para a qual $X = x_0$. Nesse caso precisamos levar em conta a variabilidade de $Y|X = x_0$ em torno de seu valor esperado $E(Y|X = x_0)$.

Como $Y_0 = \widehat{y} + e_0$ sua variância é

$$\text{var}(Y_0) = \text{var}(\widehat{y}) + \text{var}(e_0) = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x - \overline{x})^2}\right] + \sigma^2$$

Então os limites superior e inferior de um **intervalo de previsão** (aproximado) para Y_0, com $\gamma = 95\%$, são

$$\widehat{y} \pm 1,96S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

Note que se aumentarmos indefinidamente o tamanho da amostra, a amplitude do intervalo de confiança para o valor esperado tenderá para zero, porém a amplitude do intervalo de previsão correspondente a uma unidade específica tenderá para $2 \times 1,96 \times \sigma$.

Referências

Morettin, P. A. and Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC, Rio de Janeiro.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Apêndice: Quantis empíricos

- Suponha que a v.a. X tenha distribuição contínua, com f.d.a. F. Então, para $0 \leq p \leq 1$, o p - quantil de F é o valor Q_p satisfazendo $F(Q_p) = p$, ou seja,

$$F(Q_p) = P(X \leq Q_p) = p.$$

Se existir a inversa de F, então $Q_p = F^{-1}(p)$. No caso de X ser discreta, a definição tem que ser modificada: o p-quantil é o valor Q_p satisfazendo

$$\begin{aligned} P(X \leq Q_p) &\geq p, \\ P(X \geq Q_p) &\geq 1 - p. \end{aligned}$$

- Dado um conjunto de observações, podemos calcular os *quantis empíricos*. Uma maneira é considerar a *função de distribuição empírica* $\hat{F}_n$ como estimador de F, ou seja, dadas as observações $X_1, \ldots, X_n$ de X,

$$\hat{F}_n(x) = \frac{1}{n} \#\{i : 1 \leq i \leq n, X_i \leq x\}.$$

Então, o quantil Q_p é estimado pelo p-quantil de $\hat{F}_n$. Ou seja, o p-quantil estimado, q_p, seria definido por $\hat{F}_n(q_p) = p$. Contudo, usaremos um enfoque um pouco diferente.

Apêndice: Quantis empíricos

- Chamemos de $X_1, \ldots, X_T$ os valores observados e considere as estatísticas de ordem $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(T)}$. Um estimador consistente de Q_p é dado pelo p-quantil empírico, definido por

$$q_p = \begin{cases} X_{(i)}, & \text{se } p = p_i = (i - 0,5)/T,\ i = 1, \ldots, T \\ (1 - f_i)X_{(i)} + f_i X_{(i+1)}, & \text{se } p_i < p < p_{i+1} \\ X_{(1)}, & \text{se } 0 < p < p_1 \\ X_{(T)}, & \text{se } p_T < p < 1, \end{cases}$$

onde $f_i = (p - p_i)/(p_{i+1} - p_i)$.

- Ou seja, ordenados os dados, q_p é uma das estatísticas de ordem, se p for da forma $p_i = (i - 0,5)/T$ e está na reta ligando os pontos $(p_i, X_{(i)})$ e $(p_{i+1}, X_{(i+1)})$, se p estiver entre p_i e p_{i+1}. Tomamos p_i da forma escolhida e não como i/T para que, por exemplo, a mediana calculada segundo esta definição coincida com a definição usual.

Apêndice: Quantis empíricos

Há dois tipos de gráficos $Q \times Q$: teóricos e empíricos.

- O primeiro tipo é usado para verificar se um conjunto de dados vem de determinada distribuição.
- O segundo tipo é usado para verificar se dois conjuntos de dados têm uma mesma distribuição.
- Para verificar se um conjunto de dados provém de uma distribuição especificada, consideramos o gráfico em que, no eixo horizontal, colocamos os quantis teóricos da distribuição hipotetizada para os dados, e no eixo vertical, os quantis empíricos dos dados, ambos calculados nos pontos p_i, acima. Se as observações realmente são provenientes da distribuição em questão, os pontos deverão estar distribuídos ao longo de uma reta.

MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 5

30 de março de 2023

Sumário

1. Regressão Linear Múltipla

2. Regressão Logística

RLM-modelo

- Com p variáveis explicativas $X_1, \ldots, X_p$ e uma variável resposta Y, o **modelo de regressão linear múltipla** é expresso como

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i, \quad i = 1, \ldots, n. \tag{1}$$

  O coeficiente β_0 é o chamado **intercepto** e a variável explicativa associada a ele, x_{i0}, tem valor constante igual a 1. Para completar a especificação do modelo, supõe-se que os erros e_i são não correlacionados, tenham média zero e variância comum (desconhecida) σ^2.

- Se quisermos testar hipóteses a respeito dos coeficientes do modelo ou construir intervalos de confiança para eles por meio de estatísticas com distribuições exatas, a suposição de que a distribuição de frequências dos erros é Normal deve ser adicionada. O modelo (1) tem $p + 2$ parâmetros desconhecidos, a saber, $\beta_0, \beta_1 \ldots, \beta_p$ e σ^2, que precisam que ser estimados com base nos dados observados.

RLM-modelo

- Definindo $x_{i0} = 1$, $i = 1, \ldots, n$, podemos escrever (1) na forma

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + e_i, \quad i = 1, \ldots, n.$$

Minimizando a soma dos quadrados do erros e_i, *i.e.*,

$$Q(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \sum_{j=0}^{p} \beta_j x_{ij}]^2,$$

em relação a $\beta_0, \ldots, \beta_p$ obtemos os **estimadores de mínimos quadrados**(EMQ) $\widehat{\beta}_j$, $j = 1, \ldots, p$, de modo que

$$\widehat{y}_i = \sum_{j=0}^{p} \widehat{\beta}_j x_{ij}, \quad i = 1, \ldots, n$$

são os **valores estimados** (sob o modelo).

- Os termos

$$\widehat{e}_i = y_i - \widehat{y}_i, \quad i = 1, \ldots, n \tag{2}$$

são os **resíduos**, cuja análise é fundamental para avaliar se modelos da forma (1) se ajustam bem aos dados.

RLM - o modelo

Para efeitos computacionais os dados correspondentes a problemas de regressão linear múltipla devem ser dispostos como indicado na Tabela 1.

Tabela 1: Matriz de dados

Y	X_1	X_2	$\cdots$	X_p
y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Em geral, a variável correspondente ao intercepto (que é constante e igual a um) não precisa ser incluída na matriz de dados; os pacotes computacionais incluem-na naturalmente no modelo a não ser que se indique o contrário.

RLM - o modelo

Para facilitar o desenvolvimento metodológico, convém expressar o modelo na forma matricial

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}. \tag{3}$$

em que $\mathbf{y} = (y_1, \ldots, y_n)^\top$ é o vetor cujos elementos são os valores da variável resposta Y, $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p)$ é a matriz cujos elementos são os valores das variáveis explicativas, com $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top$ contendo os valores da variável X_j, $\boldsymbol{\beta} = (\beta_0, \beta_1 \ldots, \beta_p)^\top$ contém os respectivos coeficientes e $\mathbf{e} = (e_1, \ldots, e_n)^\top$ é o vetor de **erros aleatórios**.

RLM - Exemplo

- Os dados **esteira** são provenientes de um estudo cujo objetivo é avaliar o efeito do índice de massa corpórea (IMC) e da carga aplicada numa esteira ergométrica no consumo de oxigênio (VO2) numa determinada fase do exercício.
- Para associar a distribuição do consumo de oxigênio (Y) com as informações sobre carga na esteira ergométrica (X_1) e IMC (X_2), consideramos o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \tag{4}$$

$i = 1, \ldots, 28$ com as suposições usuais sobre os erros (média zero, variância constante σ^2 e não correlacionados). Aqui, o parâmetro β_1 representa a variação no VO2 esperada por unidade carga para indivíduos com o mesmo IMC. O parâmetro β_2 tem interpretação semelhante com a substituição de carga na esteira por IMC e IMC por carga na esteira.

RLM - Exemplo

- Como não temos dados para indivíduos com IMC menor que 17,50 e carga menor que 32, o parâmetro β_0 deve ser interpretado como um fator de ajuste do plano que aproxima a verdadeira função que relaciona o valor esperado da variável resposta com as variáveis explicativas na região em que há dados disponíveis.
- Se substituíssemos X_1 por $X_1 - 32$ e X_2 por $X_2 - 17.5$, o termo β_0 corresponderia ao VO2 esperado para um indivíduo com IMC $=$ 17,50 submetido a uma carga igual a 32 na esteira ergométrica.
  O modelo (4) pode ser expresso na forma matricial (3) com

$$\mathbf{y} = \begin{bmatrix} 14,1 \\ 16,3 \\ \vdots \\ 31,0 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & 24,32 & 71 \\ 1 & 27,68 & 91 \\ \vdots & \vdots & \vdots \\ 1 & 24,34 & 151 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{28} \end{bmatrix}.$$

- Para problemas com diferentes tamanhos de amostra (n) e diferentes números de variáveis explicativas (p), basta alterar o número de elementos do vetor de respostas $\mathbf{y}$ e do vetor de coeficientes $\boldsymbol{\beta}$ e modificar a matriz com os valores das variáveis explicativas, alterando o número de linhas e colunas convenientemente.

RLM - Propriedades

- Uma das vantagens da expressão do modelo de regressão linear múltipla em notação matricial é que o método de mínimos quadrados utilizado para estimar o vetor de parâmetros $\boldsymbol{\beta}$ no modelo (3) pode ser desenvolvido de maneira universal e corresponde à minimização da forma quadrática

$$Q(\boldsymbol{\beta}) = \mathbf{e}^\top \mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2. \tag{5}$$

- Por meio da utilização de operações matriciais, obtém-se a seguinte expressão para os estimadores de mínimos quadrados

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{6}$$

- Sob a suposição de que $E(\mathbf{e}) = \mathbf{0}$ e $\text{var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$, em que $\mathbf{I}_n$ denota a matriz identidade de dimensão n, temos
  - i) $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,
  - ii) $\text{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

RLM - Propriedades

- Além disso, se adicionarmos a suposição de que os erros têm distribuição Normal, pode-se mostrar que o estimador (6) tem uma distribuição Normal multivariada, o que permite a construção de intervalos de confiança para ou testes de hipóteses sobre os elementos (ou combinações lineares deles) de $\boldsymbol{\beta}$ por meio de estatísticas com distribuições exatas. Mesmo sem a suposição de normalidade para os erros, um recurso ao **Teorema Limite Central** permite mostrar que a distribuição aproximada do estimador (6) é Normal, com média a $\boldsymbol{\beta}$ e matriz de covariâncias $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.
- Um estimador não enviesado de σ^2 é

$$\begin{aligned} s^2 &= [n-(p+1)]^{-1}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= [n-(p+1)]^{-1}\mathbf{y}^\top[\mathbf{I}_n-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{y}. \end{aligned}$$

- Com duas variáveis explicativas, o gráfico de dispersão precisa ser construído num espaço tridimensional, que ainda pode ser representado em duas dimensões; para mais que 2 variáveis explicativas, o gráfico de dispersão requer um espaço com mais do que três dimensões que não pode ser representado no plano. Por isso, uma alternativa é construir gráficos de dispersão entre a variável resposta e cada uma das variáveis explicativas.

RLM - Gráficos

Para os dados **esteira**, o gráfico de dispersão com três dimensões incluindo o plano correspondente ao modelo de regressão múltipla ajustado está disposto na Figura 1.

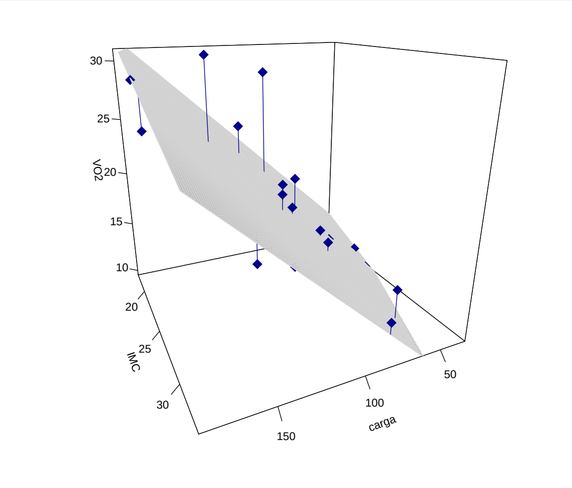


Figura 1: Gráficos de dispersão tridimensional para os dados **esteira**

RLM - Gráficos

Os gráficos de dispersão correspondentes a cada uma das duas variáveis explicativas estão dispostos na Figura 2 e indicam que a distribuição do VO2 varia positivamente com a carga na esteira e negativamente com o IMC.

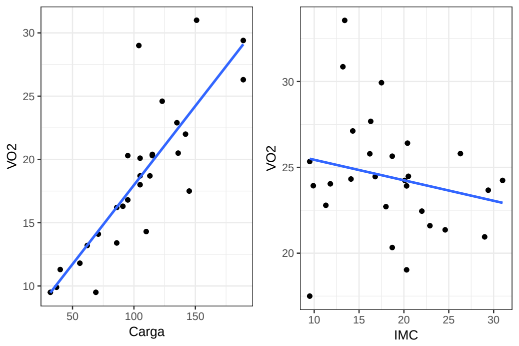


Figura 2: Gráficos de dispersão para os dados **esteira**.

RLM - Uso do R

O uso da função lm() conduz aos seguintes resultados.

Call:
lm(formula = VO2∼ IMC + carga, data = esteira)
Coefficients:

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
Intercept	15.44726	4.45431	3.468	0.00191 **
IMC	-0.41317	0.17177	-2.405	0.02389 *
carga	0.12617	0.01465	8.614	$5.95e-09^{***}$

Residual standard error: 3.057 on 25 degrees of freedom
Multiple R-squared: 0.759, Adjusted R-squared: 0.7397
F-statistic: 39.36 on 2 and 25 DF, p-value: $1.887e-08$

RLM - Uso do R

- Essa saída nos diz que os coeficientes (erro padrão) correspondentes ao ajuste do modelo (4) aos dados **esteira** são $\widehat{\beta_0} = 15,45$ $(4,45)$, $\widehat{\beta_1} = 0,13$ $(0,01)$ e $\widehat{\beta_2} = -0,41$ $(0,17)$. Então, segundo o modelo, o valor esperado do VO2 para um indivíduo (IMC fixado) aumenta de $0,13$ unidades para cada aumento de uma unidade da carga na esteira; similarmente, o valor esperado do VO2 para indivíduos submetidos à mesma carga na esteira diminui de $0,41$ unidades com o aumento de uma unidade no IMC.
- Embora o coeficiente de determinação $R^2 = 0,74$ sugira a adequação do modelo, convém avaliá-la por meio de outras ferramentas diagnósticas. No caso de regressão linear múltipla, gráficos de resíduos podem ter cada uma das variáveis explicativas ou os valores ajustados no eixo das abscissas. Para o exemplo, esses gráficos estão dispostos na Figura 3 juntamente com o gráfico contendo as distâncias de Cook.
- Os gráficos de resíduos padronizados não indicam um comprometimento da hipótese de homoscedasticidade embora seja possível suspeitar de dois ou três pontos discrepantes (correspondentes aos indivíduos com identificação 4, 8 e 28) que também são salientados no gráfico das distâncias de Cook. Veja também a Figura 1.

RLM - Uso do R

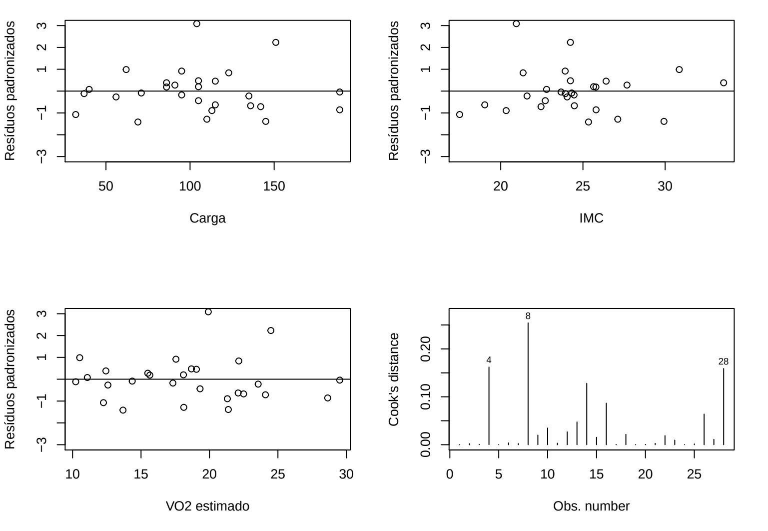


Figura 3: Gráficos de resíduos padronizados e distâncias de Cook para o ajuste do modelo (4) aos dados **esteira**.

frame

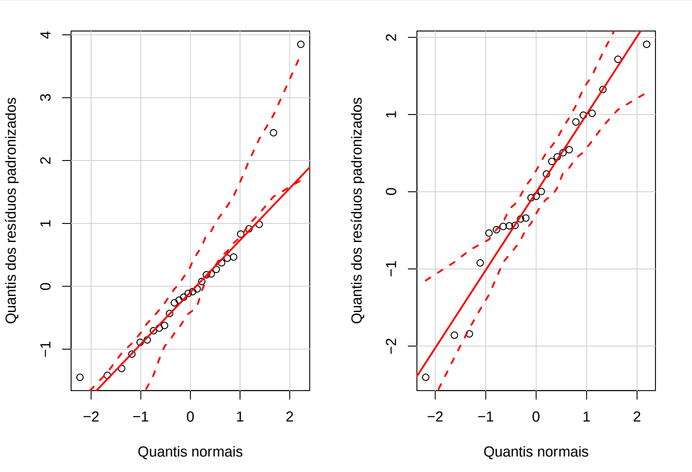


Figura 4: Gráficos QQ correspondentes ao ajuste do modelo (4) aos dados **esteira** com (painel esquerdo) e sem (painel direito) os pontos com identificação 4, 8 e 28.

RL - Regressão logística

- **Exemplo**. O conjunto de dados **inibina** foi obtido de um estudo cuja finalidade era avaliar a utilização da inibina B como marcador da reserva ovariana de pacientes submetidas à fertilização *in vitro*. A variável explicativa é a diferença entre a concentração sérica de inibina B após estímulo com o hormônio FSH (hormônio folículo estimulante) e sua concentração sérica pré estímulo e a variável resposta é a classificação das pacientes como boas ou más respondedoras com base na quantidade de oócitos recuperados.
- A diferença entre esse problema e aqueles estudados nas seções anteriores está no fato de a variável resposta ser dicotômica e não contínua. Se definirmos a variável Y com valor igual a 1 no caso de resposta positiva e igual a zero no caso de resposta negativa, a resposta média será igual à proporção $p = E(Y)$ de pacientes com resposta positiva. Essencialmente, o objetivo da análise é modelar essa proporção como função da variável explicativa.

RL - o modelo

- Em vez de modelar a resposta média, convém modelar uma função dela, a saber o logaritmo da chance de resposta positiva para evitar estimativas de proporções com valores fora do intervalo (0, 1). O modelo correspondente pode ser escrito como

$$\log \frac{P(Y_i = 1|X = x)}{P(Y_i = 0|X = x)} = \alpha + \beta x_i, \; i = 1, \ldots, n. \tag{7}$$

- De forma equivalente,

$$P(Y_i = 1|X = x) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}, \; i = 1, \ldots, n. \tag{8}$$

RL - o modelo

- Neste contexto, o parâmetro α é interpretado como o logaritmo da chance de resposta positiva para pacientes com $x_i = 0$ (concentrações de inibina pré e pós estímulo iguais) e o parâmetro β corresponde ao logaritmo da razão entre a chance de resposta positiva para pacientes com diferença de uma unidade na variável explicativa.
- O ajuste desse modelo é realizado pelo método de máxima verossimilhança. A função de verossimilhança a ser maximizada é

$$\ell(\alpha, \beta | \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} [p(x_i)]^{y_i} [1 - p(x_i)]^{1-y_i}$$

$$p(x_i) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}.$$

RL - EMV

- A maximização da verossimilhança pode ser concretizada por meio da maximização de seu logaritmo

$$L(\alpha, \beta | \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \Big\{ y_i \log[p(x_i)] + (1 - y_i) \log[1 - p(x_i)] \Big\}.$$

- Os estimadores de máxima verossimilhança de α e β correspondem à solução das **equações de estimação**

$$\sum_{i=1}^{n} \Big\{ y_i - \frac{\exp(\widehat{\alpha} + \widehat{\beta} x_i)}{1 + \exp(\widehat{\alpha} + \widehat{\beta} x_i)} \Big\} = 0 \quad \text{e} \quad \sum_{i=1}^{n} x_i \Big\{ y_i - \frac{\exp(\widehat{\alpha} + \widehat{\beta} x_i)}{1 + \exp(\widehat{\alpha} + \widehat{\beta} x_i)} \Big\} = 0.$$

- Como esse sistema de equações não tem solução explícita, deve-se recorrer a métodos iterativos como o **método de Newton-Raphson**.

RL - Uso do R

O uso da função glm() produz os resultados a seguir:

Call:
glm(formula = resposta ∼ difinib, family = binomial, data = dados)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9770	-0.5594	0.1890	0.5589	2.0631

Coefficients:

	Estimate	Std. Error	z value	Pr(> |z|)
Intercept	-2.310455	0.947438	-2.439	0.01474
inib	0.025965	0.008561	3.033	0.00242

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 43.230 on 31 degrees of freedom
Residual deviance: 24.758 on 30 degrees of freedom
AIC: 28.758
Number of Fisher Scoring iterations: 6

RL - Uso do R

- As estimativas dos parâmetros (com erro padrão entre parênteses) α e β correspondentes ao modelo ajustado aos dados **inibina** são, respectivamente,

$$\widehat{\alpha} = -2,31\ (0,95), \quad \widehat{\beta} = 0,03\ (0,01)$$

.

- Consequentemente, a chance de resposta positiva para pacientes com mesmo nível de inibina B pré e pós estímulo hormonal é $\exp(\widehat{\alpha}) = 0,10$.
- Essa chance fica multiplicada por $\exp(\widehat{\beta}) = 1,03$ para cada aumento de uma unidade na diferença entre os níveis de inibina B pré e pós estímulo hormonal.
- Os erros padrões de $\exp(\widehat{\alpha})$ e $\exp(\widehat{\beta})$ são calculados por meio do **método Delta**. Ver Nota de Capítulo 6.

RL - Uso do R

A função predict() pode ser usada para estimar a probabilidade de que a resposta seja positiva, dados os valores da variável explicativa. Algumas dessas probabilidades estão indicadas abaixo:

1	2	3	4	5	6
0.1190483	0.7018691	0.9554275	0.9988353	0.5797138	0.9588247
7	8	9	10		
0.8045906	0.8362005	0.9534173	0.8997726		

Por exemplo, o valor 0,1190483 foi obtido calculando-se

$$P(Y = 1|X = 11,90) = \frac{\exp\{-2,310455 + (0,025965)(11,90)\}}{1 + \exp\{-2,310455 + (0,025965)(11,90)\}}. \qquad (9)$$

RL - uso do R

- Para prever se a resposta vai ser positiva ou negativa, temos que converter essas probabilidades previstas em rótulos de classes, "positiva" / ou "negativa". Considerando respostas positivas como aquelas cuja probabilidade seja maior do que 0,7, digamos, podemos utilizar a função table() para obter a seguinte tabela:

	resposta	
pred	negativa	positiva
negativa	11	5
positiva	2	14

- Os elementos da diagonal dessa tabela indicam os números de observações corretamente classificadas. Ou seja, a proporção de respostas corretas será $(11+14)/32 = 78\%$. Esse valor depende do limiar fixado, $0,7$, no caso. Um *default* usualmente fixado é $0,5$, e nesse caso, a proporção de respostas corretas vai aumentar.
- A utilização de Regressão Logística nesse contexto de classificação será detalhada no Capítulo 10.

Algumas considerações

- O modelo de RLM tem dois aspectos importantes: aditividade e linearidade.
- aditividade significa que o efeito de mudanças em um preditor X_j sobre a resposta Y é independente dos valores dos demais preditores.
- linearidade significa que uma mudança em Y devida a uma mudança unitária em X_j é constante, independentemente do valor de X_j.
- Uma maneira de estender o modelo linear é incluir interações, por exemplo,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e.$$

- Outra maneira: considerar regressão polinomial. Nesse caso, temos uma função não linear, mas o modelo continua linear!
- Para verificar se há necessidade de um modelo não linear, fazer o gráfico dos resíduos *versus* x_i, no caso de RLS e de resíduos *versus* $\hat{y}_i$, no de RLM.

Referências

Morettin, P. A. and Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC: Rio de Janeiro.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 6

17 de abril de 2023

Sumário

Um exemplo

- Consideremos um exemplo [proposto em Bishop (2006)] cujo objetivo é ajustar um modelo de regressão polinomial a um conjunto de 10 pontos gerados por meio da expressão $y_i = sen(2\pi x_i) + e_i$ em que e_i segue um distribuição Normal com média nula e variância σ^2.
- Os dados estão representados na Figura 1 por pontos em azul. A curva verde corresponde a $y_i = sen(2\pi x_i)$; em vermelho estão representados os ajustes baseados em regressões polinomiais de graus, 0, 1, 3 e 9.
- Claramente, a curva baseada no polinômio do terceiro grau consegue reproduzir o padrão da curva geradora dos dados sem, no entanto, predizer os dados com total precisão. A curva baseada no polinômio de grau 9, por outro lado, tem um ajuste perfeito, mas não reproduz o padrão da curva utilizada para gerar os dados, Esse fenômeno é conhecido como sobreajuste.

Um exemplo

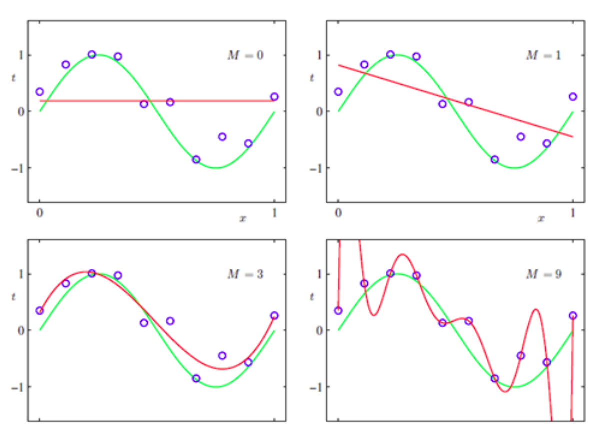


Figura 1: Ajuste de modelos polinomiais a um conjunto de dados hipotéticos.

Quando usar EMQ?

- Se a relação entre resposta e preditores for aproximadamente linear, então um modelo de regressão linear múltipla (RLM) pode ser adequado e estimadores de mínimos quadrados (EMQ) tenderão a ter baixo viés e conduzir a previsões boas.
- Se $n >> p$, EMQ terão também baixa variância.
- Se n não for muito maior do que p, EMQ apresentarão muita variabilidade (sobreajuste) e previsões ruins.
- Se $p > n$, não existirão EMQ univocamente determinados. Aumentando-se o número de variáveis, $R^2 \to 1$, o EQM de treinamento tenderá zero e o EQM de teste crescerá. Não use MQ!

Possíveis abordagens

Possíveis alternativas para remover variáveis irrelevantes de um modelo de RLM, de modo a obter maior interpretabilidade:

- Seleção de subconjuntos de variáveis (subset selection) ; vários procedimentos podem ser usados (stepwise (forward e backward), uso de critérios de informação (AIC, BIC), R^2 ajustado, validação cruzada).
- Encolhimento (shrinkage): usa todos os preditores mas os coeficentes são encolhidos para zero; pode funcionar para selecionar variáveis. Reduz variância. Também chamada regularização.
- Redução da dimensão: projetar os preditores sobre um subespaço de dimensão menor $q < p$, que consiste em obter combinações lineares (ou projeções) dos preditores. Essas q projeções são usadas como novos preditores no ajuste de MQ. ACP, AF, ICA.

Regularização

- O termo regularização refere-se a um conjunto de técnicas utilizadas para especificar modelos que se ajustem a um conjunto de dados evitando o sobreajuste (*overfitting*).
- Essencialmente, essas técnicas servem para ajustar modelos de regressão em que a função de perda contém um termo de penalização cuja finalidade é reduzir a influência de coeficientes responsáveis por flutuações excessivas.
- Embora haja várias técnicas de regularização, consideraremos apenas a regularização L_2, ou Ridge, a regularização L_1 ou Lasso (*least absolute shrinkage and selection operator*) e uma mistura dessas duas, chamada de Elastic net.

Regularização

- O termo de regularização da técnica Lasso usa uma soma de valores absolutos dos parâmetros e um coeficiente de penalização que os encolhe para zero. Essa técnica serve para seleção de modelos, porque associa pesos nulos a parâmetros não significativos.
- Isso implica uma solução esparsa (Dizemos que um modelo é esparso se a maioria dos elementos do correspondente vetor de parâmetros é nula ou desprezável).
- Na regularização L_2, por outro lado, o termo de regularização usa uma soma de quadrados dos parâmetros e um coeficiente de penalização que força alguns pesos a serem pequenos, mas não os anula e consequentemente não conduz a soluções esparsas. Essa técnica de regularização não é robusta com relação a valores atípicos, pois pode conduzir a valores muito grandes do termo de penalização.

O contexto

- Consideremos o modelo de regressão

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + e_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

ou

$$y_i = \beta_0 + \boldsymbol{\beta}_i^\top \mathbf{x}_i + e_i, \tag{2}$$

com as p variáveis preditoras reunidas no vetor $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})^\top$, y_i representando a variável resposta, e_i indicando as inovações de média zero e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ denotando o vetor de parâmetros a serem estimados.

- Vamos considerar $\mathbf{x}_i = (\mathbf{x}_i(1)^\top, \mathbf{x}_i(2)^\top)^\top$, com $\mathbf{x}_i(1) \in \mathbb{R}^s$ o vetor de variáveis relevantes e $\mathbf{x}_i(2) \in \mathbb{R}^{p-s}$ o vetor de variáveis irrelevantes, $\boldsymbol{\beta} = (\boldsymbol{\beta}(1)^\top, \boldsymbol{\beta}(2)^\top)^\top$.
- Objetivos:
  1. Selecione o conjunto de variáveis correto: $\hat{\boldsymbol{\beta}}(1) \neq 0$ e $\hat{\boldsymbol{\beta}}(2) = 0$ (seleção do modelo);
  2. Estime $\boldsymbol{\beta}(1)$ como se o conjunto de variáveis correto fosse conhecido.

Regularização Ridge

- Supomos adicionalmente que $\beta_0 = 0$ e consideremos estimadores de mínimos quadrados (EMQ) penalizados da forma

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2n} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right], \tag{3}$$

  em que λ é o coeficiente de regularização, que controla o número de parâmetros do modelo. Se $\lambda = \infty$, não há variáveis a serem incluídas no modelo e se $\lambda = 0$, obtemos os EMQ usuais.
- Dizemos que $\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda)$ é o estimador Ridge.

Ridge - Propriedades

Pode-se mostrar que

$$\widehat{\beta}_{\text{Ridge}}(\lambda) = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}, \tag{4}$$

em que $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top)^\top$ é a matriz de especificação do modelo e $\mathbf{y} = (y_1, \ldots, y_n)^\top$.

Alguns resultados sobre as propriedades dessa classe de estimadores são:

1) Em geral, o estimador *Ridge* não é consistente. Sua consistência assintótica vale quando $\lambda = v\lambda_n \to \infty$, $\lambda_n/n \to 0$ e $p < n$.
2) O estimador *Ridge* é enviesado para os parâmetros não nulos.
3) A técnica de regularização *Ridge* não serve para a seleção de modelos.
4) A escolha do coeficiente de regularização λ pode ser feita via validação cruzada ou por meio de algum critério de informação.
5) A técnica de regressão *Ridge* foi introduzida por Hoerl e Kennard (1970) para tratar do problema da multicolinearidade.

Ridge - Propriedades

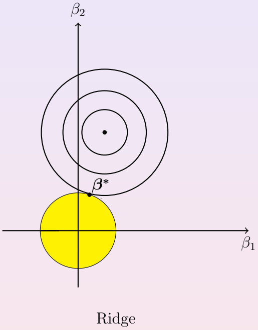
Obter o mínimo em (3) é equivalente a minimizar a soma de quadrados não regularizada sujeita à restrição

$$\sum_{j=1}^{p} \beta_j^2 \leq m, \tag{5}$$

para algum valor apropriado m, ou seja, é um problema de otimização com multiplicadores de Lagrange.

Na Figura 2 (a) apresentamos um esquema com o valor ótimo do vetor $\boldsymbol{\beta}$, a região circular correspondente à restrição (5) e os círculos representando as curvas de nível da função erro não regularizada.

Ridge - Propriedades



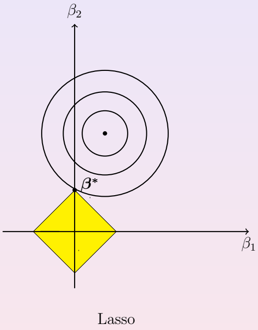


Figura 2: Esparsidade do modelo: (a) Ridge; (b) Lasso.

Regularização Lasso

- Consideremos, agora, o **estimador Lasso**, obtido de

$$\widehat{\boldsymbol{\beta}}_{\text{Lasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2n} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{6}$$

- Neste caso, a restrição (5) é substituída por

$$\sum_{j=1}^{p} |\beta_j| \leq m, \tag{7}$$

- No painel (b) da Figura 2 (b), podemos observar que a regularização Lasso pode gerar uma solução esparsa, ou seja com $\beta_1^* = 0$.

Regularização Lasso

- Existe uma correspondência 1 – 1 entre as formulações (6) e (7): para cada valor de m para a qual (7) vale, existe um valor de λ que fornece a mesma solução para (6). Reciprocamente, a solução $\hat{\beta}(\lambda)$ de (6) resolve o problema restrito, com $m = ||\hat{\beta}(\lambda)||_1$.
- Tanto no caso Ridge, como no Lasso, a constante $1/2n$ pode ser substituída por $1/2$ ou mesmo 1. Esa padronização torna os valores de λ comparáveis para diferentes tamanhos amostrais (por exemplo ao usar CV).
- Em análise convexa, a condição necessária e suficiente para a solução de (6) é

$$-\frac{1}{n}\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\beta \rangle + \lambda s_j = 0, \quad j = 1, \ldots, p, \tag{8}$$

onde s_j é uma quantidade desconhecida, igual a sinal(β_j), se $\beta_j \neq 0$ e algum valor no intervalo $[-1, 1]$, se $\beta_j = 0$ (sub-gradiente para a função valor absoluto).

- O sistema (8) é uma forma das chamadas condições de Karush-Kuhn-Tucker (KKT) para o problema (6).

Lasso - Propriedades

Algumas propriedades estatísticas do estimador Lasso:

1) O estimador Lasso encolhe para zero os parâmetros que correspondem a preditores redundantes.
2) O estimador é enviesado para parâmetros não nulos.
3) Sob certas condições, o estimador Lasso seleciona as variáveis relevantes do modelo atribuindo pesos nulos aos respectivos coeficientes.
4) O estimador não é consistente em geral.
5) Quando $p = n$, ou seja, quando o número de variáveis preditoras é igual ao número de observações, a técnica Lasso corresponde à aplicação de um limiar suave (*soft threshold*) a $Z_j = \mathbf{x}_j^\top \mathbf{y}/n$, ou seja,

$$\widehat{\beta}_j(\lambda) = \text{sinal}(Z_j)\left(|Z_j| - \lambda/2\right)_+, \tag{9}$$

em que $(x)_+ = \max\{x, 0\}$.

Threshold dado por (9)

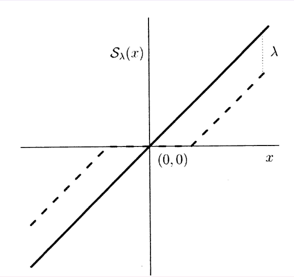


Figura 3: Threshold: MQ (linha cheia) e Lasso (linha tracejada), para o caso $p = n$.

Elastic net

- O estimador *Elastic net* é

$$\widehat{\boldsymbol{\beta}}_{\mathrm{EN}}(\lambda_1, \lambda_2) = \arg\min_{\boldsymbol{\beta}} \sum_{t=1}^{n} \frac{1}{2n}(y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda_2 \sum_{i=1}^{p} \beta_i^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i|. \quad (10)$$

- Na Figura 4 apresentamos esquematicamente uma região delimitada pela restrição $J(\boldsymbol{\beta}) \leq m$, em que $J(\boldsymbol{\beta}) = \alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|$, para algum m, com $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, além daquelas delimitadas pelas restrições *Ridge* e Lasso.
- Pode-se mostrar que sob determinadas condições, o estimador *Elastic Net* é consistente.

Elastic net

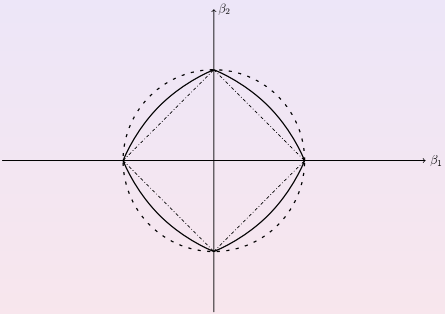


Figura 4: Geometria das restrições *Elastic Net* (curva contínua), *Ridge* (curva tracejada) e Lasso (curva pontilhada).

Lasso adaptativo

- O estimador Lasso adaptativo (adaLASSO) é dado por

$$\widehat{\boldsymbol{\beta}}_{AL}(\lambda) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{i=1}^{p} w_i |\beta_i|, \qquad (11)$$

  em que $w_1, \ldots, w_p$ são pesos não negativos pré-definidos.
- Usualmente, toma-se $w_j = |\tilde{\beta}_j|^{-\tau}$, para $0 < \tau \leq 1$ e $\tilde{\beta}_j$ é um estimador inicial (por exemplo o estimador Lasso).
- O estimador **Lasso adaptativo** é consistente sob condições não muito fortes.
- A função `adalasso` do pacote `parcor` do `R` pode ser usada para calcular esse estimador.
- O pacote `glmnet` do `R` pode ser usado para obter estimadores Lasso e *Elastic net* sob modelos de regressão linear, regressão logística e multinomial, regressão Poisson além de modelos de Cox. Para detalhes, veja Friedman et al. (2010).

Comparação entre os métodos

- Tanto Ridge como o Lasso encolhem os coeficiente para zero. No caso do Lasso, a penalidade L_1 tem a finalidade de tornar alguns dos coeficientes serem efetivamente nulos. Logo, o Lasso realiza seleção de modelos.
- Lasso resulta em modelos esparsos, ou seja, mais fáceis de interpretar.
- Tanto no Ridge, quanto no Lasso, a variância decresce e o viés cresce à medida que λ cresce.
- Ridge tem desepenho melhor que o Lasso nos caso que um número grande de preditores tem relação com a variável resposta. Em caso contrário, o Lasso tem desempenho melhor (em termos de EQM).
- Em geral, nenhum método domina os outros em todas as situações.

Viés da regularização Ridge

Supondo $p < n$, fazendo $\mathbf{R} = \mathbf{X}^\top \mathbf{X}$, e usando a expressão do estimador *ridge* (3), obtemos

$$\begin{aligned}
\widehat{\beta}_{\text{Ridge}}(\lambda) &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= (\mathbf{R} + \lambda \mathbf{I})^{-1} \mathbf{R} [\mathbf{R}^{-1} \mathbf{X}^\top \mathbf{y}) \\
&= [\mathbf{R}(\mathbf{I} + \lambda \mathbf{R}^{-1})]^{-1} \mathbf{R}((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}] \\
&= (\mathbf{I} + \lambda \mathbf{R}^{-1})^{-1} \mathbf{R}^{-1} \mathbf{R} \hat{\beta}_{\text{MQ}} \\
&= (\mathbf{I} + \lambda \mathbf{R}^{-1}) \widehat{\beta}_{\text{MQ}},
\end{aligned}$$

em que $\widehat{\beta}_{\text{MQ}}$ denota o estimador de mínimos quadrados ordinários. Tomando a esperança condicional da expressão anterior, dada $\mathbf{X}$, obtemos

$$\begin{aligned}
E[\widehat{\beta}_{\text{Ridge}}(\lambda) | \mathbf{X}] &= E[(\mathbf{I} + \lambda \mathbf{R}^{-1}) \hat{\beta}_{\text{MQ}}] \\
&= (\mathbf{I} + \lambda \mathbf{R}^{-1}) \boldsymbol{\beta},
\end{aligned}$$

de onde segue

$$E[\widehat{\beta}_{\text{Ridge}}(\lambda)] = (\mathbf{I} + \lambda \mathbf{R}^{-1}) \boldsymbol{\beta} \neq \boldsymbol{\beta}.$$

Ridge: um resultado

Pode-se provar que

$$\widehat{\beta}_{\text{Ridge}}(\lambda) = \mathbf{V}\text{diag}\left(\frac{d_1}{d_1^2+\lambda}, \frac{d_2}{d_2^2+\lambda}, \ldots, \frac{d_p}{d_p^2+\lambda}\right)\mathbf{U}^\top\mathbf{y},$$

em que $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ é a decomposição em valores singulares de $\mathbf{X}$, com $\mathbf{U}$ denotando uma matriz ortogonal de dimensão $n \times p$, $\mathbf{V}$ uma matriz ortogonal de dimensão $p \times p$ e $\mathbf{D}$ uma matriz diagonal com dimensão $p \times p$, contendo os correspondentes valores singulares $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ (raízes quadradas dos autovalores de $\mathbf{X}^\top\mathbf{X}$).

Ridge quando $\mathbf{X}$ é ortogonal

Quando $\mathbf{X}^\top \mathbf{X} = \mathbf{I}$, pode-se provar que:

1. Ridge e EMQ:

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \frac{1}{1+\lambda} \hat{\beta}_{\text{MQ}}. \tag{12}$$

2. A escolha ótima de λ minimizando o erro de previsão esperado é

$$\lambda^* = \frac{p\sigma^2}{\sum_{i=1}^{p} \beta_j^2}. \tag{13}$$

Ridge e Lasso: escolha do parâmetro λ

- A escolha do parâmetro de regularização λ pode ser baseada em validação cruzada ou em algum critério de informação.
- Seja $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ uma grade de valores para λ. No segundo caso,

$$\widehat{\lambda} = \arg\min_{\lambda \in \Lambda}[-\log \text{verossimilhança} + \text{penalização}],$$

como

$$\begin{aligned} AIC &= \log[\widehat{\sigma}^2(\lambda)] + \mathrm{gl}(\lambda)\frac{2}{n}, \\ BIC &= \log[\widehat{\sigma}^2(\lambda)] + \mathrm{gl}(\lambda)\frac{\log n}{n}, \\ HQ &= \log[\widehat{\sigma}^2(\lambda)] + \mathrm{gl}(\lambda)\frac{\log\log n}{n}, \end{aligned}$$

em que $\mathrm{gl}(\lambda)$ é o número de graus de liberdade associado a λ, nomeadamente

$$\mathrm{gl}(\lambda) = \mathrm{tr}\left[\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\right] = \sum_{j=1}^{p}\frac{d_j^2}{d_j^2 + \lambda},$$

e

$$\widehat{\sigma}^2(\lambda) = \frac{1}{n - \mathrm{gl}(\lambda)}\sum_{i=1}^{n}[y_i - \widehat{\beta}_{\mathrm{Ridge}}(\lambda)^\top\mathbf{x}_i]^2.$$

Ridge e Lasso: escolha do parâmetro λ

No caso de validação cruzada (VC):

- Calcule o erro da validação cruzada, como descrito abaixo, para cada valor de λ nessa grade.
- Escolha λ para o qual o erro da VC seja mínimo.
- O modelo é re-ajustao usando todas as observações disponíveis e o valor selecionado de λ.
- Pode-se usar o método LOOCV ou KFCV.

Ridge - Consistência

- Quando $\lambda = \lambda_n$ e $p < n$.
- O estimador Ridge pode ser escrito na foma

$$\begin{aligned}\hat{\beta}_{\text{Ridge}}(\lambda_n) &= (\mathbf{X}^\top \mathbf{X} + \lambda_n \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= \boldsymbol{\beta} - \lambda_n (n^{-1} \mathbf{X}^\top \mathbf{X} + \lambda_n \mathbf{I})^{-1} \boldsymbol{\beta} \\ &+ (n^{-1} \mathbf{X}^\top \mathbf{X} + \lambda_n \mathbf{I})^{-1} n^{-1} \mathbf{X} \mathbf{e}.\end{aligned}$$

- Quando $\lambda_n \to 0$, para $n \to \infty$, pelo Lema de Slutsky,

$$\begin{aligned}\lambda_n (n^{-1} \mathbf{X}^\top \mathbf{X} + \lambda_n \mathbf{I})^{-1} \boldsymbol{\beta} &\to 0, \\ (n^{-1} \mathbf{X}^\top \mathbf{X} + \lambda_n \mathbf{I})^{-1} n^{-1} \mathbf{X} \mathbf{e} &\to 0.\end{aligned}$$

- Pode ainda ser provado que, quando $\sqrt{n}\lambda_n \to \lambda_0 \geq 0$, quando $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_{\text{Ridge}} - \boldsymbol{\beta}) + \lambda_0 \mathbf{Q}^{-1} \boldsymbol{\beta} \to N(\mathbf{0}, \mathbf{Q}^{-1} \mathbf{V} \mathbf{Q}^{-1}),$$

onde $\mathbf{Q} = p\lim n^{-1} \mathbf{X}^\top \mathbf{X}$ e $\mathbf{V}$ é a matriz de variância assintótica de $n^{-1/2} \mathbf{X} \mathbf{e}$.

Lasso: teoria

Considere o caso de $\mathbf{X}$ ortogonal e $p < n$. Então, pode-se provar que

$$\begin{aligned}
\hat{\beta}_{\text{Lasso}}(\lambda) &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_1 \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(-\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} + \frac{1}{2}||\boldsymbol{\beta}||_2^2\right) + \lambda||\boldsymbol{\beta}||_1 \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{p} \left(-\hat{\beta}_{MQ,i}\beta_i + \frac{1}{2}\beta_i^2 + \lambda|\beta_i|\right).
\end{aligned}$$

- Como o problema é um programa de otimização quadrática com restrição convexa, a função objetivo torna-se uma soma de funções objetivos.
- Para cada i, minimiza-se $Q_i = -\hat{\beta}_{MQ,i}\beta_i + \frac{1}{2}\beta_i^2 + \lambda|\beta_i|$.

Lasso: teoria

- No modelo RLM, suponha $\mathbf{X}$ fixa ou iid e o erro normal, com média $\mathbf{0}$ e variância $\sigma^2\mathbf{I}$.
- O número de variáveis $p = p_n$ e $p >> n$.
- No caso de $\mathbf{X}$ fixa, resultados de consistência dependem da condição

$$||\boldsymbol{\beta}^0||_1 = ||\boldsymbol{\beta}_n^0||_1 = o\left(\sqrt{\frac{n}{\log p}}\right).$$

- No caso de $\mathbf{X}$ aleatória, sob condições sobre o erro e

$$||\boldsymbol{\beta}^0||_1 = o\left(\left(\frac{n}{\log n}\right)^{1/4}\right),$$

e para λ da ordem de $\sqrt{\log p/n}$,o estimador Lasso é consistente:

$$\left[\hat{\boldsymbol{\beta}}_{\text{Lasso}}(\lambda) - \boldsymbol{\beta}^0\right] \Sigma \left[\hat{\boldsymbol{\beta}}_{\text{Lasso}}(\lambda) - \boldsymbol{\beta}^0\right]^\top = o_p(1), \quad n \to \infty,$$

com $\Sigma = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ no caso fixo ou Σ sendo a matriz de covariância de $\mathbf{X}$, no caso aleatório.

Exemplo

Consideremos os dados do arquivo `antracose`, extraídos de um estudo cuja finalidade era avaliar o efeito da idade (`idade`), tempo vivendo em São Paulo (`tmunic`), horas diárias em trânsito (`htransp`), carga tabágica (`cargatabag`), classificação sócio-econômica (`ses`), densidade de tráfego na região onde habitou (`densid`) e distância mínima entre a residência a vias com alta intensidade de tráfego (`distmin`) num índice de antracose (`antracose`) que é uma medida de fuligem (*black carbon*) depositada no pulmão. Como esse índice varia entre 0 e 1, consideramos

$$logrc = \log[\text{índice de antracose}/(1 - \text{índice de antracose})]$$

Exemplo - MQ

Os estimadores de mínimos quadrados para um modelo linear podem ser obtidos por meio dos comandos

```
pulmao_lm <- lm(logrc ~ idade + tmunic + htransp + cargatabag +
                          ses + densid + distmin, data=pulmao)
summary(pulmao_lm)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.977e+00  2.459e-01 -16.169  < 2e-16 ***
idade        2.554e-02  2.979e-03   8.574  < 2e-16 ***
tmunic       2.436e-04  2.191e-03   0.111 0.911485
htransp      7.505e-02  1.634e-02   4.592 5.35e-06 ***
cargatabag   6.464e-03  1.055e-03   6.128 1.61e-09 ***
ses         -4.120e-01  1.238e-01  -3.329 0.000926 ***
densid       7.570e+00  6.349e+00   1.192 0.233582
distmin      3.014e-05  2.396e-04   0.126 0.899950
Residual standard error: 1.014 on 598 degrees of freedom
Multiple R-squared:  0.1965,Adjusted R-squared:  0.1871
F-statistic: 20.89 on 7 and 598 DF,  p-value: < 2.2e-16
```

Exemplo - Ridge

O ajuste dos modelos de regressão *Ridge*, *Lasso* ou *Elastic net* pode ser obtido com o pacote glmnet.

Utilizando esse pacote, ajustamos o modelo de regressão *Ridge* por meio de validação cruzada e obtivemos o gráfico da Figura 2 em que o erro quadrático médio (MSE) é expresso em função do logaritmo do coeficiente de regularização λ.

```
regridgecv = cv.glmnet(X, y, alpha = 0)
plot(regridgecv)
```

Exemplo - Ridge

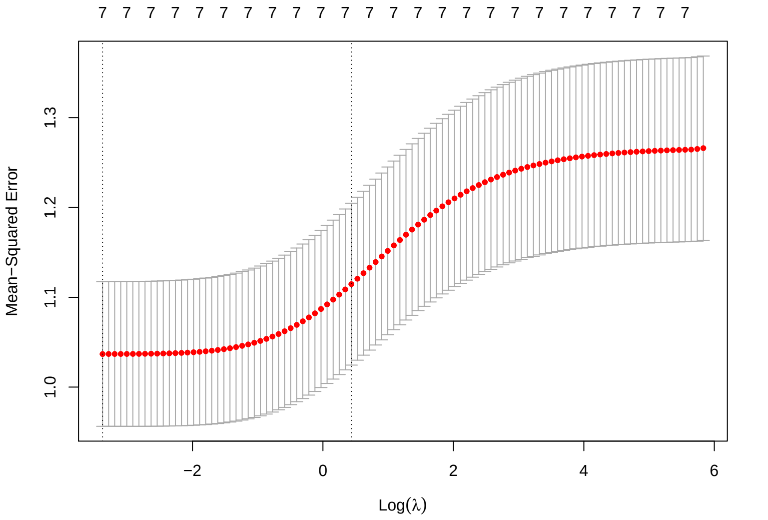


Figure 2: Gráfico para avaliação do efeito do coeficiente de regularização (*Ridge*).

Exemplo - Ridge

Os coeficientes do ajuste correspondentes ao valor mínimo do coeficiente λ, juntamente com esse valor, são obtidos com os comandos

```
coef(regridgecv, s = "lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
                        1
(Intercept) -3.905299e+00
idade        2.456715e-02
tmunic       4.905597e-04
htransp      7.251095e-02
cargatabag   6.265919e-03
ses         -3.953787e-01
densid       7.368120e+00
distmin      3.401372e-05
> regridgecv$lambda.min
[1] 0.03410028
```

Exemplo -Ridge

Com exceção das estimativas dos coeficientes das variáveis `tmunic` e `distmin` as demais foram encolhidas em direção a zero relativamente àquelas obtidas por mínimos quadrados. Os valores preditos e a correspondente raiz quadrada do *MSE* (usualmente denotada *RMSE*) são obtidos por meio de

```
predict(regridgecv, X, s = "lambda.min")
sqrt(regridgecv$cvm[regridgecv$lambda == regridgecv$lambda.min])
[1] 1.050218
```

Exemplo - Lasso

O ajuste do modelo de regressão Lasso juntamente com o gráfico para a escolha do coeficiente λ, disposto na Figura 3, são obtidos com

```
reglassocv = cv.glmnet(X, y, alpha = 1)
plot(reglassocv)
```

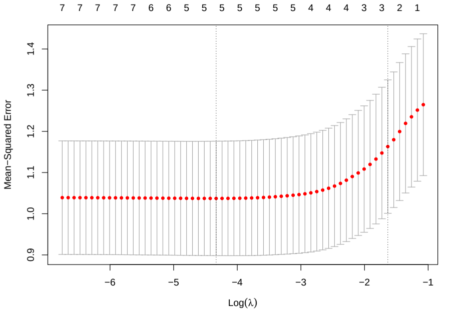


Figura 3: Gráfico para avaliação do efeito do coeficiente de regularização (*Lasso*).

Exemplo - Lasso

Os coeficientes correspondentes à regularização *Lasso*, o valor mínimo do coeficiente λ e o *RMSE* são gerados por intermédio dos comandos

```
coef(reglassocv, s = "lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept) -3.820975473
idade        0.024549358
tmunic       .
htransp      0.069750435
cargatabag   0.006177662
ses         -0.365713282
densid       5.166969594
distmin      .
reglassocv$lambda.min
[1] 0.01314064
sqrt(reglassocv$cvm[reglassocv$lambda == reglassocv$lambda.min])
[1] 1.018408
```

Neste caso, todos os coeficientes foram encolhidos em direção ao zero, e aqueles correspondentes às variáveis `tmunic` e `distmin` foram anulados.

Exemplo - Elastic Net

Para o modelo Elastic Net com $\alpha = 0,5$ os resultados são

```
regelncv = cv.glmnet(X, y, alpha = 0.5)
regelncv$lambda.min
[1] 0.02884367
coef(regelncv, s = "lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
                        1
(Intercept) -3.776354935
idade        0.024089256
tmunic       .
htransp      0.068289153
cargatabag   0.006070319
ses         -0.354080190
densid       4.889074555
distmin      .            .
sqrt(regelncv$cvm[regelncv$lambda == regelncv$lambda.min])
[1] 1.0183
```

Vemos que os 3 procedimentos resultam em EQM similares, com pequena vantagem para Elastic Net.

MAG

- Modelos lineares têm um papel muito importante na análise de dados, tanto pela facilidade de ajuste quanto de interpretação. De uma forma geral, os modelos lineares podem ser expressos como

$$y_i = \beta_0 + \beta_1 f_1(x_{i1}) + \ldots + \beta_p f_p(x_{ip}) + e_i \tag{14}$$

  $i = 1, \ldots, n$ em que as funções f_i são conhecidas. No modelo de regressão polinomial de segundo grau, por exemplo, $f_1(x_{i1}) = x_{i1}$ e $f_2(x_{i2}) = x_{ij}^2$. Em casos mais gerais, poderíamos ter $f_1(x_{i1}) = x_{ij}$ e $f_2(x_{i2}) = \exp(x_{i2})$. Em muitos problemas reais, no entanto, nem sempre é fácil especificar a forma das funções f_i e uma alternativa proposta por Hastie e Tibshirani (1996) são os chamados Modelos Aditivos Generalizados (*Generalized Additive Modelos* - GAM) que são expressos como (14) sem a especificação da forma das funções f_i.
- Quando a distribuição da variável resposta y_i pertence à **família exponencial**, o modelo pode ser considerado como uma extensão dos **Modelos Lineares Generalizados** (*Generalized Linear Models* - GLM) e é expresso como

$$g(\mu_i) = \beta_0 + \beta_1 f_1(x_{i1}) + \ldots + \beta_p f_p(x_{ip}) \tag{15}$$

  em que g é uma **função de ligação** e $\mu_i = E(y_i)$ (ver Nota de Capítulo 3).

MAG

- Existem diversas propostas para a representação das funções f_i que incluem o uso de splines naturais, splines suavizadas e regressões locais.
- A suavidade dessas funções é controlada por parâmetros de suavização, que devem ser determinados *a priori*. Curvas muito suaves podem ser muito restritivas, enquanto curvas muito rugosas podem causar sobreajuste.
- O procedimento de ajuste dos modelos aditivos generalizados depende da forma escolhida para as funções f_i. A utilização de *splines* naturais, por exemplo, permite a aplicação direta do método de mínimos quadrados, graças à sua construção a partir de funções base.
- Para *splines* penalizadas, o processo de estimação envolve algoritmos um pouco mais complexos, como aqueles conhecidos sob a denominação de retroajustamento (*backfitting*). Para detalhes sobre o ajuste dos modelos aditivos generalizados, consulte Hastie e Tibshirani (1990) e Hastie et al. (2008).

Splines

- Para entender o conceito de *splines*, consideremos o seguinte modelo linear com apenas uma variável explicativa

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n. \tag{16}$$

- A ideia subjacente aos modelos aditivos generalizados é a utilização de funções base e consiste na substituição do termo $\beta_1 x_i$ em (16) por um conjunto de transformações conhecidas $b_1(x_i), \ldots, b_t(x_i)$, gerando o modelo

$$y_i = \alpha_0 + \alpha_1 b_1(x_i) + \ldots + \alpha_t b_t(x_i) + e_i, \quad i = 1, \ldots, n. \tag{17}$$

O modelo de regressão polinomial de grau t é um caso particular de (17) com $b_j(x_i) = x_i^j$, $j = 1, \ldots, t$.

Splines

- Uma proposta para aumentar a flexibilidade da curva ajustada consiste em segmentar o domínio da variável preditora e ajustar diferentes polinômios de grau d aos dados de cada um dos intervalos gerados pela segmentação. Cada ponto de segmentação é chamado de nó e uma segmentação com k nós gera $k+1$ polinômios. Na Figura 43, apresentamos um exemplo com polinômios de terceiro grau e 4 nós.
- Nesse exemplo, a expressão (17) tem a forma

$$y_i = \begin{cases} \alpha_{01} + \alpha_{11}x_i + \alpha_{21}x_i^2 + \alpha_{31}x_i^3 + e_i, & \text{se } x_i \leq -0.5, \\ \alpha_{02} + \alpha_{12}x_i + \alpha_{22}x_i^2 + \alpha_{32}x_i^3 + e_i, & \text{se } -0.5 < x_i \leq 0, \\ \alpha_{02} + \alpha_{13}x_i + \alpha_{23}x_i^2 + \alpha_{33}x_i^3 + e_i, & \text{se } 0 < x_i \leq 0.5, \\ \alpha_{02} + \alpha_{14}x_i + \alpha_{24}x_i^2 + \alpha_{34}x_i^3 + e_i, & \text{se } 0.5 < x_i \leq 1, \\ \alpha_{05} + \alpha_{15}x_i + \alpha_{25}x_i^2 + \alpha_{35}x_i^3 + e_i, & \text{se } x_i > 1, \end{cases} \tag{18}$$

sendo que nesse caso, as funções base $b_1(X), b_2(X), ..., b_k(X)$ são construídas com a ajuda de funções indicadoras. Esse modelo é conhecido como modelo polinomial cúbico segmentado.

Splines

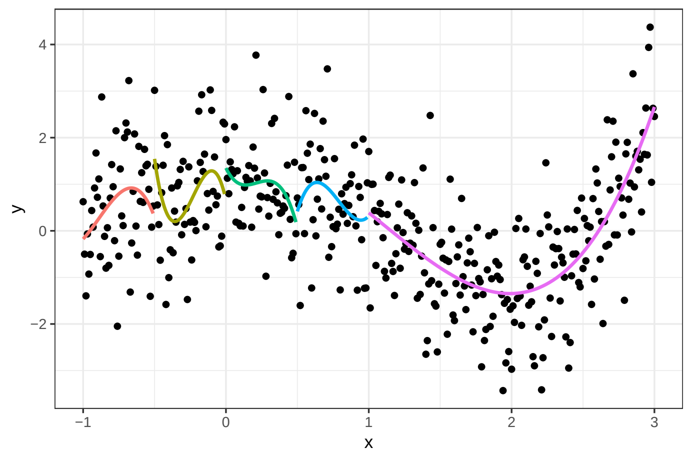


Figura 4: Polinômios de terceiro grau ajustados aos dados de cada região segmentada da variável X. Os nós são os pontos $x = -0.5$, $x = 0$, $x = 0.5$ e $x = 1$.

Splines

- A curva formada pela junção de cada um dos polinômios na Figura 43 não é contínua, apresentando "saltos" nos nós.
  Essa característica não é desejável, pois essas descontinuidades não são interpretáveis. Para contornar esse problema, podemos definir um *spline* de grau d como um polinômio segmentado de grau d com as $d-1$ primeiras derivadas contínuas em cada nó. Essa restrição garante a continuidade e suavidade (ausência de vértices) da curva obtida.
- Utilizando a representação por bases (17), um *spline* cúbico com k nós pode ser expresso como

$$y_i = \alpha_0 + \alpha_1 b_1(x_i) + \alpha_2 b_2(x_i) + \ldots + \alpha_{k+3} b_{k+3}(x_i) + e_i, \quad i = 1, ..., n, \quad (19)$$

  com as funções $b_1(x), b_2(x), \ldots, b_{k+3}(x)$ escolhidas apropriadamente.
- Usualmente, essas funções envolvem três termos polinomiais, a saber, x, x^2 e x^3 e k termos $h(x, c_1), \ldots, h(x, c_k)$ da forma

$$h(x, c_j) = (x - c_j)_+^3 = \begin{cases} (x - c_j)^3, & \text{se } x < c_j, \\ 0, & \text{em caso contrário,} \end{cases}$$

  com $c_1, \ldots, c_k$ indicando os nós.

Splines

- Com a inclusão do termo α_0, o ajuste de um *spline* cúbico com k nós envolve a estimação de $k+4$ parâmetros e, portanto, utiliza $k+4$ graus de liberdade. Mais detalhes sobre a construção desses modelos podem ser encontrados em Hastie (2008) e James et al. (2017).
- Além das restrições sobre as derivadas, podemos adicionar restrições de fronteira, exigindo que a função seja linear na região de x abaixo do menor nó e acima do maior nó. Essas restrições diminuem a variância dos valores extremos gerados pelo preditor, produzindo estimativas mais estáveis. Um *spline* cúbico com restrições de fronteira é chamado de spline natural.
- No ajuste de *splines* cúbicos ou naturais, o número de nós determina o grau de suavidade da curva e a sua escolha pode ser feita por validação cruzada conforme indicado em James et al. (2017). De uma forma geral, a maior parte dos nós é posicionada nas regiões do preditor com mais informação, isto é, com mais observações. Por pragmatismo, para modelos com mais de uma variável explicativa, costuma-se adotar o mesmo número de nós para todos os preditores.

Splines

- Os *splines* suavizados constituem uma classe de funções suavizadoras que não utilizam a abordagem por funções bases. De maneira resumida, um *spline* suavizado é uma função f que minimiza

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int f''(u)^2 du \tag{20}$$

em que f'' corresponde à segunda derivada da função f e indica sua curvatura; quanto maior for a curvatura maior a penalização.

- O primeiro termo dessa expressão garante que f se ajustará bem aos dados, enquanto o segundo penaliza a sua variabilidade, isto é, controla a suavidade de f, que é regulada pelo parâmetro λ. A função f se torna mais suave conforme λ aumenta. A escolha desse parâmetro é geralmente feita por validação cruzada.
- Outro método bastante utilizado no ajuste funções não lineares para a relação entre a variável preditora X e a variável resposta Y é conhecido como regressão local. Esse método consiste em ajustar modelos de regressão simples em regiões em torno de cada observação x_0 da variável preditora X.

Splines

- Essas regiões são formadas pelos k pontos mais próximos de x_0, sendo que o parâmetro $s = k/n$ determina o quão suave ou rugosa será a curva ajustada. O ajuste é feito por meio de mínimos quadrados ponderados, com pesos inversamente proporcionais à distância entre cada ponto da região centrada em x_0 e x_0. Aos pontos dessas regiões mais afastados de x_0 são atribuídos pesos menores.
- Lowess: ajusta retas. Veja a Nota de Capítulo 5.2.
- Para uma excelente exposição sobre *splines* e penalização o leitor pode consultar Eilers e Marx (1996) e Eilers e Marx (2021).
- Modelos aditivos generalizados podem ser ajustados utilizando-se a função gam() do pacote mgcv. Essa função permite a utilização de *splines* como funções suavizadoras. Para regressão local, é necessário usar a função `gam()` do pacote `gam`. Também é possível utilizar o pacote `caret`, a partir da função `train()` e `method = "gam"`.

MAG: exemplo

Consideremos os dados do arquivo `esforco` com o objetivo de prever os valores da variável `vo2fcpico` (VO2/FC no pico do exercício) a partir das variáveis `NYHA`, `idade`, `altura`, `peso`, `fcrep` (frequência cardíca em repouso) e `vo2rep` (VO2 em repouso). Um modelo inicial de regressão linear múltipla também pode ser ajustado por meio dos seguintes comandos da função gam

```
mod0 <- gam(vo2fcpico ~ NYHA + idade + altura + peso + fcrep
                             + vo2rep, data=esforco)
```

Como não especificamos nem a distribuição da resposta, nem a função de ligação, a função gam utiliza a distribuição normal com função de ligação logarítmica, conforme indica o resultado.

MAG: exemplo

```
Family: gaussian
Link function: identity
Formula:
vo2fcpico ~ NYHA + idade + altura + peso + fcrep + vo2rep
Parametric coefficients:
             Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -4.80229    4.43061  -1.084  0.280642
NYHA1       -0.45757    0.50032  -0.915  0.362303
NYHA2       -1.78625    0.52629  -3.394  0.000941 ***
NYHA3       -2.64609    0.56128  -4.714  6.75e-06 ***
NYHA4       -2.43352    0.70532  -3.450  0.000780 ***
idade       -0.05670    0.01515  -3.742  0.000284 ***
altura       0.09794    0.02654   3.690  0.000342 ***
peso         0.08614    0.01739   4.953  2.48e-06 ***
fcrep       -0.07096    0.01318  -5.382  3.84e-07 ***
vo2rep       0.35564    0.24606   1.445  0.151033
---
R-sq.(adj) =  0.607   Deviance explained = 63.5%
GCV =  4.075  Scale est. = 3.7542    n = 127
```

MAG: exemplo

Para avaliar a qualidade do ajuste, produzimos gráficos de dispersão entre os resíduos do ajuste e cada uma das variáveis preditoras. Esses gráficos estão dispostos na Figura 5 e sugerem relações possivelmente não lineares pelo menos em alguns casos.

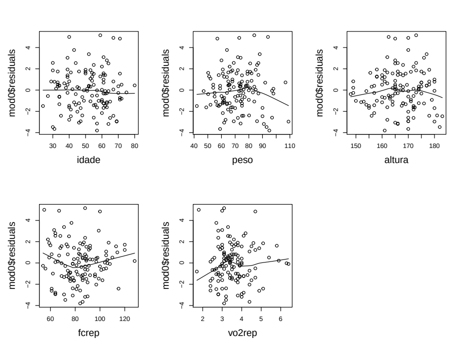


Figura 5: Gráficos de dispersão (com curva *lowess*) entre `vo2fcpico` e cada variável preditora contínua considerada no Exemplo.

MAG: exemplo

Uma alternativa é considerar modelos GAM do tipo (14) em que as funções f_i são expressas em termos de *splines*. Em particular, um modelo GAM com *splines* cúbicos para todas as variáveis explicativas contínuas pode ser ajustado por meio do comando

```
mod1 <- gam(vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) +
                        s(fcrep) + s(vo2rep), data=esforco)
```

gerando os seguintes resultados:

MAG: exemplo

```
Family: gaussian
Link function: identity
Formula:
vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) + s(fcrep) +
    s(vo2rep)
Parametric coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  10.2101     0.3207   31.841   < 2e-16 ***
NYHA1        -0.5498     0.4987   -1.103   0.272614
NYHA2        -1.8513     0.5181   -3.573   0.000522 ***
NYHA3        -2.8420     0.5664   -5.018   1.99e-06 ***
NYHA4        -2.5616     0.7031   -3.643   0.000410 ***
---
Approximate significance of smooth terms:
            edf  Ref.df    F    p-value
s(idade)   1.000  1.000 15.860  0.00012 ***
s(altura) 5.362  6.476  3.751  0.00142 **
s(peso)    1.000  1.000 22.364  6.32e-06 ***
s(fcrep)   1.742  2.185 16.236  3.95e-07 ***
s(vo2rep) 1.344  1.615  0.906  0.47319
---
R-sq.(adj) =   0.64   Deviance explained = 68.2%
GCV = 3.9107  Scale est. = 3.435     n = 127
```

MAG: exemplo

- O painel superior contém estimativas dos componentes paramétricos do modelo e o painel inferior, os resultados referentes aos termos suavizados. Neste caso apenas a variável categorizada `NYHA` não foi suavizada, dada sua natureza não paramétrica.
- A coluna rotulada `edf` contém os graus de liberdade efetivos associados a cada variável preditora. Para cada variável preditora contínua não suavizada, perde-se um grau de liberdade; para as variáveis suavizadas a atribuição de graus de liberdade é mais complexa em virtude do número de funções base e do número de nós utilizados no processo de suavização. A linha rotulada `GCV` (*Generalized Cross Validation*) está associada com a escolha (por validação cruzada) do parâmetro de suavização.
- A suavização é irrelevante apenas para a variável `vo2rep` e dado que ela também não apresentou contribuição significativa no modelo de regressão linear múltipla, pode-se considerar um novo modelo GAM obtido com a sua eliminação. Os resultados correspondentes, apresentados a seguir, sugerem que todas as variáveis preditoras contribuem significativamente para explicar sua relação com a variável resposta,

MAG: exemplo

```
Formula:
vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) + s(fcrep)
Parametric coefficients:
            Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  10.2301     0.3202  31.948   < 2e-16 ***
NYHA1        -0.5818     0.4985  -1.167   0.245650
NYHA2        -1.8385     0.5161  -3.563   0.000539 ***
NYHA3        -2.9669     0.5512  -5.382   4.04e-07 ***
NYHA4        -2.4823     0.6980  -3.556   0.000551 ***
---
Approximate significance of smooth terms:
            edf   Ref.df     F      p-value
s(idade)   1.000  1.000   16.322  9.59e-05 ***
s(altura)  5.311  6.426    3.857  0.00115 **
s(peso)    1.000  1.000   22.257  6.56e-06 ***
s(fcrep)   1.856  2.337   14.865  8.39e-07 ***
---
R-sq.(adj) =   0.64   Deviance explained = 67.8%
GCV = 3.8663  Scale est. = 3.435     n = 127
```

MAG: exemplo

- Como o número efetivo de graus de liberdade para `idade` e `peso` é igual a 1, elas se comportam de forma linear no modelo. Os gráficos dispostos na Figura 6, produzidos por meio do comando `plot(mod2, se=TRUE)` evidenciam esse fato; além disso mostram a natureza "mais não linear" da variável `altura` (com `edf`$= 5.311$).
- Uma avaliação da qualidade do ajuste pode ser realizada por meio de uma análise de resíduos e de comparação dos valores observados e preditos. Para essa finalidade, o comando `gam.check(mod2)` gera os gráficos apresentados na Figura 7 que não evidenciam problemas no ajuste.
- Além disso, é possível comparar os modelos por meio de uma **análise de desviância**, que pode ser obtida com o comando `anova(mod0, mod1, mod2, test= "F")`.

MAG: exemplo

```
Analysis of Deviance Table
Model 1: vo2fcpico ~ NYHA + idade + altura + peso + fcrep +
                     vo2rep
Model 2: vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) +
                     s(fcrep) +  s(vo2rep)
Model 3: vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) +
                     s(fcrep)
     Resid.     Df  Resid.    Dev     Df Deviance      F         Pr(>F)
1    117.00     439.24
2    109.72     383.18    7.2766     56.052        2.2425      0.03404 *
3    111.24     387.58   -1.5129     -4.399        0.8465      0.40336
```

MAG: exemplo

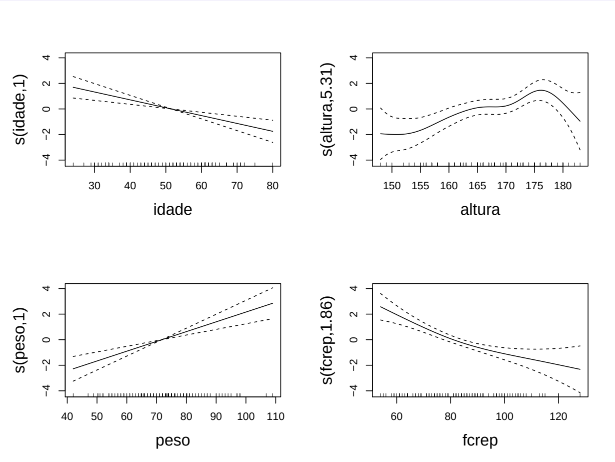


Figura 6: Funções suavizadas (com bandas de confiança) obtidas por meio do modelo GAM para os dados do Exemplo.

MAG: exemplo

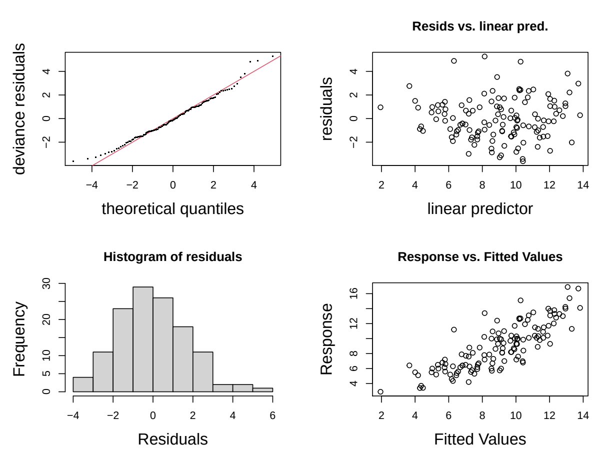


Figura 7: Gráficos diagnósticos para o ajuste do modelo GAM aos dados do Exemplo.

MAG: exemplo

- Esses resultados mostram que ambos os modelos GAM são essencialmente equivalentes ($p = 0.403$) mas significativamente mais adequados ($p = 0.034$) que o modelo de regressão linear múltipla.
- A previsão para um novo conjunto dados em que apenas os valores das variáveis preditoras estão disponíveis pode ser obtida por meio do comando `predict(mod2, newdata=esforcoprev, se=TRUE, type="response")`. Consideremos, por exemplo, o seguinte conjunto com dados de 5 novos pacientes

idade	altura	peso	NYHA	fcrep	vo2rep
66	159	50	2	86	3,4
70	171	77	4	108	4,8
64	167	56	2	91	2,5
42	150	67	2	70	3,0
54	175	89	2	91	2,9

MAG: exemplo

O resultado da previsão com o modelo adotado é

```
$fit
       1         2         3         4         5
 4.632615  5.945157  5.928703  7.577097 10.273719
$se.fit
       1         2         3         4         5
0.6747203 0.7155702 0.6255449 0.7731991 0.5660150
```

Referências

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Berlin: Springer.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparcity*. Chapman and Hall.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Morettin, P. A. e Singer, J. M. (2021). *Estatística e Ciência de Dados*. Texto Preliminar, IME-USP.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B (methodological), **58**, 267–288.

MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 7

14 de abril de 2023

Sumário

Regularização - EMQ

- **Exemplo.** Vamos considerar o conjunto de dados **esforco**, centrando o interesse na predição da variável resposta Y: VO2 (consumo de oxigênio em ml/(kg.min)) com base nas variáveis preditoras X_1: Idade (em anos), X_2: Peso (em kg), X_3: Superfície corpórea e X_4: IMC (índice de massa corpórea em kg/m^2) ($n = 126$)
- Ajustando o modelo via mínimos quadrados ordinários, obtemos os coeficientes:

Coefficients	Estimate	Std. Error	t value	P-value
Intercept	14.92204	6.69380	2.229	0.0276 *
Idade	-0.01005	0.02078	-0.483	0.6297
Peso	0.05233	0.11760	0.445	0.6571
Sup.Corp.	-1.40678	6.25353	-0.225	0.8224
IMC	-0.20030	0.16104	-1.244	0.2160

Residual standard error: 2.822 on 121 degrees of freedom
Multiple R-squared: 0.03471, Adjusted R-squared: 0.002799

Regularização - Ridge

- Os coeficientes correspondentes obtidos por meio de regularização **Ridge** são

Intercept	5.185964640
Idade	-0.000133776
Peso	-0.006946405
Sup.Corp	-0.295094364
IMC	-0.022923850

- O valor do coeficiente de regularização $\lambda = 0,82065$, mostra que as estimativas para os coeficientes de Idade e Peso foram encolhidas para zero, enquanto aquelas correspondentes à Sup.Corp tem peso maior do que as demais.
- Neste caso, a raiz quadrada do **erro quadrático médio** (*root mean squared error*) e o **coeficiente de determinação** são, respectivamente $RMSE = 0,928$ e $R^2 = 0,235$.

Regularização Lasso

- Os coeficientes correspondentes obtidos por meio de regularização Lasso são

Intercept	4.95828012
Idade	.
Peso	-0.01230145
Sup.Corp	.
IMC	-0.02011871

- O valor do coeficiente de regularização $\lambda = 0,0257$ mostra que as estimativas dos coeficientes Idade e Sup. Corp foram efetivamente encolhidas para zero.
- Neste caso RMSE e R^2 são, respectivamente, $0,927$ e $0,228$.

Regularização Elastic Net

- Os coeficientes correspondentes obtidos por meio de regularização Elastic Net são:

Intercept	4.985532570
Idade	.
Peso	-0.009099925
Sup.Corp	-0.097034844
IMC	-0.023302254

- Os parâmetros de suavização estimados foram $\alpha = 0,1$ e $\lambda = 0,227$ e também indicam que o coeficiente associado à Idade foi encolhido para zero.
- Também obtemos RMSE=0,927 e $R^2 = 0,228$ neste caso.
- Os três métodos de regularização têm desempenhos similares quando vistos pelas óticas do RMSE e do R^2.

Validação Cruzada

- Validação cruzada é a denominação atribuída a um conjunto de técnicas utilizadas para avaliar o erro de previsão de modelos estatísticos. O erro de previsão é uma medida da precisão com que um modelo pode ser usado para prever o valor de uma nova observação, ou seja diferente daquelas utilizados para o ajuste do modelo.
- Em modelos de regressão o erro de previsão é definido como

$$EP = E(y - \widehat{y})^2,$$

em que y representa uma nova observação e $\widehat{y}$ é a previsão obtida pelo modelo.

- erro quadrático médio dos resíduos pode ser usado como uma estimativa do erro de previsão (EP), mas tende, em geral, a ser muito otimista, ou seja, a subestimar o seu verdadeiro valor. Uma razão é que os mesmos dados são utilizados para ajustar e avaliar o modelo.
- No processo de validação cruzada, o modelo é ajustado a um subconjunto dos dados (**subconjunto de treinamento**) e o resultado é empregado num outro subconjunto (**subconjunto de teste**) para avaliar se ele tem um bom desempenho ou não.

Validação Cruzada

Um algoritmo utilizado nesse processo é o seguinte (Efron e Tibshirani, 1993):

1) Dadas n observações, $y_1, \ldots, y_n$, o modelo é ajustado n vezes, em cada uma delas eliminando uma observação e o valor previsto, denotado por $\widehat{y}_{-i}$, para essa observação, é calculado com base no resultados obtido com as demais $n - 1$.
2) O erro de previsão é estimado por

$$VC_{(n)} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_{-i})^2. \tag{1}$$

Esse tipo de validação cruzada é chamado validação cruzada com eliminação de uma observação (leave-one-out cross-validation - LOOCV).

Validação Cruzada

- Na chamada validação cruzada de ordem k (*k-fold cross validation*) o conjunto de dados original é subdividido em dois, sendo um deles utilizado como conjunto de treinamento e o segundo como conjunto de teste. Esse processo é repetido k vezes com conjuntos de treinamento e testes diferentes como esquematizado na Figura 5.
- O correspondente erro de previsão é estimado como

$$VC_{(k)} = \frac{1}{k} \sum_{i=1}^{k} EQM_i. \tag{2}$$

em que

$$EQM_i = \sum (y_j - \widehat{y}_j)^2 / n_i$$

é erro quadrático médio obtido no i-ésimo ajuste, $i = 1, \ldots, k$. Aqui, y_j, $\widehat{y}_j$ e n_i são, respectivamente, os valores observado e previsto para a j-ésima observação e o número de observações no i-ésimo conjunto de teste. O usual é tomar $k = 5$ ou $k = 10$.

Validação Cruzada

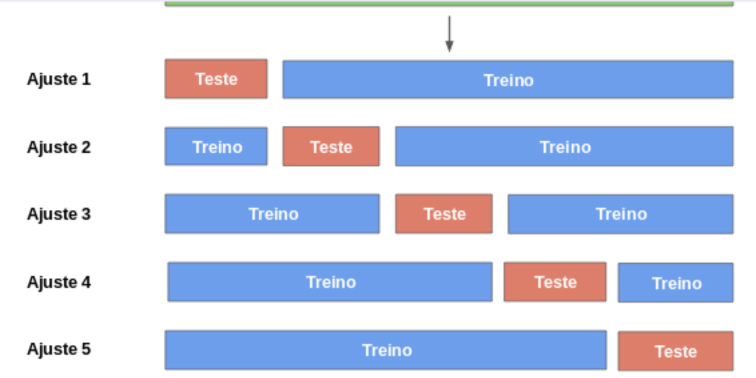


Figura 5: Representação esquemática da divisão dos dados para validação cruzada de ordem k.

Bootstrap

- Com o progresso de métodos computacionais e com capacidade cada vez maior de lidar com conjuntos grandes de dados, o cálculo de erros padrões, vieses etc, pode ser feito sem recorrer a uma teoria, que muitas vezes pode ser muito complicada ou simplesmente não existir.
- Um desses métodos é chamado bootstrap , introduzido por B. Efrom, em 1979. A ideia básica do método bootstrap é re-amostrar o conjunto disponível de dados para estimar um parâmetro θ, com o fim de criar dados replicados. A partir dessas replicações, podemos avaliar a variabilidade de um estimador proposto para θ, sem recorrer a cálculos analíticos.
- Suponha que temos dados $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ e queremos estimar a mediana populacional, Md, por meio da mediana amostral $\mathrm{md}(\mathbf{x}) = \mathrm{med}(x_1, \ldots, x_n)$.
- Escolhemos uma amostra aleatória simples, *com reposição*, de tamanho n dos dados. Tal amostra é chamada uma amostra bootstrap e denotada por $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$.

Bootstrap

- Suponha, agora, que geramos B tais amostras independentes, denotadas $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$. Para cada amostra bootstrap, geramos uma réplica bootstrap do estimador proposto, ou seja, de $\text{md}(x)$, obtendo-se

$$\text{md}(\mathbf{x}_1^*), \text{md}(\mathbf{x}_2^*), \ldots, \text{md}(\mathbf{x}_B^*). \tag{3}$$

- Definimos o estimador bootstrap do erro padrão de $\text{md}(\mathbf{x})$ como

$$\widehat{\text{e.p.}}_B(\text{md}) = \left[\frac{\sum_{b=1}^{B} (\text{md}(\mathbf{x}_b^*) - \text{md}(\cdot))^2}{B-1} \right]^{1/2}, \tag{4}$$

com

$$\text{md}(\cdot) = \frac{\sum_{b=1}^{B} \text{md}(\mathbf{x}_b^*)}{B}. \tag{5}$$

Ou seja, o estimador bootstrap do erro padrão da mediana amostral é o desvio padrão amostral do conjunto (3).

Bootstrap

- A questão que se apresenta é: qual deve ser o valor de B, ou seja, quantas amostras bootstrap devemos gerar para estimar erroes padrões de estimadores? A experiência indica que um valor razoável é $B = 200$.
- No caso geral de um estimador $\hat{\theta} = t(\mathbf{x})$, o algoritmo bootstrap para estimar o erro padrão de $\hat{\theta}$ é o seguinte:

[1] Selecione B amostras bootstrap independentes $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$, cada uma consistindo de n valores selecionados com reposição de $\mathbf{x}$. Tome $B \approx 200$.

[2] Para cada amostra bootstrap $\mathbf{x}_b^*$ calcule a réplica bootstrap

$$\hat{\theta}^*(b) = t(\mathbf{x}_b^*), \quad b = 1, 2, \ldots, B.$$

[3] O erro padrão de $\hat{\theta}$ é estimado pelo desvio padrão das B réplicas:

$$\widehat{\text{e.p.}}_B = \left[\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*(b) - \hat{\theta}^*(\cdot))^2\right]^{1/2}, \tag{6}$$

com

$$\hat{\theta}^*(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*(b). \tag{7}$$

Classificação

- De modo genérico, vamos designar a variável preditora por X (que pode ser escalar ou vetorial) e a resposta (indicadora de uma classe) por Y.
- Os dados serão indicados por (x_i, y_i), $i = 1, \ldots, n$. A ideia é usar os dados para obter agrupamentos cujos elementos sejam de alguma forma parecidos entre si (com base em alguma medida obtida a partir da variável preditora) e depois utilizar essa medida para classificar um ou mais novos elementos (para os quais dispomos apenas dos valores da variável preditora) em uma das classes.
- Se tivermos d variáveis preditoras e uma resposta dicotômica (*i.e.*, duas classes), um classificador é uma função que mapeia um espaço d-dimensional sobre $\{-1, 1\}$.
- Formalmente, seja (X, Y) um vetor aleatório, de modo que $X \in \mathbb{R}^d$ e $Y \in \{-1, 1\}$. Então, um classificador é uma função $g : \mathbb{R}^d \to \{-1, 1\}$ e a função erro ou risco é a probabilidade de erro, $L(g) = P\{g(X) \neq Y\}$.

Classificação

- A acurácia de um estimador de g, digamos $\widehat{g}$, pode ser medida pelo estimador de $L(g)$, chamado de taxa de erros, que é a proporção de erros gerados pela aplicação de $\widehat{g}$ às observações do conjunto de dados, ou seja,

$$\widehat{L}(\widehat{g}) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \widehat{y}_i), \tag{8}$$

  com $\widehat{y}_i = \widehat{g}(x_i)$ indicando o rótulo (-1 ou 1) da classe prevista por meio de $\widehat{g}$. Se $I(y_i \neq \hat{y}_i) = 0$, a i-ésima observação estará classificada corretamente.
- Sob o enfoque de aprendizado automático (AA), o objetivo é comparar diferentes modelos para identificar aquele com menor taxa de erros.

Classificação

- Nesse contexto, dispomos de um conjunto de dados de treinamento (x_i, y_i), $i = 1, \ldots, n$ e de um conjunto de dados de teste, cujo elemento típico é (x_0, y_0). O interesse é minimizar a taxa de erro de teste associada ao conjunto de observações teste que pode ser estimada por

$$\text{Média}[I(y_0 \neq \widehat{y_0})], \tag{9}$$

  em que a média é calculada relativamente aos elementos do conjunto de dados de teste.
- O classificador (ou modelo) ótimo é aquele que minimiza (9).
- Com o objetivo de classificar os elementos do conjunto de dados, deve-se ajustar o classificador ótimo ao conjunto de dados disponíveis (treinamento e teste)e utilizar a estimativa $\widehat{g}$ daí obtida para classificar os elementos do conjunto de dados para classificação.
- Quando dispomos de apenas um conjunto de dados, podemos recorrer ao processo de validação cruzada para dividi-lo em conjuntos de dados de treinamento e de dados de teste.

Classificação clássica

Neste contexto, podemos ter:

1) Classificação por regressão logística;
2) Classificação bayesiana;
3) Função discriminante linear de Fisher;
4) Classificador K-vizinho mais próximo.
5) Outras propostas

Classificação por regressão logística - RL

- **Exemplo**. Os dados da Tabela 1 são extraídos de um estudo realizado no Hospital Universitário da Universidade de São Paulo com o objetivo de avaliar se algumas medidas obtidas ultrassonograficamente poderiam ser utilizadas como substitutas de medidas obtidas por métodos de ressonância magnética, considerada como padrão áureo, para avaliação do deslocamento do disco da articulação temporomandibular (doravante referido simplesmente como disco).
- Distâncias cápsula-côndilo (em mm) com boca aberta ou fechada (referidas, respectivamente, como distância aberta ou fechada no restante do texto) foram obtidas ultrassonograficamente de 104 articulações e o disco correspondente foi classificado como deslocado (1) ou não (0) segundo a avaliação por ressonância magnética. A variável resposta é o *status* do disco (1 = deslocado ou 0 = não).
- Consideremos um modelo logístico para a chance de deslocamento do disco, tendo apenas a distância aberta como variável explicativa: Nesse contexto, o modelo

$$\log[\theta(x_i;\alpha,\beta)]/[1-\theta(x_i;\alpha,\beta)] = \alpha + x_i\beta \tag{10}$$

$i = 1, \ldots, 104$.

RL-Exemplo

Tabela: Parte dos Dados de um estudo odontológico

Dist aberta	Dist fechada	Desloc disco	Dist aberta	Dist fechada	Desloc disco	Dist aberta	Dist fechada	Desloc disco
2.2	1.4	0	0.9	0.8	0	1.0	0.6	0
2.4	1.2	0	1.1	0.9	0	1.6	1.3	0
2.6	2.0	0	1.4	1.1	0	4.3	2.3	1
3.5	1.8	1	1.6	0.8	0	2.1	1.0	0
1.3	1.0	0	2.1	1.3	0	1.6	0.9	0
2.8	1.1	1	1.8	0.9	0	2.3	1.2	0
1.5	1.2	0	2.4	0.9	0	2.4	1.3	0
2.6	1.1	0	2.0	2.3	0	2.0	1.1	0
1.2	0.6	0	2.0	2.3	0	1.8	1.2	0
1.7	1.5	0	2.4	2.9	0	1.4	1.9	0
1.3	1.2	0	2.7	2.4	1	1.5	1.3	0
1.2	1.0	0	1.9	2.7	1	2.2	1.2	0
4.0	2.5	1	2.4	1.3	1	1.6	2.0	0
1.2	1.0	0	2.1	0.8	1	1.5	1.1	0
3.1	1.7	1	0.8	1.3	0	1.2	0.7	0
2.6	0.6	1	0.8	2.0	1	1.5	0.8	0
1.8	0.8	0	0.5	0.6	0	1.8	1.1	0
1.2	1.0	0	1.5	0.7	0	2.3	1.6	1
1.9	1.0	0	2.9	1.6	1	1.2	0.4	0
1.2	0.9	0	1.4	1.2	0	1.0	1.1	0
1.7	0.9	1	3.2	0.5	1	2.9	2.4	1
1.2	0.8	0	1.2	1.2	0	2.5	3.3	1

Dist aberta: distância cápsula-côndilo com boca aberta (mm)
Dist fechada: distância cápsula-côndilo com boca fechada (mm)
Desloc disco: deslocamento do disco da articulação temporomandibular (1=sim, 0=não)

RL–Exemplo

- No modelo, $\theta(x_i; \alpha, \beta)$ representa a probabilidade de deslocamento do disco quando o valor da distância aberta é x_i, α denota o logaritmo da chance de deslocamento do disco quando a distância aberta tem valor $x_i = 0$ e β é interpretado como a variação no logaritmo da chance de deslocamento do disco por unidade de variação da distância aberta.
- Consequentemente, a razão de chances do deslocamento do disco correspondente a uma diferença de d unidades da distância aberta será $\exp(d \times \beta)$. Como não temos dados correspondentes a distâncias abertas menores que 0,5, convém substituir os valores x_i por valores "centrados", ou seja por $x_i^* = x_i - x_0$.
- Uma possível escolha para x_0 é o mínimo de x_i, que é 0,5. Essa transformação na variável explicativa altera somente a interpretação do parâmetro α que passa a ser o logaritmo da chance de deslocamento do disco quando a distância aberta tem valor $x_i = 0,5$.

RL–Exemplo–Uso do R

```
Call:
glm(formula = deslocamento ~ (distanciaAmin), family = binomial,
    data = disco)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5240  -0.4893  -0.3100   0.1085   3.1360

Coefficients:
              Estimate  Std. Error z value Pr(>|z|)
(Intercept)    -5.8593     1.1003  -5.325  1.01e-07 ***
distanciaAmin   3.1643     0.6556   4.827  1.39e-06 ***
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.11  on 103  degrees of freedom
Residual deviance:  71.60  on 102  degrees of freedom
AIC: 75.6

Number of Fisher Scoring iterations: 6
```

RL–Exemplo

- Estimativas (com erros padrões entre parênteses) dos parâmetros desse modelo ajustado por máxima verossimilhança aos dados da Tabela 1, são, $\widehat{\alpha} = -5,86$ (1, 10) e $\widehat{\beta} = 3,16$ (0, 66) e então, segundo o modelo, uma estimativa da chance de deslocamento do disco para articulações com distância aberta $x = 0,5$ (que corresponde à distância aberta transformada $x^* = 0,0$) é $\exp(-5,86) = 0,003$.
- Um intervalo de confiança (95%) para essa chance pode ser obtido exponenciando os limites (LI e LS) do intervalo para o parâmetro α, nomeadamente,

$$\begin{aligned} LI &= \exp[\widehat{\alpha} - 1,96 EP(\widehat{\alpha})] = \exp(-5,86 - 1,96 \times 1,10) = 0,000 \\ LS &= \exp[\widehat{\alpha} + 1,96 EP(\widehat{\alpha})] = \exp(-5,86 + 1,96 \times 1,10) = 0,025. \end{aligned}$$

- Os limites de um intervalo de confiança para a razão de chances correspondentes a um variação de uma unidade no valor da distância aberta podem ser obtidos de maneira similar e são 6,55 e 85,56.
- Substituindo os parâmetros α e β por suas estimativas $\widehat{\alpha}$ e $\widehat{\beta}$ em (10) podemos estimar a probabilidade de sucesso (deslocamento do disco, no exemplo sob investigação).

RL–Exemplo

- Por exemplo, para uma articulação cuja distância aberta seja 2,1 (correspondente à distância aberta transformada igual a 1,6), a estimativa dessa probabilidade é

$$\widehat{\theta} = \exp(-5,86 + 3,16 \times 1,6)/[1 + \exp(-5,86 + 3,16 \times 1,6)] = 0,31.$$

- Lembrando que o objetivo do estudo é substituir o processo de identificação de deslocamento do disco realizado via ressonância magnética por aquele baseado na medida da distância aberta por meio de ultrassonografia, podemos estimar as probabilidades de sucesso para todas as articulações e identificar um ponto de corte d_0 segundo o qual, distâncias abertas com valores acima dele sugerem decidirmos pelo deslocamento do disco e distâncias abertas com valores abaixo dele sugerem a decisão oposta.
- Obviamente, não esperamos que todas as decisões tomadas dessa forma sejam corretas e consequentemente, a escolha do ponto de corte deve ser feita com o objetivo de minimizar os erros (decidir pelo deslocamento quando ele não existe ou *vice versa*).

RL–Exemplo

Nesse contexto, um contraste entre as decisões tomadas com base em um determinado ponto de corte d_0 e o padrão áureo definido pela ressonância magnética para todas as 104 articulações pode ser resumido por meio da Tabela 2, em que as frequências da diagonal principal correspondem a decisões corretas e aquelas da diagonal secundária às decisões erradas.

Tabela: Frequência de decisões para um ponto de corte d_0

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	n_{11}	n_{12}
distância aberta d_0	não	n_{21}	n_{22}

RL–Exemplo

- O quociente $n_{11}/(n_{11} + n_{21})$ é conhecido como sensibilidade do processo de decisão e é uma estimativa da probabilidade de decisões corretas quando o disco está realmente deslocado.
- O quociente $n_{22}/(n_{12} + n_{22})$ é conhecido como especificidade do processo de decisão e é uma estimativa da probabilidade de decisões corretas quando o disco realmente não está deslocado. A situação ideal é aquela em que tanto a sensibilidade quanto a especificidade do processo de decisão são iguais a 100%.
- O problema a resolver é determinar o ponto de corte d_{max} que gere o melhor equilíbrio entre sensibilidade e especificidade. Com essa finalidade, podemos construir tabelas com o mesmo formato da Tabela 2 para diferentes pontos de corte e um gráfico cartesiano entre a sensibilidade e especificidade obtida de cada uma delas.
- Esse gráfico, conhecido como curva ROC (do termo inglês Receiver Operating Characteristic) gerado para os dados da Tabela 1 está apresentado na Figura 1.

RL–Exemplo

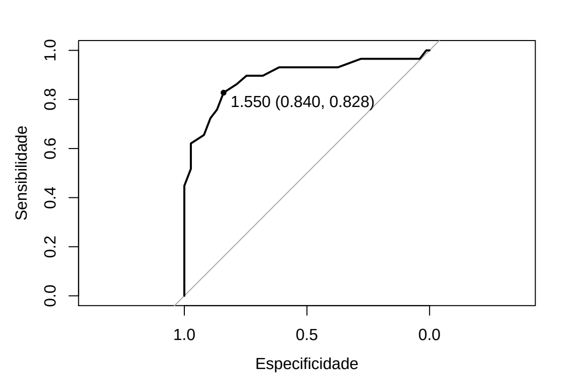


Figura: Curva ROC para os dados da Tabela 1 baseada no modelo (10) com distância aberta como variável explicativa.

RL–Exemplo

- O ponto de corte ótimo é aquele mais próximo do vértice superior esquerdo (em que tanto a sensibilidade quanto a especificidade seriam iguais a 100%).
- Para o exemplo, esse ponto está salientado na Figura 1 e corresponde à distância aberta com valor $d_{max} = 2,05$ $(= 1,55 + 0,5)$. A sensibilidade e a especificidade associadas à decisão baseada nesse ponto de corte, são, respectivamente, 83% e 84% e as frequências de decisões corretas estão indicadas na Tabela 3.

Tabela: Frequência de decisões para um ponto de corte para distância aberta $d_{max} = 2,05$

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	24	12
distância aberta $d_{max} = 2,05$	não	5	63

RL–Exemplo

- Com esse procedimento de decisão a porcentagem de acertos (**acurácia**) é 84% [$= (24+63)/104$]. A porcentagem de **falsos positivos** é 17% [$= 5/(5+29)$] e a porcentagem de **falsos negativos** é 16% [$= 12/(12+63)$].
- Um gráfico de dispersão com o correspondente ponto de corte baseado apenas na distância aberta está apresentado na Figura 2 com símbolos vermelhos indicando casos com deslocamento do disco e em preto indicando casos sem deslocamento. Os valores de ambas as distâncias foram ligeiramente alterados para diminuir a superposição nos pontos.

RL–Exemplo

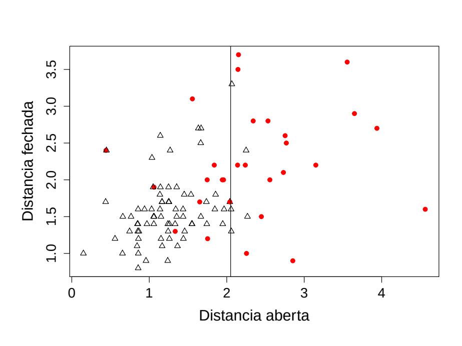


Figura: Gráfico de dispersão para os dados da Tabela 1 com ponto de corte baseado apenas na distância aberta.

RL–Exemplo

Uma análise similar, baseada na distância fechada (transformada por meio da subtração de deu valor mínimo, 0,4) gera a curva ROC apresentada na Figura 3 e frequências de decisões apresentada na Tabela 4.

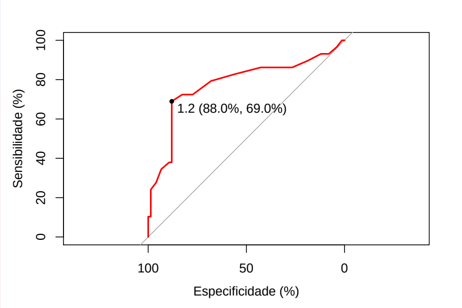


Figura: Curva ROC para os dados da Tabela 1 baseada no modelo (10) com distância fechada como variável explicativa.

RL–Exemplo

Tabela: Frequência de decisões para um ponto de corte para distância fechada $d_{max} = 1,6$

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	20	9
distância fechada $d_{max} = 1,6$	não	9	66

RL–Exemplo

- A acurácia associada a processo de decisão baseado apenas na distância fechada, 83% $[= (20 + 66)/104]$ é praticamente igual àquela obtida com base apenas na distância aberta; no entanto aquele processo apresenta um melhor equilíbrio entre sensibilidade e especificidade (83% e 84%, respectivamente, *versus* 69% e 88%).
- Se quisermos avaliar o processo de decisão com base nas observações das distâncias aberta e fechada simultaneamente, podemos considerar o modelo

$$\log[\theta(x_i; \alpha, \beta, \gamma)]/[1 - \theta(x_i; \alpha, \beta, \gamma)] = \alpha + x_i\beta + w_i\gamma \qquad (11)$$

$i = 1, \ldots, 104$ em que w_i corresponde à distância fechada observada na i-ésima articulação.

RL–Exemplo

- Neste caso, γ corresponde à razão entre a chance de deslocamento do disco para articulações com distância fechada $w+1$ e a chance de deslocamento do disco para articulações com distância fechada w para aquelas com mesmo valor da distância aberta; uma interpretação similar vale para o parâmetro β.
- Estimativas dos parâmetros (com erros padrões entre parênteses) do modelo (11) obtidas após a transformação das variáveis explicativas segundo o mesmo figurino adotado nas análises univariadas são $\widehat{\alpha} = -6,38\ (1,19)$, $\widehat{\beta} = 2,83\ (0,67)$ e $\widehat{\gamma} = 0,98\ (0,54)$.
- A estimativa do parâmetro γ é apenas marginalmente significativa, ou seja a inclusão da variável explicativa distância fechada não acrescenta muito poder de discriminação além daquele correspondente à distância aberta.

RL–Exemplo

- Uma das razões para isso é que as duas variáveis são correlacionadas (com coeficiente de correlação de Pearson igual a 0,46. A determinação de pontos de corte para modelos com duas ou mais variáveis explicativas é bem mais complexa do que no caso univariado e não será abordada neste texto.
- Para efeito de comparação com as análises anteriores, as frequências de decisões obtidas com os pontos de corte utilizados naquelas estão dispostas na Tabela 5, e correspondem a uma sensibilidade de 62%, especificidade de 97% e acurácia de 88%.

Tabela: Frequência de decisões correspondentes a pontos de corte $d_{max} = 2,05$ para distância aberta e $d_{max} = 1,6$ para distância fechada

		Deslocamento real do disco	
		sim	não
Decisão baseada em	sim	18	2
ambas as distâncias	não	11	73

RL–Exemplo

- Numa segunda análise, agora sob o paradigma de aprendizado automático (AA), a escolha do modelo ótimo é baseada apenas nas porcentagens de classificação correta (acurácia) obtidas por cada modelo num conjunto de dados de teste a partir de seu ajuste a um conjunto de dados de treinamento. Como neste caso não dispomos desses conjuntos *a priori*, podemos recorrer à técnica de **validação cruzada**.
- Neste exemplo, utilizamos validação cruzada de ordem 5 com 5 repetiçoes (VC5/5), em que o conjunto de dados é dividido em dois, cinco vezes, gerando cinco conjuntos de dados de treinamento e de teste. A análise é repetida cinco vezes em cada conjunto e a a acurácia média obtida das 25 análises serve de base para a escolha do melhor modelo.
- Comparamos quatro modelos de regressão logística, os dois primeiros com apenas uma das variáveis preditoras (distância aberta ou distância fechada), o terceiro com ambas incluídas aditivamente e o último com ambas as distâncias e sua interação. A análise pode ser concretizada por meio do pacote caret.

RL–Exemplo

Os resultados estão dispostos na Tabela 6 tanto para validação cruzada VC5/5 quanto para validação cruzada LOOCV.

Tabela: Acurácia obtida por validação cruzada para as regressões logísticas ajustados as dados do Exemplo 8.1

Modelo	Variáveis	Acurácia VC5/5	Acurácia LOOCV
1	Distância aberta	84,8 %	84,6 %
2	Distância fechada	75,2 %	74,0 %
3	Ambas (aditivamente)	85,7 %	85,6 %
4	Ambas + Interação	83,6 %	83,6 %

RL–Exemplo

- Com ambos os critérios, o melhor modelo é aquele que inclui as duas variáveis preditoras de forma aditiva. Para efeito de classificar uma nova observação (para a qual só dispomos dos valores das variáveis preditoras, o modelo selecionado deve ser ajustado ao conjunto de dados original (treinamento + teste) para obtenção dos coeficientes do classificador.
- Os comandos e a saída associada ao ajuste desse modelo aos 5 conjuntos de dados gerados para validação cruzada e no conjunto completo seguem.
- A seleção obtida por meio do AA corresponde ao modelo (11). Embora a variável Distância fechada seja apenas marginalmente significativa, sua inclusão aumenta a proporção de acertos (acurácia) de 84% no modelo que inclui apenas Distância aberta para 86%.
- A estatísca Kappa apresentada juntamente com a acurácia serve para avaliar a concordância entre o processo de classificação e a classificação observada (veja a Seção 4.2).

RL–Exemplo

```
> set.seed(369321)
> train_control =
              trainControl(method="repeatedcv", number=5, repeats=5)
> model3 = train(deslocamento ~ distanciaAmin + distanciaFmin,
          data=disco, method="glm", family=binomial,
          trControl=train_control)
> model3
Generalized Linear Model

104 samples
  2 predictor
  2 classes: 0, 1

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 5 times)
Summary of sample sizes: 83, 83, 84, 83, 83, 83, ...
Resampling results:

  Accuracy   Kappa
  0.8573333  0.6124102

> disco$predito3 = predict(model3, newdata=disco, type="raw")
> summary(model3$finalModel)
```

RL–Exemplo

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.82771  -0.45995  -0.28189   0.07403   2.82043

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.3844     1.1932  -5.351 8.76e-08 ***
distanciaAmin   2.8337     0.6676   4.245 2.19e-05 ***
distanciaFmin   0.9849     0.5383   1.830   0.0673 .


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.107  on 103  degrees of freedom
Residual deviance:  67.991  on 101  degrees of freedom
AIC: 73.991
Number of Fisher Scoring iterations: 6

> table(disco$deslocamento, disco$predito3)
     0  1
  0 72  3
  1 12 17
```

Como a divisão do conjunto original nos subconjuntos de treinamento e de teste envolve uma escolha aleatória, os resultados podem diferir (em geral de forma desprezável) para diferentes aplicações dos mesmos comandos, a não ser que se especifique a semente do processo aleatório de divisão por meio do comando set.seed().

Referências

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparcity*. Chapman and Hall.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Morettin, P. A. e Singer, J. M. (2021). *Estatística e Ciência de Dados*. Texto Preliminar, IME-USP.

MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 8

24 de abril de 2023

Sumário

ADL

Podemos ter:

1) Classificador de Bayes
2) Classificador linear de Fisher
3) Classificador do vizinho mais próximo
4) Outras propostas

ADL: classificador de Bayes

- Consideremos um conjunto de dados, $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, em que $\mathbf{x}_i$ representa os valores de p variáveis preditoras (explicativas) e y_i representa o valor de uma variável resposta indicadora da classe a que o i-ésimo elemento desse conjunto pertence.
- Seja π_k a probabilidade *a priori* de que um elemento com valor das variáveis preditoras $\mathbf{x} = (x_1, \ldots, x_p)$ pertença à classe C_k, $k = 1, \ldots, K$ e seja $f_k(\mathbf{x})$ a função densidade de probabilidade da variável preditora $\mathbf{X}$ para valores $\mathbf{x}$ associados a elementos dessa classe. Por um abuso de notação escrevemos $f_k(\mathbf{x}) = P(\mathbf{X} = \mathbf{x} | Y = k)$ (que a rigor só vale no caso discreto).
- Pelo teorema de Bayes,

$$P(Y = k | \mathbf{X} = \mathbf{x}) = p_k(\mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\mathbf{x})}, \quad k = 1, \ldots, K, \qquad (1)$$

é a probabilidade a posteriori de que um elemento com valor das variáveis preditoras igual a $\mathbf{x}$ pertença à k-ésima classe. Para calcular essa probabilidade é necessário conhecer π_k e $f_k(\mathbf{x})$; em muitos casos, supõe-se que para os elementos da k-ésima classe, os valores de $\mathbf{X}$ tenham uma distribuição $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, ou seja, com média que depende da classe k e matriz de covariâncias comum a todas as classes.

ADL: classificador de Bayes

- Suponha que $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$. Uma **regra de classificação**, R, consiste em dividir $\mathcal{X}$ em K regiões disjuntas $\mathcal{X}_1, \ldots, \mathcal{X}_K$, tal que se $\mathbf{x} \in \mathcal{X}_k$, o elemento correspondente é classificado em C_k.
- A probabilidade (condicional) de classificação incorreta, *i.e.*, de classificar um elemento com valor das variáveis preditoras $\mathbf{x}$ em C_k, quando de fato ele pertence a C_j, $j \neq k$ usando a regra R, é

$$p(C_k|C_j, R) = \int_{\mathcal{X}_k} f_j(\mathbf{x}) d\mathbf{x}. \tag{2}$$

Se $k = j$ em (2), obtemos a probabilidade de classificação correta do elemento com valor das variáveis preditoras $\mathbf{x}$ em C_k.

ADL: classificador de Bayes

- Em muitos casos é possível incluir um **custo** de classificação incorreta, denotado por $Q(C_k|C_j)$ no procedimento de classificação. Usualmente, esses custos não são iguais e admite-se que $Q(C_k|C_k) = 0$, $k = 1, \ldots, K$.
- O custo médio de classificação incorreta segundo a regra R é dado por

$$\delta(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \left[\sum_{k=1, j \neq k}^{K} p(C_j|C_k, R) Q(C_j|C_k) \right]. \tag{3}$$

- O Classificador de Bayes é obtido por meio da minimização desse custo médio, supondo os custos de classificação incorreta iguais, ou seja, o elemento com valor das variáveis preditoras $\mathbf{x}$ deve ser classificado em C_k, se

$$\delta_k(\mathbf{x}) = \sum_{k=1, j \neq k}^{K} \pi_j f_j(\mathbf{x}) \tag{4}$$

for mínima, $k = 1, \ldots K$.

ADL: classificador de Bayes

- Minimizar (4) é equivalente a classificar $\mathbf{x}$ em C_k se

$$\pi_k f_k(\mathbf{x}) = max_{1 \le j \le K}[\pi_j f_j(\mathbf{x})], \tag{5}$$

pois devemos excluir a k-ésima parcela de (4) que seja máxima, relativamente a todas as possíveis exclusões de parcelas.

- Em particular, se $K = 2$, elementos com valor das variáveis preditoras igual a $\mathbf{x}$ devem ser classificados em C_1 se

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \ge \frac{\pi_2}{\pi_1}, \tag{6}$$

e em C_2, caso contrário. Veja Johnson e Wichern (1998) e Ferreira (2011), para detalhes.

ADL: classificador de Bayes

- Suponha o caso $K = 2$ com variáveis $\mathbf{x}$ seguindo distribuições normais, com médias $\boldsymbol{\mu}_1$ para elementos da classe C_1, $\boldsymbol{\mu}_2$ para elementos da classe C_2 e matriz de covariâncias $\boldsymbol{\Sigma}$ comum.
  Usando (1), obtemos

$$P(Y = k|\mathbf{X} = xb) = \frac{\pi_k \exp\{-(\mathbf{x} - \mu_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu_k)/2\}}{\sum_{\ell=1}^{K} \pi_\ell \exp\{-(\mathbf{x} - \mu_\ell)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu_\ell)/2\}}, \quad k = 1, 2. \tag{7}$$

- Então, elementos com valores das variáveis preditoras iguais a $\mathbf{x}$ são classificados em C_1 se

$$\mathbf{d}^\top \mathbf{x} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \geq \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \log(\pi_2/\pi_1). \tag{8}$$

em que $\mathbf{d} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ contém os coeficientes da função discriminante.

ADL: classificador de Bayes

- No caso geral ($K \geq 2$), o classificador de Bayes associa um elemento com valor das variáveis preditoras igual a $\mathbf{x}$ à classe para a qual

$$\delta_k(\mathbf{x}) = \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log \pi_k \tag{9}$$

for **máxima**.

- Em particular, para $p = 1$, devemos maximizar

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k. \tag{10}$$

- Quando há apenas duas classes, C_1 e C_2, um elemento com valor da variável preditora igual a x deve ser classificado na classe C_1 se

$$dx = \frac{\mu_1 - \mu_2}{\sigma^2}x \geq \frac{\mu_1^2 - \mu_2^2}{2\sigma^2} + \log \frac{\pi_2}{\pi_1} \tag{11}$$

e na classe C_2 em caso contrário.

ADL: classificador de Bayes

- As fronteiras de Bayes [valores de $\mathbf{x}$ para os quais $\delta_k(\mathbf{x}) = \delta_\ell(\mathbf{x})$] são obtidas como soluções de

$$\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log \pi_k = \boldsymbol{\mu}_\ell^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_\ell^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell + \log \pi_\ell, \quad (12)$$

para $k \neq \ell$.

- No paradigma bayesiano, os termos utilizados para cálculo das probabilidades *a posteriori* (1) são conhecidos, o que na prática não é realista. No caso $p = 1$ pode-se aproximar o classificador de Bayes substituindo π_k, μ_k, $k = 1, \ldots, K$ e σ^2 pelas estimativas

$$\widehat{\pi}_k = n_k/n$$

em que n_k corresponde ao número dos n elementos do conjunto de dados de treinamento pertencentes à classe k,

$$\overline{x}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i, \quad \text{e} \quad S^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \widehat{\mu}_k)^2.$$

ADL: classificador de Bayes

- Com esses estimadores, a fronteira de decisão de Bayes corresponde a solução de
$$(\overline{x}_1 - \overline{x}_2)x = (\overline{x}_1^2 - \overline{x}_2^2)/2 + [\log(\widehat{\pi}_2/\widehat{\pi}_1)]S^2. \tag{13}$$
- No caso $K = 2$ e $p \geq 2$, os parâmetros $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}$ são desconhecidas e têm que ser estimadas a partir de amostras das variáveis preditoras associadas aos elementos de C_1 e C_2. Com os dados dessas amostras, podemos obter estimativas $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \mathbf{S}_1$ e $\mathbf{S}_2$, das respectivas médias e matrizes de covariâncias.
- Uma estimativa não enviesada da matriz de covariâncias comum $\boldsymbol{\Sigma}$ é
$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}. \tag{14}$$
- Quando $\pi_1 = \pi_2$, elementos com valor das variáveis preditoras igual a $\mathbf{x}$ são classificados em C_1 se
$$\widehat{\mathbf{d}}^\top \mathbf{x} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top \mathbf{S}^{-1}\mathbf{x} \geq \frac{1}{2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top \mathbf{S}^{-1}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2) \tag{15}$$

com $\widehat{\mathbf{d}} = \mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$.

Classificador de Bayes: exemplo 1

Exemplo 1. Suponha que $f_1(\mathbf{x})$ seja a densidade de uma distribuição normal padrão e $f_2(x)$ seja a densidade de uma distribuição normal com média 2 e variância 1. Supondo $\pi_1 = \pi_2 = 1/2$, elementos com valor das variáveis preditoras igual a $\mathbf{x}$ são classificados em $\mathcal{C}_1$ se $f_1(\mathbf{x})/f_2(\mathbf{x}) \geq 1$ o que equivale a

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = e^{-x^2/2} e^{(x-2)^2/2} \geq 1,$$

ou seja, se $x \leq 1$. Consequentemente, as duas probabilidades de classificação incorretas são iguais a $0,159$.

Classificador de Bayes: exemplo 2

Exemplo 2: Consideremos os dados do arquivo inibina, analisados por meio de regressão logística no Exemplo 6.9. Um dos objetivos é classificar as pacientes como tendo resposta positiva ou negativa ao tratamento com inibina com base na variável preditora `difinib = inibpos-inibpre`. Das 32 pacientes do conjunto de dados, 59, 4% apresentaram resposta positiva (classe C_1) e 40, 5% apresentaram resposta negativa (classe C_2).

Estimativas das médias das duas classes são, respectivamente, $\overline{x}_1 = 202,7$ e $\overline{x}_0 = 49,0$.
Estimativas das correspondentes variâncias são $S_1^2 = 31630,5$ e $S_0^2 = 2852,8$ e uma estimativa da variância comum é $S^2 = (18 \times S_1^2 + 12 \times S_0^2)/30 = 20119,4$.

De (11) obtemos o coeficiente da função discriminante $d = (\overline{x}_1 - \overline{x}_0)/S^2 = 0,0076$. Para decidir em que classe uma paciente com valor de `difinib` $= x$ deve ser alocada, devemos comparar d com $(\overline{x}_1^2 - \overline{x}_0^2)/(2S^2) + [\log(\widehat{\pi}_0/\widehat{\pi}_1)] = 0,58191$.

Esses resultados podem ser concretizada por meio da função lda() do pacote MASS.

Classificador de Bayes: exemplo 2

```
lda(inibina$resposta ~ inibina$difinib, data = inibina)
Prior probabilities of groups:
negativa positiva
 0.40625  0.59375
Group means:
         inibina$difinib
negativa        49.01385
positiva       202.70158
Coefficients of linear discriminants:
                        LD1
inibina$difinib 0.007050054
```

A função considera as proporções de casos negativos (41%) e positivos (59%) no conjunto de dados de treinamento como probabilidades *a priori*, dado que elas não foram especificadas no comando.
O coeficiente da função discriminante (0.00705) corresponde à combinação linear de `difinib` usada para a decisão difere daquele obtido acima $(0,0076)$ pois a função `lda()` considera uma transformação com a finalidade de deixar os resultados com variância unitária (o que não influi na classificação).

Classificador de Bayes: exemplo 2

```
lda(inibina$resposta ~ inibina$difinib, data = inibina)
Prior probabilities of groups:
negativa positiva
 0.40625  0.59375
Group means:
         inibina$difinib
negativa        49.01385
positiva       202.70158
Coefficients of linear discriminants:
                        LD1
inibina$difinib 0.007050054
```

A função considera as proporções de casos negativos (41%) e positivos (59%) no conjunto de dados de treinamento como probabilidades *a priori*, dado que elas não foram especificadas no comando.
O coeficiente da função discriminante (0.00705) corresponde à combinação linear de `difinib` usada para a decisão difere daquele obtido acima $(0,0076)$ pois a função `lda()` considera uma transformação com a finalidade de deixar os resultados com variância unitária (o que não influi na classificação).

Classificador de Bayes: exemplo 2

Uma tabela relacionando a classificação predita com os valores reais da resposta pode ser obtido por meio dos comandos

```
predito <- predict(fisher)
table(predito$class, inibina$resposta)
           negativa positiva
  negativa        9        2
  positiva        4       17
```

indicando que a probabilidade de classificação correta é 81%, ligeiramente superior ao que foi conseguido com o emprego de regressão logística (ver Exemplo 6.9). Histogramas para os valores da função discriminante calculada para cada elemento do conjunto de dados estão dispostos na Figura 1.

Classificador de Bayes: exemplo 2

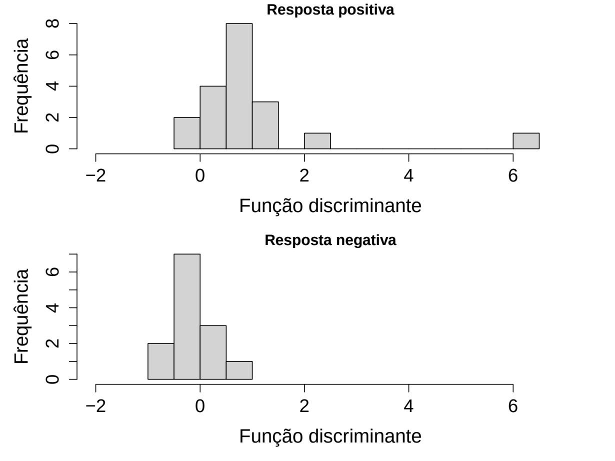


Figura 1: Histogramas para valores da função discriminante.

Função discriminante de Fisher

- Consideremos novamente o caso de duas classes (ou populações), $\mathcal{G}_1$ e $\mathcal{G}_2$ para as quais pretendemos obter um classificador com base em um vetor de variáveis preditoras, $\mathbf{X} = (X_1, \ldots, X_p)^\top$.
- A ideia de Fisher é considerar uma combinação linear $Y = \boldsymbol{\ell}^\top \mathbf{X}$, com $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_p)^\top$ de modo que o conjunto de variáveis preditoras seja transformado numa variável escalar Y.
- Sejam μ_{1Y} e μ_{2Y}, respectivamente, as médias de Y obtidas dos valores de $\mathbf{X}$ associadas aos dados $\mathcal{G}_1$ e $\mathcal{G}_2$. A regra para classificação consiste em selecionar a combinação linear que maximiza a distância quadrática entre essas duas médias, relativamente à variabilidade dos valores de Y.
- Uma suposição adicional e, às vezes, irrealista, é que as matrizes de covariâncias
$$\boldsymbol{\Sigma}_i = E(\mathbf{X} - \boldsymbol{\mu}_i)(\mathbf{X} - \boldsymbol{\mu}_i)^\top, \tag{16}$$
$i = 1, 2$, em que $\boldsymbol{\mu}_1 = E(\mathbf{X}|\mathcal{G}_1)$ e $\boldsymbol{\mu}_2 = E(\mathbf{X}|\mathcal{G}_2)$, sejam iguais para as duas classes, isto é, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$.

FDLF

- Consequentemente,

$$\sigma_Y^2 = var(\boldsymbol{\ell}^\top \mathbf{X}) = \boldsymbol{\ell}^\top \boldsymbol{\Sigma} \boldsymbol{\ell}$$

  é igual para ambas as classes.

- $$\mu_{1Y} = E(Y|\mathcal{G}_1) = \boldsymbol{\ell}^\top \boldsymbol{\mu}_1 \ \text{ e } \ \mu_{2Y} = E(Y|\mathcal{G}_2) = \boldsymbol{\ell}^\top \boldsymbol{\mu}_2$$

  e a razão

$$\begin{aligned} \frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2} &= \frac{(\boldsymbol{\ell}^\top \boldsymbol{\mu}_1 - \boldsymbol{\ell}^\top \boldsymbol{\mu}_2)^2}{\boldsymbol{\ell}^\top \boldsymbol{\Sigma} \boldsymbol{\ell}} = \frac{\boldsymbol{\ell}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\ell}}{\boldsymbol{\ell}^\top \boldsymbol{\Sigma} \boldsymbol{\ell}} \\ &= \frac{(\boldsymbol{\ell}^\top \boldsymbol{\delta})^2}{\boldsymbol{\ell}^\top \boldsymbol{\Sigma} \boldsymbol{\ell}}, \end{aligned} \tag{17}$$

  com $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ é maximizada se

$$\boldsymbol{\ell} = c\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = c\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{18}$$

  para todo $c \neq 0$.

- No caso $c = 1$, obtemos a função discriminante linear de Fisher

$$Y = \boldsymbol{\ell}^\top \mathbf{X} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \mathbf{X}. \tag{19}$$

  e o valor máximo da razão (17) é $\boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}$.

FDLF

- Para uma nova observação $\mathbf{x}_0$, sejam $y_0 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}_0$ e

$$\mu = \frac{\mu_{1Y} + \mu_{2Y}}{2} = \frac{1}{2}(\boldsymbol{\ell}^\top \boldsymbol{\mu}_1 + \boldsymbol{\ell}^\top \boldsymbol{\mu}_2) \tag{20}$$

(o ponto médio entre as médias univariadas associadas às duas classes). Em virtude de (19), esse ponto médio pode ser expresso como

$$\mu = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{2}, \tag{21}$$

- Consequentemente,

$$E(Y_0|\mathcal{G}_1) - \mu \geq 0 \text{ e } E(Y_0|\mathcal{G}_2) - \mu < 0.$$

e uma regra de classificação é

$$\begin{aligned}&\text{Classifique } \mathbf{x}_0 \text{ em } \mathcal{G}_1 \text{ se } y_0 \geq \mu,\\ &\text{Classifique } \mathbf{x}_0 \text{ em } \mathcal{G}_2 \text{ se } y_0 < \mu.\end{aligned}$$

Estimativa da FDLF

- Normalmente, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}$ são desconhecidas e têm que ser estimadas a partir de amostras de $\mathcal{G}_1$ e $\mathcal{G}_2$, denotadas por $\mathbf{X}_1 = [\mathbf{x}_{11}, \ldots, \mathbf{x}_{1,n_1}]$, uma matriz com dimensão $p \times n_1$ e $\mathbf{X}_2 = [\mathbf{x}_{21}, \ldots, \mathbf{x}_{n_2}]$, uma matriz com dimensão $p \times n_2$.
- Com os dados dessas amostras, podemos obter estimativas $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \mathbf{S}_1$ e $\mathbf{S}_2$, das médias e da matriz de covariâncias comum $\boldsymbol{\Sigma}$, para a qual um estimador não enviesado é

$$\mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}. \tag{22}$$

- A função discriminante estimada é $\widehat{\boldsymbol{\ell}}^\top \mathbf{x} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top \mathbf{S}_p \mathbf{x}$ e a regra de classificação é:

$$\begin{array}{l} \text{Classifique a observação } \mathbf{x}_0 \text{ em } \mathcal{G}_1 \text{ se } y_0 - \widehat{\mu} \geq 0, \\ \text{Classifique a observação } \mathbf{x}_0 \text{ em } \mathcal{G}_2 \text{ se } y_0 - \widehat{\mu} < 0, \end{array}$$

em que $\widehat{\mu} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top \mathbf{S}_p^{-1}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)$.
- Nas demais expressões, os parâmetros são substituídas pelas respectivas estimativas.
- Outra suposição comumente adotada é que as variáveis preditoras têm distribuição Normal multivariada. Nesse caso, a solução encontrada por meio da função discriminante linear de Fisher é ótima.

Classificador KNN

- Vimos que o classificador de Bayes associa cada observação de teste com o valor do preditor x_0 à classe j de forma que

$$P(Y = j|X = x_0) \tag{23}$$

  seja a maior possível.

- No caso de duas classes, a observação será associada à Classe 1 se $P(Y = 1|X = x_0) > 0,5$ e à Classe 2, se $P(Y = 0|X = x_0) < 0,5$. A fronteira de Bayes é $P(Y = 1|X = x_0) = 0,5$.
- A taxa de erro de Bayes global é $1 - E(\max_j P(Y = j|X))$, obtida com base na média de todas as taxas de erro sobre todos os valores possíveis de j.
- Na prática como não conhecemos a distribuição condicional de Y, dado X, precisamos estimar essa probabilidade condicional, o que pode ser efetivado por meio de um método conhecido por K-ésimo vizinho mais próximo (*K-nearest neighbor*, KNN).

Classificador KNN

O algoritmo associado a esse método é:

i) Fixe K e uma observação teste x_0;

ii) Identifique K pontos do conjunto de dados de treinamento que sejam os mais próximos de x_0 segundo alguma medida de distância; denote esse conjunto por $\mathcal{V}_0$;

iii) Estime a probabilidade condicional de que a observação teste pertença à Classe j como a fração dos pontos de $\mathcal{V}_0$ cujos valores de Y sejam iguais a j, ou seja, como

$$P(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{V}_0} I(y_i = j). \tag{24}$$

iv) classifique x_0 na classe asociada à maior probabilidade.

A função knn() do pacote caret pode ser utilizada com essa finalidade.

Classificador KNN-Exemplo

- **Exemplo.** Consideremos, novamente, os dados do arquivo inibina utilizando a variável difinib como preditora e adotemos a estratégia de validação cruzada por meio do método LOOCV. Além disso, avaliemos o efeito de considerar entre 1 e 5 vizinhos mais próximos no processo de classificação.
- Os comandos necessários para a concretização da análise são

```
set.seed(2327854)
trControl <- trainControl(method  = "LOOCV")

fit <- train(resposta ~ difinib, method = "knn",
             tuneGrid   = expand.grid(k = 1:5),
             trControl  = trControl, metric= "Accuracy",
             data = inibina)
fit
```

Classificador KNN-Exemplo

Os resultados correspondentes são:

```
k-Nearest Neighbors

32 samples
 1 predictor
 2 classes: 'negativa', 'positiva'

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 31, 31, 31, 31, 31, 31, ...
Resampling results across tuning parameters:

  k  Accuracy  Kappa
  1  0.71875   0.4240000
  2  0.78125   0.5409836
  3  0.81250   0.6016598
  4  0.78125   0.5409836
  5  0.81250   0.6016598

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

Classificador KNN-Exemplo

- Segundo o esse processo, o melhor resultado (com K=5 vizinhos) gera uma acurácia (média) de 81.3%. A tabela de classificação obtida por meio do ajuste do modelo final ao conjunto de dados original, juntamente com estatísticas descritivas pode ser obtida por meio dos comandos:

```
predito <- predict(fit)
confusionMatrix(predito, inibina\$resposta)
```

- que geram os seguintes resultados:

```
Confusion Matrix and Statistics
          Reference
Prediction negativa positiva
  negativa        9        1
  positiva        4       18

               Accuracy : 0.8438
                 95% CI : (0.6721, 0.9472)
    No Information Rate : 0.5938
    P-Value [Acc > NIR] : 0.002273
                  Kappa : 0.6639
 Mcnemar's Test P-Value : 0.371093
```

Classificador KNN-Exemplo

```
         Sensitivity : 0.6923
         Specificity : 0.9474
      Pos Pred Value : 0.9000
      Neg Pred Value : 0.8182
          Prevalence : 0.4062
      Detection Rate : 0.2812
Detection Prevalence : 0.3125
   Balanced Accuracy : 0.8198
```

A acurácia é de $84,4\%$, sensibilidade de $69,2\%$ e especificidade de $94,7\%$.

Algumas medidas

Considere a tabela (matriz de confusão):

		Condição Verdadeira	
		Positiva	Negativa
Condição	Positiva	n_{11} (TP)	n_{12} (FP)
Prevista	Negativa	n_{21}(FN)	n_{22} (TN)
		CP	CN

TP=True positive, FP= False positive, FN= False negative, TN= True negative

TPR=$\frac{TP}{TP+FN} = \frac{n_{11}}{n_{11}+n_{21}}$ (sensibilidade) (True positive rate)

TNR=$\frac{TN}{FP+TN} = \frac{n_{22}}{n_{12}+n_{22}}$ (especificidade) (True negative rate)

Prevalence=$\frac{CP}{CP+CN}$

PPV=$\frac{TP}{TP+FP}$ (Positive predictive value, precision)

NPV=$\frac{TN}{FN+TN}$ (Negative predictive value)

FDR=$\frac{FP}{TP+FP}$ (False discovery rate)

Accuracy=ACC=$\frac{TP+TN}{CP+CN}$, Balanced accuracy= BA=$\frac{TPR+TNR}{2}$

Teste de McNemar

- É um teste para dados nominais pareados, dispostos numa tabela de contingência 2×2 (McNemar, 1947).
- Considere a tabela, com resultados de dois testes para n indivíduos:

		Teste 2		
		Positivo	Negativo	
Teste 1	Positivo	n_{11}	n_{12}	$n_{1.}$
Teste 1	Negativo	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	1

- Sejam $p_{ij} = n_{ij}/n$, $i, j = 1, 2$, $p_{i.}$ as soma das linhas, $i = 1, 2$ $p_{.j}$ as somas das colunas, $j = 1, 2$ (com os p_{ij} substituindo os n_{ij} na tabela). A hipótese nula de homogeneidade marginal afirma que as duas probabilidades marginais de cada resultado são iguais, isto é, $p_{11} + p_{12} = p_{11} + p_{21}$ e $p_{21} + p_{22} = p_{12} + p_{22}$, ou seja, $p_{1.} = p_{.1}$ e $p_{2.} = p_{.2}$, ou ainda

$$H_0 : p_{12} = p_{21},$$

$$H_1 : p_{12} \neq p_{21}.$$

- A estatística de MacNemar é

$$M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}},$$

que, sob H_0, tem uma distribuição qui-quadrado com 1 grau de liberdade.

Teste de McNemar

- Se n_{12} ou n_{21} for pequeno (soma< 25), então M não é bem aproximada pela distribuição qui-quadrado. Um teste binomial exato pode ser usado para n_{21}. Para $n_{12} > n_{21}$, o valor-p exato é

$$p = 2 \sum_{i=n_{12}}^{N} \binom{N}{i} (1/2)^i (1/2)^{N-i},$$

  com $N = n_{12} + n_{21}$.

- Edwards (1948) propôs a seguinte correção de continuidade para M:

$$M = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}}.$$

- No exemplo, $M = (|1 - 4| - 1)^2/5 = 0,8$ e valor-p é $P(\chi_1^2 > 0,8|H_0) \approx 0,37$, logo não rejeitamos H_0.

Aplicação do teste em ML

- O teste de McNemar pode ser usado para comparar técnicas de classificação, para aqueles algoritmos que são usados em conjuntos de dados grandes e não podem ser repetidos via algum método de reamostragem, como CV.
- Dietterich (1998) considerou 5 testes para determinar se um algoritmo de classificação é melhor do que um outro, em um particular conjunto de dados. O objetivo era determinar a probabilidade de erro de tipo I de cada teste.
- Testes que não devem ser usados:(a) teste para a diferença de duas proporções; (b) teste t pareado (diferenças) baseado em partições aleatórias dos conjuntos de treinamento/teste; (c) teste t pareado baseado em 10-fold CV. Todos exibem probabilidades de erro de tipo I altas.
- O teste de McNemar tem baixa probabilidade de erro do tipo I.
- O autor introduziu um teste, 5×2 CV, baseado em 5 iterações de uma 2-fold CV, que tem uma probabilidade de erro de tipo I aceitável e poder maior do que o teste de McNemar.

Aplicação do teste em ML

- Questão: Dados dois classificadores C_1 e C_2 e dados suficientes para aplicá-los em um conjunto de teste, determinar qual classificador será mais acurado em novos conjuntos de testes.
- Essa questão pode ser respondida medindo-se a acurácia de cada classificador no conjuto teste a aplicando o teste de McNemar.
- Dietterich (1998) considera 9 questões , algumas ainda não respondidas, e foca seu artigo na seguinte questão: Dados dois algoritmos de aprendizagem A e B, e um conjuntos de dados pequeno S, qual algoritmo produzirá classificadores mais acurados quando treinados em conjuntos de dados do mesmo tamanho que S?
- Para isso, é necessário usar métodos de reamostragem. Ele compara vários testes estatísticos para responder a questão.
- O primeiro passo é identificar as fontes de variação que podem ser controladas por cada teste.
- 4 fontes de variação: (a) variação aleatória na seleção do conjunto de teste que será usado para avaliar os algoritmos; (b) variação na escolha dos dados de treinamento (instabilidade); (c) aleatoriedade interna do algoritmo de aprendizagem. Por exemplo, o algoritmo backpropagation depende dos pesos (aleatórios) iniciais; (d) erro de classificação aleatório.

Aplicação do teste em ML

- Um teste deve concluir que dois algoritmos são diferentes se, e somente se, suas taxas de classificação corretas sejam diferentes, em média, quando treinados em um conjunto de treinamento de tamanho fixo e testado em todos os dados da população.
- Para tanto, o teste deve considerar o tamanho do teste (probabilidade do erro de tipo I) e excutar o algoritmo múltiplas vezes e medir a variação da acurácia dos classificadores resultantes
- Para aplicar o teste de McNemar, dividimos a amostra de dados S em um conjunto de treinamento T_0 (com n observações) e um conjunto teste T_1 (com m observações). Treinamos os algoritmos C_1 e C_2 no conjunto T_0, obtendo-se classificadores $\hat{C}_1$ e $\hat{C}_2$. Então, testamos esses classificadores no conjunto T_1. Para cada $\mathbf{x} \in T_1$, registramos como esse ponto foi classificado e construímos a seguinte tabela 2×2:

		Classificador 2	
		Class. Correta	Class. Errônea
Classificador 1	Class. correta	n_{11} = Sim/Sim	n_{12}=Sim/Não
Classificador 1	Class. errôna	n_{21} =Não/Sim	n_{22}=Não/Não

- $\sum_i \sum_j n_{ij} = m$.

Aplicação do teste em ML

- Rejeitandos-e H_0, os dois algoritmos terão desempenho diferentes quando treinados em T_0.
- Note que esse teste tem dois problemas: primeiro, não mede diretamente a variabilidade devida à escolha de T_0, nem a aleatoriedade interna do algoritmo, pois um único conjunto de treinamento é escolhido. Segundo, ele não compara os desempenhos dos algoritmos em conjuntos de treinamento de tamanho $|S|$, mas sobre conjuntos de tamanho n, que deve ser menor do que $|S|$, para que tenhamos um conjunto de teste grande.

AD quadrática

- O classificador obtido por meio de Análise Discriminante Quadrática supõe, como na Análise Discriminante Linear usual, que as observações são extraídas de uma distribuição gaussiana multivariada, mas não necessita da suposição de homocedasticidade (matrizes de covariâncias iguais), ou seja, admite que a cada classe esteja associada uma matriz de covariância, $\boldsymbol{\Sigma}_k$.
- Como no caso da Análise Discriminante Linear, não é difícil ver que o classificador de Bayes associa um elemento com valor das variáveis preditora igual a $\mathbf{x}$ à classe para a qual a função quadrática

$$\begin{aligned}
\delta_k(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \pi_k \qquad (25)\\
&= -\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \pi_K.
\end{aligned}$$

é **máxima**.

AD quadrática

- Como os elementos de (25) não são conhecidos, pode-se estimá-los por meio de

$$\overline{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_i, \quad \mathbf{S}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^\top$$

  e $\widehat{\pi}_k = n_k/n$, em que n_k é o número de observações na classe k e n é o número total de observações no conjunto de dados. O primeiro termo do segundo membro da primeira igualdade de (25) é a distância de Mahalanobis.

- O número de parâmetros a estimar, $Kp(p+1)/2$, é maior que no caso de Análise Discriminante Linear, na qual a matriz de covariâncias é comum. Além disso, a versão linear apresenta variância substancialmente menor mas viés maior do que a versão quadrática. A Análise Discriminante Quadrática é recomendada se o número de dados for grande; em caso contrário, convém usar Análise Discriminante Linear.

AD regularizada

- A Análise Discriminante Regularizada foi proposta por Friedman (1989) e é um compromisso entre Análise Discriminante Linear e Análise Discriminante Quadrática.
- O método proposto por Friedman consiste em "encolher" (*shrink*) as matrizes de covariâncias da Análise Discriminante Quadrática em direção a uma matriz de covariâncias comum.
- Friedman (1989) propõe o seguinte procedimento de regularização

$$\boldsymbol{\Sigma}_k(\lambda,\gamma) = (1-\gamma)\boldsymbol{\Sigma}_k(\lambda) + \frac{\gamma}{p}\mathrm{tr}[\boldsymbol{\Sigma}_k(\lambda)]\mathbf{I}, \tag{26}$$

em que $\mathrm{tr}(\mathbf{A})$ indica o traço da matriz $\mathbf{A}$, $\mathbf{I}$ é a matriz identidade e

$$\boldsymbol{\Sigma}_k(\lambda) = \lambda\boldsymbol{\Sigma}_k + (1-\lambda)\boldsymbol{\Sigma} \tag{27}$$

com $\boldsymbol{\Sigma} = \sum n_k \boldsymbol{\Sigma}_k / n$.

AD regularizada

- O parâmetro $\lambda \in [0,1]$ controla o grau segundo o qual a matriz de covariâncias ponderada pode ser usada e $\gamma \in [0,1]$ controla o grau de encolhimento ao autovalor médio. Na prática, λ e γ são escolhidos por meio de LOOVC para cada ponto de uma grade no quadrado unitário.
- Quando $\lambda = 1$ e $\gamma = 0$, a Análise Discriminante Regularizada reduz-se à Análise Discriminante Linear. Se $\lambda = 0$ e $\gamma = 0$, o método reduz-se à Análise Discriminante Quadrática. Se $p > n_k$, para todo k e $p < n$, $\mathbf{\Sigma}_k$ é singular, mas nem a matriz ponderada $\mathbf{\Sigma}$ nem $\mathbf{\Sigma}_k(\lambda)$ em (27) o são. Se $p > n$, todas essas matrizes são singulares e a matriz $\mathbf{\Sigma}_k(\lambda, \gamma)$ é regularizada por (26).
- Essas análises podem ser concretizadas por meio das funções `qda()` do pacote `MASS` e `rda()` do pacote `klaR`.

ADQ e ADR - exemplo

Vamos considerar novamente o conjunto de dados `disco` do Exemplo 8.1. Na segunda análise realizada por meio de regressão logística, a classificação foi concretizada via validação cruzada VC5/5 e LOOCV tendo como variáveis preditoras a distância aberta, a distância fechada ou ambas.
A melhor acurácia, $85,7\%$ foi obtida com ambas as variáveis preditoras e VC5/5.
Agora, consideramos a classificação realizada por intermédio de Análises Discriminantes Linear, Quadrática e Regularizada, separando os dados em um conjunto de treinamento contendo 80% (83) dos elementos, selecionados aleatoriamente, e em um conjunto de validação com os restantes 21 elementos.
Lembremos que $y = 1$ corresponde a discos deslocados e $y = 0$ a discos não deslocados.
As acurácias obtidas por meio de Análise Discriminante Linear e Análise Discriminante Quadrática foram, respectivamente, 90% e 85%.

Referências

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, **10**, 1895–1923.

Edwards, A (1948). Note on the correction for continuity in testing the significance of the difference between correlated proportions. *Psychometrika*, **13**, 185–187.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, 153–157.

Morettin, P. A. e Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC: Rio de Janeiro.

MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/$\sim$ pam

Aula 9

5 de maio de 2023

Sumário

ASV/SVM

- Algoritmos de Suporte Vetorial (ASV), conhecidos na literatura anglo-saxônica como Support Vector Machines (SVM) foram introduzidos por Cortes and Vapnik (1995), que desenvolveram essa classe de algoritmos para classificação binária e englobam técnicas úteis para classificação, com inúmeras aplicações, dentre as quais destacamos reconhecimento de padrões, classificação de imagens, reconhecimentos de textos escritos à mão, expressão de gens em DNAs etc.
- Vapnik and Chervonenkis (1964, 1974) foram, talvez, os primeiros a usar o termo Aprendizado com Estatística (Statistical Learning) em conexão com problemas de reconhecimento de padrões e inteligência artificial.
- Algoritmos de suporte vetorial são generalizações não lineares do algoritmo *Generalized Portrait*, desenvolvido por Vapnik e Chervonenkis (1964).
- Embora a tradução literal do termo proposto por Vapnik seja Máquinas de Suporte Vetorial, optamos por utilizar Algoritmos de Suporte Vetorial para que não se pense que algum tipo de máquina esteja ligado a essas técnicas. Aparentemente, Vapnik utilizou esse termo para enfatizar o aspecto computacional intrínseco à aplicação dos algoritmos.

ASV/SVM

- Uma propriedade importante dos ASV é que a determinação dos parâmetros do modelo corresponde a um problema de otimização convexa, de modo que qualquer solução local é também uma solução global.
- Fazem uso extensivo de Multiplicadores de Lagrange.
- Os algoritmos de suporte vetorial competem com outras técnicas bastante utilizadas, como Modelos Lineares Generalizados (MLG), Modelos Aditivos Generalizados (MAG), Redes Neuronais (Neurais), modelos baseados em árvores etc.
- A comparação com esses métodos é baseada em três fatores: interpretabilidade do modelo usado, desempenho na presença de valores atípicos e poder preditivo.
- Por exemplo, os MLG têm baixo desempenho na presença de valores atípicos, valor preditivo moderado e boa interpretabilidade.
- Por outro lado, os ASV têm desempenho moderado na presença de valores atípicos, alto poder preditivo e baixa interpretabilidade.
- ASV são usados para AE Supervisionado (regressão e classificação).

ASV/SVM

- O princípio operacional fundamental dos ASV é que um kernel (núcleo) é usado para mapear os dados de entrada (ou padrões) em um espaço de dimensão mais alta (feature space), de tal sorte que o problema de classificação, por exemplo, torna-se separável.
- O sucesso da aplicação dos ASV depende da escolha, a priori, desse kernel.
- Os kernels mais populares são:

  Gaussiano
  Polinomial
  Exponential radial basis
  Splines
- Recentemente, têm sido usados kernels baseados em ondaletas.
- Essencialmente, um ASV é implementado por um código computacional que realiza essas tarefas. No Repositório R há pacotes como e1071 e a função svm desenvolvidos com essa finalidade. Outras alternativas são o pacote kernlab e a função ksvm.

ASV/SVM

A abordagem de Cortes and Vapnik (1995) para o problema de classificação baseia-se nas seguintes premissas:

a) Separação de classes: procura-se o melhor hiperplano separador entre as classes, maximizando-se a margem entre os pontos mais próximos das duas classes. Os pontos sobre as fronteiras dessas classes são chamados vetores suporte (**support vectors**).

b) Superposição de classes: pontos de uma classe que estão no outro lado do hiperplano separador são ponderados com baixo peso para reduzir sua influência.

c) Não linearidade: quando não pudermos encontrar um separador linear, utilizamos um **kernel** para mapear os dados de entrada em um espaço de dimensão mais alta (feature space) de tal forma que nesse espaço, são construídos os hiperplanos separadores.

d) Solução do problema: o problema envolve otimização quadrática e pode ser resolvido com técnicas conhecidas.

Fundamentação dos ASV

- Apresentaremos as ideias básicas sobre algoritmos de suporte vetorial (ASV), concentrando-nos no problema de classificação dicotômica, *i.e.*, em que as unidades amostrais devem ser classificadas em uma de duas classes possíveis. Para ideias sobre o caso de mais de duas classes, veja o Texto.
- Adotaremos uma abordagem heurística, mais próxima daquela usualmente empregada em Estatística, deixando para as Notas de Capítulo do Texto a abordagem original (e mais formal) dos ASV.
- Seja $\mathcal{X}$ o espaço dos dados (ou dos padrões); em geral, $\mathcal{X} = \mathbb{R}^d$ e seja a resposta $y \in \{-1, 1\}$.
- Por exemplo, podemos ter dados de várias variáveis explicativas (idade, peso, taxa de colesterol etc.) e uma variável resposta (doença cardíaca, com $y = 1$ em caso afirmativo e $y = -1$ em caso negativo) observadas em vários indivíduos (o conjunto de treinamento). O problema de classificação consiste na determinação de dois subconjuntos (classes) de $\mathcal{X}$, um das quais estará associado a indivíduos com doença cardíaca. O classificador indicará em qual das classes deveremos incluir novos indivíduos (o conjunto de teste) para os quais conhecemos os valores das variáveis explicativas.

Fundamentação dos ASV

Vamos considerar três situações:

1) As classes são perfeitamente separáveis por uma fronteira linear; nesse caso, o separador (hiperplano) é conhecido como classificador de margem máxima (CMM).

- Para duas variáveis, o separador é uma reta; para três variáveis, o separador é um plano. No caso de p variáveis, o separador é um hiperplano de dimensão $p-1$. A Figura 1 é um exemplo. Note que podemos ter mais de uma reta separando as duas classes.

2) Não há um hiperplano que separe as duas classes, como no exemplo apresentado na Figura 2, que corresponde à Figura 2 com pontos trocados de lugar. O separador, neste caso é o classificador de margem flexível (CMF).

3) Um separador linear não conduz a resultados satisfatórios exigindo a definição de fronteiras de separação não lineares. Para isso, recorremos ou a funções não lineares das observações ou a kernels, para mapear o espaço dos dados em um espaço de dimensão maior. O separador, neste caso é o classificador de margem não linear (CMNL).

Fundamentação dos ASV

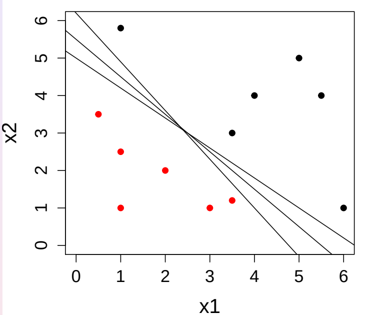


Figura 1: Dois conjuntos de pontos perfeitamente separáveis por um hiperplano (reta).

Fundamentação dos ASV

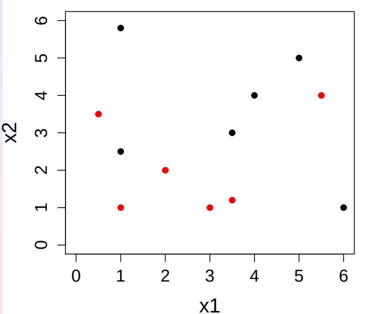


Figura 2: Dois conjuntos de pontos não separáveis por um hiperplano (reta).

Margem e Vetores Suporte

- No caso de duas variáveis, o hiperplano é uma reta com equação $\alpha + \beta_1 X_1 + \beta_2 X_2 = 0$.
- Essa reta separa o plano em duas regiões, uma em que $\alpha + \beta_1 X_1 + \beta_2 X_2 > 0$ e outra em que $\alpha + \beta_1 X_1 + \beta_2 X_2 < 0$.
- Consideremos n observações das variáveis $X_1, \ldots, X_p$, dispostas na forma de uma matriz $\mathbf{X}$, de ordem $n \times p$. Seja $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, o vetor correspondente à i-ésima coluna de $\mathbf{X}$.
- Além disso, sejam $y_1, \ldots, y_n \in \{-1, 1\}$, definindo o conjunto de treinamento $\mathcal{T} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ e seja $\mathbf{x}_0 = (x_{10}, \ldots, x_{p0})^\top$ um **vetor de teste**.

Margem e Vetores Suporte

- Queremos desenvolver um classificador usando um hiperplano separador no espaço $\mathbb{R}^p$ com base no conjunto de treinamento.
- Definindo $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, teremos

$$\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i > 0, \quad \text{se} \quad y_i = 1, \tag{1}$$
$$\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i < 0, \quad \text{se} \quad y_i = -1. \tag{2}$$

- Chamemos

$$f(\mathbf{x}) = \alpha + \boldsymbol{\beta}^\top \mathbf{x}. \tag{3}$$

Então, classificaremos $\mathbf{x}_0$ a partir do sinal de $f(\mathbf{x}_0) = \alpha + \boldsymbol{\beta}^\top \mathbf{x}_0$; se o sinal for positivo, $\mathbf{x}_0$ será classificado na Classe 1 (para a qual $y = 1$, digamos), e se o sinal for negativo, na Classe 2 (para a qual $y = -1$). Em qualquer situação, $y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \geq 0$.

Margem e Vetores Suporte

- Como vimos, podem existir infinitos hiperplanos separadores, se os dados de treinamento estiverem perfeitamente separados.
- A sugestão de Vapnik e colaboradores é escolher um hiperplano que esteja o mais afastado das observações de treinamento, chamado de hiperplano de margem máxima.
- A margem é a menor distância entre o hiperplano e os pontos de treinamento.
- O classificador de margem máxima (CMM) é a solução (se existir) do seguinte problema de otimização:

$$\text{maximizar}_{(\alpha,\boldsymbol{\beta})} \quad m(\alpha, \boldsymbol{\beta}) \tag{4}$$

$$\text{sujeito a}$$

$$\sum_{i=1}^{p} \beta_i^2 = 1, \tag{5}$$

$$y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \geq m(\alpha, \boldsymbol{\beta}), \quad i = 1, \ldots, n. \tag{6}$$

- Dizemos que $m = m(\alpha, \boldsymbol{\beta})$ é a margem do hiperplano e cada observação estará do lado correto do hiperplano se $m > 0$.

Margem e Vetores Suporte

- Os chamados vetores suporte são definidos pelos pontos cujas distâncias ao hiperplano separador sejam iguais à margem e se situam sobre as fronteiras de separação, que são hiperplanos "paralelos" cujas distâncias ao hiperplano separador é igual à margem.
- O classificador depende desses vetores, mas não das demais observações.
- A distância m do hiperplano separador a um ponto do conjunto de treinamento é

$$m = |f(\mathbf{x})|/||\boldsymbol{\beta}||,$$

em que o denominador indica a norma do vetor $\boldsymbol{\beta}$.

Margem e Vetores Suporte

- Como o interesse está nos pontos que são corretamente classificados, devemos ter $y_i f(\mathbf{x}_i) > 0$, $i = 1, \ldots, n$. Então

$$\frac{y_i f(\mathbf{x}_i)}{||\boldsymbol{\beta}||} = \frac{y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i)}{||\boldsymbol{\beta}||}, \tag{7}$$

  e queremos escolher α e $\boldsymbol{\beta}$ de modo a maximizar essa distância.

- A margem máxima é encontrada resolvendo

$$\text{argmax}_{\alpha,\boldsymbol{\beta}} \left\{ \frac{1}{||\boldsymbol{\beta}||} \min_i \left[y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \right] \right\}. \tag{8}$$

- A solução de (8) é complicada e sua formulação canônica pode ser convertida num problema mais fácil por meio do uso de Multiplicadores de Lagrange.

Exemplo CMM

Consideremos os 12 pontos dispostos na Figura 1, sendo 6 em cada classe. Usando a função svm do pacote e1071 e o comando summary(svm.model) obtemos o seguinte resultado:

```
Call:

svm(formula = type ~ ., data = my.data, type = "C-classification",
kernel = "linear", scale = FALSE)

Parameters:

   SVM-Type:  C-classification

 SVM-Kernel:  linear

       cost:  1

      gamma:  0.5

Number of Support Vectors:  3

 ( 1 2 )

Number of Classes:  2

Levels:

 -1 1
```

Exemplo CMM

- Observe que a função usa o kernel linear, que corresponde ao CMM.
- As opções *cost* e *gamma* serão explicadas adiante.
- Os coeficientes do hiperplano separador, que nesse caso é uma reta, podem ser obtidos por meio dos comandos

```
\alpha = svm.model$ rho

\beta = t(svm.model$coefs) %*% svm.model $ SV



e são


> alpha

[1] 5.365853

> beta

          x1           x2

[1,] -0.8780489    -1.097561
```

Exemplo CMM

- A equação do hiperplano separador, disposto na Figura 3 é $5,366 - 0,878X_1 - 1,098X_2 = 0$.
- Na mesma figura, indicamos as fronteiras de separação e os vetores suporte, dados pela solução de (4).Note que os coeficientes $\beta_1 = 0,8780489$ e $\beta_2 = -1,097561$ não satisfazem a restrição indicada em (4), pois foram obtidos por meio da formulação canônica do problema em que a restrição é imposta ao numerador de (7). Para detalhes, consulte a Nota de Capítulo 3.
- Neste caso há três vetores suporte (indicados por círculos azuis), um na Classe 1 (ponto em vermelho) e dois na Classe 2 (pontos em preto). Os demais pontos estão em lados separados, delimitados pelas fronteiras de separação (não há pontos entre as fronteiras).
- A margem é $m = 0,71$.

Exemplo CMM

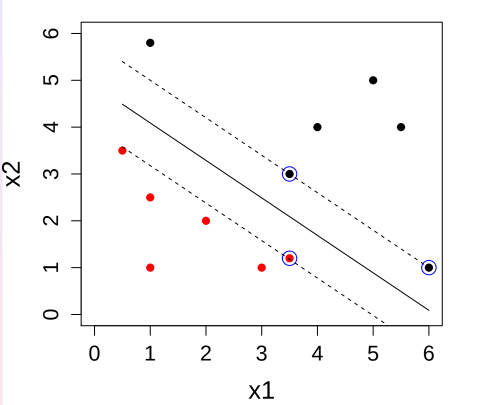


Figura 3: Hiperplano (reta) separador, margem, fronteiras e vetores suporte.

Exemplo CMM

- Consideremos agora dois pontos, $\mathbf{x}_0^*$ e $\mathbf{x}_1^*$, o primeiro na Classe 1 e o segundo na Classe 2 e vamos classificá-los, usando o algoritmo.
- Por meio da função predict, obtemos a Figura 4, que mostra a classificação correta de ambos os pontos (representados nas cores verde e azul).
- Se o problema acima não tiver solução não existirá hiperplano separador, como é o caso apresentado na Figura 2. Nesse caso precisamos recorrer a um classificador que quase separa as duas classes. É o que veremos a seguir.

Exemplo CMM

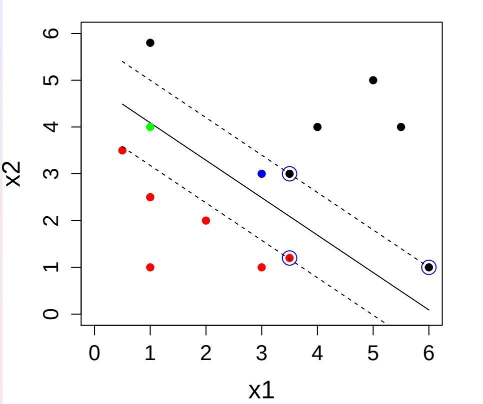


Figura 4: Classificação dos pontos indicados pelas cores verde e azul.

CMF

- Se não existir um hiperplano separador, como aquele do Exemplo anterior, observações podem estar do lado errado da margem ou mesmo do hiperplano, correspondendo nesse caso a classificações erradas.
- O classificador de margem flexível (CMF), também conhecido como classificador baseado em suporte vetorial, é escolhido de modo a classificar corretamente a maioria das observações, o que se consegue com a introdução de variáveis de folga, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$, no seguinte problema de otimização:

$$\text{maximizar}_{(\alpha, \boldsymbol{\beta}, \boldsymbol{\xi})} \quad m(\alpha, \boldsymbol{\beta}, \boldsymbol{\xi}), \tag{9}$$

sujeito a

$$\sum_{i=1}^{p} \beta_i^2 = 1, \tag{10}$$

$$y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \geq m(\alpha, \boldsymbol{\beta}, \boldsymbol{\xi})(1 - \xi_i), \tag{11}$$

$$\xi_i \geq 0, \quad \sum_{i=1}^{n} \xi_i \leq C.$$

em que C é uma constante positiva. Veja abaixo para mais detalhes sobre C.

CMF

- Embora esse tipo de classificador seja conhecido como support vector classifier ou soft margin classifier, optamos por denominá-lo "classificador de margem flexível" para diferenciá-lo do "classificador de margem máxima", que também é baseado em vetores suporte.
- As variáveis de folga permitem que observações estejam do lado errado da margem ou do hiperplano. Pontos tais que $\xi_i = 0$ são corretamente classificados e estão sobre a fronteira de separação ou do lado correto da fronteira. Pontos para os quais $0 < \xi_n \leq 1$ estão dentro da fronteira da margem, mas do lado correto do hiperplano, e pontos para os quais $\xi_n > 1$ estão do lado errado do hiperplano e serão classificados erroneamente. Veja a Figura 5, extraída de Bishop (2006). Nessa figura, m está normalizada apropriadamente, veja as Notas de Capítulo 3 e 4.
- O objetivo é maximizar a margem e, então, minimizamos

$$C \sum_{i=1}^{n} \xi_i + \frac{1}{2} ||\boldsymbol{\beta}||^2, \tag{12}$$

em que $C > 0$ controla o balanço entre a penalidade das variáveis de folga e a margem.

CMF

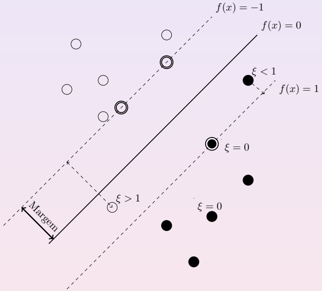


Figura 5: Detalhes sobre o classificador de margem flexível.

CMF

- Como qualquer ponto classificado erroneamente satisfaz $\xi_i > 1$, segue-se que $\sum_{i=1}^{n} \xi_i$ é um limite superior para o número de classificações errôneas. No limite, quando $C \to \infty$, obtemos o CMM.
- Queremos minimizar(12) sujeito a (9). Veja a Nota de Capítulo 4.
- A constante $C \geq 0$ deve ser escolhida apropriadamente e determina o número de violações (classificações erradas) permitidas pelo algoritmo. Se $C = 0$, então não há violações e $\xi_1 = \ldots = \xi_n = 0$. Se C aumenta, a margem fica mais larga e o contrário ocorre se C decresce. O valor de C tem a ver com a relação viés–variância: quando a constante C é pequena, o viés é pequeno e a variância é grande; se C é grande, o viés é grande e a variância é pequena.
  Pode-se dizer que C representa o custo do classificador.
- A constante C normalmente é escolhida por **validação cruzada** O pacote e1071 tem uma função, tune(), que realiza esse procedimento para escolher o melhor modelo, para diferentes valores de C.

Exemplo 1 CMF

Exemplo 1. Consideremos agora os dados dispostos na Figura 3 em que as duas classes não são perfeitamente separáveis. Nesse caso, a utilização da função tune() do pacote e1071 gera o seguinte resultado (editado) indicando que a melhor opção é considerar $C = 4$ e gamma $= 0.5$.

```
Parameter tuning of svm:
- sampling method: 10-fold cross validation
- best parameters:
 gamma   cost
   0.5    4
- best performance: 0.5


- Detailed performance results:
  gamma cost error dispersion
1   0.5    4  0.50  0.4714045
2   1.0    4  0.60  0.4594683
3   2.0    4  0.70  0.4216370
4   0.5    8  0.65  0.4743416
5   1.0    8  0.65  0.4743416
6   2.0    8  0.70  0.4216370
7   0.5   16  0.65  0.4743416
8   1.0   16  0.65  0.4743416
9   2.0   16  0.70  0.4216370
```

Exemplo 1 CMF

Com esses parâmetros, as funções svm e summary geram o seguinte resultado, indicando que há 8 vetores suporte, 4 em cada classe.

```
svm(formula = type ~ ., data = my.data, type = "C-classification",
kernel = "linear", gamma = 0.5, cost = 4, scale = FALSE)
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  4
      gamma:  0.5
Number of Support Vectors:  8
 ( 4 4 )
Number of Classes:  2
Levels:
 -1 1
```

Exemplo 1 CMF

Um gráfico indicando os vetores suporte e as regiões de classificação correspondentes está apresentados na Figura 6.

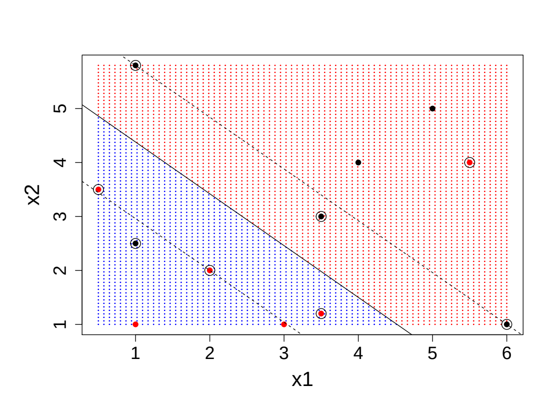


Figura 6: Vetores suporte para os dados da Figura 3.

Exemplo 1 CMF

- A equação do hiperplano classificador é $3,760 - 0,676x_1 - 0,704x_2 = 0$ ou equivalentemente, $x_2 = 3,760/0,704 - 0,676/0,704x_1 = 5,339 - 0,960x_1$. A margem correspondente é $m = (0,676^2 + 0,704^2)^{1/2} = 0,976$. Para detalhes, consulte as Notas de Capítulo 3 e 4.
- Com os comandos svm.pred $< -$ predict(svm.model, my.data) e table(svm.pred, ys) podem-se obter uma tabela com as classificações certas e erradas assim como as classificações determinadas pelo algoritmo. No exemplo, há 2 classificações erradas conforme indicado na Tabela 1.

Exemplo 1 CMF

Tabela 1: Coordenadas e classificação dos pontos do Exemplo 8.1 com classificação predita pelo algoritmo

observação	x1	x2	y	y predito
1	0.5	3.5	1	1
2	1.0	1.0	1	1
3	1.0	2.5	-1	**1**
4	2.0	2.0	1	1
5	3.0	1.0	1	1
6	3.5	1.2	1	1
7	1.0	5.8	-1	-1
8	3.5	3.0	-1	-1
9	4.0	4.0	-1	-1
10	5.0	5.0	-1	-1
11	5.5	4.0	1	**-1**
12	6.0	1.0	-1	-1

Exemplo 2 CMF

- Os dados do arquivo tipofacial foram extraídos de um estudo odontológico realizado pelo Dr. Flávio Cotrim Vellini. Um dos objetivos era utilizar medidas entre diferentes pontos do crânio para caracterizar indivíduos com diferentes tipos faciais, a saber, braquicéfalos, mesocéfalos e dolicocéfalos.
- O conjunto de dados contém observações de 11 variáveis em 101 pacientes. Para efeitos didáticos, utilizaremos apenas a altura facial e a profundidade facial como variáveis preditoras.
- A Figura 7 mostra os três grupos (correspondentes à classificação do tipo facial).

Exemplo 2 CMF

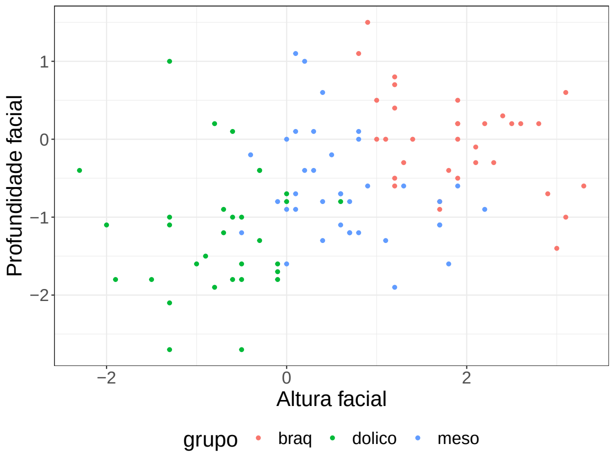


Figura 7: Gráfico de dispersão com identificação dos três tipos faciais.

Exemplo 2 CMF

Utilizando a função tune.svm() do pacote e1071 por meio dos seguintes comandos

```
> escolhaparam <- tune.svm(grupo ~ altfac + proffac, data = face,
                  gamma = 2^(-2:2), cost = 2^2:5,
                  na.action(na.omit(c(1, NA))))
> summary(escolhaparam)
```

obtemos os resultados, apresentados abaixo, que indicam que as melhores opções para os parâmetros C e gamma (obtidas por meio de validação cruzada de ordem 10) para o classificador de margem flexível são $C = 4$ e gamma=2.

Exemplo 2 CMF

```
Parameter tuning of svm:
- sampling method: 10-fold cross validation
- best parameters:
 gamma cost
     2    4
- best performance: 0.1281818
- Detailed performance results:
   gamma cost     error dispersion
1   0.25    4 0.1481818  0.1774759
2   0.50    4 0.1681818  0.1700348
3   1.00    4 0.1681818  0.1764485
4   2.00    4 0.1281818  0.1241648
5   4.00    4 0.1581818  0.1345127
6   0.25    5 0.1481818  0.1774759
7   0.50    5 0.1681818  0.1700348
8   1.00    5 0.1481818  0.1503623
9   2.00    5 0.1281818  0.1148681
10  4.00    5 0.1772727  0.1453440
```

Exemplo 2 CMF

Por intermédio da função svm com os parâmetros $C = 4$ e gamma=2 obtemos o seguinte resultado com o classificador de margem flexível:

```
svm.model <- svm(grupo ~ altfac + proffac, data = face,
                 kernel = "linear", gamma=2, cost=4)
summary(svm.model)

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  4
Number of Support Vectors:  43
 ( 12 10 21 )
Number of Classes:  3
Levels:
 braq dolico meso
```

Exemplo 2 CMF

A tabela de classificação obtida com os comandos apresentados abaixo, indica o número de classificações certas e erradas.

```
svm.pred  <- predict(svm.model, face)
table(pred = svm.pred, true = face\$grupo)
```

```
             true
pred    braq  dolico  meso
braq     26      0     2
dolico    0     28     4
meso      7      3    31
```

Acurácia=0,84

Exemplo 2 CMF

- Na Figura 8 apresentamos o gráfico de classificação correspondente, obtido por meio do comando

  ```
  plot(svm.model, face, proffac ~ altfac, svSymbol = 4,
  dataSymbol = 4,  cex.lab=1.8, main="",
  color.palette = terrain.colors)
  ```

- Uma das características importantes dos classificadores baseados em vetores suporte é que apenas as observações que se situam sobre a margem ou do lado errado da mesma afetam o hiperplano.
- Observações que se situam no lado correto da margem podem ser alteradas (mantendo-se suas classificações) sem que o hiperplano separador seja afetado.

Exemplo 2 CMF

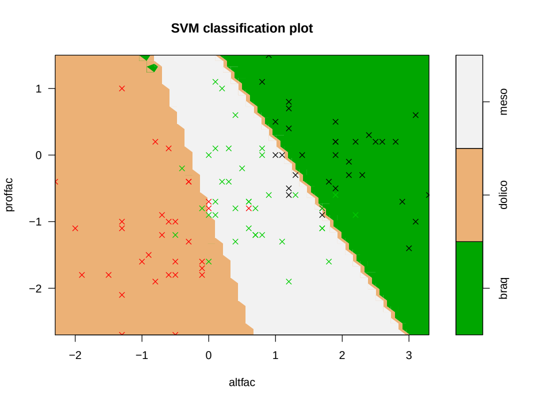


Figura 8: Classificação do tipo facial obtida pelo classificador de margem flexível.

Referências

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20**, 273-297.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Morettin, P. A. e Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC.

Vapnik, V. and Chervonenkis, A. (1964). A note on a class of perceptrons. *Automation and Remote Control*, **25**.

Vapnik, V. and Chervonenkis, A. (1974). *Theory of Pattern recognition* [in Russian]. Moskow: Nauka.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.

MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 10

8 de maio de 2023

Sumário

CMNL

- Na seção anterior apresentamos um algoritmo de classificação (CMF), usado quando as fronteiras são lineares. Para fronteiras não lineares, precisamos aumentar a dimensão do espaço de dados por meio de outras funções, polinomiais ou não, para determinar as fronteiras de separação.
- Pode-se demonstrar que um classificador linear como aquele definido anteriormente (CMF) depende somente dos vetores suporte e pode ser escrito na forma
$$f(\mathbf{x}) = \sum_{i \in S} \gamma_i < \mathbf{x}, \mathbf{x}_i > + \delta, \tag{1}$$
em que S indica o conjunto dos vetores suporte, os γ_i são funções de α e $\boldsymbol{\beta}$ e $< \mathbf{x}, \mathbf{y} >$ indica o produto interno dos vetores $\mathbf{x}$ e $\mathbf{y}$.
- Uma das vantagens de se utilizar kernels na construção de classificadores é que eles dependem somente dos vetores suporte e não de todas as observações o que implica uma redução considerável no custo computacional.

CMNL

- O classificador CMF usa um *kernel* linear, da forma

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} x_{ik} x_{jk} = \mathbf{x}_i^\top \mathbf{x}_j.$$

- Se quisermos usar um CMF em um espaço característico de dimensão maior, podemos incluir polinômios de grau maior ou mesmo outras funções na definição do classificador.
- Os *kernels* mais utilizados na prática são:

  a) lineares: $K(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^\top \mathbf{x}_2$;
  b) polinomiais: $K(\mathbf{x}_1, \mathbf{x}_2) = (a + \mathbf{x}_1^\top \mathbf{x}_2)^d$;
  c) radiais: $K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\gamma ||\mathbf{x}_1 - \mathbf{x}_2||^2\right)$, com $\gamma > 0$ constante.
  d) tangentes hiperbólicas: $K(\mathbf{x}_1, \mathbf{x}_2) = \tanh(\theta + k\mathbf{x}_1^\top \mathbf{x}_2)$.

CMNL

Os classificadores CMNL são obtidos combinando-se CMF com kernels não lineares, de modo a obter

$$f(\mathbf{x}) = \alpha + \sum_{i \in S} \gamma_i K(\mathbf{x}, \mathbf{x}_i) + \delta. \tag{2}$$

em que os γ_i são funções de α e $\boldsymbol{\beta}$.

Exemplo CMNL

- **Exemplo**. Consideremos uma análise alternativa para dados do exemplo anterior, utilizando um *kernel* polinomial, de grau 3.
- Os comandos e resultados da reanálise dos dados por meio do classificador de margem não linear são:

```
escolhaparam <- tune.svm(grupo ~ altfac + proffac, data = face,
                         kernel = "polynomial", degree=3,
                         gamma = 2^(-1:2), cost = 2^2:6)
> summary(escolhaparam)

Parameter tuning of svm:
- sampling method: 10-fold cross validation

- best parameters:
 degree gamma cost
      3   0.5    4
- best performance: 0.1681818
```

Exemplo CMNL

```
- Detailed performance results:

   degree  gamma  cost   error dispersion
1       3    0.5     4 0.1681818 0.09440257
2       3    1.0     4 0.1772727 0.12024233
3       3    2.0     4 0.1872727 0.11722221
4       3    4.0     4 0.1872727 0.11722221
5       3    0.5     5 0.1972727 0.11314439
6       3    1.0     5 0.1772727 0.12024233
7       3    2.0     5 0.1872727 0.11722221
8       3    4.0     5 0.1872727 0.11722221
9       3    0.5     6 0.1872727 0.12634583
10      3    1.0     6 0.1772727 0.12024233
11      3    2.0     6 0.1872727 0.11722221
12      3    4.0     6 0.1872727 0.11722221
```

Exemplo CMNL

```
svm.model <- svm(grupo ~ altfac + proffac, data=face,
                 type='C-classification', kernel='polynomial',
                 degree=3, gamma=1, cost=4, coef0=1, scale=FALSE)
summary(svm.model)

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  4
     degree:  3
     coef.0:  1
Number of Support Vectors:  40
 ( 11 10 19 )Number of Classes:  3
Levels:
 braq dolico meso
```

A tabela de classificação é

```
        true
pred     braq dolico meso
  braq     29      0    4
  dolico    0     26    3
  meso      4      5   30
```

acurácia=0,84

Exemplo CMNL

O gráfico correspondente está apresentado na Figura 1.

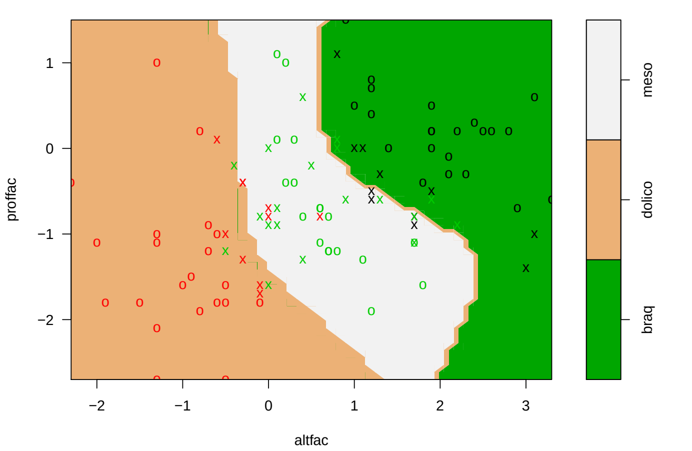


Figura: Classificação do tipo facial obtida pelo classificador de margem não linear.

Exemplo CMNL

- Neste caso, o número de classificações erradas (16) é igual ao caso do classificador de margem flexível. A TEC é 0,16.
- Com base nesses resultados, podemos classificar indivíduos para os quais dispomos apenas dos valores das variáveis preditoras. Com essa finalidade, consideremos o seguinte conjunto de previsão com 4 indivíduos:

```
  paciente altfac proffac
1      102    1.4     1.0
2      103    3.2     0.1
3      104   -2.9    -1.0
4      105    0.5     0.9
```

CMNL

Por meio dos seguintes comandos

```
svm.model <- svm(grupo ~ altfac + proffac, data=face, type='C-classification',
                 kernel='polynomial', degree=3, gamma=1, cost=4, coef0=1,
                 scale=FALSE, probability=TRUE)
prednovos  <- predict(svm.model, teste, probability=TRUE)
```

obtemos a tabela com as probabilidades de classificação de cada um dos 4 indivíduos

```
     1      2      3       4
   braq   braq   dolico   meso
attr(,"probabilities

      braq         dolico        meso
1 0.954231749 0.0193863931 0.0263818582
2 0.961362058 0.0006154201 0.0380225221
3 0.008257919 0.9910764215 0.0006656599
4 0.254247666 0.1197179567 0.6260343773

Levels: braq dolico meso
```

O processo classifica os indivíduos 102 e 103 como braquicéfalos, o indivíduo 104 como dolicocéfalo e o 105, como mesocéfalo.

Hiperplano separador

Um hiperplano definido num espaço de dimensão p é um **subespaço** de dimensão $p-1$ definido por

$$\alpha + \beta_1 X_1 + \ldots + \beta_p X_p = 0. \tag{3}$$

Um ponto com coordenadas $(x_1, \ldots, x_p)$ satisfazendo (3) situa-se no hiperplano. Se $\alpha + \beta_1 x_1 + \ldots + \beta_p x_p > 0$, esse ponto situa-se num lado do hiperplano e se $\alpha + \beta_1 x_1 + \ldots + \beta_p x_p < 0$, o ponto situa-se no outro lado desse hiperplano. Dessa forma, o hiperplano separa o espaço p dimensional em duas metades.

Teoria–CMM

- Consideremos o espaço característico $\mathcal{T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ e as respostas $y_1, \ldots, y_n$ com $y_i \in \{-1, 1\}$, definindo o conjunto de treinamento. Novos dados $\mathbf{x}_0$ são classificados de acordo com o sinal de $f(\mathbf{x}_0)$.
- Suponha que exista um hiperplano separador, de modo que α e $\boldsymbol{\beta}$ são tais que $f(\mathbf{x}) > 0$, para pontos com $y = +1$ e $f(\mathbf{x}) < 0$, para pontos com $y = -1$, de modo que $yf(\mathbf{x}) > 0$, para qualquer dado de treinamento.
- O CMM tem como objetivo maximizar a margem que é a menor distância entre o hiperplano e qualquer ponto do conjunto de treinamento.
- Para entender o procedimento de otimização, considere a distância de um ponto $\mathbf{x}$ ao hiperplano cuja equação é $f(\mathbf{x}) = 0$, nomeadamente

$$d = |f(\mathbf{x})|/||\boldsymbol{\beta}||,$$

em que denominador indica a norma do vetor $\boldsymbol{\beta}$.

Teoria–CMM

- Como o interesse está nos pontos que são corretamente classificados, devemos ter $y_i f(\mathbf{x}_i) > 0$, $i = 1, \ldots, n$. Logo, a distância entre qualquer ponto $\mathbf{x}_i$ e o hiperplano é

$$\frac{y_i f(\mathbf{x}_i)}{||\boldsymbol{\beta}||} = \frac{y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i)}{||\boldsymbol{\beta}||}. \tag{4}$$

- A margem é a distância do hiperplano ao ponto $\mathbf{x}$ mais próximo e queremos escolher α e $\boldsymbol{\beta}$ de modo a maximizar essa distância. A margem máxima é obtida por meio da resolução de

$$\text{argmax}_{\alpha,\boldsymbol{\beta}} \left\{ \frac{1}{||\boldsymbol{\beta}||} \min \left[y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \right] \right\}. \tag{5}$$

- A solução de (5) é complicada mas é possível obtê-la por meio da utilização de Multiplicadores de Lagrange. Note que se multiplicarmos α e $\boldsymbol{\beta}$ por uma constante, a distância de um ponto $\mathbf{x}$ ao hiperplano separador não se altera.

Teora–CMM

- Logo podemos considerar a transformação $\alpha^* = \alpha/f(\mathbf{x})$ e $\boldsymbol{\beta}^* = \boldsymbol{\beta}/f(\mathbf{x})$ e para o ponto mais próximo do hiperplano, digamos $\mathbf{x}^*$, obtendo

$$y^*(\alpha + \boldsymbol{\beta}^\top \mathbf{x}^*) = 1, \tag{6}$$

  e consequentemente, $d = ||\boldsymbol{\beta}||^{-1}$.

- Desse modo, todos os pontos do conjunto de treinamento satisfarão

$$y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \geq 1, \quad i = 1, \ldots, n. \tag{7}$$

  Esta relação é chamada representação canônica do hiperplano separador.

- Dizemos que há uma restrição ativa para os pontos em que há igualdade; para os pontos em que vale a desigualdade, dizemos que há uma restrição inativa. Como sempre haverá um ponto que está mais próximo do hiperplano, sempre haverá uma restrição ativa.

Teoria–CMM

- Então, o problema de otimização implica maximizar $||\boldsymbol{\beta}||^{-1}$, que é equivalente a minimizar $||\boldsymbol{\beta}||^2$.
- Na linguagem de Vapnik (1995), isso equivale a escolher $f(\mathbf{x})$ de maneira que seja a mais achatada (*flat*) possível, que por sua vez implica que $\boldsymbol{\beta}$ deve ser pequeno.
- Isso corresponde à resolução do problema de programação quadrática

$$\text{argmin}_{\alpha,\boldsymbol{\beta}} \left\{ \frac{1}{2} ||\boldsymbol{\beta}||^2 \right\}, \tag{8}$$

sujeito a (7). O fator $1/2$ é introduzido por conveniência.

- Com esse objetivo, para cada restrição em (7), introduzimos os Multiplicadores de Lagrange $\lambda_i \geq 0$, obtendo a função lagrangeana

$$L(\alpha, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} - \sum_{i=1}^{n} \lambda_i [y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) - 1], \tag{9}$$

em que $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^\top$. O sinal negativo no segundo termo de (9) justifica-se por que queremos minimizar em relação a α e $\boldsymbol{\beta}$ e maximizar em relação a $\boldsymbol{\lambda}$.

Teoria–CMM

- Derivando L em relação a $\boldsymbol{\beta}$ e a $\boldsymbol{\lambda}$, obtemos

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i \quad \text{e} \quad \sum_{i=1}^{n} \lambda_i y_i = 0. \tag{10}$$

- Eliminando α e $\boldsymbol{\beta}$ em (9) e usando (10), obtemos a chamada representação dual do problema da margem máxima, no qual maximizamos

$$\tilde{L}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{11}$$

com respeito a $\boldsymbol{\lambda}$, sujeito às restrições

$$\lambda_i \geq 0, \quad i = 1, \ldots, n, \tag{12}$$

$$\sum_{i=1}^{b} \lambda_i y_i = 0. \tag{13}$$

- Em (11), $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ é um *kernel* linear, que será estendido para algum *kernel* mais geral com a finalidade de ser aplicado a espaços característicos cuja dimensionalidade excede o número de dados. Esse *kernel* deve ser positivo definido.

Teoria–CMM

- Para classificar um novo dado $\mathbf{x}_0$ usando o modelo treinado, avaliamos o sinal de $f(\mathbf{x}_0)$, que por meio de (10), pode ser escrito como

$$f(\mathbf{x}_0) = \alpha + \sum_{i=1}^{n} \lambda_i y_i K(\mathbf{x}_0, \mathbf{x}_i). \tag{14}$$

- Pode-se demonstrar (veja Bishop, 2006), que esse tipo de otimização restrita satisfaz certas condições, chamadas de condições de Karush-Kuhn-Tucker (KKT) que implicam

$$\begin{aligned} \lambda_i &\geq 0, \\ y_i f(\mathbf{x}_i) - 1 &\geq 0, \\ \lambda_i (y_i f(\mathbf{x}_i) - 1) &= 0. \end{aligned} \tag{15}$$

Teoria–CMM

- Para cada ponto, ou $\lambda_i = 0$ ou $y_i f(\mathbf{x}_i) = 1$. Um ponto para o qual $\lambda_i = 0$ não aparece em (14) não tem influência na classificação de novos pontos.
- Os pontos restantes são chamados vetores suporte e satisfazem $y_i f(\mathbf{x}_i) = 1$; logo esses pontos estão sobre as fronteiras do espaço separador, como na Figura 3 da Aula 9.
- O valor de α pode ser encontrado a partir de

$$y_i \left(\sum_{j \in S} \lambda_j y_j K(\mathbf{x}_j \mathbf{x}_j) + \alpha \right) = 1, \tag{16}$$

em que S é o conjunto dos vetores suporte.
- Multiplicando essa expressão por y_i, observando que $y_i^2 = 1$ e tomando a média de todas as equações sobre S, obtemos

$$\alpha = \frac{1}{n_S} \sum_{i \in S} \left(y_i - \sum_{j \in S} \lambda_i y_i K(\mathbf{x}_i, \mathbf{x}_j) \right), \tag{17}$$

em que n_S é o número de vetores suporte.

Teoria–CMF

- Vamos considerar agora, o caso em que as duas classes podem se sobrepor. Precisamos modificar o CMM para permitir que alguns pontos do conjunto de treinamento sejam classificados erroneamente. Para isso introduzimos uma penalidade, que cresce com a distância ao hiperplano separador.
- Isso é conseguido pela introdução de variáveis de folga (*slack*) $\xi_i \geq 0, i = 1, \ldots, n$, uma para cada dado.
- Então, $\xi_i = 0$ para pontos sobre ou dentro da fronteira correta [delimitada por $f(\mathbf{x}) = -1$ e $f(\mathbf{x}) = 1$] e ξ_i dado pela distância do ponto à fronteira, para os outros pontos.
- Assim, um ponto que estiver sobre o hiperplano $f(\mathbf{x}) = 0$ terá $\xi_i = 1$ e pontos com $\xi_i > 1$ são classificados erroneamente.
- Nesse caso, a restrição para o caso CMM será substituída por

$$y_i(\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i) \geq 1 - \xi_i, \quad i = 1, \ldots, n, \tag{18}$$

com $\xi_i \geq 0$.

Teoria - CMF

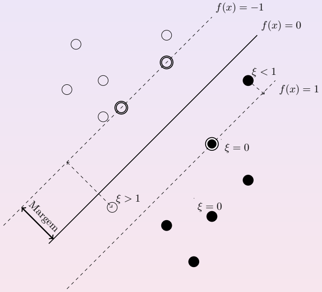


Figura: Detalhes sobre o classificador de margem flexível.

Teoria–CMF

- Pontos para os quais $0 < \xi_i \leq 1$ estão dentro da fronteira da margem, mas do lado correto do hiperplano, e pontos para os quais $\xi_i > 1$ estão do lado errado do hiperplano e são classificados erroneamente. Pontos para os quais $\xi_i = 0$ são corretamente classificados e estão sobre a fronteira da margem ou do lado correto da fronteira da margem.
- Nesse contexto, estamos diante de uma margem flexível ou suave. O objetivo é maximizar a margem e, para isso, minimizamos

$$C \sum_{i=1}^{n} \xi_i + \frac{1}{2}||\boldsymbol{\beta}||^2, \tag{19}$$

em que $C > 0$ controla o balanço entre a penalidade das variáveis de folga e a margem.
- Como qualquer ponto classificado erroneamente satisfaz $\xi_i > 1$, segue-se que $\sum_{i=1}^{n} \xi_i$ é um limite superior do número de classificações errôneas. No limite, quando $C \to \infty$, obtemos o CMM.

Teoria–CMF

- Para minimizar (19) sujeito a (18) e $\xi_i > 0$ consideramos o lagrangeano

$$L(\alpha, \boldsymbol{\beta}, \mathbf{x}_i, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}||\boldsymbol{\beta}||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i[y_i f(\mathbf{x}_i) + \xi_i - 1] - \sum_{i=1}^{n}\mu_i\xi_i, \qquad (20)$$

  em quel $\lambda_i \geq 0, \mu_i \geq 0$ são multiplicadores de Lagrange.
- Derivando (21) com relação a $\boldsymbol{\beta}, \alpha, \xi_i$, obtemos

$$\boldsymbol{\beta} = \sum_{i=1}^{n}\lambda_i y_i \mathbf{x}_i, \quad \sum_{i=1}^{n}\lambda_i y_i = 0 \qquad (21)$$

  e

$$\lambda_i = C - \mu_i. \qquad (22)$$

Teoria - CMF

- Substituindo (21) - (22) em (21), temos

$$\widetilde{L}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i} \sum_{j} \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{23}$$

  que é uma expressão idêntica ao caso separável, com exceção das restrições, que são diferentes.

- Como $\lambda_i \geq 0$ são multiplicadores de Lagrange e como $\mu_i \geq 0$, de (22) segue que $\lambda_i \leq C$. Logo, precisamos maximizar (23) com respeito às variáveis duais λ_i, sujeito a

$$0 \leq \lambda_i \quad \leq \quad C, \tag{24}$$

$$\sum_{i=1}^{n} \lambda_i y_i \quad = \quad 0, \quad i = 1, \ldots, n. \tag{25}$$

- Novamente, estamos diante de um problema de programação quadrática.

Teoria - CMF

- A previsão para um novo ponto $\mathbf{x}$ é obtida avaliando o sinal de $f(\mathbf{x})$ na equação do hiperplano(eq. 3, Aula 9). Substituindo (21) nessa mesma equação, obtemos

$$f(\mathbf{x}) = \alpha + \sum_{i=1}^{n} \lambda_i y_i K(\mathbf{x}, \mathbf{x}_i). \qquad (26)$$

- Dados para os quais $\lambda_i = 0$ não contribuem para (26). Os dados restantes formam os vetores de suporte. Para esses, $\lambda_i > 0$ e, por (28) abaixo, devem satisfazer

$$y_i f(\mathbf{x}_i) = 1 - \xi_i. \qquad (27)$$

- No caso de CMF, as condições de KKT são dadas por

$$\lambda_i \geq 0, \quad y_i f(\mathbf{x}_i) - 1 + \xi_i \geq 0,$$

$$\lambda_i (y_i f(\mathbf{x}_i) - 1 + \xi_i) = 0, \qquad (28)$$

$$\mu_i \geq 0, \quad \xi_i \geq 0,$$

$$\mu_i \xi_i = 0, \quad i = 1, \ldots, n. \qquad (29)$$

Teoria - CMF

- Procedendo como no caso de CMM, obtemos

$$\alpha = \frac{1}{N_{\mathcal{M}}} \sum_{i \in \mathcal{M}} \left(y_i - \sum_{j \in S} \lambda_j y_j K(\mathbf{x}_i, \mathbf{x}_j) \right), \tag{30}$$

em que $\mathcal{M}$ é o conjunto do pontos tais que $0 < \lambda_i < C$.

- Se $\lambda_i < C$, então, por (22), $\mu_i > 0$ e por (29), temos $\xi = 0$ e tais pontos estão na fronteira de separação. Pontos com $\lambda_i = C$ estão dentro da fronteira de separação e podem ser classificados corretamente se $\xi_i \leq 1$ e erroneamente se $\xi_i > 1$.

Teoria-CMNL

- Seja $\mathcal{X}$ o conjunto de dados (ou de padrões). A função $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ é um kernel se existir um espaço vetorial com produto interno, $\mathcal{H}$ (usualmente um espaço de Hilbert) e uma aplicação $\Phi : \mathcal{X} \to \mathcal{H}$, tal que, para todos $x, y \in \mathcal{X}$, tivermos

$$K(x, y) =< \Phi(x), \Phi(y) > . \tag{31}$$

  Φ é a aplicação característica e $\mathcal{H}$, o espaço característico.

- Por exemplo, tomemos $\mathcal{X} = \mathbb{R}^2$ e $\mathcal{H} = \mathbb{R}^3$ e definamos

$$\begin{aligned} \Phi : \mathbb{R}^2 &\to \mathbb{R}^3, \\ (x_1, x_2) &\to (x_1^2, x_2^2, \sqrt{2}x_1x_2). \end{aligned}$$

  Então, se $x = (x_1, x_2)$ e $y = (y_1, y_2)$, é fácil verificar que $< \Phi(x), \Phi(y) >=< x, y >$; logo $K(x, y) =< \Phi(x), \Phi(y) >=< x, y >$ é um *kernel*.

Teoria-CMNL

- Para tornar o algoritmo de suporte vetorial não linear, notamos que ele depende somente de produtos internos entre os vetores de $\mathcal{X}$; logo, é suficiente conhecer $K(\mathbf{x}, \mathbf{x}^\top) =< \Phi(\mathbf{x}), \Phi(\mathbf{x}^\top) >$, e não Φ explicitamente. Isso permite formular o problema de otimização, substituindo a derivada do Lagrangeano no caso de CMF por

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i \Phi(\mathbf{x}_i). \tag{32}$$

- Agora, $\boldsymbol{\beta}$ não é mais dado explicitamente como antes. Também, o problema de otimização é agora realizado no espaço característico e não em $\mathcal{X}$.
- Os *kernels* a serem usados têm que satisfazer certas condições de admissibilidade. Veja Smola e Schölkopf (2004) para detalhes. Os *kernels* mencionados anteriormente são admissíveis.

Regressão via SVM

- Dado um conjunto de treinamento, $\mathcal{T} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, o objetivo é obter uma função $f(\mathbf{x}_i)$, a mais achatada (*flat*) possível tal que $|y_i - f(\mathbf{x}_i)| < \epsilon$, $i = 1, \ldots, n$ em que $\epsilon > 0$ é o maior erro que estamos dispostos a cometer. Por exemplo, ϵ pode ser a máxima perda que admitimos ao negociar com ações dadas certas características obtidas do balanço de um conjunto de empresas.
- No caso de funções lineares, o objetivo é determinar α e $\boldsymbol{\beta}$ tais que $|f(\mathbf{x}_i)| = |\alpha + \boldsymbol{\beta}^\top \mathbf{x}_i| \leq \epsilon$. A condição de que $f(\mathbf{x})$ seja a mais achatada possível corresponde a que $\boldsymbol{\beta}$ seja pequeno, ou seja o problema a resolver pode ser expresso como

$$\text{minimizar } \frac{1}{2}||\boldsymbol{\beta}||^2 \text{ sujeito a } \left\{ \begin{array}{l} y_i - \boldsymbol{\beta}^\top \mathbf{x}_i - \alpha \leq \epsilon, \\ \alpha + \boldsymbol{\beta}^\top \mathbf{x}_i - y_i \leq \epsilon \end{array} \right. . \qquad (33)$$

Regressão via SVM

- Nem sempre as condições (33) podem ser satisfeitas e nesse caso, assim como nos modelos de classificação, podemos introduzir variáveis de folga ξ_i e ξ_i^*, $i = 1, \ldots, n$, que permitem flexibilizar a restrição de que o máximo erro permitido seja ϵ. O problema a resolver nesse contexto é

$$\text{minimizar } \frac{1}{2}||\boldsymbol{\beta}||^2 + \sum_{i=1}^{n} C(\xi + \xi^*) \text{ sujeito a } \begin{cases} y_i - \boldsymbol{\beta}^\top \mathbf{x}_i - \alpha \leq \epsilon + \xi_i, \\ \alpha + \boldsymbol{\beta}^\top \mathbf{x}_i - y_i \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i* > 0. \end{cases} \tag{34}$$

- A constante $C > 0$ determina um compromisso entre o achatamento da função f e o quanto estamos dispostos a tolerar erros com magnitude maior do que ϵ.

Regressão via SVM

- As soluções de (33) ou (34) podem ser encontradas mais facilmente usando a formulação dual (ver Nota de Capítulo 3). No caso de modelos lineares, a previsão para um elemento com valor das variáveis preditoras igual a $\mathbf{x}_0$ é obtida de

$$f(\mathbf{x}_0) = \sum_{i=1}^{n} \widehat{\boldsymbol{\lambda}}_i K(\mathbf{x}_0, \mathbf{x}_i) + \widehat{\alpha},$$

  em que $\widehat{\boldsymbol{\lambda}}_i$ são multiplicadores de Lagrange, $K(\mathbf{x}_0, \mathbf{x}_i)$ é um *kernel*, $\widehat{\alpha} = y_i - \varepsilon - \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i$ e $\widehat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \widehat{\boldsymbol{\lambda}}_i \mathbf{x}_i$.
- Os vetores suporte são aqueles para os quais os multiplicadores de Lagrange $\widehat{\boldsymbol{\lambda}}_i$ são positivos.
- Se optarmos por um *kernel* linear, $K(\mathbf{x}, \mathbf{x}_i) =< \mathbf{x}_0, \mathbf{x}_i >$.

Regressão via SVM - Exemplo

Consideremos os dados de distancia com o objetivo de estudar a relação entre a distância com que motoristas conseguem distinguir um certo objeto e sua idade. O diagrama de dispersão e a reta de mínimos quadrados ajustada ($y = 174,2 - 1,0x$) correspondentes estão apresentados na Figura 3.

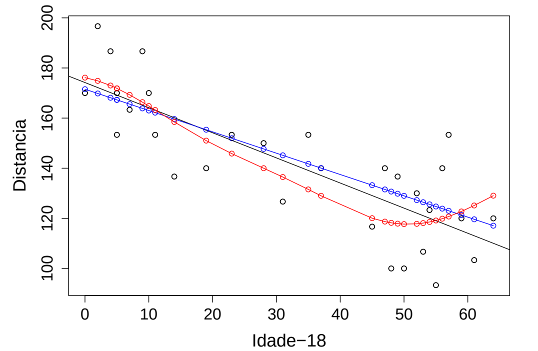


Figura 3: Regressão SVM para os dados de distância.

Regressão via SVM - Exemplo

O ajuste de uma regressão com suporte vetorial baseada num *kernel* linear com os parâmetros *default* pode ser obtido por meio dos comandos

```
model1<- svm(x, y, kernel="linear")
summary(model1)
Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  linear
       cost:  1
      gamma:  1
    epsilon:  0.1
Number of Support Vectors:  23

betahat <- model1$rho
[1] -0.08572489
coef1 <- sum(model1$coefs*x[model1$index])
alfahat <- coef1/model1$rho
[1] 172.8264
```

de forma que a função previsora corresponde à $f(x) = 172,9 - 0,09x$.

Regressão via SVM - Exemplo

- A previsão para as distâncias segundo esse modelo pode ser obtida por meio do comando `yhat1 <- predict(model1, x)`. O *RMSE* correspondente pode ser obtido por meio do comando `rmse(yhat1, y)` é $16,51464$ (maior do que o *RMSE* associado ao ajuste por meio de mínimos quadrados, que é $16,02487$).
- Um modelo mais flexível pode ser ajustado com um *kernel* radial do tipo $K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\gamma||\mathbf{x}_1 - \mathbf{x}_2||^2\right)$ com $\gamma > 0$ constante. Nesse caso, convém realizar uma análise de sensibilidade com validação cruzada para a seleção da melhor combinação dos valores do máximo erro ϵ que estamos dispostos a cometer e do custo de penalização, C. Isso pode ser concretizado por meio dos comandos

```
sensib <- tune(svm, y ~ x,  ranges = list(epsilon = seq(0,1,0.1),
            cost = 2^(2:9)))
Parameter tuning of svm:
- sampling method: 10-fold cross validation
- best parameters:
 epsilon cost
     0.8    8
- best performance: 275.8086
```

Regressão via SVM - Exemplo

Com esses resultados, realizamos um ajuste por meio de um *kernel* radial com parâmetros $C = 8$ e $\epsilon = 0.8$, obtendo

```
model2 <- svm(x, y, kernel="radial", cost=8, epsilon=0.8)
summary(model2)
Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  8
      gamma:  1
    epsilon:  0.8
Number of Support Vectors:  6
```

O *RMSE* para esse modelo é $15,84272$, menor do que aqueles obtidos por meio dos demais ajustes. Um gráfico com os ajustes por mínimos quadrados (em preto) e por regressões com suporte vetorial baseadas em *kernels* linear (em azul) e radial (em vermelho) está apresentado na Figura 3.

Regressão via SVM - Observações

- Algoritmos de suporte vetorial no contexto de regressão também podem ser utilizados com o mesmo propósito de suavização daquele concretizado pelo método **Lowess** (veja a Nota de Capítulo 2 do Capítulo 5).
- Nesse contexto, a suavidade do ajuste deve ser modulada pela escolha do parâmetro ϵ. Valores de ϵ pequenos (próximos de zero) geram curvas mais suaves e requerem muitos vetores suporte, podendo produzir sobreajuste. Valores de ϵ grandes (próximos de 1,0, por exemplo) geram curvas menos suaves e requerem menos vetores suporte.
- O parâmetro C tem influência no equilíbrio entre as magnitudes da margem e das variáveis de folga. Em geral, o valor desse parâmetro deve ser selecionado por meio de uma análise de sensibilidade concretizada por validação cruzada.

Referências

Morettin, P. A. e Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC: Rio de Janeiro.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, **14**, 199–222.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.