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 $\mathbf{x} = (+1, x_1, x_2, ..., x_p)^\top$ contém as entradas, $\mathbf{w} = (b, w_1, w_2, ..., w_p)^\top$ 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

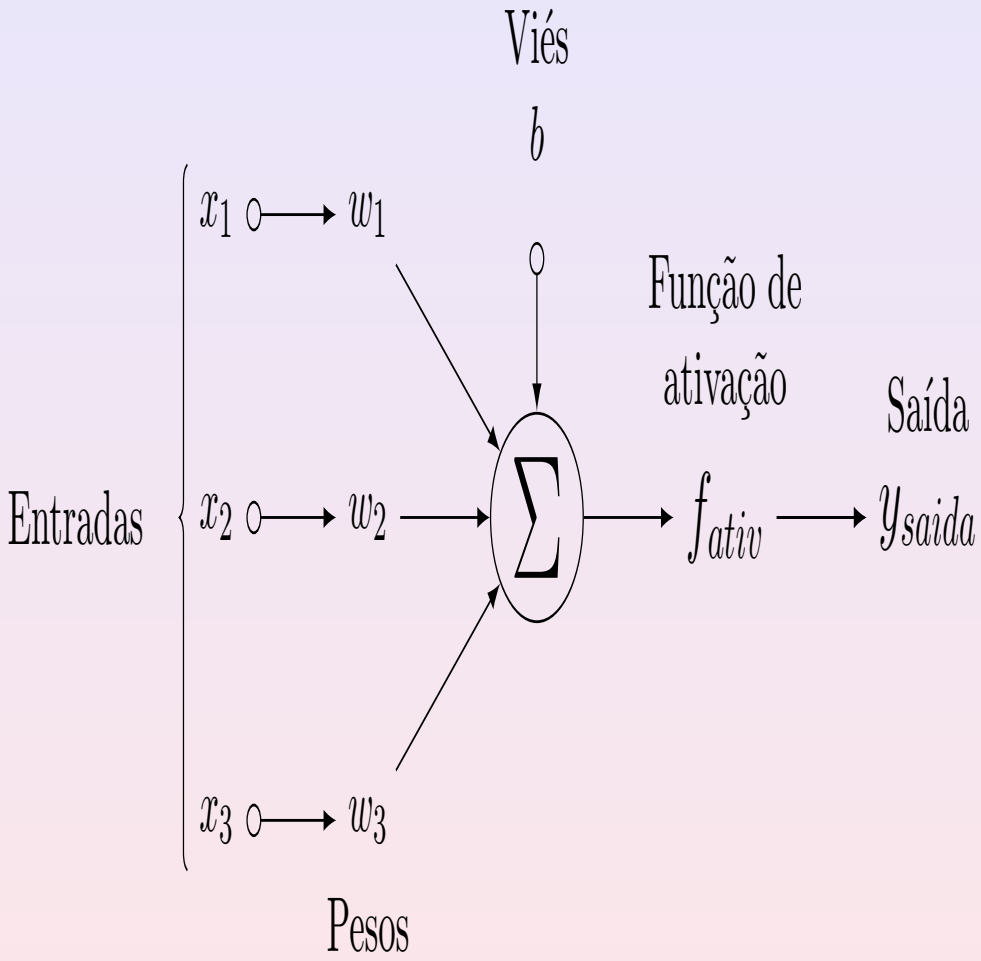
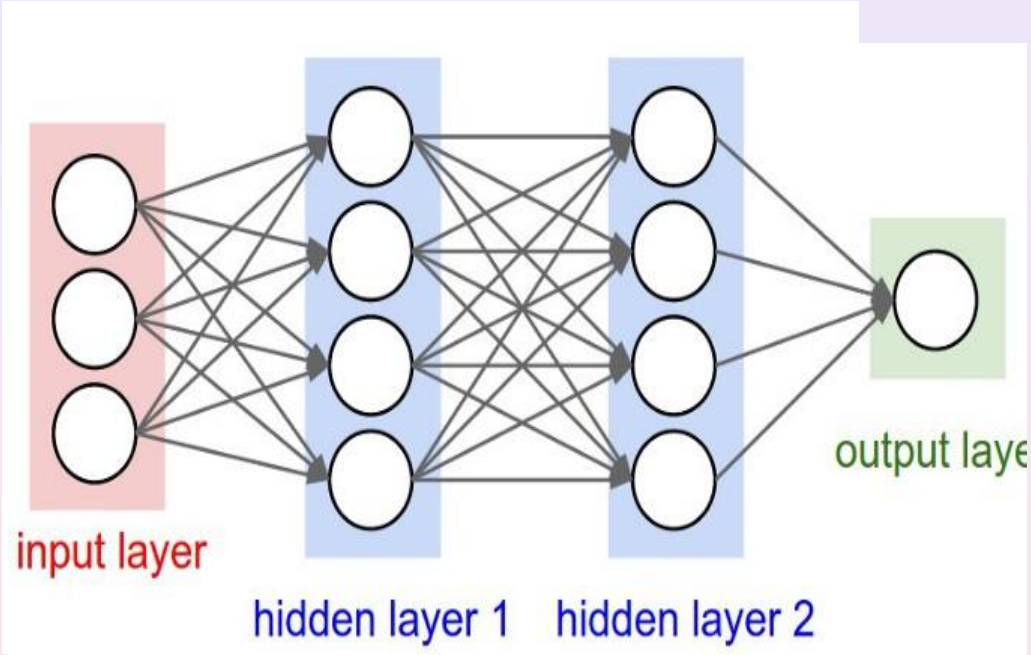


Figura: Perceptron de Rosenblatt

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).