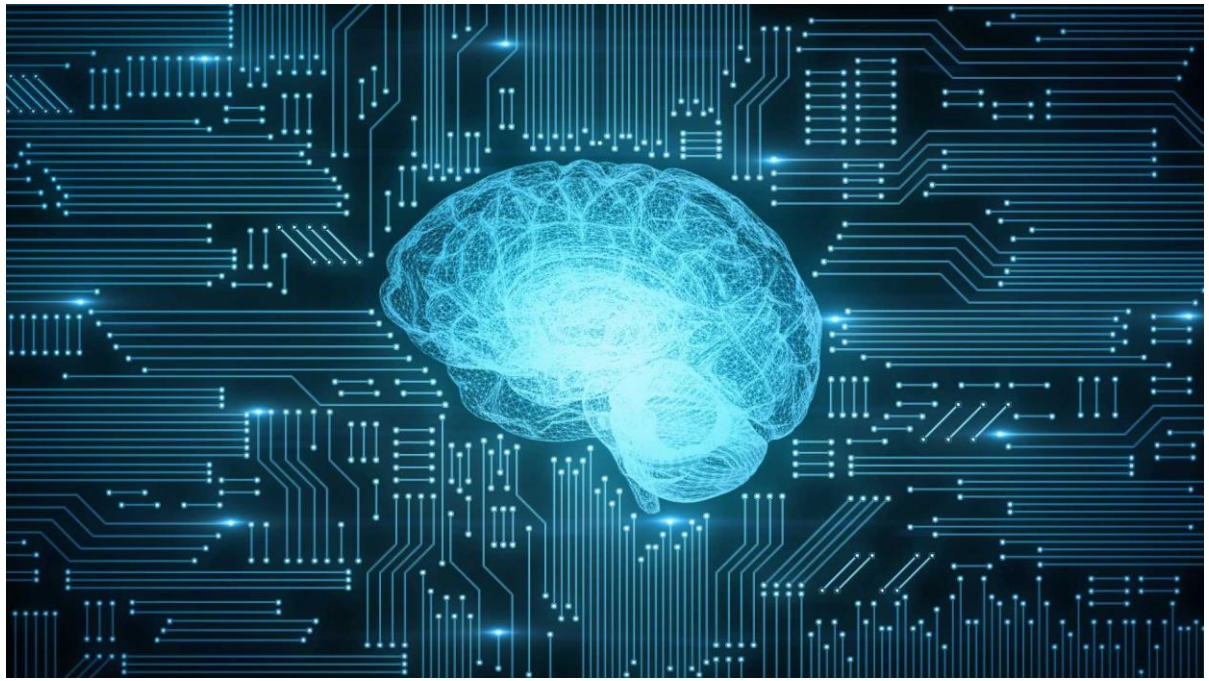
FGV EESP
ESCOLA DE ECONOMIA DE SÃO PAULO

INTELIGÊNCIA COMPUTACIONAL



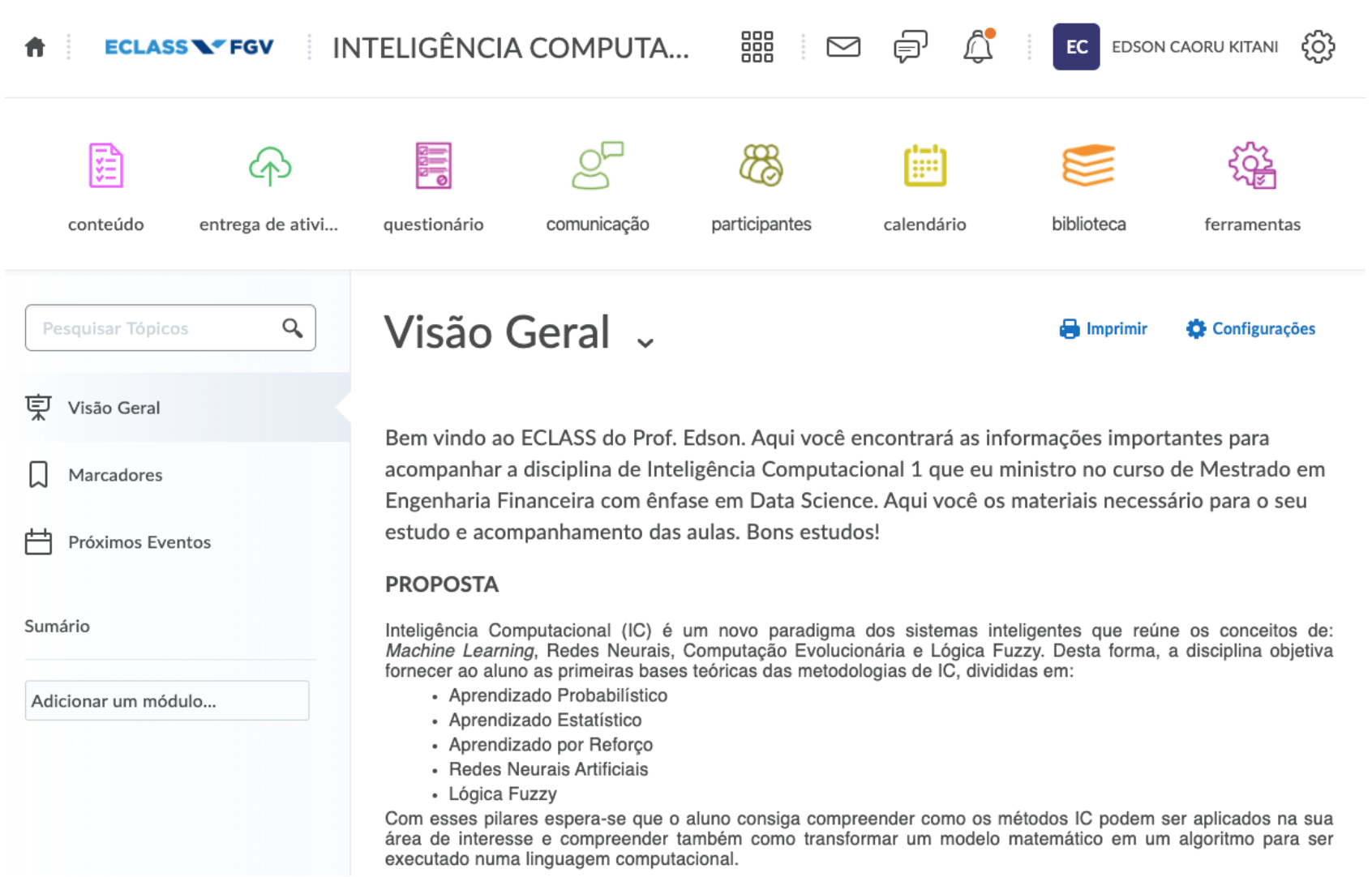
PROF. Dr. EDSON C. KITANI

2022

Sumário da Aula

- Apresentação da disciplina
- Histórico
- Vetores Aleatórios e Espaço de Entrada
- Tipos de Aprendizado
- Teoria do Aprendizado
- Aprendizado Probabilístico

Acesso aos Materiais da Disciplina de IC



Regras de Avaliação

A avaliação será composta de um conjunto de listas de exercícios, trabalho de conclusão e duas provas. A nota final será composta por:

$$Média = Exame \times 0{,}6 \; + \left(\frac{1}{n}\sum_{i=1}^{n} Listas_i\right) \times 0{,}4$$

Para ser aprovado, a Média deve ser maior ou igual a 6,0 e ter mais de 75% de presença nas aulas teóricas.

O conteúdo do Exame Final será baseado nas listas de exercícios entregues ao longo do curso.

As lista serão compostas de exercícios numéricos, computacionais e/ou resumo de artigos científicos.

Bibliografia

Livros Texto

- Tom Mitchell, **Machine Learning,** McGraw Hill, 1997
- Simon Haykin. **Neural Networks: A comprehensive foundation** 2nd Ed. 1999, Prentice Hall
- Richard Duda, Peter Hart, David Stork. **Pattern Recognition** 2nd Ed. 2001, John Wiley & Sons
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, **The Elements of Statistical Learning**, Springer 2nd Ed.
- Ian Goodfellow, Yoshua Bengio, Aaron Courville, **Deep Learning**, MIT Press 2016
- Satish Kumar, **Neural Networks A classroom approach**, McGraw Hill, 2004
- John Hertz, Anders Krogh, Richard Palmer**, Introduction to the theory of neural computation**, Westview Press, 1991
- Teuvo Kohone, **Self-Organizing Maps, Springer** 3rd Ed. 2001
- Charu C. Aggarwal, **Neural Networks and Deep Learning**, Springer, 2018
- Christpher M. Bishop, **Neural Networks for Pattern Recognition**, Oxfor, 2008
- Kevin P. Murphy, **Machine Learning: A Probabilistic Perspective**, MIT PRESS 2012

Dicas Importantes

- Preste atenção na numeração no rodapé de cada *slide* e anote os pontos mais importantes de cada aula.
- As listas de exercícios são individuais e devem ser entregues manuscritas (exceto quando indicadas em contrário) nas datas solicitadas.
- O conteúdo da prova será baseado nas listas de exercício.

TAXONOMIA

INTELIGÊNCIA COMPUTACIONAL

Taxonomia

Inteligência Artificial

É a concepção mais antiga sobre a possibilidade de inteligência não biológica. Formalmente, começou como disciplina em 1956 e trabalha conceitos complexos da inteligência e cognição humana.

Inteligência Computacional

Foi definido em 1990 pelo IEEE para englobar as áreas de **ANN**, **Fuzzy Logic**, **Computação Evolucionária**, Teoria de Aprendizado e Probabilística. É considerado como a Inteligência Artificial baseado nos paradigmas da computação que é conhecido como *Soft Computing.*

Aprendizado de Máquina

ML é o conjunto dos algoritmos que usam a estatística e probabilidade para aprender a partir de exemplos/dados. Iniciou no passado como **Reconhecimento de Padrões,** agregando tópicos como aprendizado supervisionado, não supervisionado, aprendizado por reforço e etc.

Computação Cognitiva

Essa área combina AI com ML e IC para tentar reproduzir o comportamento do cérebro humano. É um sistema que combina também *hardware* especializado e *software*. Um bom exemplo é o Watson da IBM.

ML – Machine Learning AI – Artificial Intelligence IEEE - Institute of Electrical and Electronics Engineering ANN – Artificial Neural Networks

Definição de IC por IEEE

Computational Intelligence (CI) is the theory, design, application and development of biologically and linguistically motivated computational paradigms. Traditionally the three main pillars of CI have been ***Neural Networks, Fuzzy Systems and Evolutionary Computation****. However, in time many nature inspired computing paradigms have evolved.*

Thus, CI is an evolving field and at present in addition to the three main constituents, it encompasses computing paradigms like ambient intelligence, artificial life, cultural learning, artificial endocrine networks, social reasoning, and artificial hormone networks. CI plays a major role in developing successful intelligent systems, including games and cognitive developmental systems. Over the last few years there has been an explosion of research on Deep Learning, in particular deep convolutional neural networks. Nowadays, deep learning has become the core method for artificial intelligence. ***In fact, some of the most successful AI systems are based on CI.***

Main areas of CI: Artificial Neural Networks, Learning Theory, Probabilistic Methods, Fuzzy Logic and Evolutionary Computation,

https://cis.ieee.org/about/what-is-ci

O que estudaremos em IC

O objetivo da disciplina de IC é fornecer as bases **teóricas** e **aplicadas** dos principais modelos matemáticos que darão os conceitos primários para compreender e modelar problemas em *Data Science*. Será a base para compreender outras metodologias que compõe o *Machine Learning*.

- Conceitos Básicos
- Aprendizado Probabilístico (Bayes, EM)
- Aprendizado Estatístico (LDA, PCA, SVM)
- Princípios de Redes Neurais (Perceptron, MLP, RBF, SOM)
- Kernel Models (Não Linear SVM, Autoencoders)
- Redes Neurais Recorrentes (LSTM)

Portanto, as técnicas de *Machine Learning* são também as base teóricas das Redes Neurais estudas em IC.

Tipos de Aprendizado de Máquina

Aprendizado Supervisionado

Trabalha com dados rotulados, exigindo sempre um professor para mostrar os erros e acertos. É muito utilizado para classificar dados quando temos as classes previamente definidas.

Aprendizado não Supervisionado

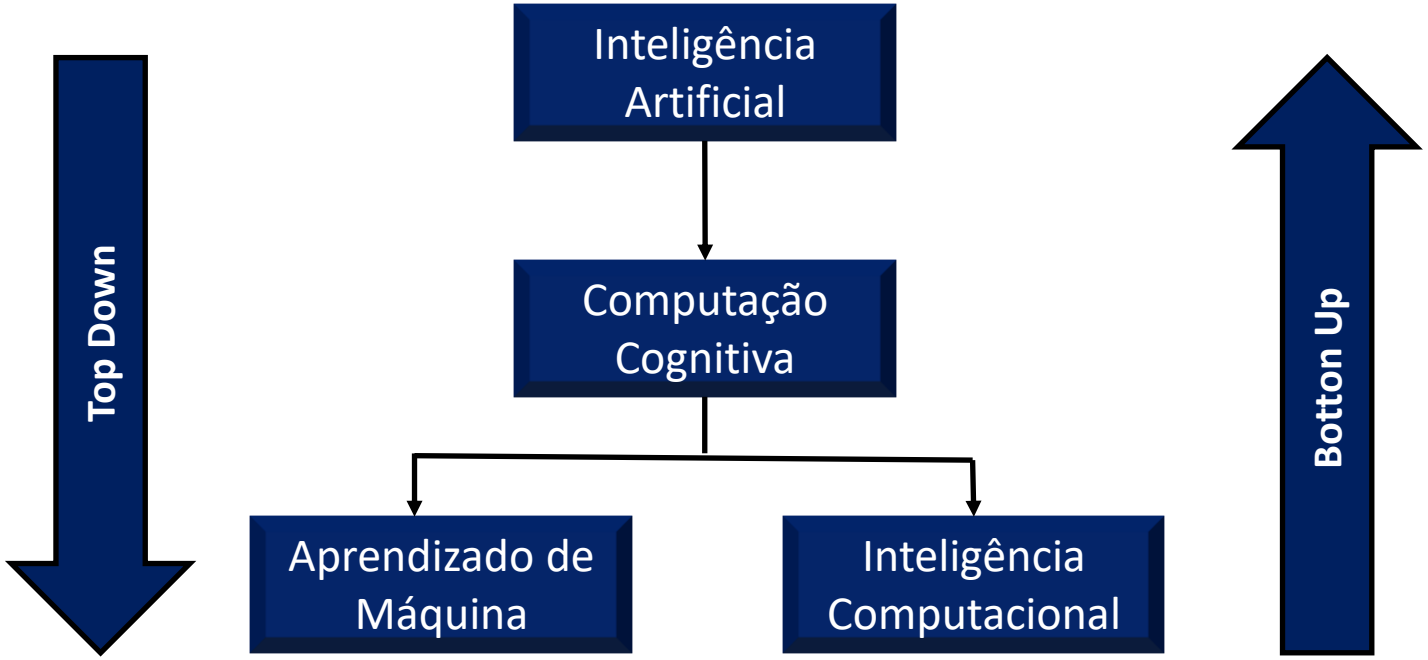
Trabalha com dados não rotulados e é independente de professor. Procura estrutura de dados ocultas. É muito utilizado para categorizar dados e agrupar informações semelhantes. Usado também para visualização de estrutura dos dados.

Aprendizado por Reforço

Trabalha para tomada de ação. É um misto de supervisionado e não supervisionado. Aprende com os próprios erros a partir de recompensas e punições. É muito utilizado em jogos par tomada de decisões autônomas.

É importante entender a diferença entre classificar e categorizar. O primeiro trabalha com a identificação do dado e a determinação estatística da pertinência. O segundo caso é quando temos dados e precisamos categorizar em grupos, pois não temos as classes definidas.

Hierarquia


Top Down
Inteligência Artificial
Computação Cognitiva
Aprendizado de Máquina
Inteligência Computacional
Botton Up

Hierarquia

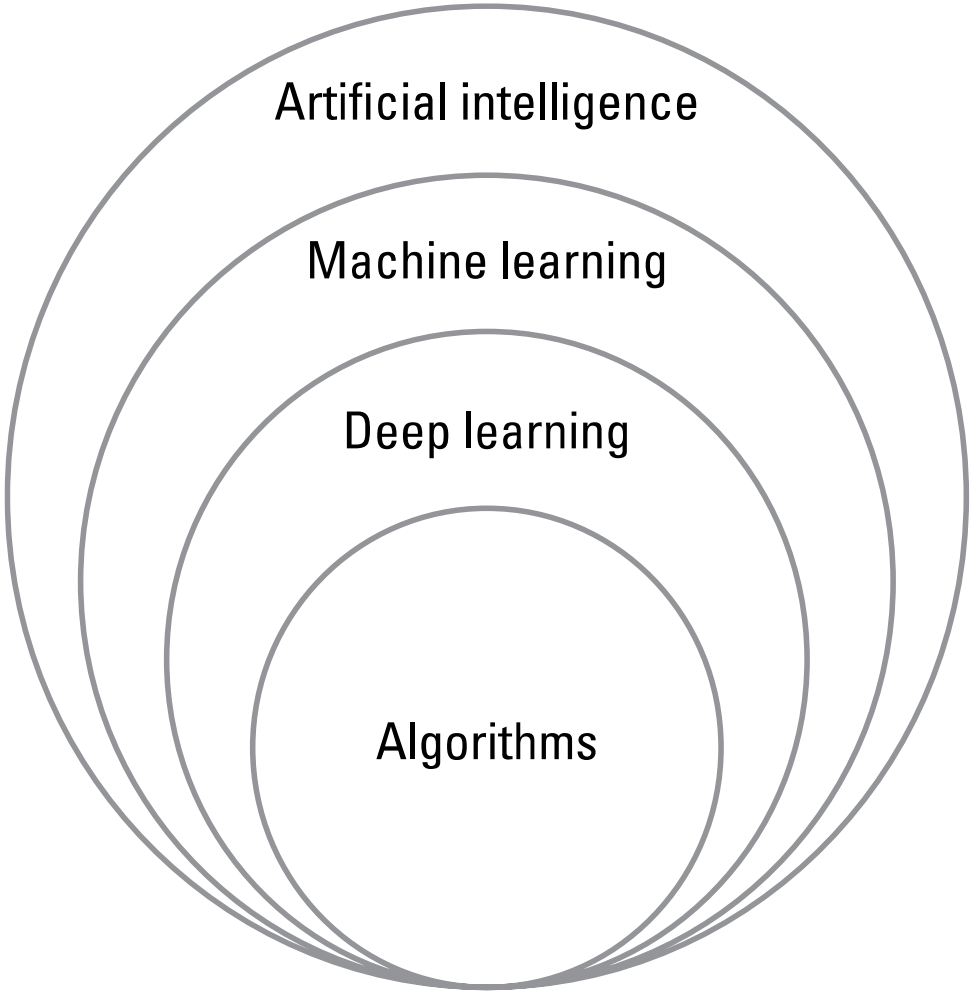
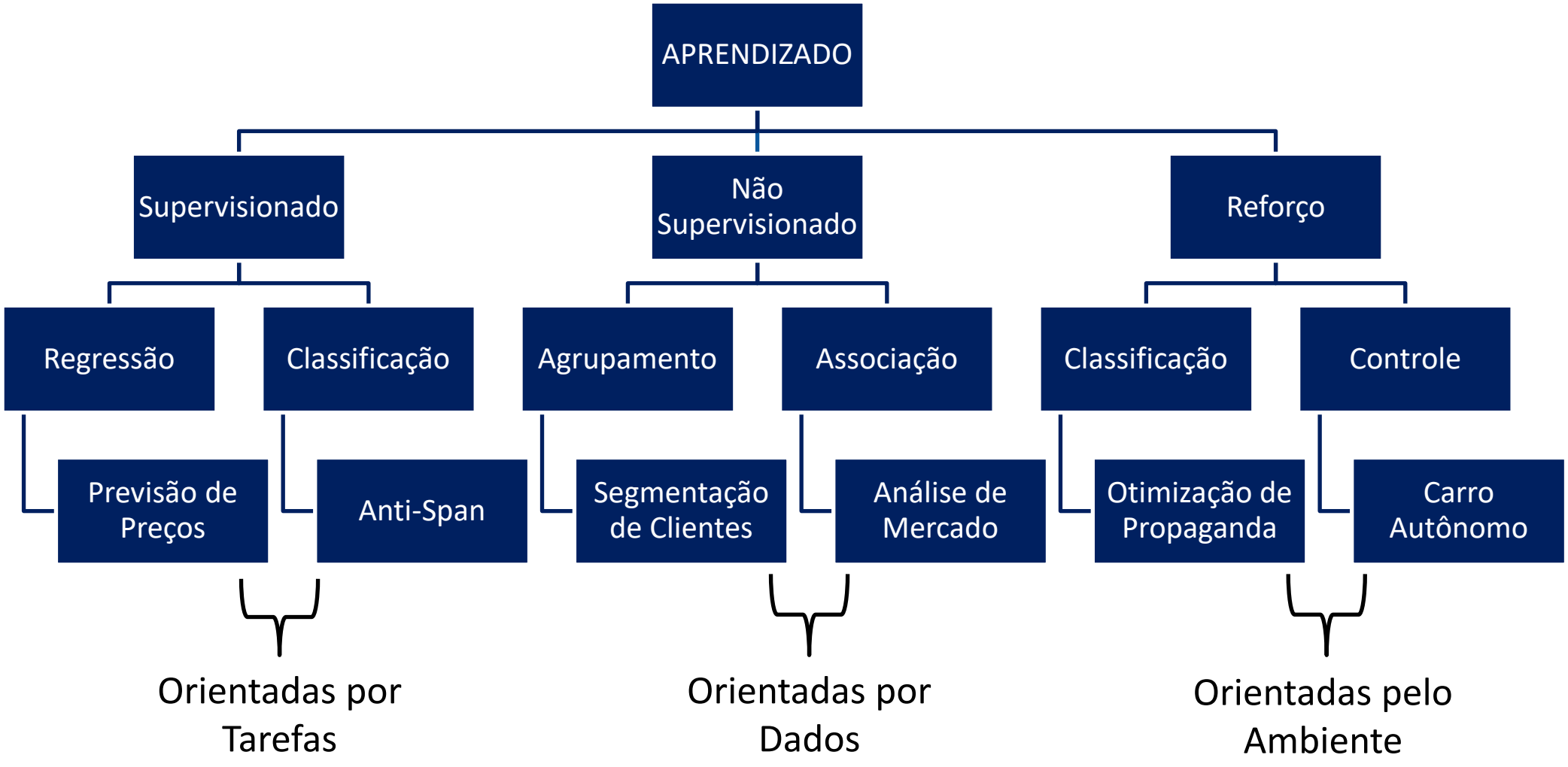


FIGURE 1-1: Machine learning is a subset of artificial intelligence.

Taxonomia


APRENDIZADO
Supervisionado
Não Supervisionado
Reforço
Regressão
Classificação
Agrupamento
Associação
Classificação
Controle
Previsão de Preços
Anti-Span
Segmentação de Clientes
Análise de Mercado
Otimização de Propaganda
Carro Autônomo
Orientadas por Tarefas
Orientadas por Dados
Orientadas pelo Ambiente

Definições sobre Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed."
Arthur Samuel(1959).

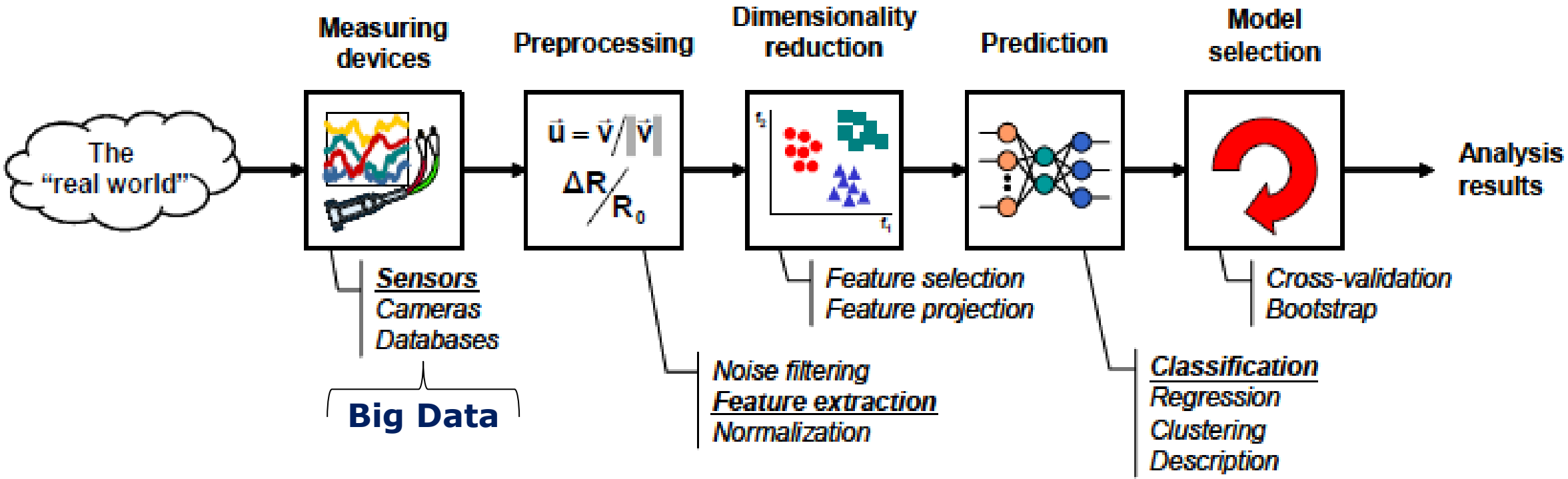
"In particular, we define machine learning as a set of methods that can automatically detect pattern in data, and then use the uncovered patterns to predict future data or perform other kind of decision making under uncertainty." Kevin Murphy

"Machine Learning is the study of data-driven methods capable of mimicking, understanding and aiding human and biological information processing tasks." David Barber

"Machine learning is defined as an automated process that extract patterns from data." Kelleher et al.

"Financial Machine Learning methods do not replace theory. They guide it. An ML algorithm learns patterns in a high dimensional space without specifically directed. Once we understand what features are predictive of a phenomenon, we can build a theoretical explanation, which can be tested on an independent dataset." Marcos López de Prado

Arquitetura Geral do Processo de ML



Ricardo Gutierrez-Osuna, 2002

GIGO

https://medium.com/nyc-design/gigo-garbage-in-garbage-out-concept-for-ux-research-7e3f50695b82

CONCEITOS SOBRE APRENDIZADO DE MÁQUINA

Aprendizado - Definições

O que é aprendizado humano?

"É a aquisição de conhecimento ou habilidade através da experiência, do estudo ou do que for ensinado."

O que é aprendizado de máquina?

*"A computer program is said to **learn** from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**."*
Tom Mitchell

Portanto:

É necessário criar um modelo matemático que defina o que é o resultado da tarefa **T** baseado nos dados **E**, e que otimiza (minimiza erro ou maximiza desempenho) **P** a medida que novos dados **E** são fornecidos.

Exemplo: Tarefa **T** $\Longrightarrow$ Aprender a jogar Damas
Métrica **P** $\Longrightarrow$ Ganhar o jogo
Dados **E** $\Longrightarrow$ Jogadas realizadas e recebidas

Outros exemplos:

Exemplo: Tarefa **T** ⟹ Reconhecer palavras com caracteres manuscritos
Métrica **P** ⟹ Número de palavras corretamente reconhecidas
Dados **E** ⟹ Conjunto de diferentes palavras manuscritas por diferentes pessoas

Exemplo: Tarefa **T** ⟹ Analisar risco de crédito
Métrica **P** ⟹ Número de riscos corretamente avaliadas
Dados **E** ⟹ Conjunto do perfil de diferentes clientes e histórico de créditos

Exemplo: Tarefa **T** ⟹ Detecção de fraudes em cartão de crédito
Métrica **P** ⟹ Número de fraudes corretamente detectadas
Dados **E** ⟹ Conjunto de dados de consumo e perfil dos clientes

Exemplo: Tarefa **T** ⟹ Analisar preferências de consumo
Métrica **P** ⟹ Aumentar o volume de vendas
Dados **E** ⟹ Perfil de consumo, preferências, sazonalidade, crédito, estimulo, etc.

Algumas considerações importantes sobre “Aprendizado de Máquina”:

- Devemos restringir a essência da palavra “aprendizado” ao que foi definido por Tom Mitchell, e não se preocupar em definir o significado da palavra “aprendizado” ou filosofar sobre ela.

- O objetivo é definir claramente uma classe de problemas que podem ser solucionadas por algoritmos que trabalham baseados nos paradigmas computacionais, e explorar as estruturas fundamentais do aprendizado de máquina para resolver esses problemas.
- Diferenças entre *Business Intelligence* (BI) e *Machine Learning* (ML): BI trabalha com dados do passado para auxiliar na tomada de decisões e mostrar o desempenho do que foi realizado ou decidido. ML trabalha com dados do passado para realizar predições, categorizações, classificações e auxiliar na tomada de decisão do BI.

- “*Conhecimento refere-se às informações ou modelos armazenados por pessoas ou máquinas para interpretar, predizer e responder apropriadamente ao mundo exterior*” (Fischler & Firschein, 1987)

Todos os problemas de ML tem aos menos um conjunto de entradas, também chamados de: preditores, variáveis independentes, características ou simplesmente, variáveis. Um conjunto de saída que é formando por variáveis de resposta, variável dependente, ou simplesmente, saída. Portanto, podemos escrever formalmente que um modelo dependente das entradas seria:

$$y = f(x)$$

na qual $y \in \Re$ (escalar) é uma resposta quantitativa para diferentes valores de $x \in \Re^n$, f é uma função fixa, mas desconhecida que descreve a relação entre y e x, ou seja:

$$f: x \to y$$

Leia-se, f é uma aplicação de x em y. Contudo, nas aplicações reais é sempre possível encontrar um ruído gaussiano ε que está associado às medições, então teremos:

$$y = f(x) + \varepsilon$$

Apesar dos exemplos acima indicarem que $y \in \Re$ é um escalar, podemos ter aplicações cujas saídas sejam vetores, $y \in \Re^n$.

Quando discutimos aplicações com ML, veremos que temos apenas o conjunto de dados de entrada e saída:

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\} \mid (\mathbf{x}_i) \in \mathbb{R}^n, y_i \in \mathbb{R} \ e \ i = 1, 2, \dots, N \ ,$$

ou seja, precisamos encontrar uma função $g: \mathbf{x} \rightarrow y$, tal que aproximemos $g(\mathbf{x})$ de $f(\mathbf{x})$. Observe que buscar $g(\mathbf{x})$ é escolher uma função dentre várias possíveis funções $g \in \mathcal{H}$ que é o conjunto de hipóteses para a solução do nosso problema.

Dessa forma, a partir de $g(\mathbf{x})$ estimamos o valor $\hat{y} = g(\mathbf{x})$ tal que minimizemos o erro médio quadrático total:

$$\mathcal{L}(g) = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2} \, .$$

Este é apenas um modelo de uma função de perda (*loss function*), erro ou custo de escolher $g_a \in \mathcal{H}$. Os modelos de ML não conhecem o espaço de dados completos, nem a função $f(x)$, assim, o erro não é diretamente disponível, mas extraído a partir do conjunto de treinamento.

Outra aplicação para o aprendizado de máquina é a classificação de padrões de entrada tal que:

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\} \mid (\mathbf{x}_i) \in \mathbb{R}^n, y_i \in \mathbb{R} \; e \; i = 1, 2, \dots, N \times \mathcal{C},$$

significando que teremos amostras multidimensionais $(\mathbf{x}_i, y_i)$ $i = 1, 2, \dots N$ associadas à d classes $(c_1, c_2, \dots c_d) \in \mathcal{C}$.

A função custo para um classificador pode ser tanto o erro médio quadrático como também um sistema de contagem, representando o custo do erro de classificação.

$$\mathcal{L}(g) = \frac{1}{N}\sum_{i=1}^{N} \delta_{g(\mathbf{x}_i) \neq y_i} \quad na \; qual \quad \delta_{g(\mathbf{x}_i) \neq y_i} = \begin{Bmatrix} 1 \; se \; g(\mathbf{x}_i) \neq y_i \\ 0 \; se \; g(\mathbf{x}_i) = y_i \end{Bmatrix}.$$

Assim, buscamos uma hipótese que minimize o custo da decisão, ou seja:

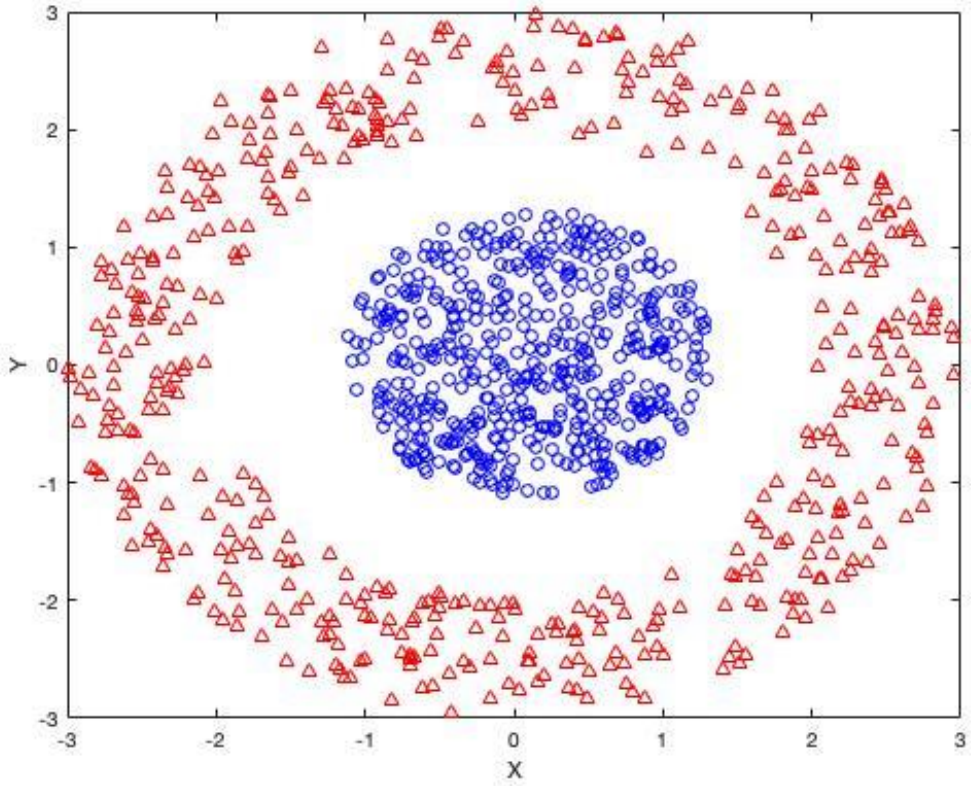
$$g_a = argmin_{g_a \in \mathcal{G}}\{\mathcal{L}(g)\}.$$

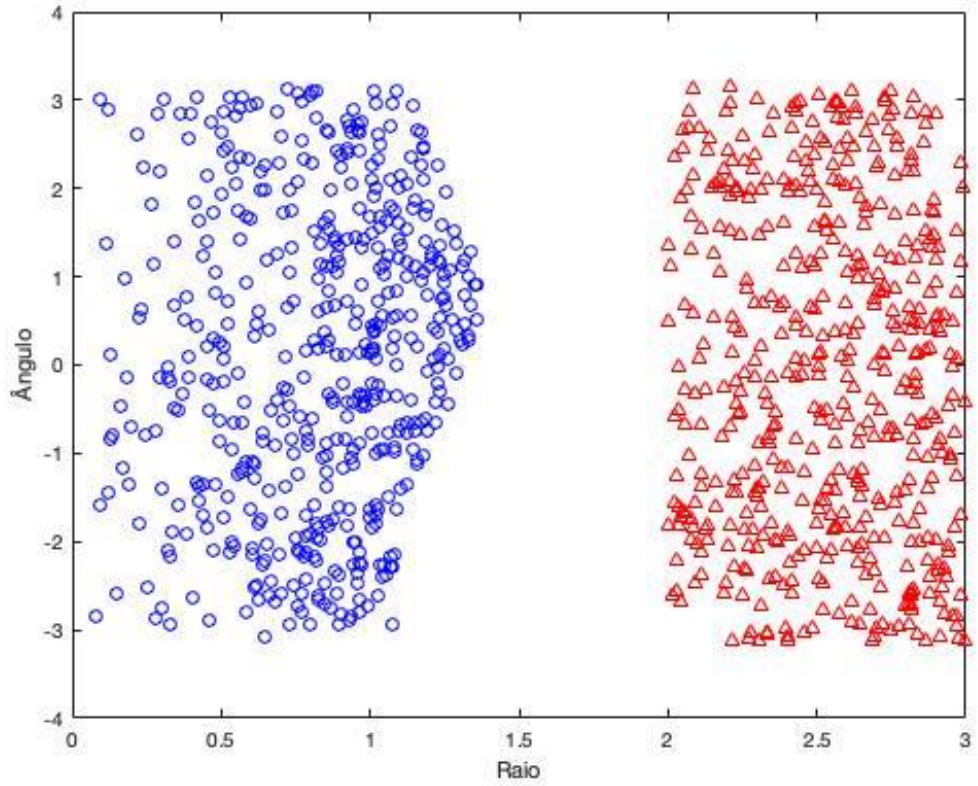
"A critical skill in data science is the ability to decompose a data- analytics problem into pieces such that each piece matches a known task for which tools are available. Recognizing familiar problems and their solutions avoids wasting time and resources reinventing the wheel. It also allows people to focus attention on more interesting parts of the process that require human involvement—parts that have not been automated, so human creativity and intelligence must come in- to play."

Foster Provost and Tom Fawcett – Data Science for Business

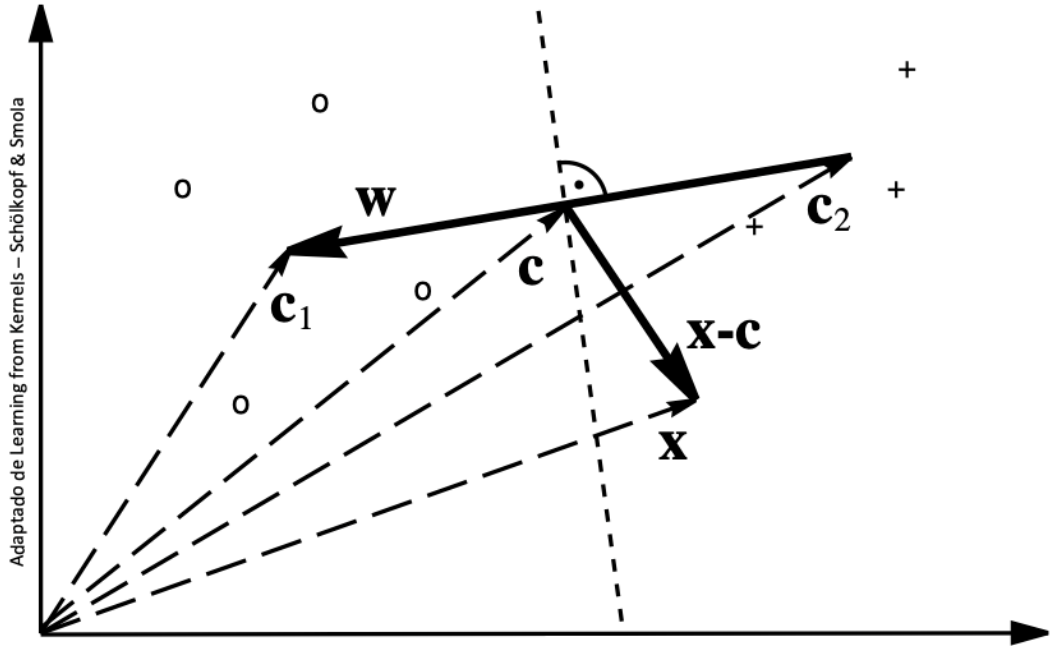
"There's a joke that says a data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."
Joel Grus

Como extrair informação a partir dos dados? Qual o impacto da representação dos dados no aprendizado?





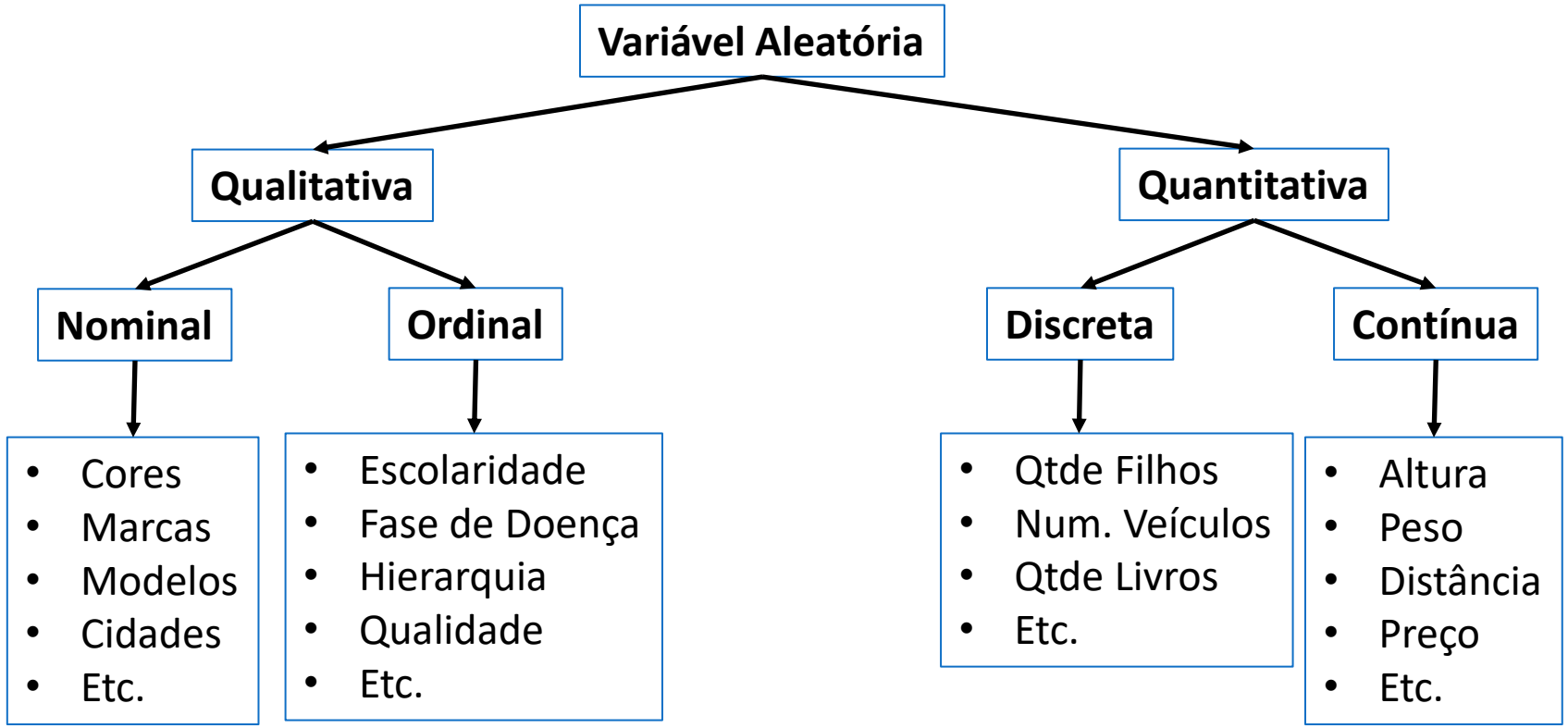
Então, o formato dos dados é muito importante para trabalhar com *Machine Learning e Computational Intelligence*.



Classificador linear geométrico

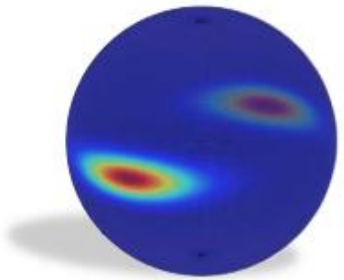
Seja um espaço de características que contém círculos e cruzes. Considere o vetor $\mathbf{c}_1$ e $\mathbf{c}_2$ serem as médias dos círculos e cruzes. Então, um novo padrão $\mathbf{X}$ pertencerá a dado conjunto quando mais próximo ele estiver de uma das médias. A reta $\mathbf{W}$ representa a distância entre as médias, e o plano que separa os dois conjuntos é a reta pontilhada que é ortogonal à reta $\mathbf{W}$. Essa reta pontilhada é conhecido como limiar ou fronteira de decisão. Caso o espaço vetorial seja maior do que 4, então ela será chamada de hiperplano de separação.

Assim, para relembrar, os dados obtidos por medição ou observação são chamadas também de variáveis aleatórias, e são formalmente classificadas como:



Quando trabalhamos em espaços vetoriais multidimensionais é preciso tomar cuidado com o tipo de métrica que será usado. As características dos dados não necessariamente obedecerão os axiomas dos espaços Euclidianos, pois podemos ter espaços de dados elípticos, hiperbólicos e mesmo não convexos.

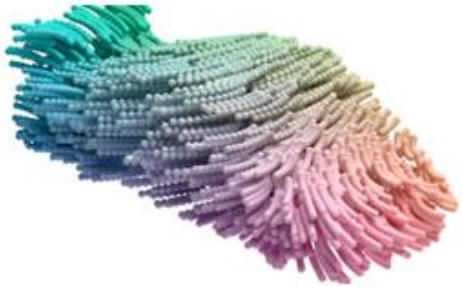

Surfaces

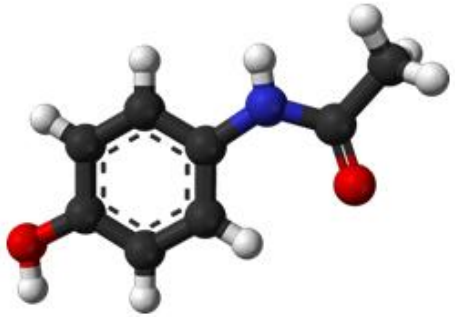

Distributions

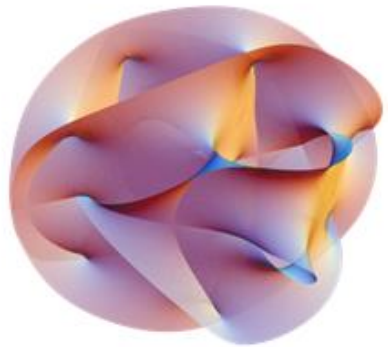

Graphs / Networks


Functions on Manifolds


Hyperbolic spaces


Hyper-surfaces


Molecules


General manifolds

Cronograma das Aulas Teóricas

Técnicas de Inteligência Computacional - 1				
Área	**Tópico**	**Aula**		**Conteúdo**
Conceitos Básicos	Introdução	1	26/abril	Apresentação da ementa. Teoria sobre aprendizado de máquina. Principais paradigmas de aprendizado. Aprendizado supervisionado, não supervisionado, por reforço e *deep learning,* Classificador de Bayes
	Aprendizado Probabilístico	2	03/maio	Estimação paramétrica e não paramétrica, Knn, Distância de Minkowski,, Problemas de Dimensionalidade, Visualização de Dados, Manifold Learning, Grades de Voronoi
Aprendizado de Máquina	Aprendizado Estatístico	3	10/maio	Perceptron de Rosemblat, LMS, Critério de Fisher, LDA, Métricas de Desempenho , SVM
	Sistemas Lineares e não lineares	4	17/maio	Árvores de Decisão, Randon Forest
		5	24/maio	Redução de Dimensionalidade, PCA, Kmeans, EM
Inteligência Computacional	Redes Neurais	6	31/maio	Back Propagation Algorithm, processo de Treinamento e Validação Perceptron multi-camadas (MLP)
		7	07/junho	SOM, RBF
		8	14/junho	Autoencoder e Variational Autoencoder
	Kernel models, Neuro dinâmica	9	21/junho	Deep Neural Networks, SVM não Linear
		10	24/junho	Redes Recorrentes e LSTM
		11	25/junho	Revisão para Prova Final
Avaliação	Individual	12	07/julho	Prova

APRENDIZADO PROBABILÍSTICO

Aprendizado Probabilístico – Conceitos Básicos

Em *Data Science* os dados coletados são o resultado da observação de qualquer ensaio ou evento. Consequentemente, há incertezas associadas aos dados, tais como:

a) Saber se todas as variáveis envolvidas representam o ensaio/evento;

b) Dados ausentes, insuficientes ou errados;

c) Complexidade do processo que impede uma análise precisa do efeito combinado de n variáveis;

d) Imprecisão na medição das variáveis.

Desta maneira, consideramos que os dados obtidos são conhecidos como "variáveis aleatórias" e eles (os dados) se caracterizam por:

a) Pertencer a um espaço amostral Ω de todos os possíveis resultados de um processo. Assim, cada elemento enumerável é denominado ponto ou elemento de Ω, tal que $\omega \in \Omega$;

b) Tem um subconjunto (A) de elementos $\omega \in A$ que está contido ao espaço de resultados $A \subset \Omega$;

c) Para conjunto de elementos que mapeiam algum resultado no espaço Ω deve existir um valor $p \in \mathbb{R}^+$ que descreve o quão provável é a ocorrência de um elemento contido no conjunto de entrada.

Portanto, as premissas acima definem um espaço de probabilidades formados pela tupla $\{\Omega, \mathcal{F}, P\}$.

Aprendizado Probabilístico – Conceitos Básicos

As principais interpretações que podemos dar para Probabilidade são a "**Frequentista**" e a "**Bayesiana**".

Na abordagem frequentista a probabilidade é usada para determinar a possível ocorrência de um evento, como no caso do jogo de dados.

Na abordagem Bayesiana a probabilidade é interpretada como um grau de "incerteza" da ocorrência de um evento, como por exemplo nas previsões de tempo ou resultado de jogo de futebol.

Em ambos os casos a teoria é a mesma, apenas que na f**requentista** usamos a contagem dos eventos possíveis para determinar a probabilidade de um evento, enquanto que no caso **Bayesiano** usamos informações para estimar na incerteza sobre o evento. Entretanto, a definição mais precisa e atual para a probabilidade é aquela que satisfaz os **Axiomas de Kolmogorov:**

$(A1)\quad P(\Omega) = 1$

$(A2)\quad Para\ todo\ subconjunto\ A\ \in\ \mathcal{F}\ |\ \mathcal{F} \subset\ \Omega, P(A) \geq 0$

$(A3)\quad Para\ A_1, A_2, \ldots . \in\ \mathcal{F}, mutualmente\ exclusivos, então\ P(\bigcup_{i=1}^{\infty} A_i = \sum_{i=1}^{\infty} P(A_i)$

Portanto, como trabalhamos com dados, cujo comportamento são muitas vezes desconhecidos, precisamos de um ferramental probabilístico para lidar com eles, principalmente em ***Soft Computing***.

PROBABILIDADE À PRIORI

Se um evento (A) pode ocorre de h maneiras diferentes em um total de n maneiras prováveis e possíveis, então a probabilidade do evento (A) é dado por :

$$P(A) = \frac{h}{n} = \frac{Evento\ Desejado}{Total\ de\ Eventos}$$

E se tivermos 2 ou mais eventos **"independentes"**, a probabilidade do evento "A" ocorrer junto com o evento "B" será dado por:

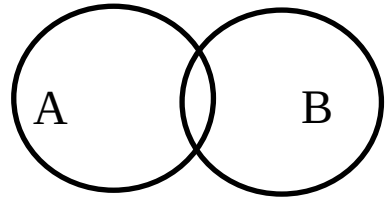
$$P(A \wedge B) = P(A).P(B)$$

Note que, como consideramos os eventos independentes, a probabilidade deles ocorrerem simultaneamente é muito pequena, por isso o resultado do produto diminui.

Por exemplo: a queda de um meteoro em SP e você ganhar na Mega Sena.

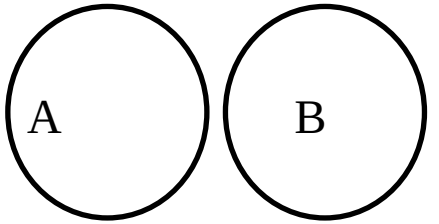
REGRA DE ADIÇÃO DE PROBABILIDADES

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Se os eventos são independentes $(A \cap B = \emptyset)$ então:

$$P(A \cup B) = P(A) + P(B)$$



COMPLEMENTO DO EVENTO

Podemos escrevemos P(A) como sendo a probabilidade para os eventos verdadeiros e $\mathrm{P}(\neg A)$ ou $P(\bar{A})$ para indicar os eventos falsos.

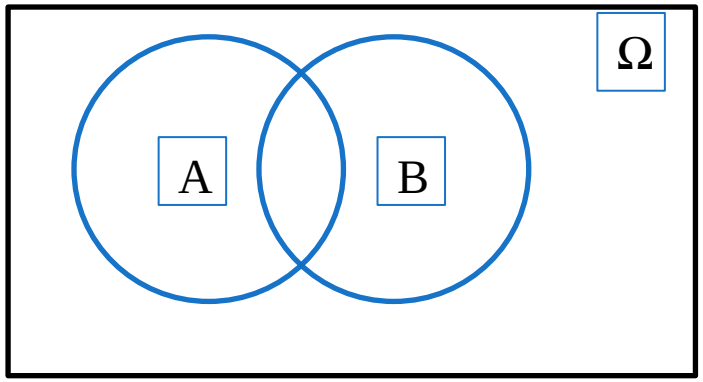
Assim, $\mathrm{P}(A) =$ "vai chover amanhã" com um incerteza que estará entre $0 \leq P(A) \leq 1$.

Logo, $P(\neg A) = 1 - P(A)$ é o complemento de $P(A)$.

PROBABILIDADE CONJUNTA

A probabilidade de ocorrer A e B é igual a probabilidade da ocorrência de A multiplicado pela probabilidade da ocorrência de B na hipótese de A ter ocorrido.

$$P(A \cap B) = P(A.B) = P(B)P(A|B)$$



PROBABILIDADE CONDICIONAL

É a probabilidade de um evento A ocorrer dado a ocorrência de um evento B. Essa probabilidade é a intersecção entre os dois conjuntos de possíveis ocorrências dividida pela probabilidade de B.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{ou} \quad P(A \cap B) = P(A|B).P(B)$$

Note que agora a ocorrência de B torna-se o novo espaço amostral.

TEOREMA OU REGRA DE BAYES

O Teorema de Bayes nos auxilia a determinar a probabilidade de um evento baseado nas observações à priori. O Teorema auxilia a revisar a probabilidade à posteriori a partir das novas evidências. É um conceito simples, mas muito importante na área de Aprendizado de Máquinas.

Seja $C_1, C_{2,\ldots} C_N \in (\Omega, \mathcal{F}, P)$ *que formam partições de Ω e têm probabilidades positivas. Seja A um evento tal que P(A)>0.*
Logo, para todo j = 1, 2,...N, temos que:

$$P(C_j|A) = \frac{P(A|C_j).P(C_j)}{\sum_{i=1}^{N} P(A|C_i).P(C_i)}$$

O Teorema de Bayes permite recalcular as probabilidades das causas $P(C_i|A), i = 1,2,..N$ *ou à posteriori, indicando o quanto cada causa é responsável pela ocorrência do evento A.*

(Magalhões-2011)

MÉDIA DE UMA AMOSTRA E ESPERANÇA

Dois termos que aparecem com frequência na literatura são o valor esperado de uma variável $E(X)$ e a média $\mu(X)$.

É muito importante compreender uma sutil diferença entre elas. No caso da Esperança Matemática $E(X)$, calcula-se a **média da probabilidade da distribuição** tal que:

$$E(X) = \sum_{i=1}^{N} x_i p(x_i)$$

E no caso da média $\mu(X)$, calcula a média aritmética das amostras:

$$\mu(X) = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Observe que $E(X) = \mu(X)$, somente se considerarmos que todas as variáveis **são equiprováveis**, $p(x_1) = p(x_2) = \cdots p(x_N)$, o que não necessariamente pode ser verdadeiro.

Aprendizado Probabilístico – Conceitos Básicos

Considere a tabela abaixo:

Gols num Jogo (x_i)	Probabilidade $p(x_i)$
0	0,10
1	0,20
2	0,20
3	0,20
4	0,20
5	0,08
6	0,001
7	0,001

A esperança matemática $E(X) = \sum_{i=1}^{N} x_i p(x_i)$ do número de gols num jogo (baseado na tabela acima) será de:

$$E(X) = 0.(0{,}1) + 1.(0{,}20) + 2.(0{,}20) + 3.(0{,}2) + 4.(0{,}2) + 5.(0{,}08) + 6.(0.01) + 7.(0.01) = 2{,}53$$

Aprendizado Probabilístico – Conceitos Básicos

Estatística do Campeonato Paulista de 2022:

2022

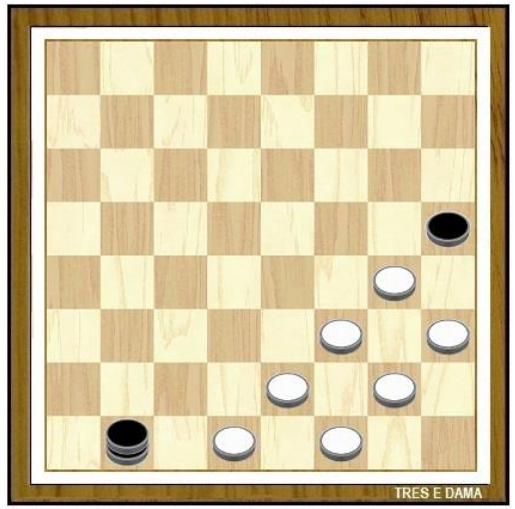
GLOBAIS

Jogos:	109	Vitórias em Casa:	52 (48%)	Jogos com -3 gols:	59 (54%)
Gols:	259	Vitórias Fora:	26 (24%)	Jogos com 3 ou mais gols:	50 (46%)
Média:	2,38	Empates:	31 (28%)	Resultado Típico:	1-0 (14 J)

https://www.ogol.com.br/edition_stats.php?id=160302

Aprendizado Probabilístico – Conceitos Básicos

Nos primórdios das pesquisas com IA aplicado em robótica o conceito de raciocínio era baseado em lógica de primeira ordem. O conceito funciona bem em ambientes controlados e com problemas limitados. Para ambiente mais complexos o método torna-se muito impraticável.



Algoritmos de busca em árvores para tomar decisões de movimento.

Aprendizado Probabilístico – Conceitos Básicos

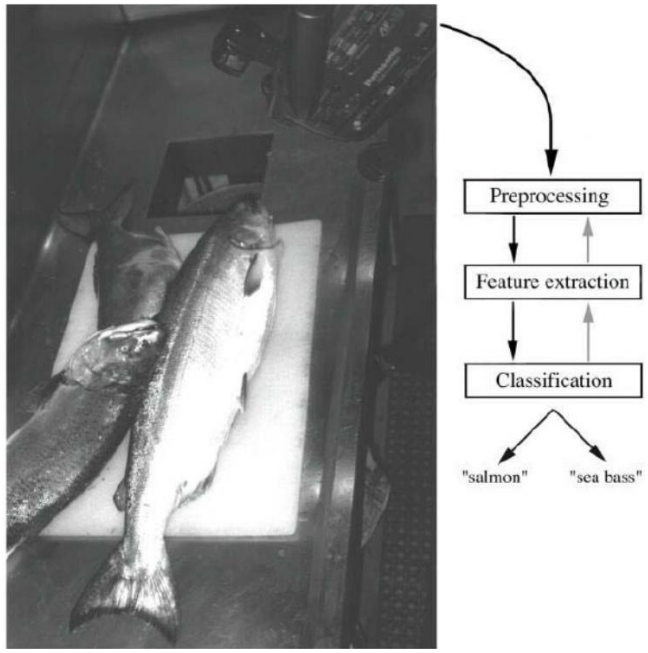
Em um ambiente real, a tomada de decisões vai muito além do conjunto de possíveis movimentos ou jogadas num jogo de tabuleiro. As incertezas são do agente que está no mundo e não necessariamente do domínio. Suponha que alguém vai ao dentista porque está com uma dor de dente. Qual seria o diagnóstico?

O raciocínio lógico de primeira ordem falha porque o agente tem:

- Preguiça: É trabalhoso listar todas as possibilidades
- Ignorância teórica: Não conhece todas as possibilidades
- Ignorância prática: Pode não ter todos os dados e testes para avaliar

Aprendizado Probabilístico – Conceitos Básicos

Suponha que você trabalha num empresa pesqueira em alto mar. Os peixes pescados numa rede são armazenados no porão refrigerados e posteriormente processados, ainda dentro do navio. Uma esteira traz os peixes do porão para a área de processamento via uma esteira rolante. Você nota que há dois tipos de peixes, robalo e salmão.



Duda, Hart & Stork, Pattern Classification

Você conclui, após várias observações, que:

$$Total = Qtde(salmao) + Qtde(robalo)$$

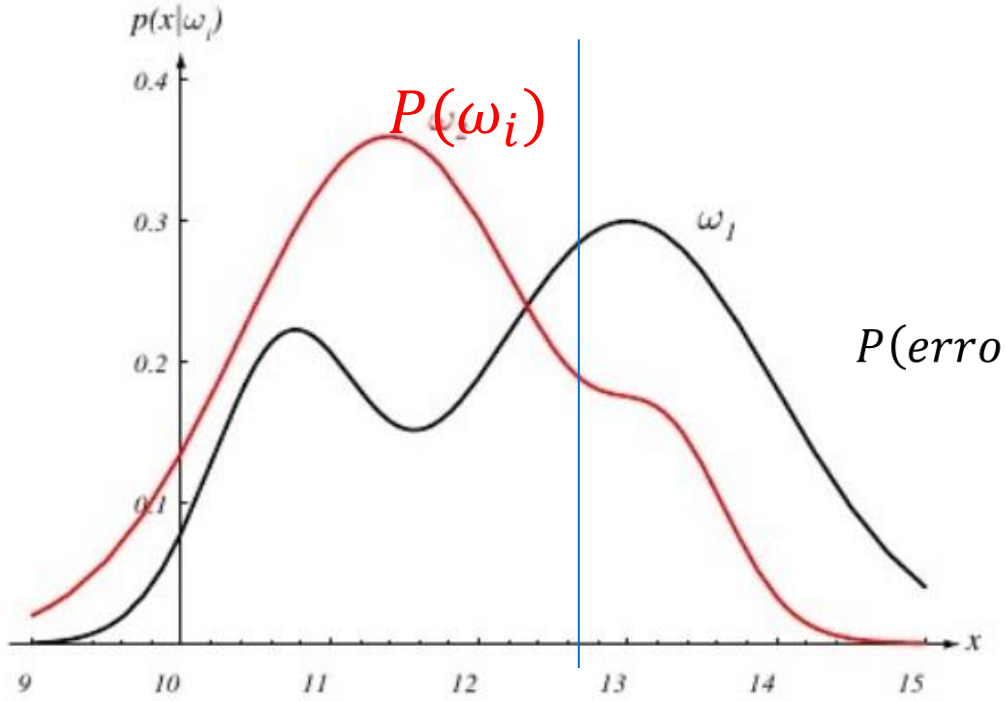
É possível predizer qual espécie de peixe virá pela esteira usando somente essas probabilidades?

$$P(salmao) = \frac{Qtde(salmao)}{Total} \text{ e } P(robalo) = \frac{Qtde(robalo)}{Total}$$

$$P(salmao) + P(Robalo) = 1$$

Aprendizado Probabilístico – Conceitos Básicos

Agora, vamos formalizar o exemplo dos peixes definindo que o estado da natureza $\omega = \omega_1$ se for robalo e $\omega = \omega_2$ se for salmão, e sendo o estado da natureza imprevisível, logo só teremos as probabilidades $P(\omega_1) + P(\omega_2) = 1$.



$$P(erro|peixe) = \begin{cases} P(salmao|peixe)\ se\ decidiu\ robalo \\ P(robalo|peixe)\ se\ decidiu\ salmao \end{cases}$$

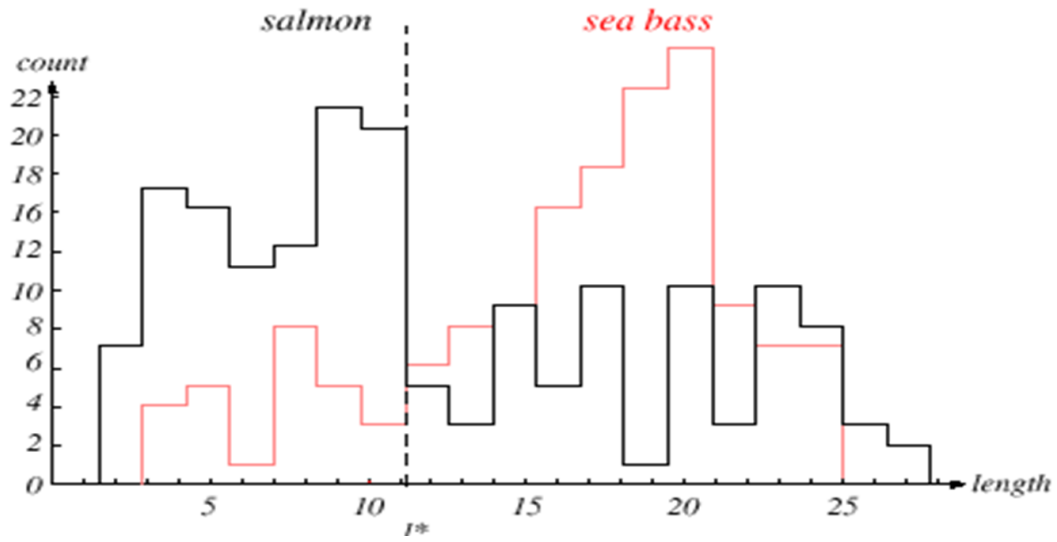
As curvas são as funções densidades de probabilidade à priori da característica x de cada peixe que foi capturado.

Duda, Hart & Stork, Pattern Classification

Aprendizado Probabilístico – Conceitos Básicos

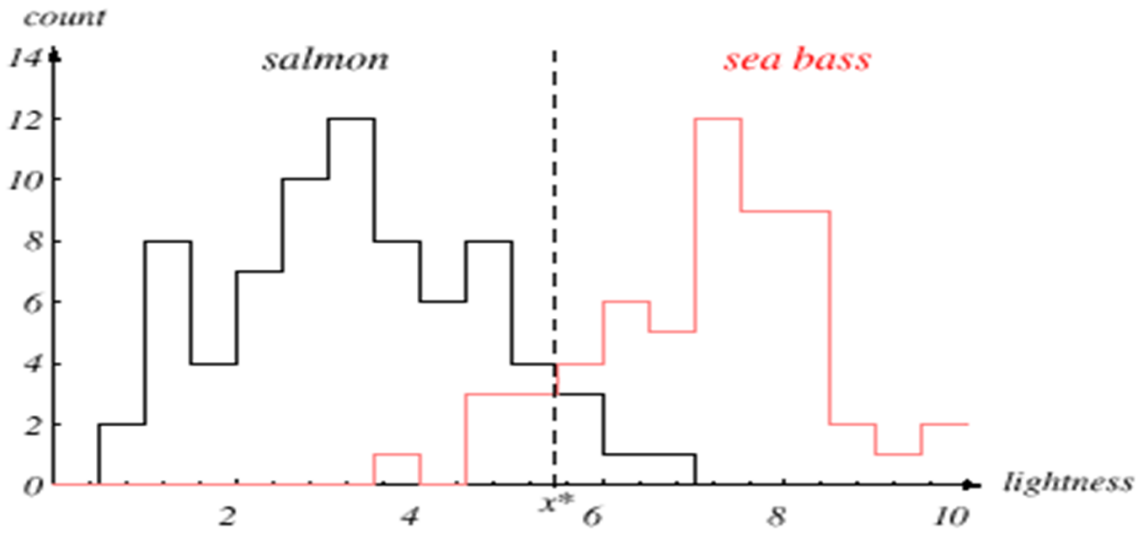
O problema com a probabilidade à priori (incondicional) é que ela somente dá uma ideia sobre as observações passadas e é usada quando somente temos as observações, como no caso de um jogo de Cara ou Coroa ou de Dados.

A probabilidade condicional (posteriori) é quando o agente usa outra observação que não somente aquela do tipo de peixe classificado $P(c_i|\mathrm{x}_k)$.



Duda, Hart & Stork, Pattern Classification

Comprimento

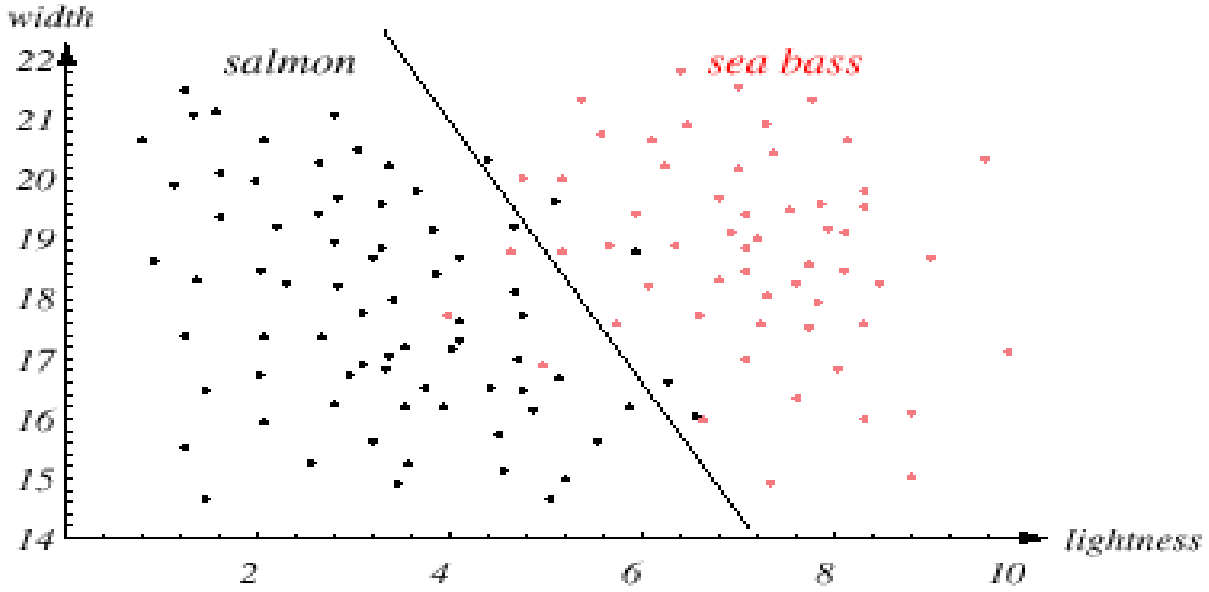


Luminosidade

Aprendizado Probabilístico – Conceitos Básicos

Assim, podemos montar uma expressão que pode nos ajudar a estimar o tipo de peixe que vem na esteira, baseado no comprimento e na luminosidade. É claro que ainda com possibilidade de errar na escolha.

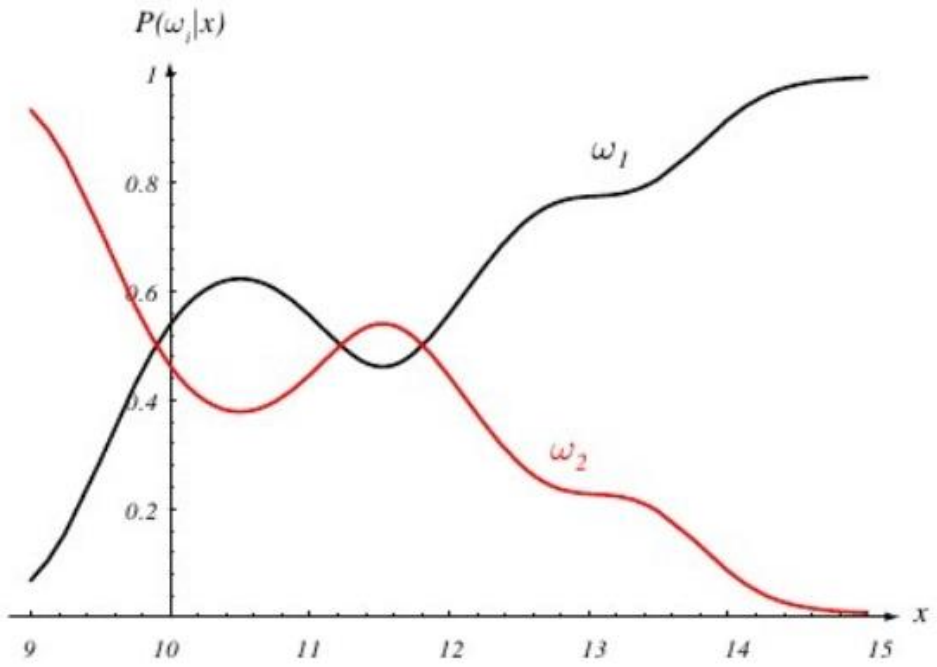
$$P(Peixe) = \begin{cases} p(salmao|comprimento, brilho) > p(robalo|comprimento, brilho) \\ p(salmao|comprimento, brilho) < p(robalo|comprimento, brilho) \end{cases}$$



Duda, Hart & Stork, Pattern Classification

Aprendizado Probabilístico – Conceitos Básicos

Podemos então usar o conhecimento à priori para criar uma regra de classificação baseada na característica que foi observada, ou seja, daremos o tipo de peixe dado a luminosidade medida.



Duda, Hart & Stork, Pattern Classification

A equação abaixo é conhecida como Regra de Bayes.

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{\sum_{j=1}^{2} p(x|\omega_j)P(\omega_j)}$$

E diz que podemos converter a probabilidade à priori de $P(\omega_j)$ em uma probabilidade à posteriori $P(\omega_j|x)$. Note que agora temos a seguinte probabilidade de erro:

$$P(erro|x) = \begin{cases} P(\omega_1|x) \; se \; decidiu \; por \; \omega_2 \\ P(\omega_2|x) \; se \; decidiu \; por \; \omega_1 \end{cases}$$

PROBABILIDADE FREQUENTISTA – À POSTERIORI

Vamos imaginar a seguinte situação. Uma tabela que mostra se houve um determinado jogo de vôlei de praia conforme algumas condições climáticas (Adaptado de Mitchel).

Amostras	Clima	Temperatura	Umidade	Vento	Jogar Volei
1	Ensolarado	Quente	Alta	Fraco	Não
2	Ensolarado	Quente	Alta	Forte	Não
3	Nublado	Quente	Alta	Fraco	Sim
4	Chovendo	Agradável	Alta	Fraco	Sim
5	Chovendo	Fria	Normal	Fraco	Sim
6	Chovendo	Fria	Normal	Forte	Não
7	Nublado	Fria	Normal	Forte	Sim
8	Ensolarado	Agradável	Alta	Fraco	Não
9	Ensolarado	Fria	Normal	Fraco	Sim
10	Chovendo	Agradável	Normal	Fraco	Sim
11	Ensolarado	Agradável	Normal	Forte	Sim
12	Nublado	Agradável	Alta	Forte	Sim
13	Nublado	Quente	Normal	Fraco	Sim
14	Chovendo	Agradável	Alta	Forte	Não

Quiz: Qual a probabilidade condicional de ter jogo se o clima estiver ensolarado?

PROBABILIDADE FREQUENTISTA – À POSTERIORI

Observou-se pelo exemplo do jogo de vôlei de praia que a ocorrência de um evento pode ou não disparar outro. Assim, a probabilidade condicional calcula esses eventos dados as observações que foram obtidas (tempo = ensolarado) para estimar duas hipóteses (h1 = joga ou h2 = não joga).

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Podemos também escrever que:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Então:

$$P(A \cap B) = P(B \cap A)$$

$$P(B).P(A|B) = P(A).P(B|A)$$

Aprendizado Probabilístico – Conceitos Básicos

$$P(Clima = ensolarado) = \frac{5}{14} = 0{,}357$$

$$P(Volei = joga \cap Clima = ensolarado) = \frac{2}{14} = 0{,}143$$

$$PP(Volei = não\ joga \cap Clima = ensolarado) = \frac{3}{14} = 0{,}214$$

Assim, a **probabilidade condicional** para a observação clima = ensolarado será:

$$P(Volei|Clima) = \frac{P(Volei \cap Clima)}{P(Clima)}$$

Para a hipótese Vôlei = joga

$$P(Volei|Clima) = \frac{0{,}143}{0{,}357} = 0{,}4$$

Para a hipótese Vôlei = não joga

$$P(Volei|Clima) = \frac{0{,}214}{0{,}357} = 0{,}6$$

CLASSIFICADOR NAÏVE BAYES

REGRA DE BAYES

A Regra de Bayes estabelece que dado uma hipótese h e uma evidência E para essa hipótese, então existe uma relação entre os vários eventos de E que podem ser a causa da hipótese h. Formalmente é definido como:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = P(A_j|B) = \frac{P(B|A_j).P(A_j)}{\sum_{i=1}^{N} P(B|A_i).P(A_i)} \mid A_1 \cap \cdots A_n = \emptyset \; e \; A_1 \cup \cdots A_N = \Omega$$

Que pode ser lido como:

$$P(Hipótese|Evento) = \frac{P(Evento|Hipótese)P(Hipótese)}{P(Evento)}$$

A principal pergunta que a Regra de Bayes procura responder é: se observarmos o valor de um evento podemos converter o conhecimento a priori *P(A)* em conhecimento à posteriori?

MAP – MAXIMUM A POSTERIORI

A Regra de Bayes ajuda responder qual a maior probabilidade (h_{MAP}), dentre as várias possíveis h_i ocorrer dado que temos um conjunto de observações E_j e algumas probabilidades à priori.

Um dos modelos usados em ML é o conceito de MAP (*Maximum A Posteriori*) que busca o resultado que maximiza as chances.

$$h_{MAP} = argmax_{h \in H} P(H|E)$$

Contudo, lembre-se que sempre há mais de um fator que afeta uma decisão, como no exemplo anterior de jogar ou não jogar vôlei de praia.

Ou seja, precisamos avaliar as probabilidades condicionais de todas as variáveis.

$$P(Volei = joga | Clima = c, T = t, U = u, V = v)$$

$$P(Volei = não\ joga | Clima = c, T = t, U = u, V = v)$$

O problema é calcular essa probabilidade

CLASSIFICADOR NAÏVE BAYES

O classificador NB considera que as características de um vetor aleatório são descorrelacionadas entre as classes. Mesmo se essas características dependam entre si ou pela existência de outras características, ele considera que as características contribuem independentemente para a probabilidade de cada classe. E é essa a característica que dá o nome "naïve".

A parte de Bayes considera que as quantidades de interesse são governadas pela distribuição de probabilidades e que as decisões ótimas podem ser feitas pelo raciocínio sobre essas probabilidades $P(\omega_j|\mathbf{x})$ juntamente com os dados observados $(\mathcal{D})$.

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i, \mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^{C} p(x|\omega_j, \mathcal{D})P(\omega_j)}$$

O problema é sempre estimar $p(\mathbf{x}|\omega_j, \mathcal{D})$ justamente pelo número de amostras limitado.

$$P(x_1, \ldots, x_n|\omega_j) = \prod_i P(x_i|\omega_j)$$

CLASSIFICADOR NAÏVE BAYES

Com Bayes é possível superar o problema anterior simplificando a hipótese e usando o conhecimento à priori que temos a partir da base de dados

$$\omega_{NB} = argmax_{\omega_j \in \Omega} P(\omega_j) \prod_i P(a_i | v_j)$$

Por exemplo, suponha que tenhamos a seguinte tarefa de classificar a seguinte entrada:

Clima = ensolarado, Temperatura = Fria, Umidade = alta, Vento = forte

CLASSIFICADOR NAÏVE BAYES

$$\omega_{NB} = argmax_{\omega_j \in \Omega} P(\omega_j) \prod_i P(a_i | v_j)$$

Calculando $\omega_{NB} = \underset{\omega_j \in \Omega}{argmax} P(\omega_j)$

$$P(Clima = ensolarado|\omega_j).P(Temperatura = fria|\omega_j).P(Umidade = Alta|\omega_j).P(Vento = forte|\omega_j)$$

Amostras	Clima	Temperatura	Umidade	Vento	Jogar Volei
1	Ensolarado	Quente	Alta	Fraco	Não
2	Ensolarado	Quente	Alta	Forte	Não
3	Nublado	Quente	Alta	Fraco	Sim
4	Chovendo	Agradável	Alta	Fraco	Sim
5	Chovendo	Fria	Normal	Fraco	Sim
6	Chovendo	Fria	Normal	Forte	Não
7	Nublado	Fria	Normal	Forte	Sim
8	Ensolarado	Agradável	Alta	Fraco	Não
9	Ensolarado	Fria	Normal	Fraco	Sim
10	Chovendo	Agradável	Normal	Fraco	Sim
11	Ensolarado	Agradável	Normal	Forte	Sim
12	Nublado	Agradável	Alta	Forte	Sim
13	Nublado	Quente	Normal	Fraco	Sim
14	Chovendo	Agradável	Alta	Forte	Não

Calcule a probabilidade para jogar e não jogar.

CLASSIFICADOR NAÏVE BAYES

Calculando a probabilidade de cada evento dado o conjunto das observações, temos:

$$P(Joga|\mathbf{X}) = P(Jogar).P(Clima = Ensol.|Joga).P(Temp = Fria|Joga).P(Umid. = Alta|Joga).P(Vento = Forte|Joga) = 0.0053.$$

$$P(\neg Joga|\mathbf{X}) = P(\neg Jogar).P(Clima = Ensol.|\neg Joga).P(Temp = Fria|\neg Joga).P(Umid. = Alta|\neg Joga).P(Vento = Forte|\neg Joga) = 0.0206.$$

Normalizando as duas probabilidades tal que:

$$P(Jogar) = \frac{P(C_1)}{\sum_{i=1}^{N} P(C_i|\mathbf{X})} = 0{,}205 \;\; e \quad P(\neg Jogar) = \frac{P(C_2)}{\sum_{i=1}^{N} P(C_i|\mathbf{X})} = 0{,}795$$

Logo, decidimos que não haverá jogo com um probabilidade de 79,5%, considerando as condições climáticas.

$$P(Clima = ensolarado|c_j).P(Temperatura = fria|c_j).P(Umidade = Alta|c_j).P(Vento = forte|c_j)$$

CLASSIFICADOR NAÏVE BAYES EM PYTHON NO GOOGLE COLAB