Resumos ADI

2021/22

Índice

Sistemas de Aprendizagem **3**
Aprendizagem com supervisão . . . 3
Aprendizagem sem supervisão . . . 3
Aprendizagem por reforço . . . 4

Metodologias de Análise de Dados **5**
CRISP-DM . . . 5
SEMMA . . . 6
PMML . . . 7

Preparação de Dados **8**
Tarefas na preparação de dados . . . 9
Discretização/Enumeração . . . 10
Limpeza de dados . . . 11
Integração dos dados . . . 11
Transformação de dados . . . 12
Redução de dados . . . 12

Avaliação de Modelos **13**
Árvores de Decisão . . . 13
Modelos de Decisão . . . 13
Ciclo de Execução . . . 14
Avaliação de Modelos . . . 15
Hold-out Validation . . . 15
Cross Validation . . . 15
Leave-one-out Cross Validation (k=N) . . . 16

Técnicas de Regressão **17**
Regressão Linear . . . 17
Regressão Linear Múltipla . . . 18
Regressão Logística . . . 18

Métricas de Qualidade **19**
Matrizes de Confusão 19
Modelos de Classificação 20

Árvores de Decisão **21**
Tipos de Árvores de Decisão 21
Entropia 22

Segmentação (Clustering) **23**
Utilização da Segmentação 23
Tipos de dados para análise 24
Atributos contínuos 24
Atributos binários 24
Atributos nominais 25
Atributos ordinais 25
Atributos mistos 25
Principais Métodos de Segmentação 25
Algoritmos de Particionamento 26
Método k-means 26
Método k-medoids 27
Algoritmos de Hierarquização 27
AGNES: Agglomerative Nesting 28
DIANA: Divisive Analysis 28
Vantagens e desvantagens 28

Redes Neuronais Artificiais **29**
Conceitos e definições: Neurónio 29
Conceitos e definições: Axónio 30
Conceitos e definições: Sinapses 30
Conceitos e definições: Ativação 31
Conceitos e definições: Transferência 31
Organização dos neurónios 32
Aprendizagem 32

Sistemas de Aprendizagem

Paradigma de computação em que a característica essencial do sistema se revela pela sua capacidade de aprender de modo autónomo e independente.

Existem 3 tipos de aprendizagem:

Aprendizagem com supervisão

Paradigma de aprendizagem em que os casos que se usam para aprender contêm informação acerca dos resultados pretendidos, sendo possível estabelecer uma relação entre os valores pretendidos e os valores produzidos pelo sistema.

- Pode ser dividida em:
  1. Classificação: Quando os resultados são discretos (preto, branco, cinza...);
  2. Regressão: quando os resultados são contínuos (variação da temperatura ou da luz solar ao longo do dia).

Aprendizagem sem supervisão

Paradigma de aprendizagem em que não são conhecidos resultados sobre os casos, apenas os enunciados dos problemas, tornando necessário a escolha de técnicas de aprendizagem que avaliem o funcionamento interno do sistema.

- Pode ser dividida em:
  1. Segmentação: quando se pretende organizar os dados em grupos coerentes (agrupar clientes que compram bebidas açucaradas);
  2. Associação: quando se pretende conhecer regras que associem o comportamento demonstrado pelos dados (pessoas que comprar bebidas açucaradas não compram bebidas alcoólicas).

Aprendizagem por Reforço

Paradigma de aprendizagem que, apesar de não ter informação sobre os resultados pretendidos, permite efetuar uma avaliação sobre se os resultados produzidos são bons ou maus.

- Algoritmos de Reinforcement Learning usam técnicas de auto-alimentação de sinais, com vista a melhorar os resultados, por influência da noção de recompensa/penalização;
- Não se pode comparar com Aprendizagem Supervisionada uma vez que a "opinião" sobre os resultados não é dada por um professor/treinador;
- Também não se pode considerar Aprendizagem não Supervisionada, uma vez que não existe ausência absoluta de informação sobre os resultados;

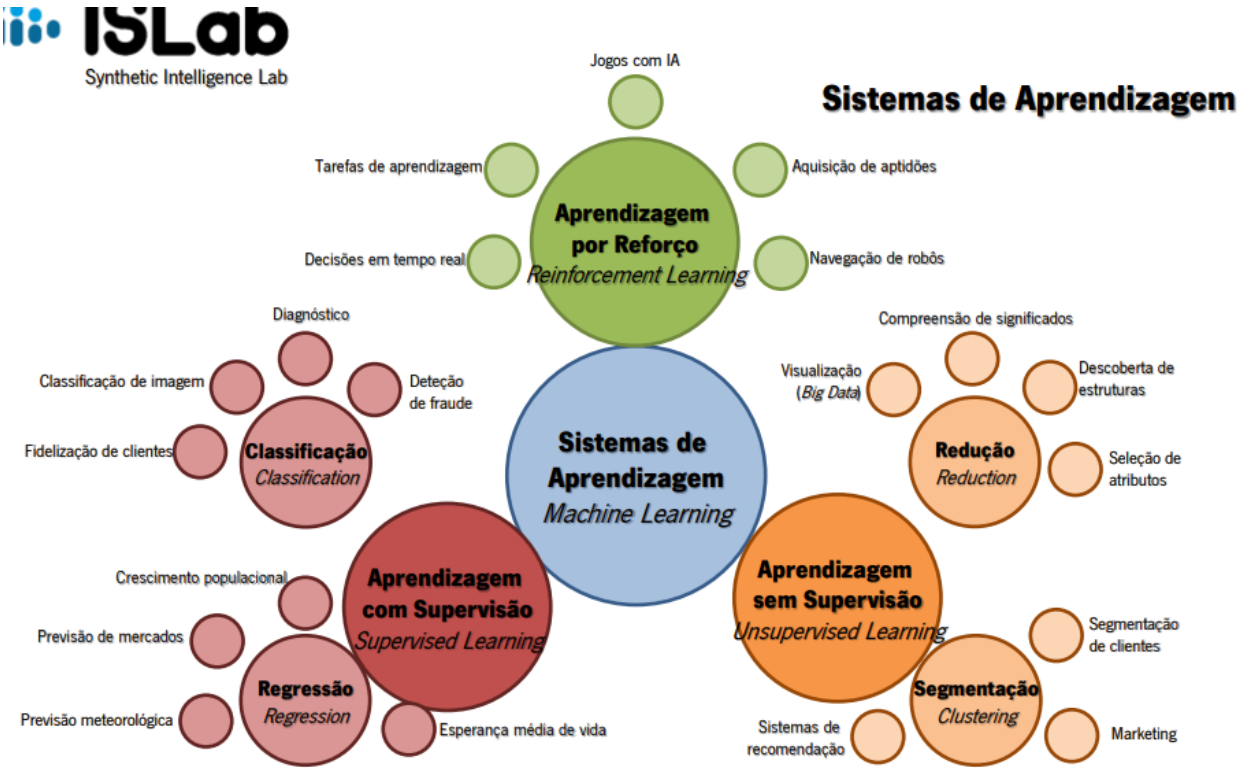


Figura 1: Sistemas de Aprendizagem

Metodologias de Análise de Dados

Uma Metodologia para Análise de Dados descreve e cria um conjunto de passos pelos quais deverá passar o desenvolvimento de um Projeto de Aprendizagem Automática (Machine Learning) para a resolução de problemas.

CRISP-DM

CRoss **I**ndustry Standard **P**rocess for **D**ata **M**ining

O CRISP-DM é um modelo de processos com vista a definir um "guião" para o desenvolvimento de projetos de AD, que se desenrola em 6 etapas:

1. **Estudo do negócio**: Compreensão dos objetivos do projeto e definição do problema de AD;
2. **Estudo dos dados**: Obter os dados e identificar a qualidade dos dados;
3. **Preparação dos dados**: Seleção de atributos e limpeza dos dados;
4. **Modelação**: Experimentação com as ferramentas de AD;
5. **Avaliação**: Comparação dos resultados com os objetivos do negócio;
6. **Desenvolvimento**: Colocação do modelo em produção

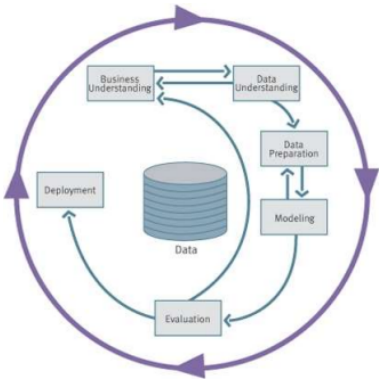


Figura 2: CRISP-DM

SEMMA

Sample, **E**xplore, **M**odify, **M**odel and **A**ssess

Divide o processo de Data Mining em 5 etapas:

1. **Sample/Amostragem**: Extração de dados do universo do problema (Amostra pequena e significativa);
2. **Explore/Exploração**: Exploração visual e/ou numérica das tendências. Utiliza técnicas estatísticas e é efetuado o refinamento do processo de descoberta.
3. **Modify/Modificação**: Concentração de todas as modificações necessárias;
4. **Model/Modelação**: Definição das técnicas de construção de modelos de Data Mining: redes neuronais artificiais, árvores de decisão, regressão linear, etc.;
5. **Assess/Avaliação**: Aferição do desempenho do modelo construído para Data Mining.

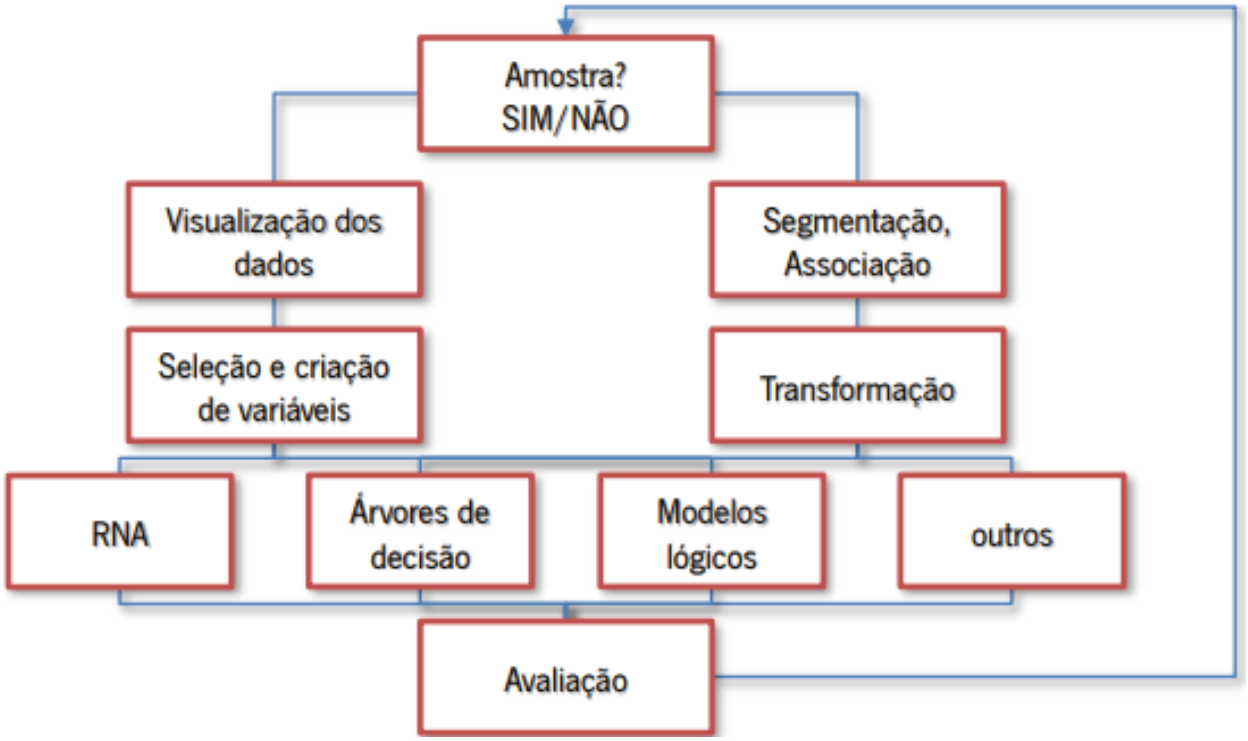


Figura 3: SEMMA

PMML

Predictive **M**odel **M**arkup **L**anguage;

O PMML utiliza XML para descrever modelos de Data Mining.

1. Permite que aplicações utilizem diversas fontes de dados sem se preocuparem com as diferenças entre elas;
2. Permite a utilização combinada e/ou cooperativa de modelos de Data Mining;
3. Permite a administração de modelos de Data Mining baseados em áreas de negócio.

Preparação de Dados

O principal objetivo da preparação dos dados consiste em transformar os data sets por forma a que a informação neles contida esteja adequadamente exposta à ferramenta de extração de conhecimento;

Os dados recolhidos no "mundo real":

- são incompletos;
  - falta de valores em alguns atributos;
  - falta de alguns atributos;
  - etc...
- contêm lixo;
  - identificam valores impossíveis;
  - Salário: -1.000EUR;
  - etc...
- podem conter inconsistências.
  - encontram-se discrepâncias entre valores ou nomes;
  - Idade = 35; Data de nascimento = 31/maio/1969;
  - etc...

Tarefas na preparação de dados

1. **Discretização/Enumeração**
    - Redução de dados com importante aplicação a dados numéricos;
2. **Limpeza**
    - Preenchimento de valores de atributos;
    - Remoção de lixo dos dados;
    - Remoção de valores impossíveis;
    - Resolução de inconsistências;
3. **Integração**
    - Múltiplas fontes de dados (BD's, ficheiros, papel, web, etc.);
4. **Transformação**
    - Normalização e agregação de dados;
5. **Redução**
    - Obtenção de representações de dados menos volumosas, mas com capacidade para produzir idênticos resultados analíticos;
    - Redução de dimensões;
    - Compressão de dados.

Discretização/Enumeração

Utiliza-se a discretização (ou enumeração) para reduzir o número de valores de um atributo contínuo, dividindo-o em intervalos.

- **Discretização de igual largura**:
  - Divide a gama de valores em N intervalos de igual largura, resultando numa grelha uniforme;
  - Sendo A e B os limites da gama de valores, a largura dos intervalos será L = (B - A) / N:

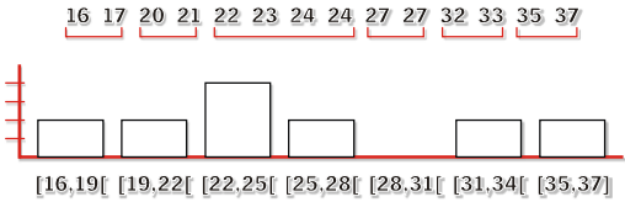


Figura 4: Discretização de igual largura

- **Discretização de igual altura**:
  - Divide a gama de valores em N intervalos, contendo, cada um, aproximadamente a mesma quantidade de valores:

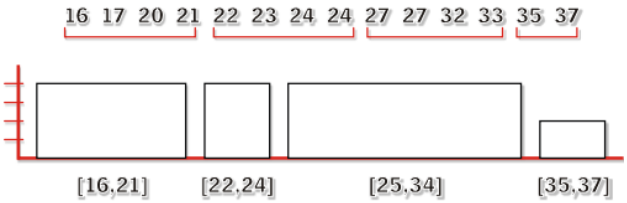


Figura 5: Discretização de igual altura

Limpeza de dados

- **Como tratar a ausência de dados?**
  - Ignorar os registos onde faltam os dados e lidar, apenas com os dados conhecidos;
  - Ignorar os atributos onde faltam os dados;
  - Preencher (manualmente) os dados em falta;
  - Preencher com o valor médio do atributo;
  - Preencher com o valor mais frequente do atributo;
  - Quantos mais valores "inventados", maior o desvio dos dados que caracterizam o problema face à realidade que o problema ilustra! **Deve-se evitar adicionar distorção aos dados**.

Integração dos dados

Os dados que caracterizam o problema podem ter proveniências diversas. O objetivo da integração é o de compor um conjunto de peças de informação numa coleção coerente e integrada de dados.

Integração exige "conhecimento do negócio".

Transformação de dados

1. **Alisamento (smoothing)**
    - Remover lixo/ruído dos dados (binning, regressão, clustering);
2. **Agregação**
    - Pressupõe que o resultado sumaria os dados iniciais;
      (resumo de vendas trimestrais, durante 5 anos, em valores anuais)
3. **Generalização**
    - Hierarquização de conceitos ($distrito \rightarrow cidade \rightarrow rua$).
4. **Construção de Atributos**
    - Construção de novos atributos a partir de outros
      (cálculo do preço líquido baseado no preço ilíquido e no IVA);
5. **Uniformização**
    - Pretende evitar que atributos com uma gama alargada de valores sobressaiam em relação a outros atributos com menor quantidade de valores:
6. **Deteção de valores atípicos**
    - A parir de Box plots ou de desvio padrão.

Redução de dados

A Redução de dados pretende obter uma representação reduzida do volume de dados, mas produzindo os mesmos (ou quase os mesmos) resultados analíticos.

- Construção de cubos de dados;
- Redução de dimensões;
- Compressão de dados;
- etc...

Avaliação de Modelos

Árvores de Decisão

Uma Árvore de Decisão é um grafo hierarquizado (árvore!) em que:

- Cada ramo representa a seleção entre um conjunto de alternativas
- Cada folha representa uma decisão

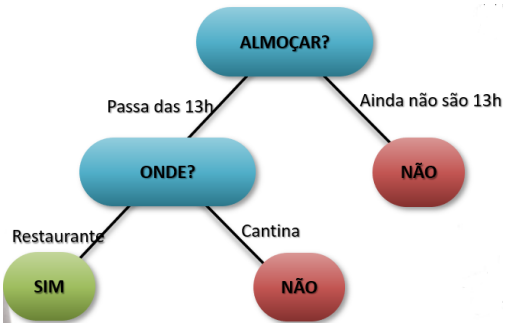


Figura 6: Árvore de decisão

Modelos de Decisão

Existem dois tipos de paradigmas de criação de modelos de decisão:

- **Top-down**:
  - O modelo é construído a partir do conhecimento de especialistas;
  - O "todo" é dividido em "partes";
- **Bottom-Up**:
  - O modelo é construído pela identificação de relações entre os atributos do dataset;
  - O modelo é induzido por "generalização" dos dados;

Ciclo de Execução

Dada uma árvore de decisão (treinada), o processo de decisão desenvolve-se do seguinte modo:

1. Começar no nodo correspondente ao atributo "raiz";
2. Identificar o valor do atributo;
3. Seguir pelo ramo correspondente ao valor identificado,
4. Alcançar o nodo relativo ao ramo percorrido;
5. Voltar a 2. até que o nodo seja uma "folha";
6. O nodo alcançado indica a decisão para o problema.

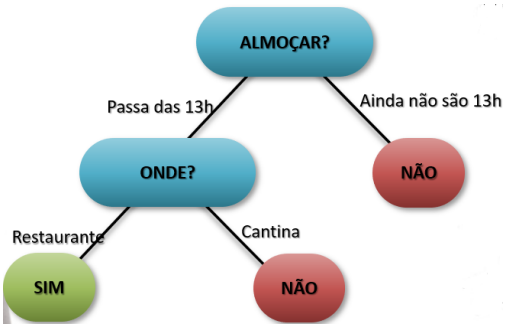


Figura 7: Árvore de decisão

Notas:

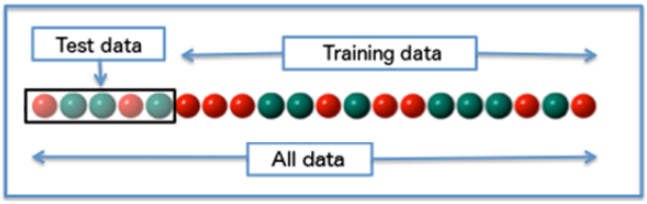
- Uma Árvore de Decisão pode ser utilizada para fazer classificação;
- Uma Árvore de Decisão pode ser utilizada para fazer regressão.

Avaliação de Modelos

Após a criação (treino) de um modelo usando uma técnica de aprendizagem (machine learning), é necessário avaliar o seu desempenho. A medição do desempenho de um modelo é feita com dados não apresentados durante o treino.

Hold-out Validation

- Método de particionamento de dados;
- Divide o conjunto de dados em dados de treino e dados de teste;



- Separa-se uma parte *(hold-out)* do conjunto de dados para treino/teste (80/20; 75/25; ...)

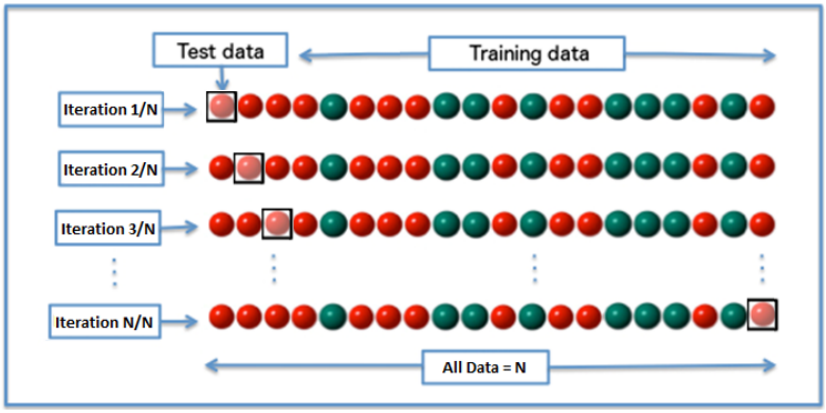
Cross Validation

- Método de validação por cruzamento de dados;
- Consiste em dividir o conjunto de dados em k partes (k *folds*);
  - A cada iteração, o método utiliza k-1 partes (*folds*) para treino e 1 parte (*fold*) para teste;
  - O processo repete-se durante k vezes;

  - O erro final é dado pela média dos valores parciais dos erros.

Leave-one-out Cross Validation (k=N)

- Método de validação por cruzamento de dados;
- Caso particular em que o número de casos N é igual ao número de *folds* k;

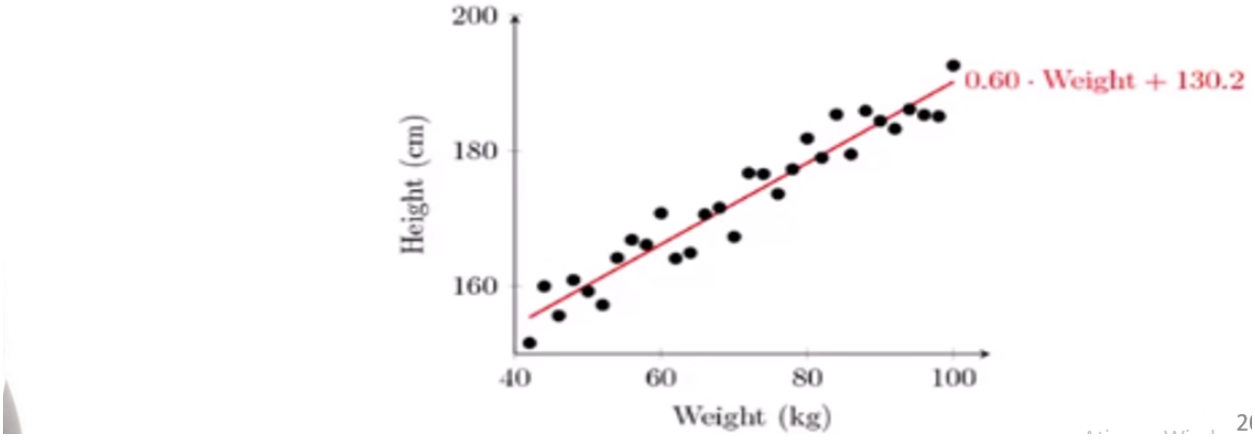


Técnicas de Regressão

A regressão é um procedimento estatístico que determina a equação para a linha reta que melhor se ajusta a um conjunto específico de dados.

Regressão Linear

- Tem como objetivo prever o valor de um resultado, Y, com base no valor de uma variável de previsão, X;
  - Como "encaixar" uma linha reta num conjunto de dados;
  - Usar esta linha para estimar a resolução de problemas.

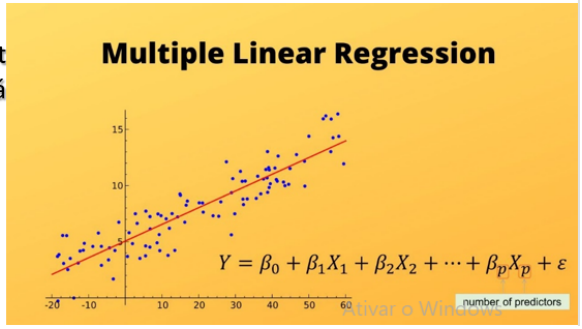


Regressão Linear Múltipla

- A regressão múltipla é usada para determinar o efeito de diversas variáveis independentes, $x_1, x_2, x_3, \ldots$ numa variável dependente, y;
- As diferentes variáveis x_i são combinadas de forma linear e cada uma tem seu próprio coeficiente de regressão:

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + \cdots + a_n \cdot x_n + b + \varepsilon$$

- Os parâmetros a_i refletem a contribuição independent de cada variável independente x_i, para o valor da variá dependente, y.

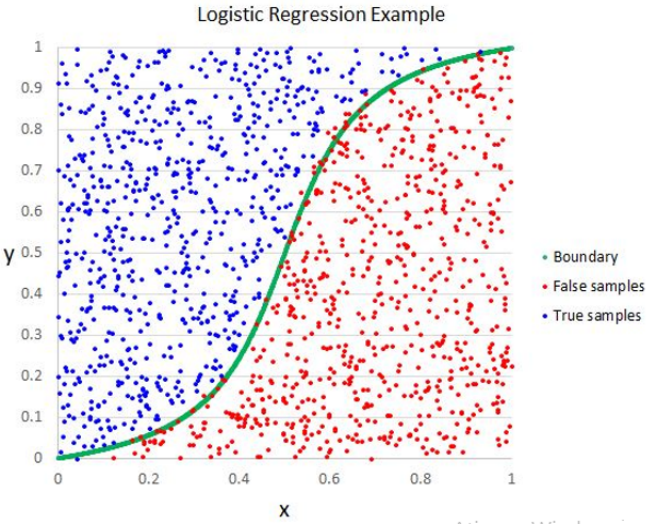


Regressão Logística

- A diferença essencial entre regressão linear e **regressão logística** é que esta é usada quando a variável dependente é de natureza binária.

- Em contraste, a **regressão linear** é usada quando a variável dependente é contínua e a natureza da linha de regressão é linear.

- A Regressão Logística é uma técnica de **classificação**:
  - Empréstimo (SIM/NÃO)
  - Diagnóstico (São/Doente)
  - Vinho (Branco/Rosé/Tinto)

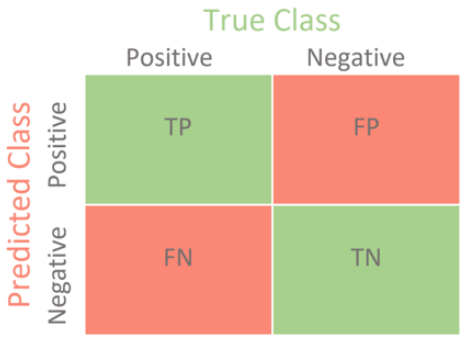


Métricas de Qualidade

Matrizes de Confusão

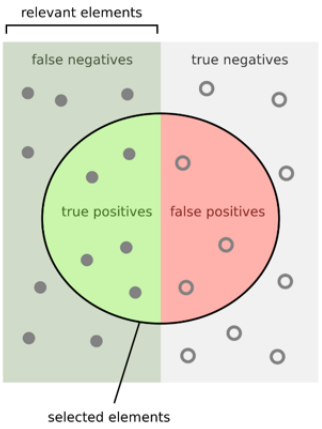
- Matrizes de Confusão
  - Tabela utilizada para descrever o desempenho de um modelo de classificação.

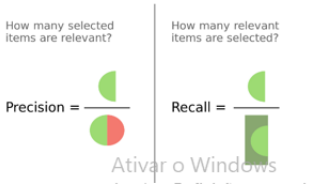
- *Accuracy*
  - Quantidade de previsões corretas dividido pela quantidade total de observações:
  - $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$



Modelos de classificação

- Matrizes de Confusão
  - Tabela utilizada para descrever o desempenho de um modelo de classificaç
- Precisão (*Precision aka Sensitivity*)
  - É uma medida da exatidão;
  - Determina a proporção de itens relevantes entre todos os itens:
    - $Precision = \frac{TP}{TP+FP}$
- *Recall* (*aka Specificity*)
  - É uma medida de completude;
  - Determina a proporção de itens relevantes obtidos:
    - $Precision = \frac{TP}{TP+FN}$





Árvores de Decisão

Tipos de Árvores de Decisão

As Árvores de Decisão podem ser:

- **Contínuas:**
  - O atributo de decisão representa uma sequência, conjunto ou intervalos de possíveis valores;
  - As folhas da árvore de decisão identificam intervalos ou conjuntos de valores;

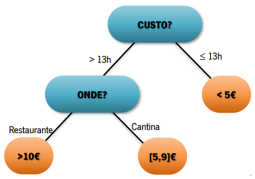


Figura 8: Árvore de Decisão Contínua

- **Discretas:**
  - O atributo de decisão representa uma categoria ou uma classe;
  - Os valores representados nas folhas da árvore de decisão são as categorias ou classes;

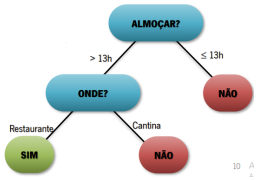


Figura 9: Árvore de Decisão Contínua

Entropia

Entropia é uma medida da incerteza associada a um conjunto de objetos e permite-nos identificar o grau de desorganização dos dados.

- A entropia é 0 (zero) quando todos os objetos de S são do mesmo valor (positivo ou negativo).
- Ganho de informação:
  - Esta métrica mede a redução esperada na entropia;
  - Decisão sobre qual o atributo que será selecionado para ser nodo;
  - O atributo com a maior redução de entropia é a melhor escolha para ser nodo (para reduzir a profundidade da árvore);

Segmentação (Clustering)

A Segmentação/Clustering de dados é um processo através do qual se particiona um conjunto de dados em segmentos/clusters de menor dimensão, que agrupam conjuntos de dados similares.

- **Um Segmento/Cluster é uma coleção de valores/objetos que**:
  - são similares entre si, dentro de um mesmo segmento;
  - são diferentes dos valores/objetos de outros segmentos.

- **Medidas de similaridade**:
  - distância Euclidiana ou de Manhattan, para atributos contínuos;
  - coeficiente de Jacqard, para atributos discretos/binários;

Utilização da Segmentação

A deteção de segmentos é útil quando:

- quando se suspeita da existência de agrupamentos "naturais", que podem representar grupos de clientes, de produtos ou de bens que partilhem (muita) informação;
- quando existam muitos padrões diferentes nos dados, dificultando a tarefa de identificar um determinado padrão;

Tipos de dados para análise

1. Atributos contínuos;
2. Atributos binários;
3. Atributos nominais;
4. Atributos ordinais;
5. Atributos mistos.

Atributos contínuos

- normalizar os dados: evita que os resultados dependam das unidades de medida;
- normalmente, utilizam-se medidas de distância para calcular a proximidade (similaridade) entre objetos.

- distância Euclidiana: é a medida de distância geométrica no espaço (a mais usada):

$$d(x,y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad \text{(para 2 dimensões)}$$

- distância *Manhattan*: mede a distância pela diferença entre os pontos (função não quadrática):

$$d(x,y) = |x_1 - x_2| + |y_1 - y_2| \quad \text{(para 2 dimensões)}$$

Atributos binários

- **Podem ser:**
  - Simétricos: significado de ser 0 é o mesmo de ser 1 (coeficiente simples)
  - Assimétricos: significado de ser 0 é diferente de ser 1 (coeficiente Jaccard)
- A similaridade calculada com base em atributos simétricos é designada similaridade invariante; no caso oposto diz-se similaridade não-invariante;

- coeficiente simples (simétricos):

$$d(i,j) = \frac{b + c}{a + b + c + d}$$

- coeficiente Jaccard (assimétricos):

$$d(i,j) = \frac{b + c}{a + b + c}$$

Atributos nominais

- trata-se de uma generalização dos atributos binários, em que os dados podem assumir mais do que 2 valores;
- Método 1:

  - *matching* simples;
  - $d(i,j) = \frac{nº\ variáveis - nº\ matches}{nº\ variáveis}$

- Método 2:
  - Utilizar variáveis binárias;
  - Criar uma variável binária para cada valor nominal.

Atributos ordinais

- a ordem é relevante;
- podem ser tratados como atributos contínuos, sendo que a ordenação dos valores define uma classificação;
- as similaridades devem ser calculadas utilizando os mesmos métodos que para os atributos contínuos.

Atributos mistos

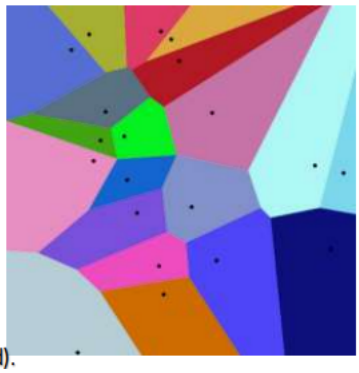
- o conjunto de dados pode conter diversos tipos de atributos
- tipicamente, utiliza-se uma função pesada para ponderar e medir os efeitos de cada atributo.

Principais Métodos de Segmentação

- **Particionamento**: criar várias partições e adotar um critério de avaliação;
- **Hierarquização**: decompor hierarquicamente o conjunto de dados;

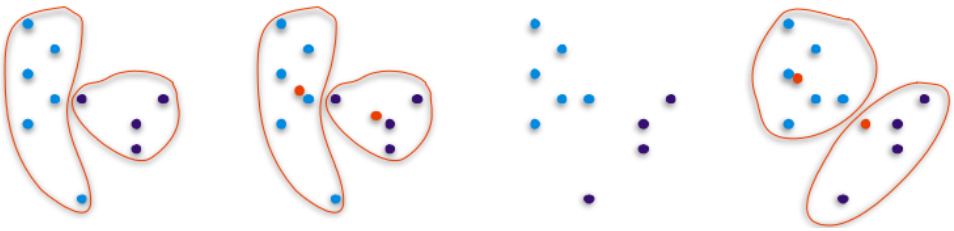
Algoritmos de Particionamento

- Particionar um conjunto de dados 'D' contendo 'n' objetos num conjunto de 'k' segmentos/ *clusters*;
- Sendo dado 'k', particionar 'D' em 'k' segmentos de forma a otimizar o critério de particionamento:
  - Ótimo Global: enumeração exaustiva de todas as partições;
  - Métodos heurísticos:
    - k-means:
      cada segmento é representado pelo **centro** do segmento (centroid);
    - k-medoids:
      cada segmento é representado por **um dos elementos** do segmento (medoid).



Método k-means

- Sendo dado 'k' (número de segmentos), seguir os 4 passos:
  1. Dividir os objetos em 'k' subconjuntos não vazios;
  2. Calcular o centro de cada segmento (centroid);
  3. Atribuir cada objeto ao centroid mais próximo;
  4. Voltar ao ponto 2.;
     parar quando não houver mais possibilidades de atribuição.



- Vantagens:
  - Relativamente eficiente:
    sendo 'n' o número de objetos, 'k' o número de segmentos e 'i' o número de iterações, normalmente acontece k,i << n;
  - Termina com ótimos locais.

- Desvantagens:
  - Aplicável, apenas, quando é possível calcular a média (*mean*);
  - É necessário identificar o número de segmentos *a priori*;
  - Incapacidade de lidar com ruído nos dados;
  - Inadequado para determinar segmentos côncavos.

Método k-medoids

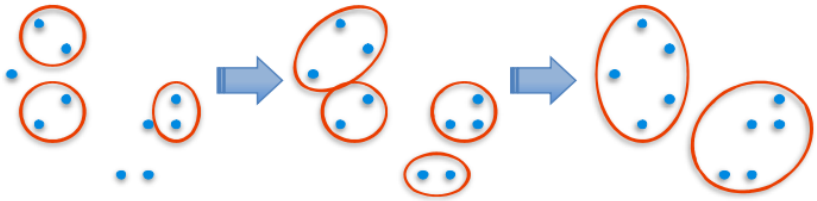
- Medoids são objetos **representativos** do conjunto de dados;
- Inicia-se com um conjunto de medoids que, iterativamente, vão sendo substituídos por outros não-medoids desde que a distância do segmento resultante seja melhorada.
- Vantagens e Desvantagens:
  - É mais robusto do que o método k-means na presença de dados ruidosos, uma vez que os objetos selecionados são menos influenciáveis por valores extremos do que a média (*mean*);
  - Produz bons resultados para conjuntos de dados de pequenas dimensões;
  - Não se comporta tão bem quando se pretende a sua aplicação em conjuntos de dados de grandes dimensões.

Algoritmos de Hierarquização

- Utilizam a matriz de distâncias como critério de segmentação;
- Os dados são agrupados em árvores de segmentos;
- Não requerem a definição do número de segmentos a procurar;
- Exigem a definição de uma condição de paragem:
  - quantidade de segmentos;
  - distância mínima entre objetos;
  - etc.
- Existem dois tipos de algoritmos de hierarquização:
  - Aglomeração: estratégia *bottom-up*;
  - Divisão: estratégia *top-down*.

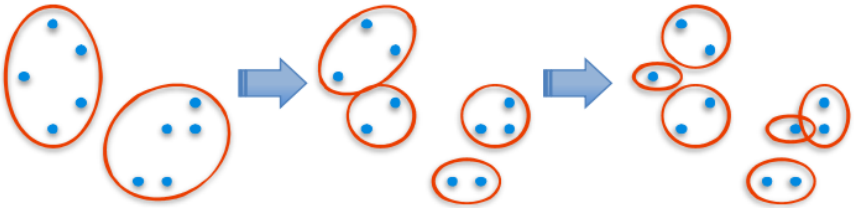
AGNES: Agglomerative Nesting

- Iterativamente, vai juntando objetos que apresentam menores valores de dissemelhança: os conjuntos C1 e C2 são juntos se os objetos de C1 e de C2 produzem o menor valor de distância Euclidiana entre quaisquer dois objetos de segmentos distintos.



DIANA: Divisive Analysis

- Iterativamente e partindo de um segmento composto por todos os objetos, dividir em segmentos menores que maximizam a distância Euclidiana entre objetos vizinhos de segmentos diferentes.



Vantagens e desvantagens

- Dificuldades com o aumento de atributos ou de objetos:
  - à medida que aumentam os objetos a agrupar, aumenta o tempo necessário para procurar tais grupos;
- Não é necessário especificar o número de segmentos 'k';
  basta "cortar" a árvore no nível 'k-1';
- Produz melhores resultados do que os algoritmos k-means;
- Uma hierarquia traduz alguma organização dos segmentos, ao contrário de um simples conjunto de segmentos.

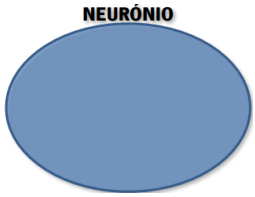
Redes Neuronais Artificiais

Uma Rede Neuronal Artificial (RNA) é um sistema computacional de base conexionista para a resolução de problemas. É concebida com base num modelo simplificado do sistema nervoso central dos seres humanos.

Uma RNA é definida por uma estrutura interligada de unidades computacionais, designadas neurónios, com capacidade de aprendizagem

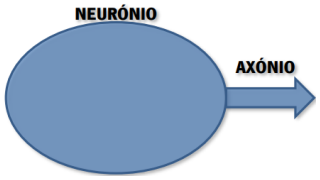
Conceitos e definições: Neurónio

- Unidade computacional de composição da RNA.
- Identificado pela sua posição na rede.
- Caracterizado pelo valor do estado.



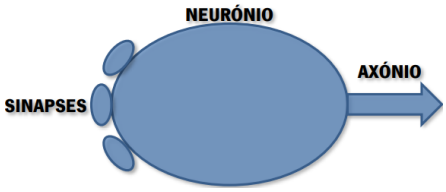
Conceitos e definições: Axónio

- Via de comunicação entre os neurónios.
- Pode ligar qualquer neurónio, incluindo o próprio.
- As ligações podem variar ao longo do tempo.
- A informação circula em um só sentido.



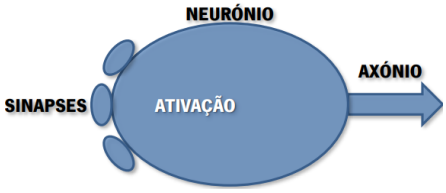
Conceitos e definições: Sinapses

- Ponto de ligação entre axónios e neurónios
- O valor da sinapse determina o peso (importância) do sinal a entrar no neurónio: excitativo, inibidor ou nulo.
- A variação no tempo determina a aprendizagem da RNA.



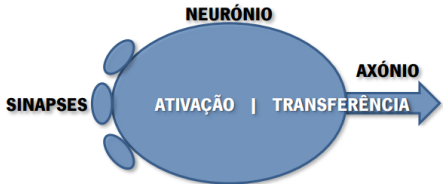
Conceitos e definições: Ativação

- O valor de ativação é representado por um único valor.
- O valor de ativação varia com o tempo.
- A gama de valores varia com o modelo adotado.



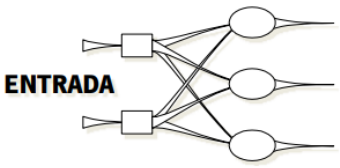
Conceitos e definições: Transferência

- O valor de transferência de um neurónio determina o valor que é colocado na saída.
- É calculado como uma função do valor de ativação.

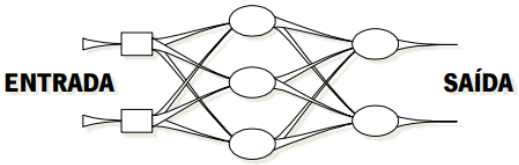


Organização dos neurónios

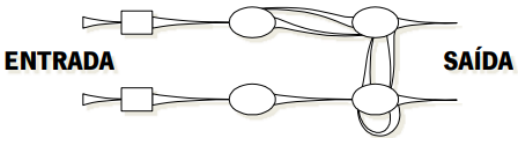
- Arquitetura *Feed forward*, de uma só camada: (*Perceptron*)



- Arquitetura *Feed forward*, multi-camada: (*Multi-layer Perceptron*)

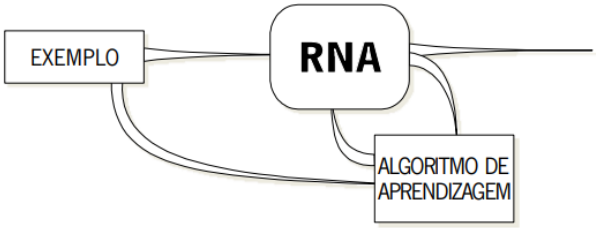


- Arquitetura Recorrente



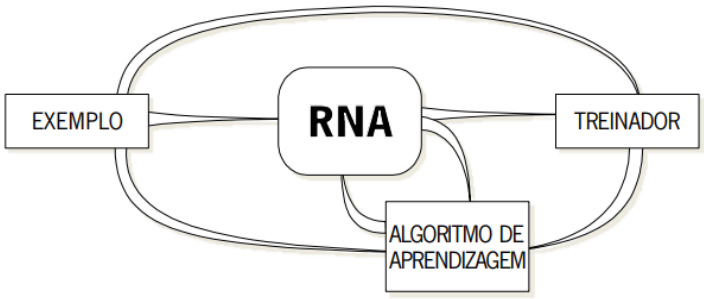
Aprendizagem

- Sem supervisão:



(p.ex., quando dois neurónios adjacentes têm variações da ativação no mesmo sentido, então o peso da ligação deve ser progressivamente aumentado.)

- Com supervisão:



(p.ex., os ajustes nos pesos das ligações são efetuados por forma a minimizar o erro produzido pelos resultados da RNA.)

- De reforço: o exemplo contém, apenas, uma indicação sobre a correção do resultado.