# Métricas em Machine Learning

**Equipa docente:**

- Rui Pereira - Departamento de Matemática - Azurém;
- Pedro Patrício - Departamento de Matemática - Gualtar;

**Principais temas da UC:**

- 1. O dados : tipos, representação, operações;
- 2. Métricas ponto a ponto;
- 3. Subset metrics; Representantes;
- 4. Principal Components Analysis (PCA);
- 5. LDA+ Kernel PCA.
- 5. Clustering: LLoyd, Hierarchical. Avaliação do clustering; Implementações.
- 6. Trabalho prático. (Temas a 6/10; Escolha: 24/11; Entrega: 2/1; Apresentação: 4/1)

# Métricas em Machine Learning

**Avaliação:**

Teste (50%) + Trabalho em grupo (50%) $\rightarrow$ grupos de 3 a 4 alunos;
(14 de dezembro)

ou

Exame final : parte teórica + parte prática

**Bibliografia:**

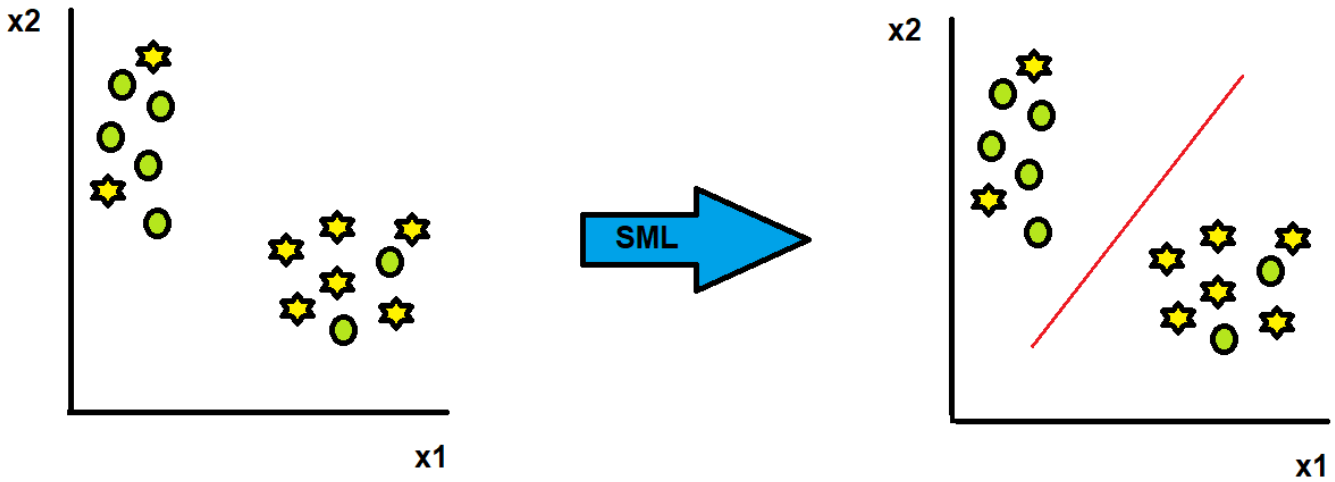
- Metric Learning, Bellet, Habrard, Sebban, 2015.
- Pattern recognition and ML, Bishop, 2006.
- Introduction to ML, Alapaydin, 2010.

## 1. Os dados

Machine Learning (ML):

É uma sub área da IA que se dedica ao estudo de algoritmos que tem por objetivo 'ensinar os computadores a aprender'.

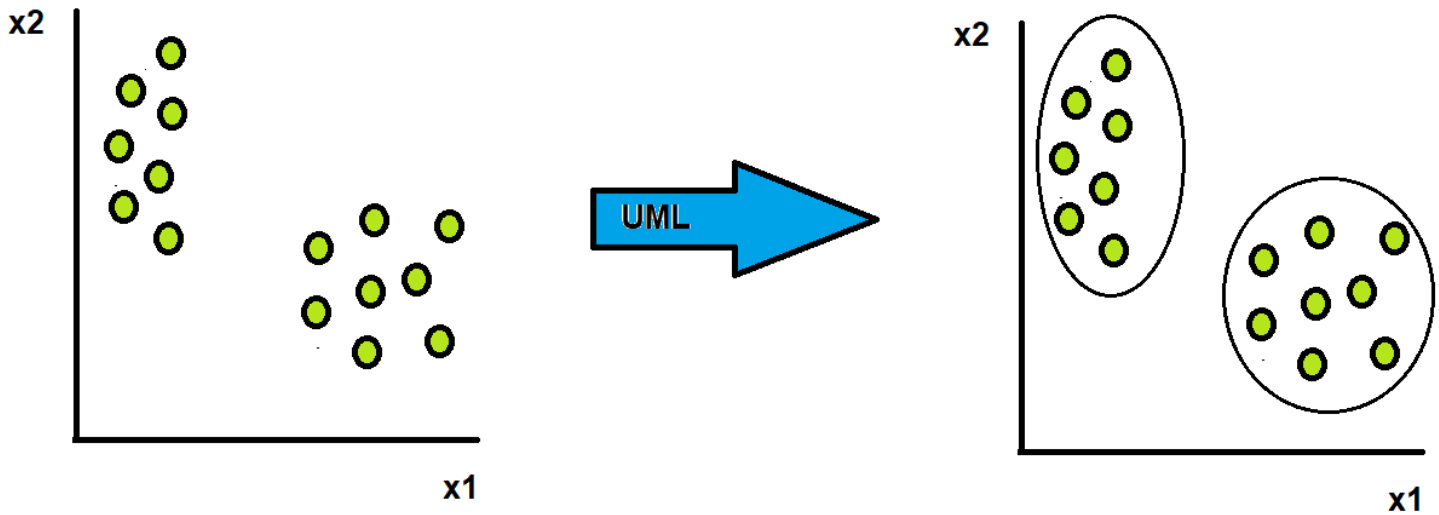
Há essencialmente 2 tipos de filosofia - Supervised ML (SML) /Unsupervised ML (UML).




**SML** : Dados rotulados (training set) $\rightarrow$ constrói função $\rightarrow$ determina valor var. dep. (classif./regressão)

# Métricas em Machine Learning

**UML:** algoritmos que têm por objetivo identificar padrões em dados não rotulados (ex: clustering); Depois de criados os clusters, e usando métricas adequadas poderão ser usados para classificação (se os dados o permitirem).




**IMPORTANTE:** Tanto em SML como em UML, são necessárias métricas. É importante usarmos as métricas adequadas.

# Métricas em Machine Learning

**Dados**:

Podem ser classificados de diferentes formas. Por exemplo: nominativos, binários, ordenados, monóide, discretos, reais, etc.

De acordo com o tipo de dados, as operações possíveis são diferentes.

**Notações**

**Def. 1:** Seja $A$ um conjunto. Será usado para definir atributos. $|A|$ é a cardinalidade do conjunto.

**Def. 2:** Sejam $A_1, A_2, \ldots, A_d$ atributos e $C_1, C_2, \ldots, C_p$ classes.
Então $\mathcal{D} = (A_1 \times A_2 \times \ldots A_d) \times (C_1 \times C_2 \times \ldots C_p) = \mathcal{A} \times \mathcal{C}$ representa o espaço admissível para os dados.

Nota: $A \times B = \{(a, b) : a \in A \wedge b \in B\}$.

Um elemento $e = (x, y) \in \mathcal{D}$ é caracterizado por:
$x = (x_1, x_2, ..., x_d)$ $x_i \in A_i$, e , $y = (y_1, y_2, ..., y_p)$ $x_j \in C_j$

**Def. 3:** Um conjunto de dados $D$ é uma lista de eventos que pode representar uma BD.

$$D = \{e^n = (x^n, y^n),\ n = 1, ...N\},$$

onde $N$ é o número de eventos, e $x_i^n \in A_i$ representa o valor do atributo $i$ do evento $n$ e $y_j^n \in C_j$ representa a classe j do evento $n$.

Nota: $D$ é um conjunto com $N$ eventos, os quais são pares ordenados $(x^n, y^n)$, onde $x^n \in \mathcal{A}$ e $y^n \in \mathcal{C}$.

**Exemplo 1:** Seja $A_1 = \mathbb{N}_0$, $A_2 = \{azul, vermelho, laranja\}$, $A_3 = \{1, 2, 3\}$, $C_1 = \{yes, no\}$ e $C_2 = \mathbb{R}$. Considere ainda a vase de dados $D$ com os seguintes eventos:

| n | A1 | A2 | A3 | C1 | C2 |
|---|---|---|---|---|---|
| 1 | 1 | azul | 1 | yes | 1.32 |
| 2 | 17 | vermelho | 3 | no | 4.17 |
| 3 | 0 | laranja | 3 | yes | 2.22 |
| 4 | 4 | laranja | 2 | yes | 7.44 |

Represente o conjunto $D$ em extensão.

## 1.1 Dados nominativos ou categóricos

Constituem uma lista finita de objetos sem regras particulares entre eles. A **única** operação que podemos fazer é dizer se 2 destes dados são iguais ou diferentes.

Se $A$ é um conjunto de dados categóricos. Para 2 elementos $x, x' \in A$ temos que $x = x'$ ou $x \neq x'$.

Seja agora $A = \{a_1, a_2, ..., a_I\}$ é um conjunto com $I$ dados categóricos. Seja $D = \{x^1, x^2, ..., x^N\}$ um conjunto, onde $x^n \in A$.

Podemos definir as frequências absolutas e relativas, a saber,

$N_i = |\{x \in D : x = a_i\}|$ representa o número de ocorrências do atributo $a_i$ em $D$

$f_i = \frac{N_i}{N} \quad i = 1, ...I$ representa o rácio entre o número de ocorrências do atributo $a_i$ em D e o número de elementos de $D$.

**Exemplo 2:**

Considere que $A = \{banana, gato, rato, caixa, caneta\}$ define um atributo $A$. Considere ainda a base de dados
$D = \{banana, gato, rato, banana, banana, caixa, caneta, caixa, caneta, gato\}$ onde os seus eventos são objetos de $A$.

Temos que $N = |D| = 10$. $N_1 = 3$, $N_2 = 2$, $N_3 = 1$, $N_4 = 2$ e $N_5 = 2$.
Pelo que $f_1 = 0.3$, $f_2 = 0.2$, $f_3 = 0.1$, $f_4 = 0.2$ e $f_5 = 0.2$.

## 1.2 Dados binários

- $A = \{0, 1\}$.
- Complementaridade, $x \in A \wedge \neg x \in A$, $\{x\} U \{\neg x\} = A$, $x \neq \neg x$.
- Pode-se aplicar lógica binária: $x \wedge x'$, $x \vee x'$, $x \oplus x'$, ...
- Uma vez que $A = \{0, 1\}$ também pode ser considerado dados categóricos, para $D = \{x^1, x^2, ..., x^N\}$ com $x^i \in A$, podemos definir,
  $N_0 = |\{x \in D : x = 0\}|$, $N_1 = |\{x \in D : x = 1\}|$, $f_0 = \frac{N_0}{N}$ e $f_1 = \frac{N_1}{N}$.
- Se tivermos 2 conjuntos $D$ e $D'$ com $N$ elementos, respectivamente $x \in A$ e $x' \in A$, podemos definir,
  $M00 = |\{x^n = 0 \wedge x'^n = 0\}|$ com $n = 1, ..N$
  $M01 = |\{x^n = 0 \wedge x'^n = 1\}|$ com $n = 1, ..N$
  $M10 = |\{x^n = 1 \wedge x'^n = 0\}|$ com $n = 1, ..N$
  $M11 = |\{x^n = 1 \wedge x'^n = 1\}|$ com $n = 1, ..N$

# Métricas em Machine Learning

Se o conjunto $D$ representar os dados reais e $D'$ o resultado duma classificação dos dados, temos aqui os ingredientes duma **Tabela de Confusão**.

Uma das avaliações é $Accuracy = \frac{M00+M11}{M00+M01+M10+M11}$.

| D \ D' | 0 | 1 |
|---|---|---|
| 0 | M00 | M01 |
| 1 | M10 | M11 |

Binarização de dados: Por vezes podemos binarizar os dados, mas, temos que ter o cuidado de usar **princípio de exclusão**.

**Exemplo 3:** $A = \{red, blue, green\}$. Se $D = \{x^1, x^2, ..., x^N\}$ onde $x^i \in A$ com $i = 1, ..., N$.
Cada evento de $D$ pode ser red ou blue ou green. Se for red, não é blue ou green.

# Métricas em Machine Learning

Binarizar os dados, significa representar os mesmos dados duma forma 'binária'. Assim, podemos considerar $A_1 = \{0, 1\}$, $A_2 = \{0, 1\}$, $A_3 = \{0, 1\}$.

Se $x = \textit{blue} \rightarrow x' = (1, 0, 0)$

Se $x = \textit{green} \rightarrow x' = (0, 1, 0)$

Se $x = \textit{red} \rightarrow x' = (0, 0, 1)$

**NOTA:** $x \in A \rightarrow x' \in A_1 \times A_2 \times A_3$ + **P. exclusão**

Ou seja, não podemos ter por exemplo $x' = (1, 1, 1)$.

## 1.3 Dados Ordenados.

Vamos agora considerar $A = \{a_1, a_2, a_3, ...., A_l\}$ um conjunto ordenado. Ou seja, existe um operador $\leq$ que verifica as propriedades:

i) $\forall x \in A\ , x \leq x$;

ii) $\forall x, x' \in A\ [x \leq x' \wedge x' \leq x \rightarrow x = x']$.

iii) $\forall x, x', x'' \in A\ [x \leq x' \wedge x' \leq x'' \rightarrow x \leq x'']$.

iv) $\forall x, x' \in A\ [x \leq x' \vee x' \leq x]$.

Se A fôr conjunto ordenado então:

- não podemos quantificar diferença - não faz sentido $grande - pequeno$.
- não podemos aritemizar dados - $grande - pequeno = medio$?
- podemos definir operações MAX e MIN entre 2 objetos de $A$.
- podemos definir estatísticas a partir dum conjunto de dados $A$.

# Métricas em Machine Learning

Seja $D = \{x^1, x^2, x^3, ..., x^n\}$ onde $x^i \in A$. Podemos definir,

$N_i = |\{x \in D : x = a^i\}|$ com $i = 1, ..., N$.

$P_i = \sum_{j \leq i} N_j$ que representa a frequência absoluta acumulada.

para $\alpha \in [0, 1]$ podemos definir $E_\alpha = floor(\alpha.N)$ tal que,

$E_\alpha \leq \alpha.N \leq E_\alpha + 1$

**Nota:** O modo $a \in A$ associado a $\alpha$ é o valor de $a_i \in A$ tal que $P_{i-1} < E_\alpha \leq P_i$ (com $P_0 = 0$). A mediana corresponde a $\alpha = 1/2$.

**Exemplo 4:** Seja $A = \{small, medium, large, huge\}$.
$D = \{large, large, medium, huge, small, large, small, medium, large, small, medium, medium\}$. Qual a mediana?

$N_1 = 3$ , $N_2 = 4$ , $N_3 = 4$ , $N_4 = 1$;
A mediana corresponde ao valor $a_i \in A$ tal que $P_{i-1} < floor(0.5 \times 12) \leq P_i$.

# Métricas em Machine Learning

$P_0 = 0,\ P_1 = 3,\ P_2 = 7,\ P_3 = 11,\ P_4 = 12.$

Se $i = 1 \rightarrow P_0 < floor(0.5 \times 12) \leq P_1$ (falso).

Se $i = 2 \rightarrow P_1 < floor(0.5 \times 12) \leq P_2$ (verdadeiro) $\rightarrow$ mediana$=a_2 = medium$.

### 1.4 Dados tipo monóide - string.

Seja $\varepsilon = \{\epsilon, a, b, c, ..., z\}$ a que chamamos alfabeto. Um monóide livre é $\varepsilon^*$ constituido por todas as listas finitas e ordenadas $l$, baseadas em $\varepsilon$ com a operação de concatenação $+$, tal que,

i) $\forall l \in \varepsilon^*,\ l + \epsilon = \epsilon + l = l$, onde $\epsilon$ é o elemento neutro.

ii) $\forall l, l', l'' \in \varepsilon^*,\ (l + l') + l'' = l + (l' + l'')$.

Um monóide livre $\varepsilon^*$ corresponde a todas as palavras que se conseguem definir com as letras de $\varepsilon$. $|l|$ representa o número de elementos da lista $l$. $l_k$ corresponde à letra $k$ da lista $l$.

Podem-se definir operações tais como:

Eliminação: $E(l, k)$, elimina a letra $k$ da lista $l$.

Inserção: $I(l, k, \alpha$, que insere o elemento $\alpha$ na posição $k$ da lista $l$.

Outras como Substituição, permutação, poderiam ser definidas...

# Métricas em Machine Learning

O português usa monóides de monóides, onde,
$\varepsilon = \{\epsilon, a, b, c, ..., z\}$ é o alfabeto, *palavras* $\subset \epsilon^*$, $S = \{palavras, \epsilon, \ \}$, $S^*$ poderá ser todos os textos que se podem escrever.

**1.5 Números discretos.**

Por exemplo os números inteiros $\mathbb{Z}$.

Propriedades: ordenados, operçẽs $+$, $-$, $\times$, divisão euclidiana %.

Proposição: Sejam $a, b \in A$ com $b > 0$. Existe $q \in A$ com $0 \leq r < b$ tal que $a = a.q + r$. $q$ é o quciente da divisão e $r$ é o seu resto.

Como temos atributos ordenados, o representante dum conjunto cujos elementos são inteiros poderá ser a mediana.

Como temos divisão Euclidiana o representante dum conjunto cujos elementos são inteiros poderá ser a média.

**Exemplo 5:** Considere $A = \mathbb{Z}$, $D = \{7, 8, 3, 4, 1, 9, 7, 0\}$. Calcule os representantes de $D$.

$S = 7 + 8 + 3 + 4 + 1 + 9 + 7 + 0 = 39, \rightarrow$ 39 % 8 tem quociente $q = 4$ e resto $r = 7$. Se Representante fôr média, então é 4.

Ordenando $D$, temos $DO = \{0, 1, 3, 4, 7, 7, 8, 9\}$. Se pretendemos a mediana,

callculamos $\frac{N}{2} = 4$. Se $N$ é ímpar escolhemos elemento do meio. Senão o utilizador pode definir critério, por exemplo o elemento $ceil(\frac{N}{2})$ ou $floor(\frac{N}{2})$ ou mesmo o elemento que corresponde média entre este 2 últimos.

### 1.6 Números Reais e seus derivados.

$A = \mathbb{R}$ ou afins. É o tipo mais usado em ML:

Temos a noção de continuidade entre os valores.

Propriedades: aritemética $(+, -, \times, \%)$, logo podemos calcular média; ordenado logo podemos calcular mediana, quartis, etc.

**Exemplo 6:** Vetores de $\mathbb{R}^2$.

Podemos por exemplo definir $A_i = \mathbb{R}$ , $i = 1, 2 \ \rightarrow \mathcal{A} = \mathbb{R}^2$.

Aritemética: vetores $(+, -)$ como soma e subtração de vetores. para os valores (componentes) $(\times, \%)$ a multiplicação e divisão reais.

Não há ordem natural em $\mathbb{R}^2$. Poderá ser introduzida usando preferências (regras a definir) .

Muitas outros exemplos poderiam ser considerados.