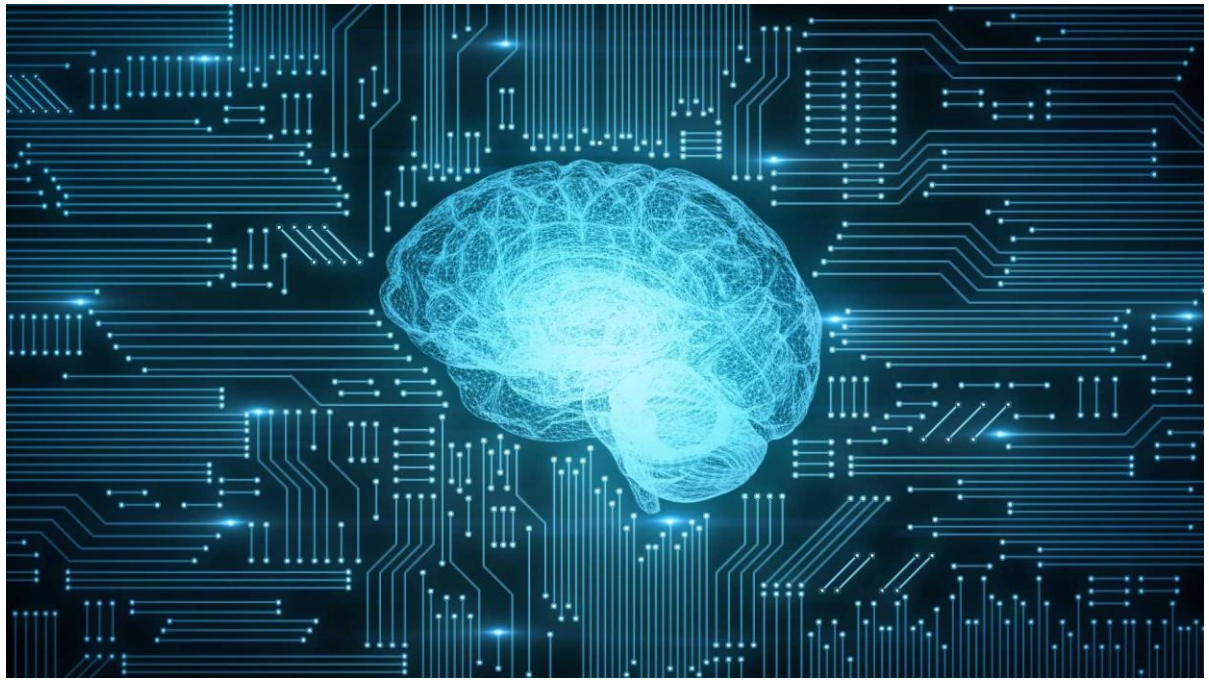
FGV EESP
ESCOLA DE ECONOMIA DE SÃO PAULO

INTELIGÊNCIA COMPUTACIONAL



PROF. Dr. EDSON C. KITANI

2022

Sumário da Aula

- Distâncias (Duda et al. Cap. 4, section 4.6)
- Manifold Learning (John Lee & Michel Verleysen – Nonlinear Dimensionality Reduction)
- Árvores de Decisão (Duda et al. Cap. 8, section 8.1)

Medidas de Distâncias

Métricas de Distâncias

Vamos primeiro fazer uma definição básica sobre o significado matemático da função distância. Seja X um conjunto, na qual temos uma aplicação $d: X \times X \rightarrow \mathbb{R}^{+}$. Chamamos de distância em X se, para todo $x, y \in X$, as seguintes propriedades são respeitadas:

I. $d(x, y) \geq 0$, ou seja nenhuma distância é negativa
II. $d(x, y) = d(y, x)$, a distância é simétrica
III. $d(x, x) = 0$, para o mesmo ponto ou se $x = y$, que é a propriedade da reflexividade ou similaridade perfeita
IV. $d(x, y) + d(y, z) \geq d(x, z)$, conceito da desigualdade triangular

Observe que sobre o conceito de função de distância pode trazer a interpretação de algo ser "verdadeiro" se a distância é zero e "falso" se for infinito. Outro aspecto importante é que as distâncias métricas são consideradas como retas que unem dois pontos, assim ele tem significado nos espaços vetoriais, e é conhecido também como "norma" de um vetor, $d(x, y) = \|x - y\|$.
Portanto, medir a distância de maneira métrica ou não métrica em IC é determinar a dissimilaridades envolvidas entre as amostras.

Distância Euclidiana

A primeira métrica de distância, e a mais conhecida e utilizada, é a **distância Euclidiana**, também conhecida como **norma Euclidiana** ou L_2. Fundamentalmente, ela se baseia na distância numérica de uma reta que une dois pontos no espaço $\mathbb{R}^n$.

A distância entre dois pontos a e b:

Em $\mathbb{R}^1$ é definido como: $d(a,b) = |a-b| = \sqrt{(a-b)^2}$.

Em $\mathbb{R}^2$ é definido como: $d(a,b) = \|a-b\| = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2}$.

Em $\mathbb{R}^3$ é definido como: $d(a,b) = \|a-b\| = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2 + (a_3-b_3)^2}$.

Em $\mathbb{R}^n$ é definido como: $d(a,b) = \|a-b\| = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2 +, \ldots . + (a_n-b_n)^2}$.

Como um caso geral podemos escrever: $d(a,b) = \|a-b\| = \sqrt{\sum_{i=1}^{n}(a_i-b_i)^2}$ onde n é a dimensionalidade.
Essa métrica dá mais ênfase às *features* com maior dissimilaridade.

Distância de Manhattan

A distância de Manhattan, também conhecida como City Block ou norma L_1, é semelhante a distância Euclidiana. A distância de Manhattan tem a característica de fornecer a medida do comprimento dos degraus que levam um ponto a até o ponto b em movimentos ortogonais aos pontos dados. É definido como:

$$d(a,b) = \|a-b\|_1 = \sum_{i=1}^{n} |a_i - b_i|$$

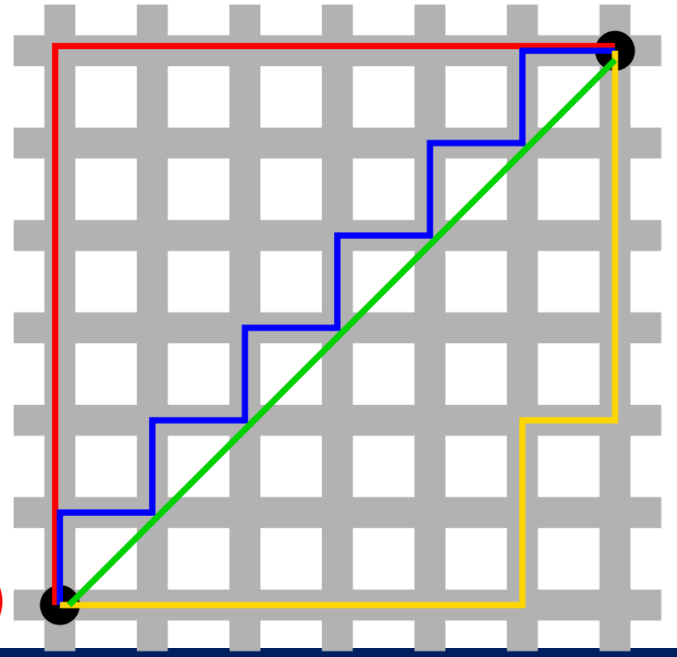
Tal que $a = (a_1, \ldots, a_n)$ e $b = (b_1, \ldots, b_n)$

No exemplo ao lado suponha que as coordenadas dos pontos a e b são respectivamente (0, 0) e (6, 6). Logo:

$L1 = |6-0| + |6-0| = 12$ (Trajetos vermelho, amarelo e azul)

$L2 = \sqrt{(6-0)^2 + (6-0)^2} = 6\sqrt{2}$ (Trajeto verde)

Esta distância tem aplicação na área de visão computacional, planejamento de rotas de robôs e etc. Entretanto, tem uma importância grande na área de Data Science devido ao fenômeno do "*encolhimento*" dos espaços em altas dimensões vetoriais.



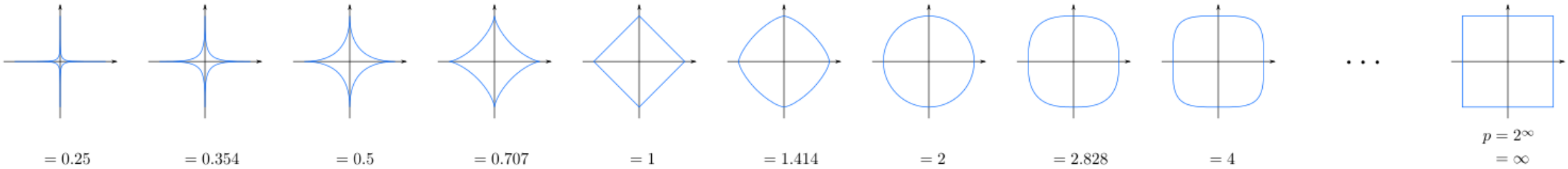
Distância de Minkowski

A distância de Minkowski é considerada uma generalização da distância Euclidiana, porque dependendo do valor do parâmetro p podemos ter a norma $L_1, L_2, \ldots L_p$.

$$d(a,b) = \|a-b\|_p = \left(\sum_{i=1}^{n} |a_i - b_i|^p\right)^{1/p}$$

Outra derivação importante da distância de Minkowski é conhecida como distâncias fracionárias. Nesse caso o que muda é o parâmetro $0 < p < 1$. Acontece que quando $0 < p < 1$, a distância deixa de ser métrica porque viola o princípio da desigualdade triangular $d(x,y) + d(y,z) \geq d(x,z)$.

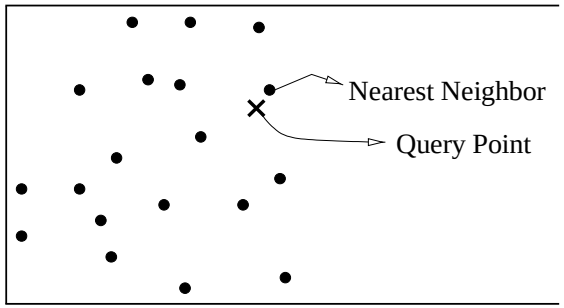
A figura ilustra exemplos da distância entre dois pontos $a = (0,0)$ e $b = (1,1)$ para alguns valores de p.

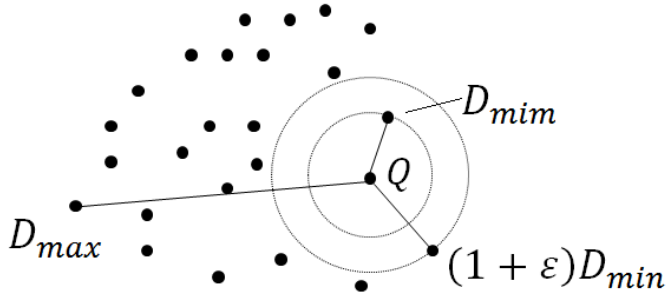


Praga da Dimensionalidade

Simulações conduzidas pro Beyer, Goldtein, et al. 1999, indicaram que a diferença entre as distâncias de D_{min} e D_{max} diminuem rapidamente nas primeiras 20 dimensões. O teorema de Beyer et al*. postula então que:

"A nearest neighbor query is unstable for a given ε if the distance from the query point to most data points is less than $(1+\varepsilon)$times the distance from the query point to its nearest neighbor."





*When Is "Nearest Neighbor" Meaningful? Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft

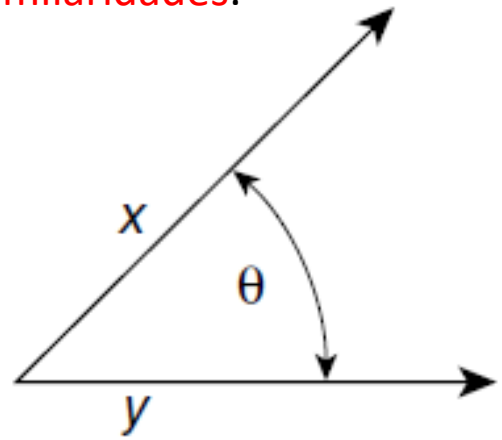
Se poucos pontos caírem dentro da esfera alargada $(1+\varepsilon)D_{min} \mid \varepsilon > 0$, significa que o ponto mais próximo do ponto de consulta está separado dos outros pontos de uma maneira significativa. Contudo, se muitos pontos caírem dentro região alargada por ε pode tornar o conceito de vizinho mais próximo sem sentido se ε for muito pequeno. Isso cria uma instabilidade em algoritmos do tipo K-NN que usa distâncias como métrica de classificação e espaços de alta dimensionalidade.

Distância Cosseno

A distância Cosseno é uma medida vetorial baseada no ângulo θ formado entre dois vetores $x\ e\ y\ \in \mathbb{R}^n$, e nada mais é do que a razão entre o produto interno dos vetores $(x.y)$ pelo produto dos seus módulos.

$$d_{cos}(x,y) = \frac{x.y^T}{\|x\|\|y\|}$$

O resultado será o cosseno do ângulo θ formado pelos vetores $x\ e\ y$. Como normalizamos o produto interno pelos comprimentos de cada vetor, então teremos valores entre 0 e 1, e isso indica que a distância de cosseno não considera a magnitude das amostras para avaliar as dissimilaridades.



Medida de Similaridade de Jaccard

A medida de similaridade de Jaccard é na verdade uma métrica usada para medir dissimilaridades baseado na Teoria de Conjuntos. Formalmente é definida como:

$$s_J(A,B) = 1 - J(A,B)$$

Onde $J(A,B)$ é a medida do coeficiente de Jaccard. O coeficiente de Jaccard é a medida de similaridade entre dois conjuntos finitos de amostras, e é a razão entre os valores da intersecção das amostras pela união das amostras.

$$J(A,B) = \frac{(A \cap B)}{(A \uplus B)} = \frac{(A \cap B)}{(A) + (B) - (A \cap B)}$$

Desta forma, a distância de Jaccard pode ser escrita como:

$$s_J(A,B) = 1 - \frac{(A \cap B)}{(A) + (B) - (A \cap B)}$$

Assim, a dissimilaridade entre duas amostras será maior quando $s_J \to 1$.

Medida de Similaridade de Tanimoto

A medida de similaridade de Tanimoto é bastante semelhante à distância de Jaccard, e é bastante adequado para atributos que representam a presença ou ausência de características. Formalmente, Tanimoto (1958) definiu o coeficiente de similaridade como:

$$s_T = \frac{(A \cap B)}{(A \cup B)} = \frac{A \wedge B}{A + B - A \wedge B}$$

O conceito discreto também pode ser convertido para contínuo como um coeficiente de Tanimoto:

$$c_T(X, Y) = \frac{X.Y}{\|X\|^2 + \|Y\|^2 - X.Y}$$

Que é a razão do produto interno dos vetores pela soma das normas menos o produto interno.

Distância de Hamming

A distância de Hamming é uma técnica usada para avaliar as similaridades entre duas informações digitais transmitidas por algum meio de comunicação. Ele tem a função de detectar erros na transmissão, sendo 0 sem erro e 1 completamente diferente

$$d_H = \frac{1}{n}\sum_{i=1}^{n}\begin{Bmatrix}1\ se\ x_i = y_i \\ 0 = se\ x_i \neq y_i\end{Bmatrix}$$

Distância de Mahalanobis

A distância Euclidiana dá igual importância para todas as direções, e casos duas classes tenham a mesma distância teremos um distribuição em esfera ou hiper esfera. Contudo, as vezes é melhor termos as distâncias definidas pelas estatísticas das classes, ou seja pela média e variância, tornando-a invariante à escala.
Assim, a distância de um ponto de consulta X para uma dada média m_j é definida por:

$$d(X, m_j) = \left((X - m_j)^T \Sigma^{-1} (X - m_j) \right)^{1/2}$$

E para a distância de Mahalanobis entre dois vetores $a\ e\ b$ da mesma distribuição:

$$d_{Mh}(a, b) = \left((a - b)^T \Sigma^{-1} (a - b) \right)^{1/2}$$

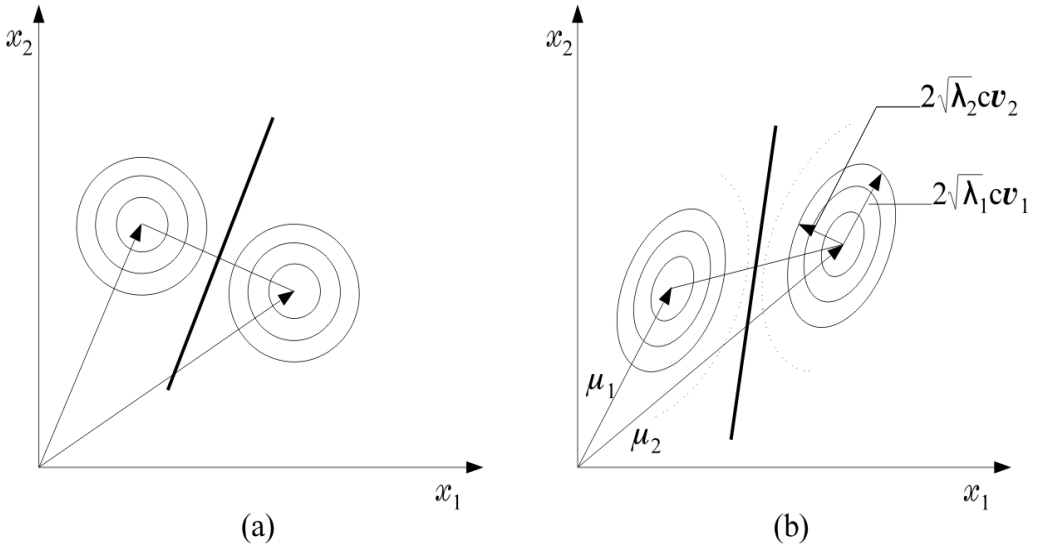
O problema é sempre determinar a verdadeira matriz de covariância do espaço de dados. Normalmente usamos a estimativa da covariância baseada nos dados disponíveis.

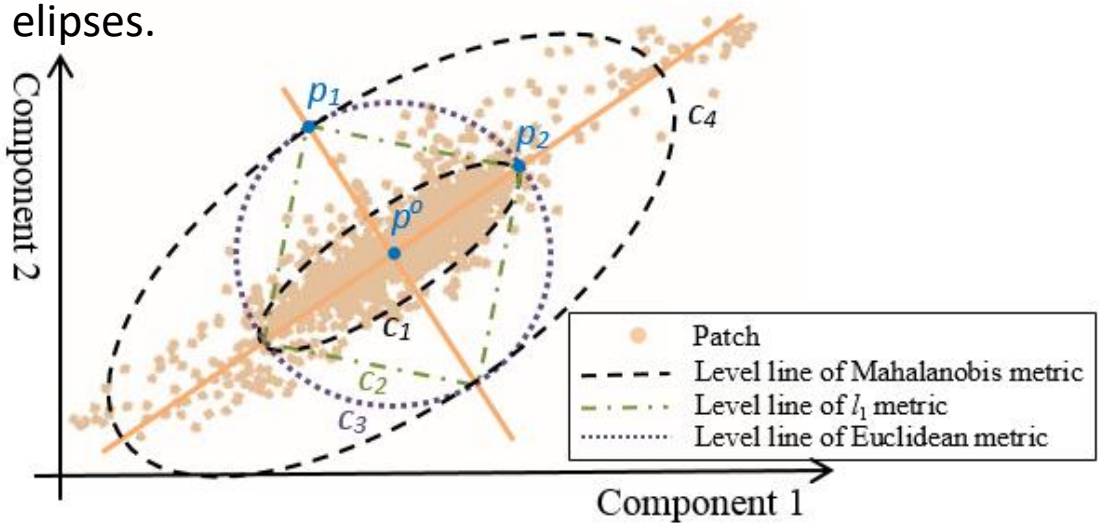
Distância de Mahalanobis

Como a matriz de covariância é simétrica, podemos decompô-la em autovetores e autovalores tal que: $\Sigma = \Phi\Lambda\Phi^T$ onde $\Phi^T = \Phi^{-1}$ e Λ é a matriz diagonal com os autovalores de Σ, e claro, considerando que para duas classes a covariância é igual. Assim, podemos projetar X numa nova base vetorial sem a covariância original entre as características.

$$(d_{Mh})^2 = \left(X - m_j\right)^T \Phi\Lambda^{-1}\Phi^T\left(X - m_j\right)$$

Faça $X^{'} = \Phi^T X$, assim as coordenadas de $X^{'}$ serão iguais a $\varphi_k^T x,\ k = 1, 2, \ldots, n$, o que implica na projeção dos pontos de x num novo sistema de coordenadas cujos eixos são determinados pelos autovetores φ_k. Assim, todos os pontos tem a mesma distância de Mahalanobis e estarão localizados dentro das elipses.





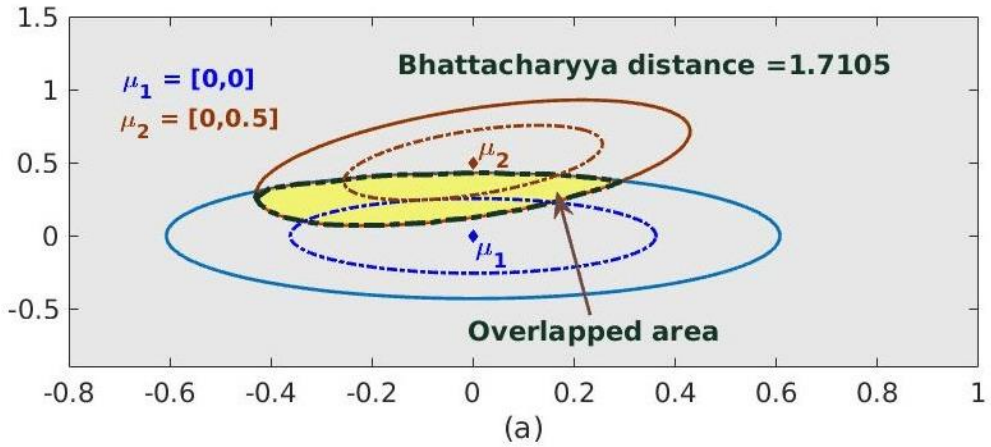
https://www.researchgate.net/figure/Diagrammatic-explanation-of-why-Bhattacharyya-distance-is-not-a-perfect-metric-to-model_fig1_326008135

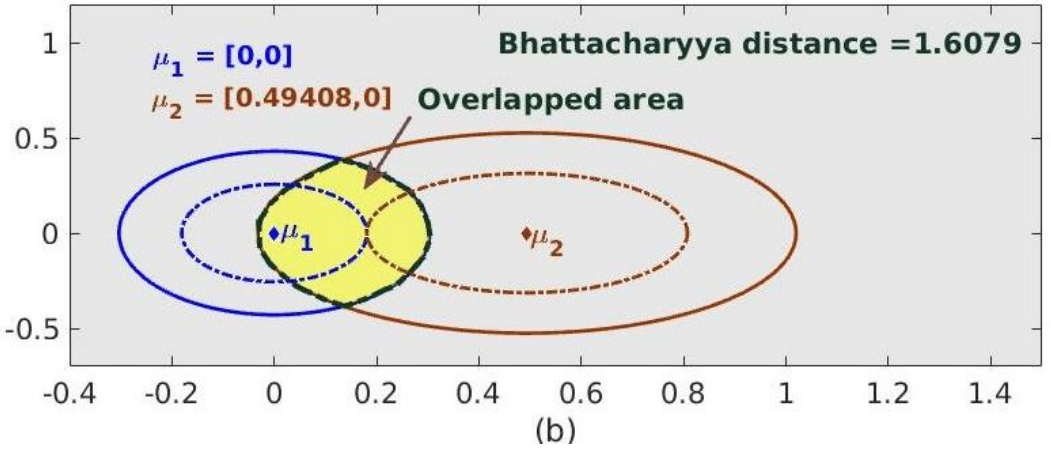
Distância de Bhattacharyya

A distância de Bhattacharyya é usado para determinar uma medida de separabilidade entre classes. O Mahalanobis é uma derivação do Bhattacharyya, pois considera a distância entre as variância de duas classes iguais.

$$m_{bat} = \frac{1}{8}(M_2 - M_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (M_2 - M_1) + \frac{1}{2} ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1|}\sqrt{|\Sigma_2|}}$$

Observe que o 1ª ou o 2ª termo desaparecem quando $M_1 = M_2$ ou $\Sigma_1 = \Sigma_2$. O primeiro termo fornece a separabilidade das classes por causa da diferença entre as médias e o segundo termo por causa da diferença na covariância.

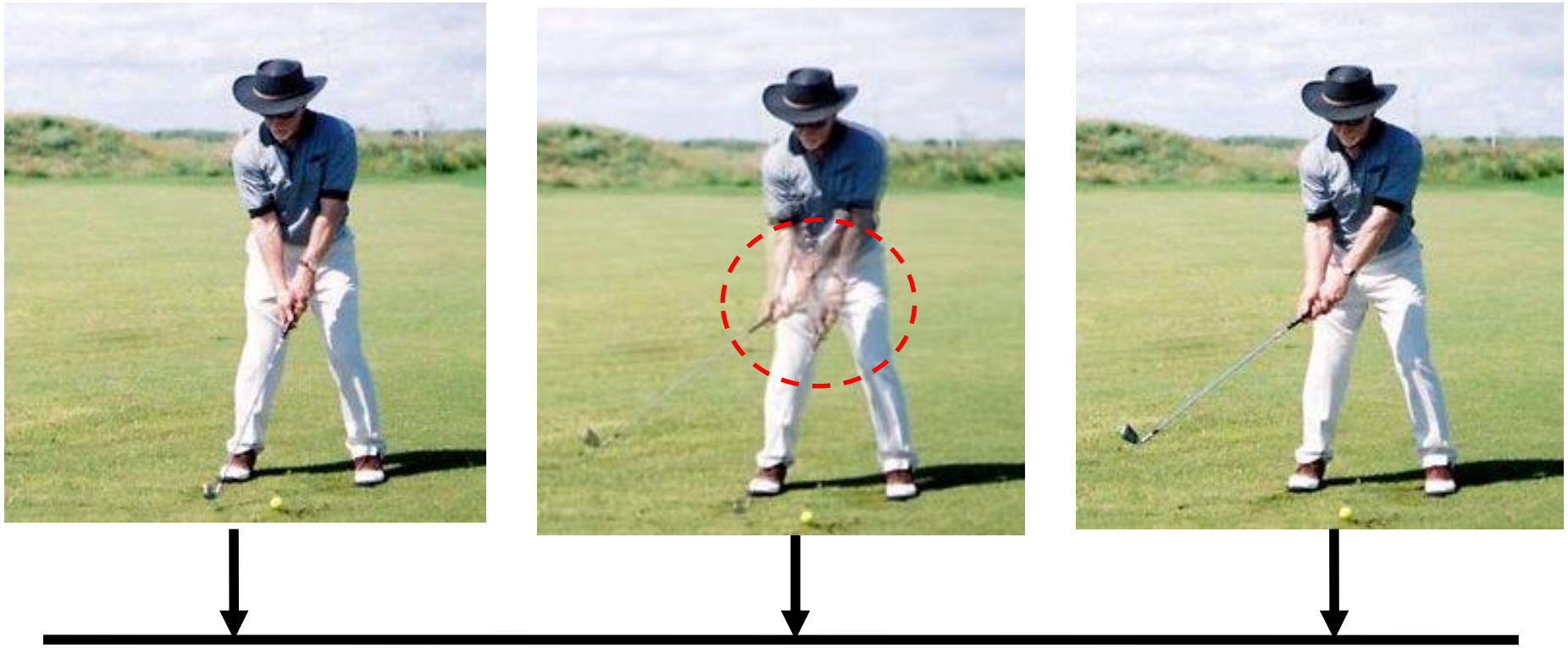




Manifold Learning

Manifold Learning

Imagine que você tem uma situação em que não há informações sobre o quadro do meio.

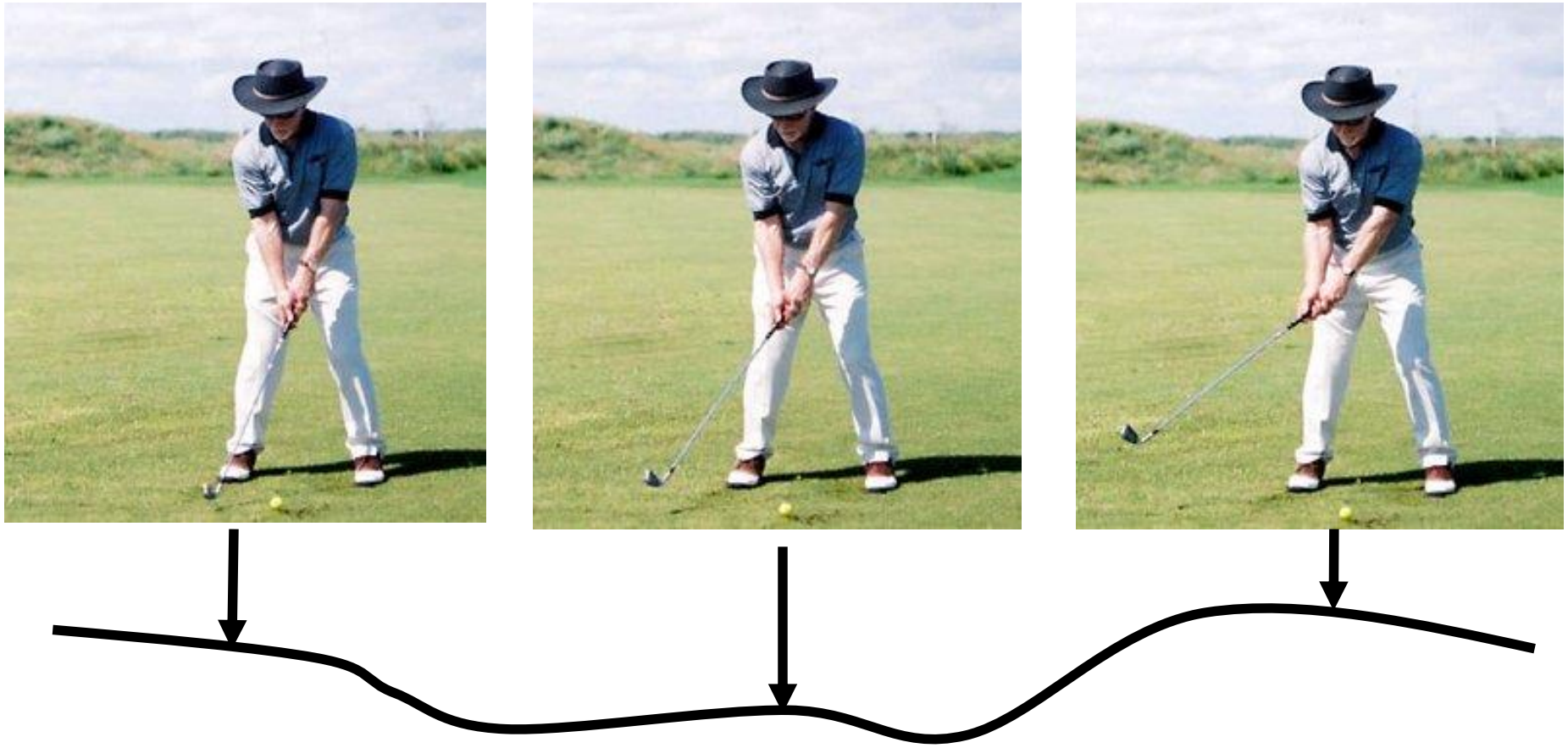


Interpolação linear

THOMPSON, David, disponível em
http://www.cs.cmu.edu/~efros/courses/AP06/presentations/ThompsonDimensionalityReduction.pdf

Manifold Learning

Nesta situação, a interpolação não linear é sempre preferível do que a interpolação linear.

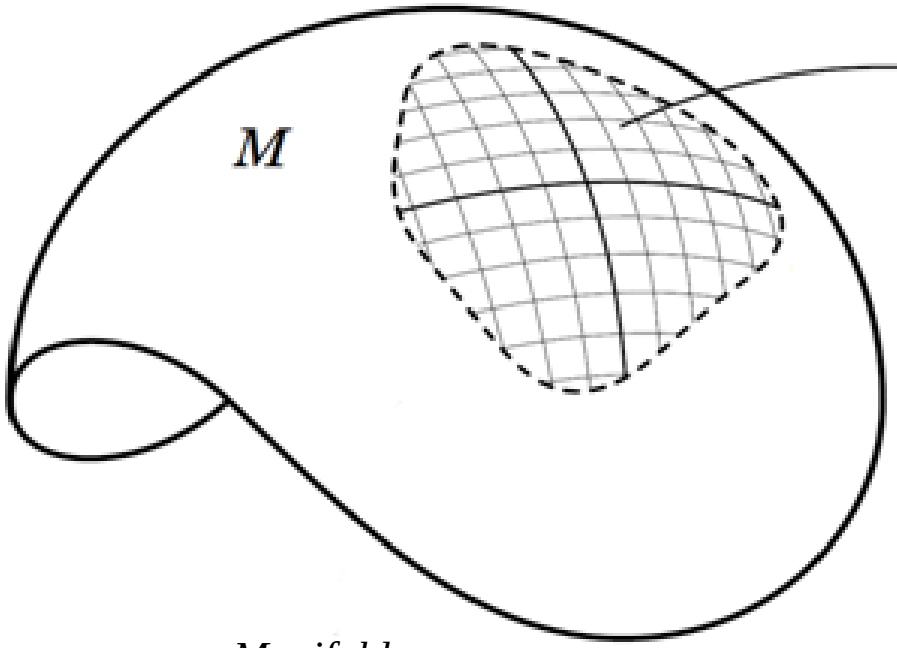


Interpolação por Manifold

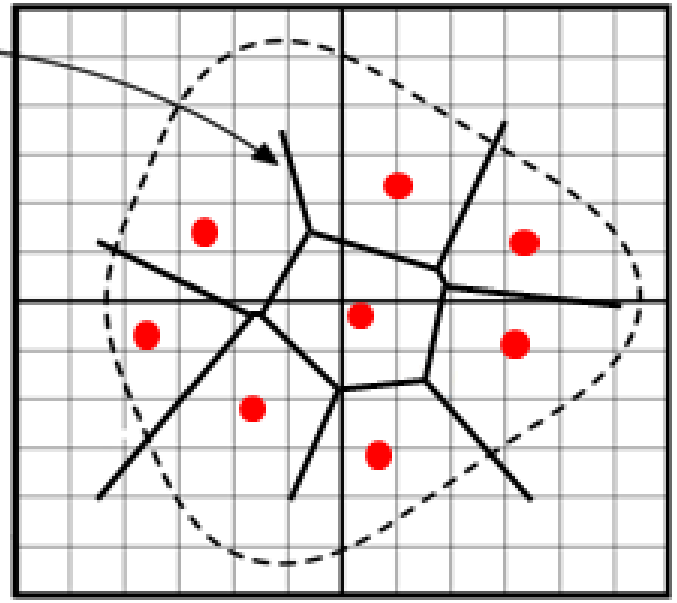
THOMPSON, David, disponível em
http://www.cs.cmu.edu/~efros/courses/AP06/presentations/ThompsonDimensionalityReduction.pdf

Manifold Learning

Os *manifolds* ou variedades geométricas, são espaços topológicos onde localmente as propriedades Euclidianas são preservadas, e conhecer esses espaço ajuda a entender a distribuição dos dados em $\mathbb{R}^n$.



Manifold



Regiões de Voronoi V_j definidos pelos círculos vermelhos.

Árvores de Decisão

Árvores de Decisão - Conceito

O aprendizado por Árvore de Decisão é uma técnica largamente utilizada, dada a sua simplicidade e robustez. É um método baseado na inferência indutiva*, e serve tanto para classificação quanto para aproximação de funções (regressão). As AD são técnicas desenvolvidas pelos grupos que trabalham com Mineração de Dados.

Um dos grandes benefícios das ADs é que elas podem trabalhar com informações categóricas e sistemas não métricos e, principalmente, não exige normalização dos dados, pelo fato das características serem avaliadas independentemente.

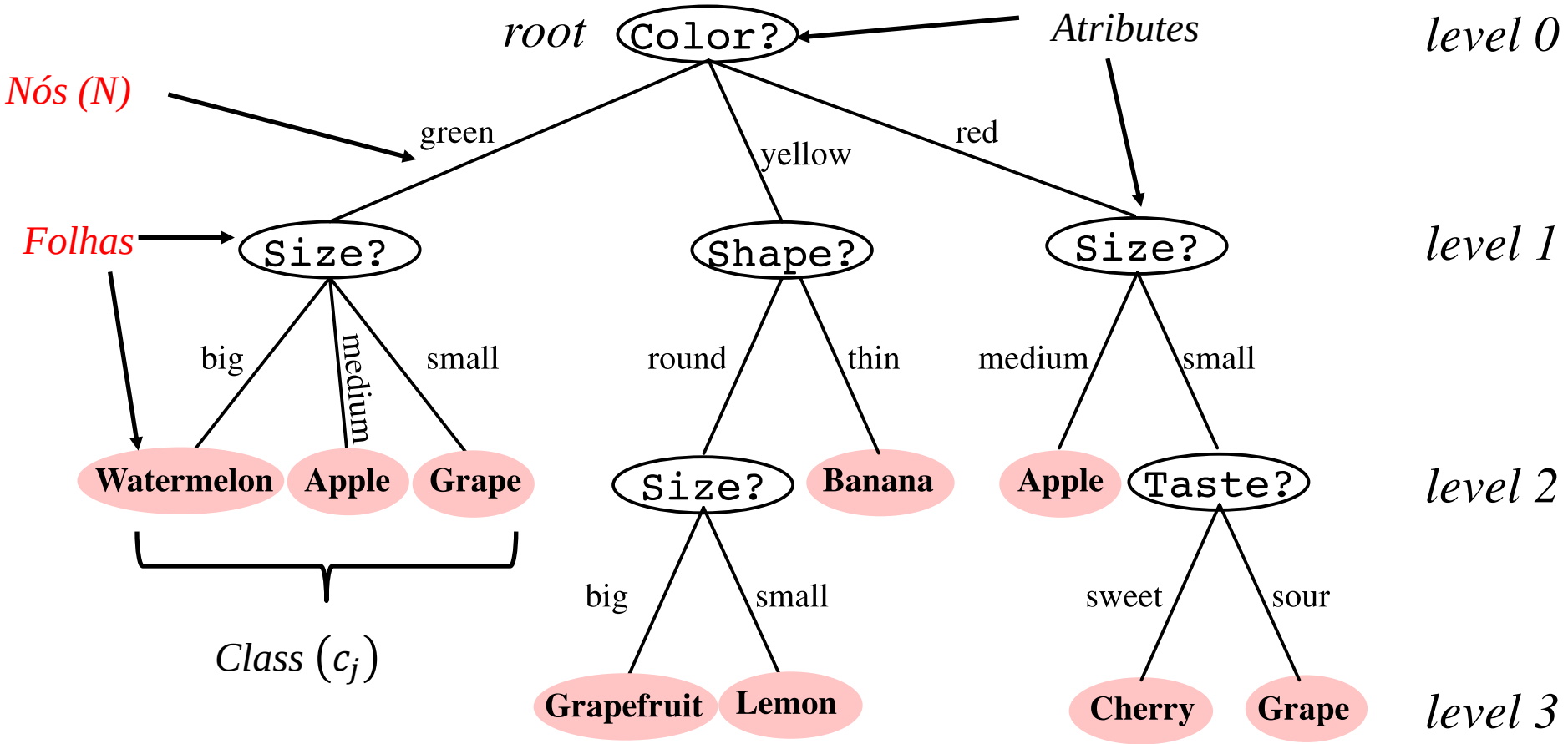
Atualmente, os principais algoritmos de AD encontrados nos pacotes são: CART (Classification And Regression Tree), ID-3 (*Iterative Dichotomiser* 3), C4.5 (é uma extensão do ID-3).

Suponha que você precisar criar um modelo para classificar frutas. Quais atributos você utilizaria para realizar a classificação?

Frutas = {Melão, Maçã, Uva, Laranja, Limão, Banana, Cereja}

Árvores de Decisão – Abstração

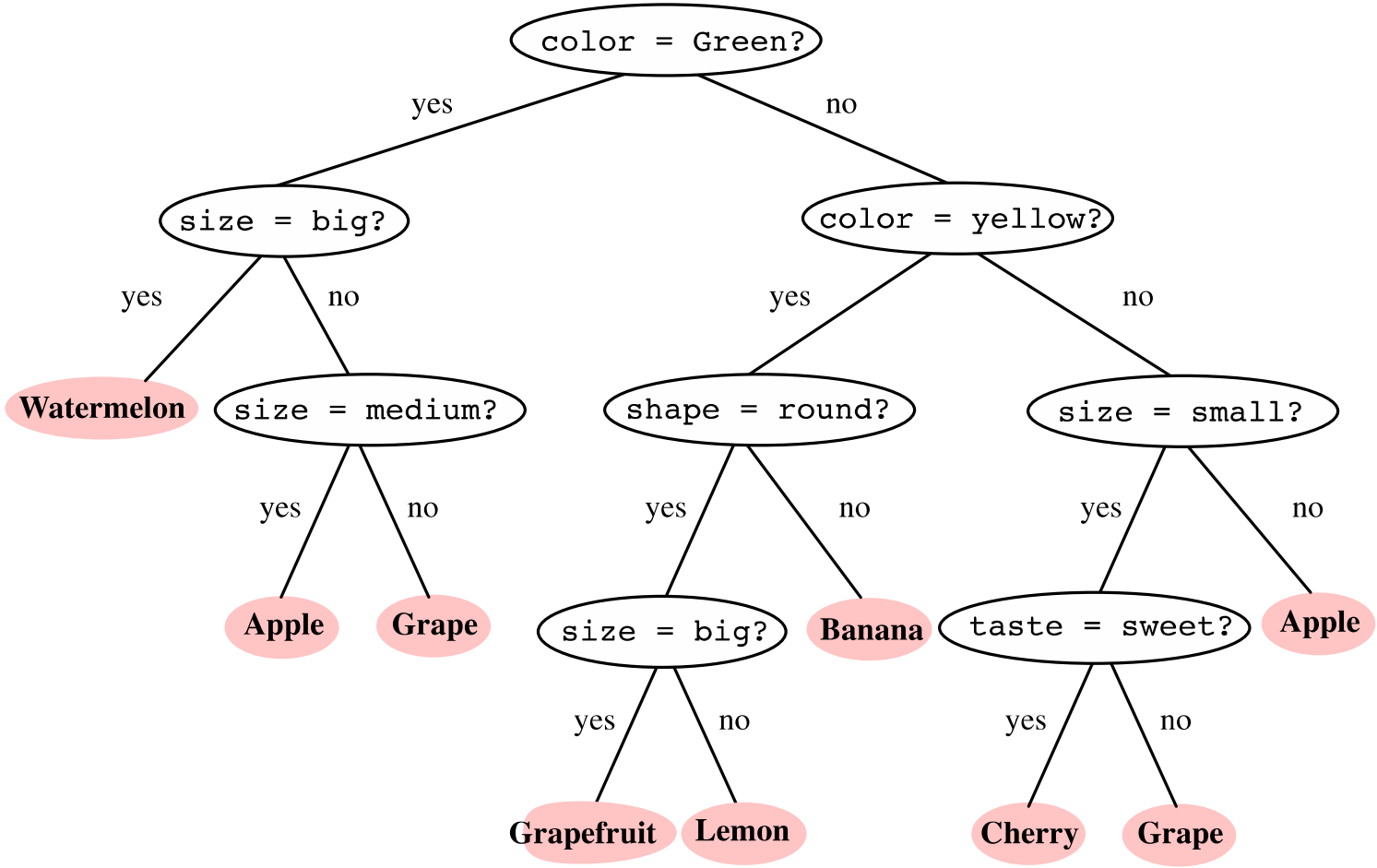
Exemplo de AD para classificação de frutas por categorização de características.



Duda, Hart & Stork

Árvores de Decisão - Abstração

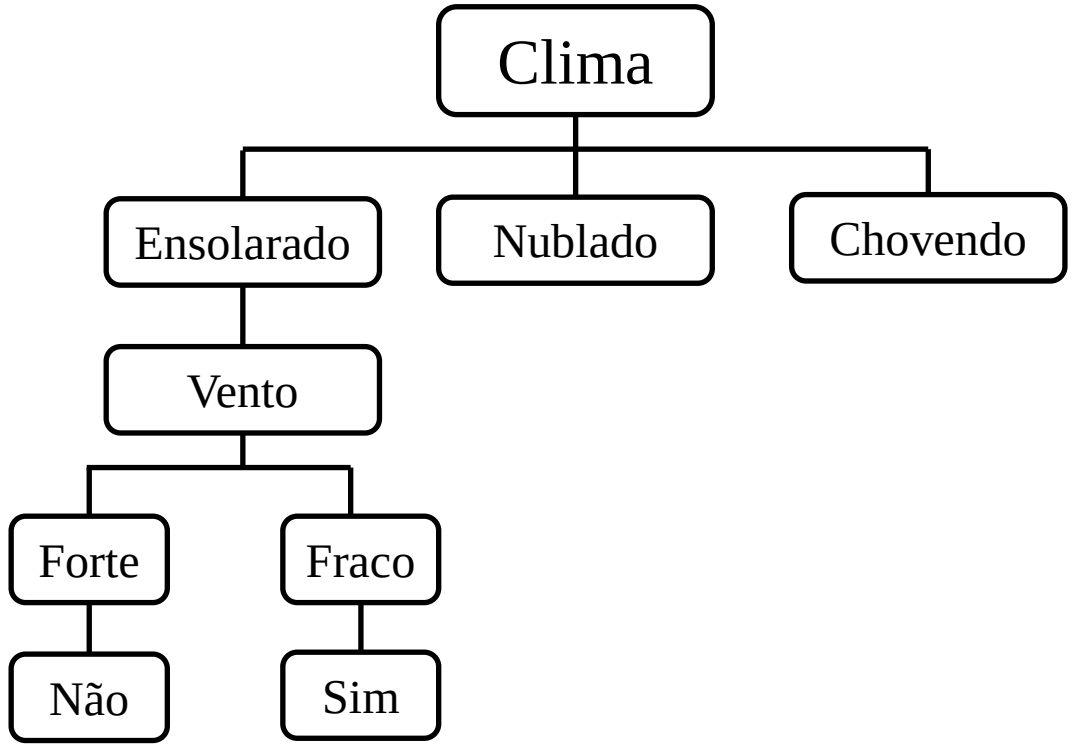
Exemplo de AD para classificação de frutas por atributos binários e baseado no conceito if then.



Árvores de Decisão – Exemplo com o Jogo de Vôlei

Você observou que diferentes inícios podem produzir classificações diferentes. Portanto, é necessário alguma métrica que auxilie na determinação das variáveis que otimizam a decisão.

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



Árvores de Decisão – Entropia

A primeira dificuldade ao se construir uma AD é definir qual atributo é a raiz da árvore quais são as folhas. Uma escolha adequada do atributo raiz e das folhas leva a um classificador otimizado. Naturalmente, quanto menor o número de níveis, mais otimizado será o modelo.

Como determinar qual atributo é relevante para compor a hierarquia da árvore?

$$Entropia(S) = -\sum_{j=1}^{d} P(c_j) log_2 P(c_j)$$

Onde $P(c_j)$ é a fração dos padrões de N que pertencem a categoria c_j, $j = 1, 2, \ldots, d$ classes.

A entropia mede a homogeneidade entre as amostras e permite determinar a pureza ou impureza da informação de um dado atributo (A). Para um caso de **duas classes**, poderíamos escrever a equação acima como:

$$Entropia(S) = -p_{\oplus} log_2(p_{\oplus}) - p_{\ominus} log_2(p_{\ominus})$$

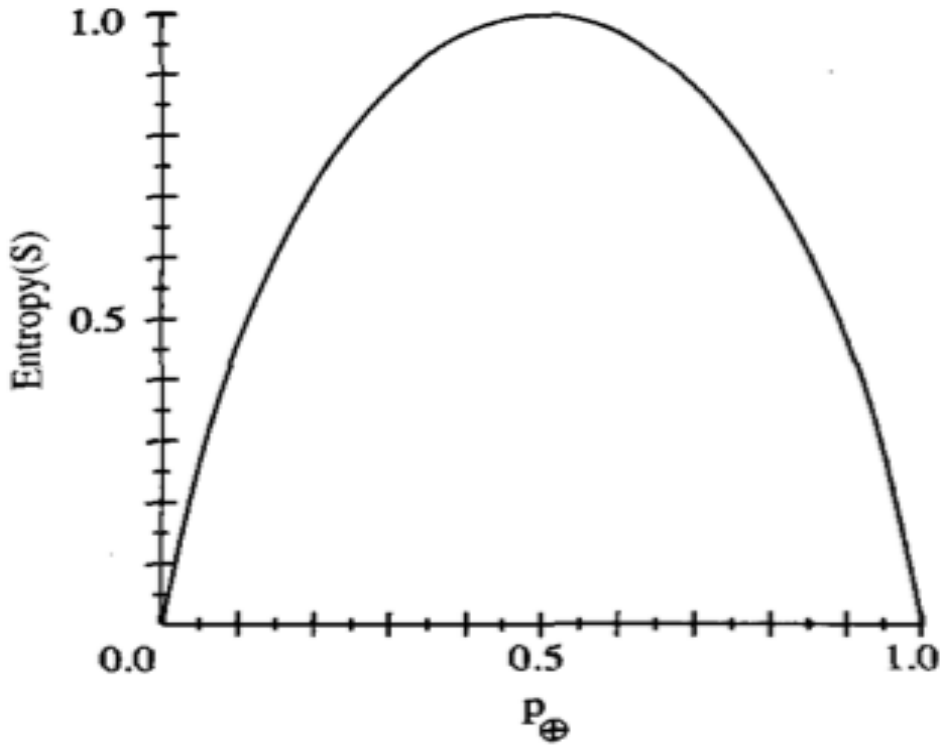
Na qual $p_{\oplus}$ é a proporção do exemplos positivos de S e $p_{\ominus}$ a proporção dos exemplos negativos de (S). Assim, se a $Entropia(S) = 0$, significa que todas as amostras pertencem a mesma classe. No caso de amostras equiprováveis a $Entropia(S) = 1$. O $log_2 x$ foi mantido por Quinlan, assim como está originalmente na Teoria da Informação de Shannon. Esse conceito é núcleo do algoritmo ID3 desenvolvido por J. R. Quinlan.

Nos cálculos com entropia de AD, costuma se definir $0 log_2 0 = 0$.

Árvores de Decisão - Entropia

Forma da Entropia (S) $quando$ $p_{\oplus}$ varia de 0 a 1 está representada na figura abaixo. O algoritmo ID3 usa a entropia para determinar a homogeneidade do conjunto. Se o conjunto é homogêneo, então a entropia é zero (0), e se o conjunto é equiprovável a entropia é 1.

$$Entropia(S) = -p_{\oplus} log_2(p_{\oplus}) - p_{\ominus} log_2(p_{\ominus})$$



Árvores de Decisão – Cálculo do Ganho de Informação

Dado a medida da entropia ou impureza das amostras, podemos calcular a eficácia de cada atributo na classificação. Essa medida é conhecida como ganho de informação e é simplesmente a redução da entropia causada pelo particionamento das amostras de acordo com cada atributo.

$$G(S, Atributo) = Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

Onde S é o conjunto das amostras, A o atributo a se avaliado, S_v é o subconjunto dos atributos de A que ocorrem em S.

O conceito de entropia em AD está associado à variabilidade de classes presentes no conjunto S. Quanto mais classes, maior entropia e mais impura ela será.
Além disso, busca-se uma propriedade de consulta T em cada ramo, tal que a informação buscada esteja imediatamente abaixo do ramo N, e tão puro quando possível.

A Entropia é uma das medidas utilizadas por AD para determinar a impureza de conjunto. Outra medida muito comum é a medida Gini, que avalia a variância entre as classes.

Árvores de Decisão – Cálculo do Ganho de Informação

Para entender o funcionamento da AD, vamos utilizar novamente o exemplo de Mitchel para estimar a possibilidade de ter jogo de praia dado as condições meteorológicas. Usaremos o algoritmo ID3 que foi desenvolvido por Quinlan (1987).

Amostras	Tempo	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

a) Calcule a entropia para todos os atributos A do conjunto de dados S.

b) Particione o conjunto S em subconjuntos usando o atributo que maximiza o ganho.

c) Monte a árvore de decisão baseado no atributo encontrando em (b).

d) Crie os ramos e folhas com os atributos restantes.

Árvores de Decisão – Cálculo do Ganho de Informação

Amostras	Tempo	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

Valores(Vento) = {Fraco, Forte}

Classes = {Sim, Não}

Podemos determinar que o $S = [9_{\oplus}, 5_{\ominus}]$.

$$S^{Fraco} = [6_{\oplus}, 2_{\ominus}]$$

$$S^{Forte} = [3_{\oplus}, 3_{\ominus}]$$

E o ganho de informação pode ser determinado como:

$$G(S, Vento) = Entropia(S) - \sum_{v \in \{Fraco, Forte\}} \frac{|S_v|}{|S|} Entropia(S_v)$$

Primeiro calculamos a entropia de S, nesse caso, consideramos duas classes.

$$Entropia(S) = \left[-\frac{9}{14} log_2\left(\frac{9}{14}\right)\right] - \left[\frac{5}{14} log_2\left(\frac{5}{14}\right)\right] = 0{,}940$$

Agora calculamos para o valor do atributo $Vento = Fraco$:

$$Entropia(S^{Fraco}) = \left[-\frac{6}{8} log_2\left(\frac{6}{8}\right)\right] - \left[\frac{2}{8} log_2\left(\frac{2}{8}\right)\right] = 0{,}811$$

Para $Vento = Forte$:

$$Entropia(S^{Forte}) = \left[-\frac{3}{6} log_2\left(\frac{3}{6}\right)\right] - \left[\frac{3}{6} log_2\left(\frac{3}{6}\right)\right] = 1$$

Observe que neste caso, o conjunto é equiprovável e a entropia é igual a 1.

Finalmente, o **ganho de informação** para o atributo $Vento$:

$$G(S, Vento) = Entropia(S) - \sum_{v \in \{Fraco, Forte\}} \frac{|S_v|}{|S|} Entropia(S_v)$$

$$G(S, Vento) = 0{,}94 - \frac{8}{14} \times (0{,}811) - \frac{6}{14} \times (1) = 0{,}048$$

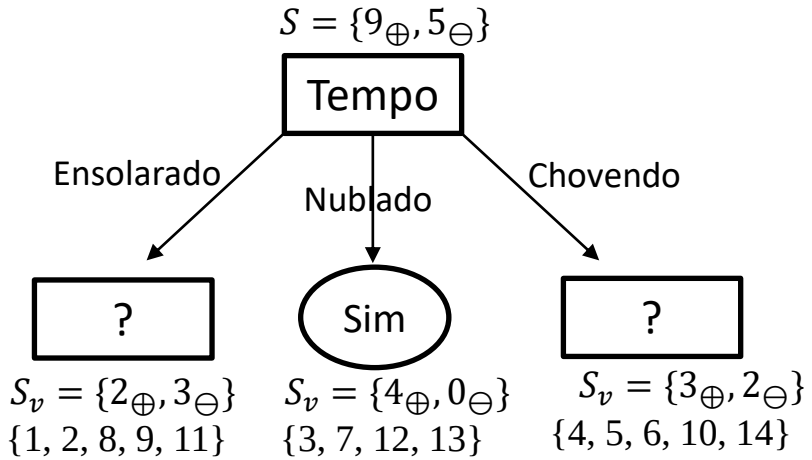
Calculando o ganho de informação para cada atributo, e comparando os resultados, podemos determinar qual atributo contribui mais para otimizar a classificação do conjunto de treinamento.

Árvores de Decisão

Calcule os ganhos de informação gerados de cada atributo (Fazer em casa):

$$G(S, Vento) = 0{,}048$$
$$G(S, Tempo) = 0{,}246$$
$$G(S, Humidade) = 0{,}151$$
$$G(S, Temperatura) = 0{,}029$$

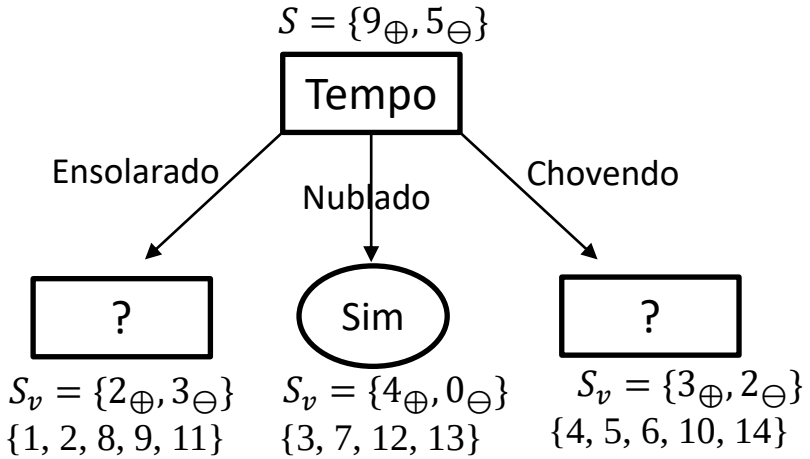
O atributo que melhor fornece informação para o resultado “Jogar Vôlei” é o atributo Tempo. Assim, ele será definido como raiz da árvore de decisão.



Agora devemos determinar os atributos restantes para definir as folhas (Temperatura, Vento, Humidade). Use o mesmo critério anterior, determinando o ganho de informação.

Árvores de Decisão

Olhando o resultado anterior da folha mais a esquerda temos que para o atributo ensolarado temos as amostras {1, 2, 8, 9, 11}. Calcule o ganho de informação para cada atributo desse sub-conjunto, menos o atributo do ramo.



Assim, a entropia é calculada sobre todas amostras cujo atributo é "Ensolarado", que nesse caso é um subconjunto de S.

$$Entropia(Ensolarado) = -\frac{2}{5} log_2 \left(\frac{2}{5}\right) - \frac{3}{5} log_2 \left(\frac{3}{5}\right) = 0{,}970$$

Árvores de Decisão

Calculando o ganho de informação para cada atributo dependente do atributo “ensolarado”.

$$G(Ensolado, Humidade) = 0{,}970 - \frac{3}{5}(0{,}0) - \frac{2}{5}(1{,}0) = 0{,}97$$

Calculando o ganho de informação para o atributo Vento:

$$G(Ensolado, Vento) = 0{,}970 - \frac{2}{5}(1{,}0) - \frac{3}{5}(0{,}918) = 0{,}019$$

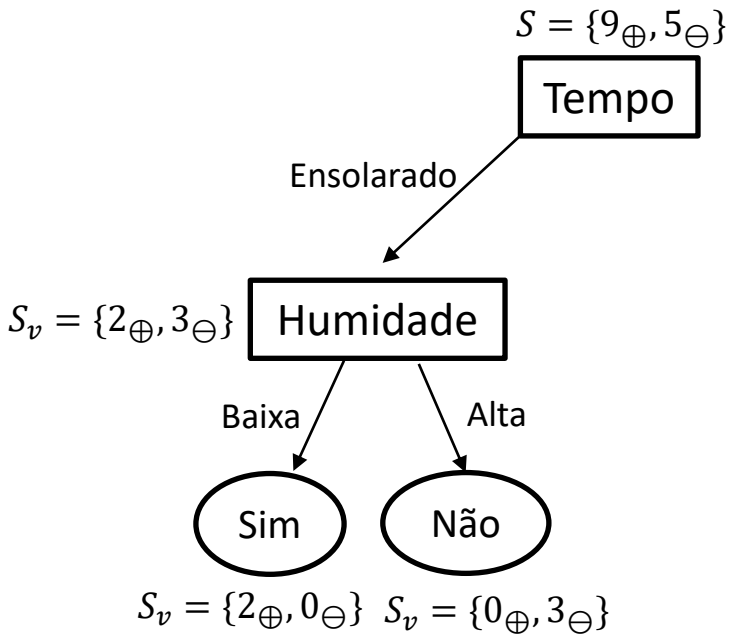
Calculando o ganho de informação para o atributo Temperatura:

$$G(Ensolado, Temperatura) = 0{,}970 - \frac{2}{5}(0{,}0) - \frac{2}{5}(1{,}0) - \frac{1}{5}(0{,}0) = 0{,}57$$

Assim, a folha é definido pelo atributo de maior ganho, que nesse caso é a humidade.

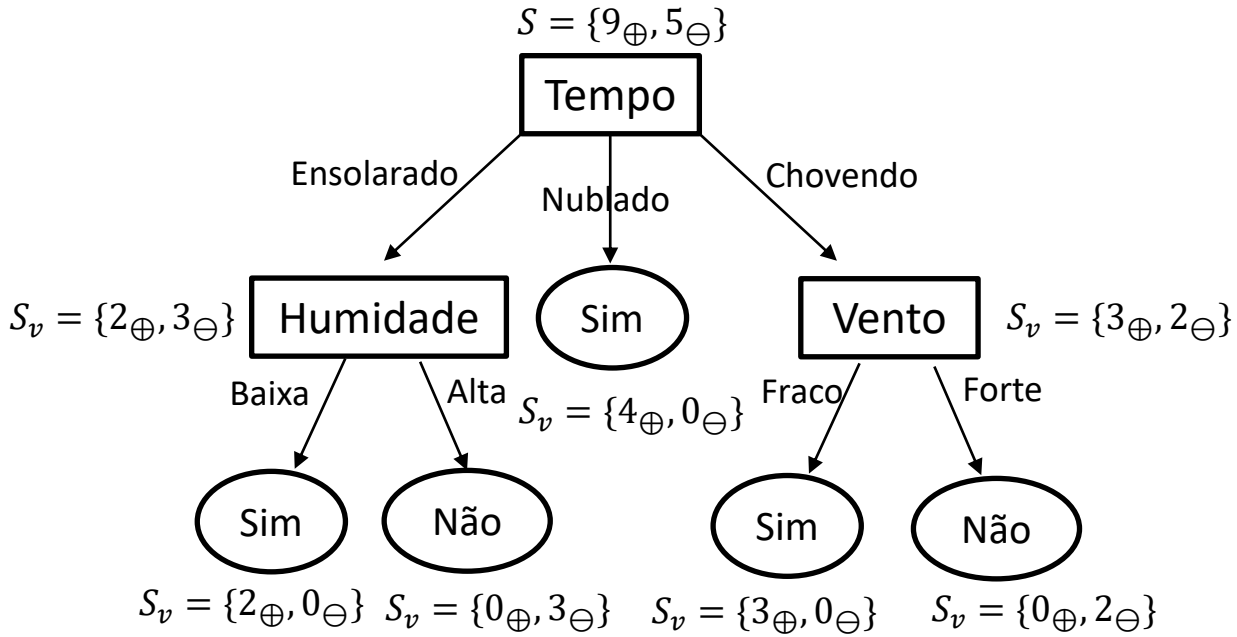
Árvores de Decisão

Finalmente, temos a folha esquerda definida. Repita o processo para os atributos restantes.



Árvores de Decisão

Finalmente, temos:



O resultado da árvore indica que o atributo Temperatura **não** influencia na decisão de "jogar vôlei".

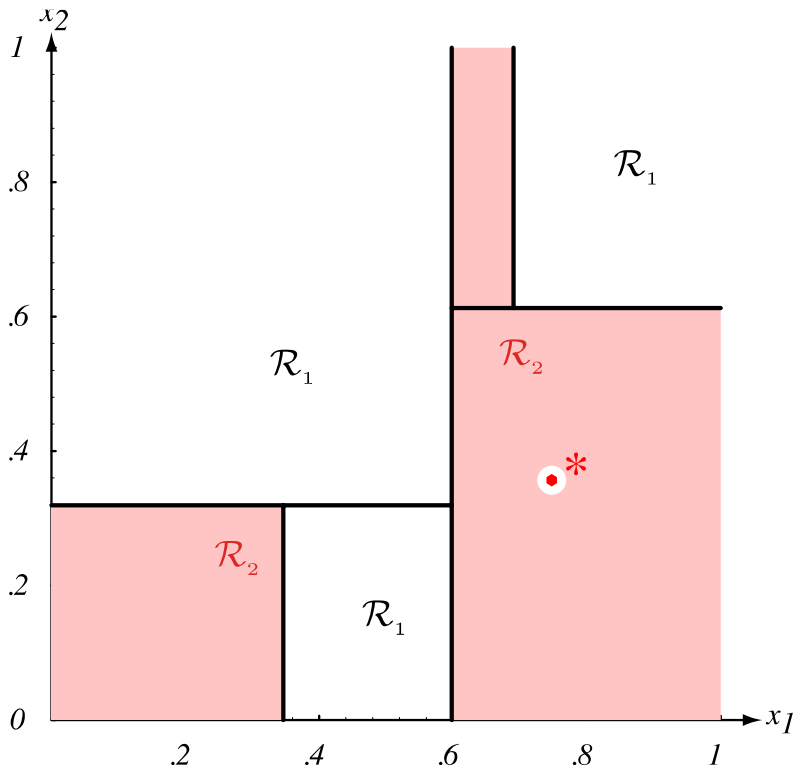
Tom Mitchell

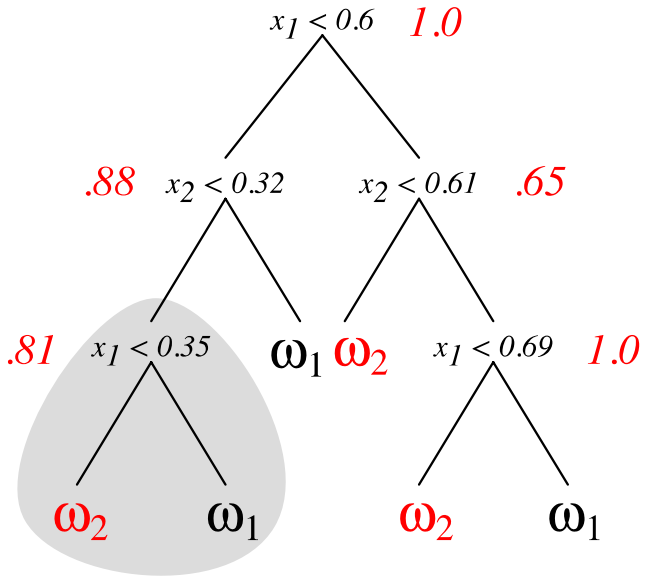
Árvores de Decisão

As ADs também pode ser usada com dados não categóricos, tais como valores numéricos. No exemplo abaixo temos duas classes (w_1, w_2), cujo espaço de entrada está sendo particionado de maneira não linear em regiões $\mathcal{R}_1$ *e* $\mathcal{R}_2$. O método é conhecido como Split Point.

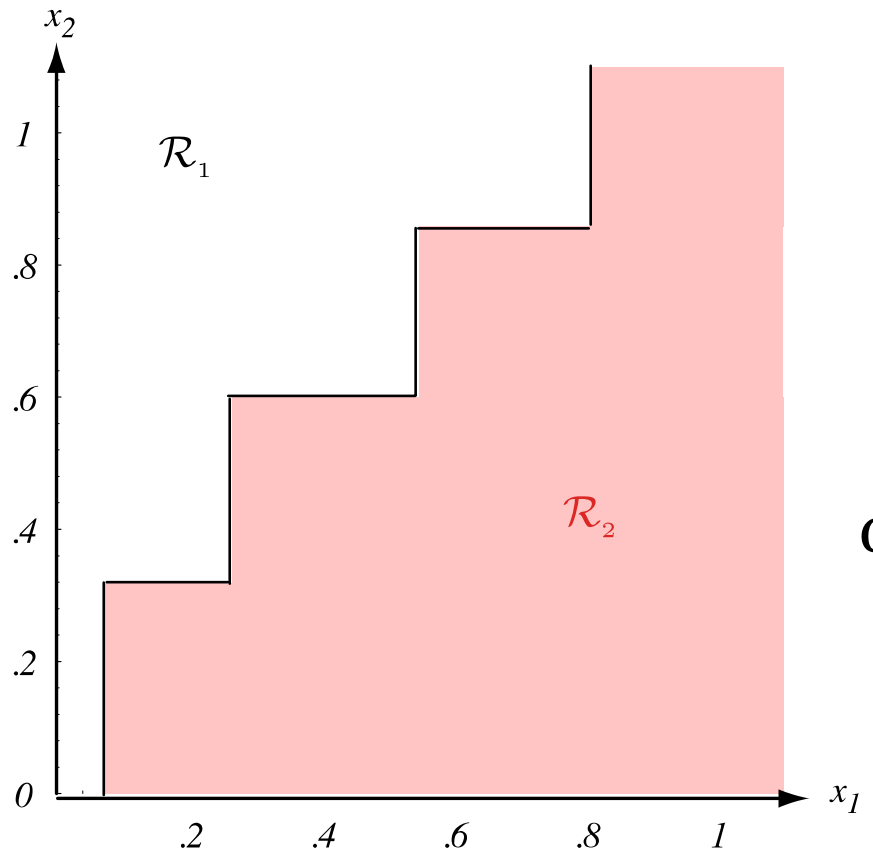
ω_1 (black)	
x_1	x_2
.15	.83
.09	.55
.29	.35
.38	.70
.52	.48
.57	.73
.73	.75
.47	.06

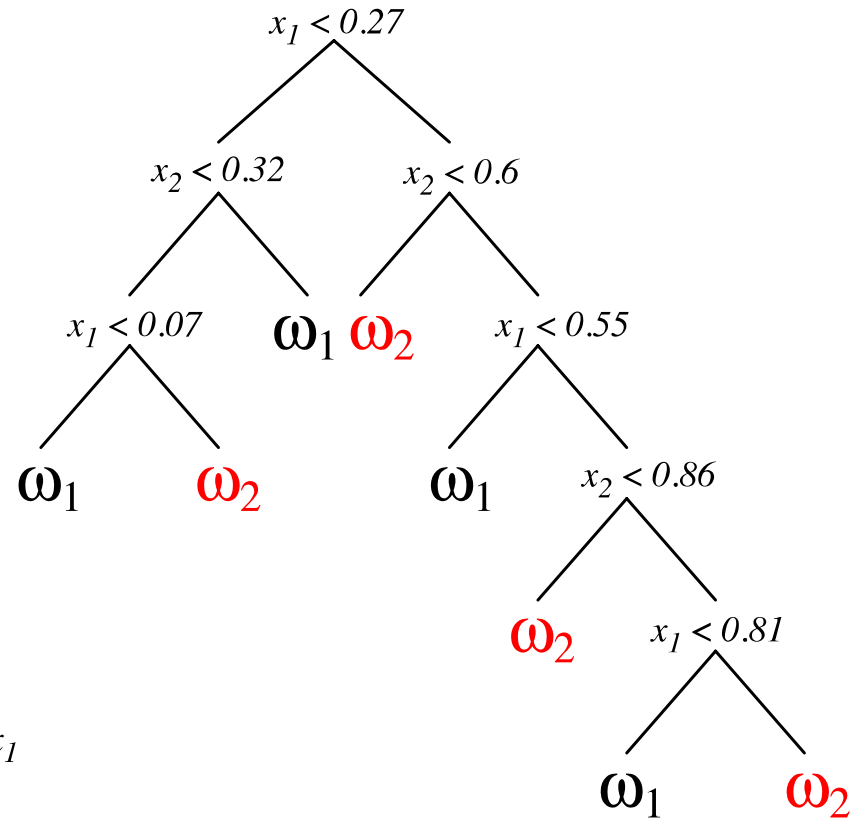
ω_2 (red)	
x_1	x_2
.10	.29
.08	.15
.23	.16
.70	.19
.62	.47
.91	.27
.65	.90
.75	.36* (.32†)





Árvores de Decisão





Duda Hart & Strok

Árvores de Decisão