MAC 0460 / 5832

Aprendizagem Computacional
Modelos, algoritmos e aplicações

Nina Hirata (nina@ime.usp.br)
Sala 6 - bloco C

Monitor: Igor

Aula 19 (2012)

ÁRVORES DE DECISÃO

Exemplo: Akinator

http://en.akinator.com/

Árvores de decisão/classificação

- **Seja S um conjunto de itens em $\mathbb{R}^d$ e $i \in \{1, 2, \ldots, d\}$. Considere, por exemplo, os subconjuntos:**

$$S_1 = \{\mathbf{x} \in S : x_i \leq 0\}$$

$$S_2 = \{\mathbf{x} \in S : x_i > 0\}$$

A pergunta "$x_i \leq 0$?" admite duas respostas, SIM e NÃO, e particiona o conjunto S em exatamente dois subconjuntos. Mais que isso, particiona o espaço de características em dois subespaços.

Árvores de decisão/classificação

- **Cada um desses subconjuntos pode ser sucessivamente particionado fazendo-se outras perguntas.**

  Dependendo da pergunta, o espaço pode ser particionado em mais de dois subespaços.

- **A sequência de perguntas e possíveis respostas podem ser organizadas em uma estrutura do tipo árvore.**

  As perguntas são associadas aos nós da árvore, enquanto o número de possíveis respostas define o número de ramificações de um nó.

- **Essa idéia é usada pelos classificadores conhecidos por árvores de decisão.**

Classificação usando árvores de decisão

- Na fase de crescimento (treinamento) da árvore, começa-se associando o conjunto S de itens (o espaço todo) ao nó raiz.

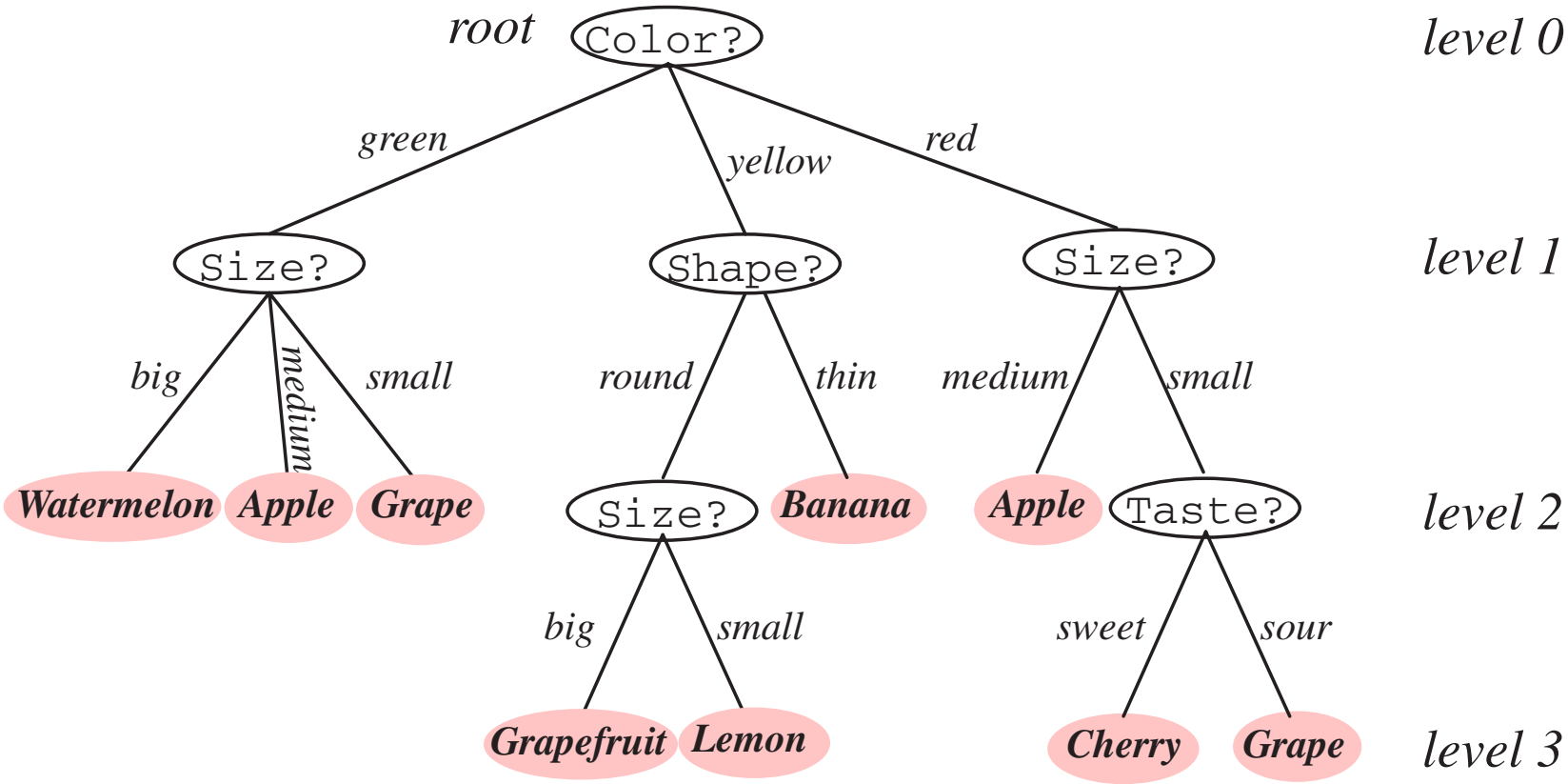
  Os conjuntos associados aos nós são sucessivamente subdivididos caso não sejam puros (contenham itens de diferentes classes), usando perguntas adequadas.

  Nós folhas correspondem a subespaços puros.

- Dada uma árvore de decisão, para classificar um item, percorre-se a árvore do nó raiz até um nó folha, sempre seguindo o ramo correspondente à resposta para a pergunta dos nós visitados.

Exemplo de uma árvore de decisão

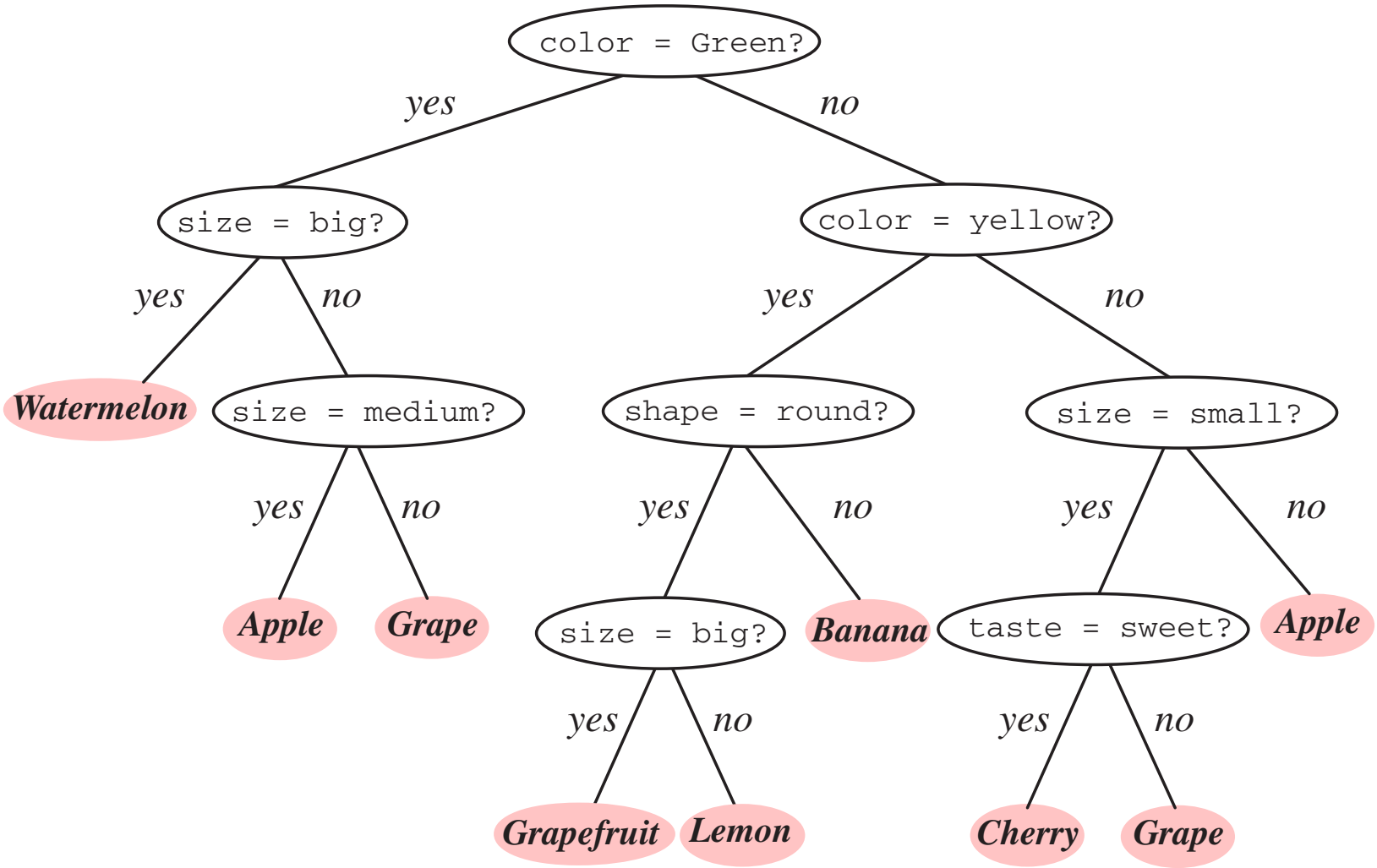
Classificação de frutas baseada em cor, tamanho, forma e sabor



Fonte: Duda et al

Árvores de decisão

Toda árvore de decisão pode ser expressa por uma árvore binária.



Construção de árvores de decisão

Crescimento de uma árvore

- Como escolher a pergunta a ser associada a um nó ?
- Quando devemos parar de dividir um subespaço? O que é um subespaço puro?

A maioria dos algoritmos considera árvores binárias (por causa de sua expressividade e simplicidade no treinamento)

Árvores de decisão

Perguntas possíveis

- **Características categóricas**: basta considerarmos $x_i = v?$, na qual v é um possível valor de x_i.

  Essa pergunta tem resposta binária

- **Características numéricas**: ordenar os valores de x_i e para quaisquer dois valores v_k e v_{k+1} de x_i adjacentes nessa ordenação tal que v_k e v_{k+1} são de classes distintas, considera-se a pergunta $x_i \leq \frac{v_k + v_{k+1}}{2}?$.

  (o número de possíveis perguntas pode ser beeeem grande ...)

- Pode-se ainda considerar uma **combinação linear de várias características**. Neste caso, a superfície que particiona o espaço em dois subespaços é um hiperplano não necessariamente ortogonal a uma das características.

Árvores de decisão: como escolher uma pergunta?

- escolher perguntas simples e que resultem em árvores simples e compactas
- escolher uma pergunta cujas respostas gerem subconjuntos mais "puros" (menos mistura de classes) possíveis
- para isso, definem-se índices de impureza, que são calculados sobre os exemplos associados aos nós
- para dividir o conjunto de exemplos associados a um nó, calcula-se a impureza do nó a ser dividido e, para cada possível pergunta, dos nós que resultariam após a divisão. Escolhe-se a pergunta que resulta em maior redução de impureza

Árvores de decisão

Redução de impureza:

- Seja N um nó. O índice de impureza do subconjunto asociado a esse nó é denotado $i(N)$

- sejam N_L e N_R os nós filhos que resultariam de uma divisão dos itens em N

  Seja P_L (P_R) a proporção de itens em N que iriam para o nó N_L (N_R)

- o decréscimo de impureza dessa divisão pode ser calculado por

$$\Delta i(N) = i(N) - P_L\, i(N_L) - (1 - P_L)\, i(N_R)$$

Generalização para partições não binárias

Suponha que um nó N, associado a um conjunto S de itens é subdividido em subconjuntos S_v (v é um índice), associados aos nós filhos N_v. O **ganho de informação** dessa divisão é dado por:

$$Gain(N) = i(N) - \sum_v \frac{|S_v|}{|S|} i(N_v)$$

MAS: um número grande de subconjuntos $S_v \longrightarrow$ melhor ganho!

Generalização para partições não binárias

Suponha que um nó N, associado a um conjunto S de itens é subdividido em subconjuntos S_v (v é um índice), associados aos nós filhos N_v. O **ganho de informação** dessa divisão é dado por:

$$Gain(N) = i(N) - \sum_v \frac{|S_v|}{|S|} i(N_v)$$

MAS: um número grande de subconjuntos $S_v \longrightarrow$ melhor ganho!

Melhor considerar algo como:

$$GainRatio(N) = \frac{Gain(N)}{SplitInfo(N)}$$

onde

$$SplitInfo(N) = -\sum_v \frac{|S_v|}{|S|} \log \frac{|S_v|}{|S|}$$

Árvores de decisão: tipos de índices de impureza

Entropy impurity:

$$i(N) = -\sum_{i=1}^{c} P(\omega_i|N) log_2 P(\omega_i|N)$$

Gini impurity:

$$i(N) = \sum_{i \neq j} P(\omega_i|N) P(\omega_j|N) = \frac{1}{2}\left[1 - \sum_{j} P^2(\omega_j|N)\right]$$

Misclassification impurity:

$$i(N) = 1 - max_j P(\omega_j|N)$$

Árvores de decisão: índices de impureza

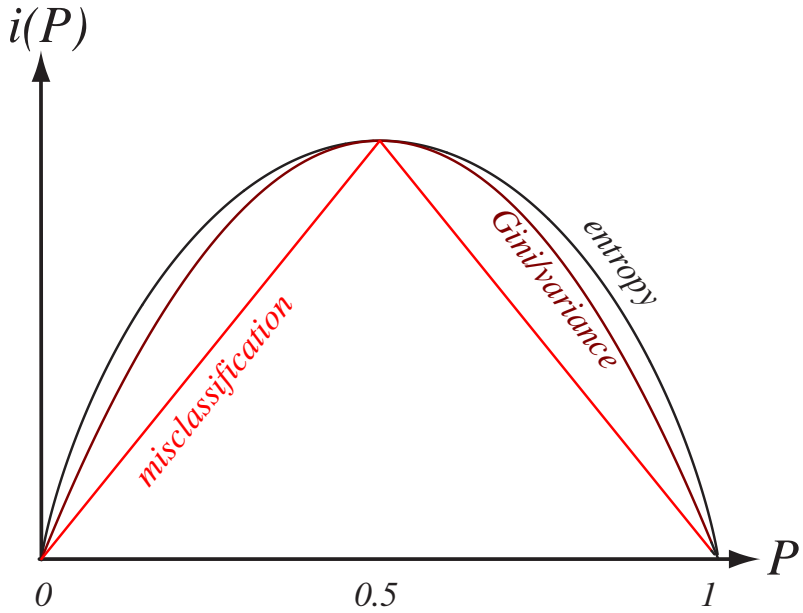


FIGURE 8.4. For the two-category case, the impurity functions peak at equal class frequencies and the variance and the Gini impurity functions are identical. The entropy, variance, Gini, and misclassification impurities (given by Eqs. 1–4, respectively) have been adjusted in scale and offset to facilitate comparison here; such scale and offset do not directly affect learning or classification. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Árvores de decisão: índices de impureza

Critério twoing

A escolha de perguntas que minimizam a impureza pode levar a uma divisão na qual elementos de uma mesma classe ficam divididos, ficando parte deles num subconjunto e outra noutro subconjunto.

twoing: procura escolher uma divisão que melhor separa elementos das c classes em dois subconjuntos, de forma que elementos de uma classe não fiquem divididos.

Algoritmo: seja $C_1 = \{\omega_{i_1}, \ldots, \omega_{i_k}\}$ e $C_2 = C \setminus C_1$ na qual C denota o conjunto de todas as classes

Considerar todas as dicotomias C_1/C_2

Calcular um dos índices anteriores, porém considerando-se que as classes são C_1 e C_2

Árvores de decisão: quando finalizar as divisões?

Árvores de decisão: quando finalizar as divisões?

- quando o nó estiver **puro**
  - é comum resultar em nós folha com uma amostra apenas
  - pode ocorrer *overfitting*

Árvores de decisão: quando finalizar as divisões?

- quando o nó estiver **puro**
  - é comum resultar em nós folha com uma amostra apenas
  - pode ocorrer *overfitting*

- quando o **decréscimo de impureza for desprezível**
  Vantagem de ter um critério único para todos os nós
  Difícil quantificar o que é um decréscimo desprezível

Árvores de decisão: quando finalizar as divisões?

- quando o nó estiver **puro**
  - é comum resultar em nós folha com uma amostra apenas
  - pode ocorrer *overfitting*

- quando o **decréscimo de impureza for desprezível**
  Vantagem de ter um critério único para todos os nós
  Difícil quantificar o que é um decréscimo desprezível

- quando o **número de amostras** associadas ao nó ficar menor que um limiar (valor absoluto ou porcentagem)
  similaridade com vizinhos próximos (volume pequeno para regiões densas e volume grande para regiões esparsas)

Árvores de decisão: quando finalizar as divisões?

- usar técnica de **validação ou validação cruzada** (usar uma parte dos dados de treinamento para avaliar o desempenho da árvore; se a divisão resulta em melhora de desempenho no conjunto de validação, divide-se; caso contrário, pára)
  Desvantagem: parte da amostra deve ser usado para validação (menor conjunto de treinamento)

Árvores de decisão: quando finalizar as divisões?

- usar técnica de **validação ou validação cruzada** (usar uma parte dos dados de treinamento para avaliar o desempenho da árvore; se a divisão resulta em melhora de desempenho no conjunto de validação, divide-se; caso contrário, pára)
  Desvantagem: parte da amostra deve ser usado para validação (menor conjunto de treinamento)

- usar algum **critério global** (por exemplo, que leva em consideração a impureza de todos os nós folhas)

Árvores de decisão: quando finalizar as divisões?

- usar técnica de **validação ou validação cruzada** (usar uma parte dos dados de treinamento para avaliar o desempenho da árvore; se a divisão resulta em melhora de desempenho no conjunto de validação, divide-se; caso contrário, pára)
  Desvantagem: parte da amostra deve ser usado para validação (menor conjunto de treinamento)

- usar algum **critério global** (por exemplo, que leva em consideração a impureza de todos os nós folhas)

- etc

Árvores de decisão: poda

Motivação: a escolha de uma divisão não leva em consideração possíveis escolhas nos nós descendentes. Se isso fosse levado em consideração, divisões globalmente melhores poderiam ser feitas.

Árvores de decisão: poda

Motivação: a escolha de uma divisão não leva em consideração possíveis escolhas nos nós descendentes. Se isso fosse levado em consideração, divisões globalmente melhores poderiam ser feitas.

Poda: deixar crescer a árvore totalmente, até termos todos os nós folhas puros.

Em seguida qualquer par de nós folhas adjacentes (filhos de um mesmo nó pai N) cuja eliminação resulta em "pequeno" acréscimo de impureza são eliminados e N é considerado folha. Esse processo é aplicado sucessivamente até que não mais existam tais pares de nós.

Árvores de decisão: rótulos de classe a serem associados aos nós folhas

Esta é a parte mais simples na construção de árvores de decisão.
A escolha mais natural para o rótulo de classe para um nó folha da árvore é o da classe mais frequente entre as amostras associadas ao nó.

Árvore de decisão: superfície de decisão

Para decisões simples, divisões correspondem a hiperplanos ortogonais a um dos eixos.

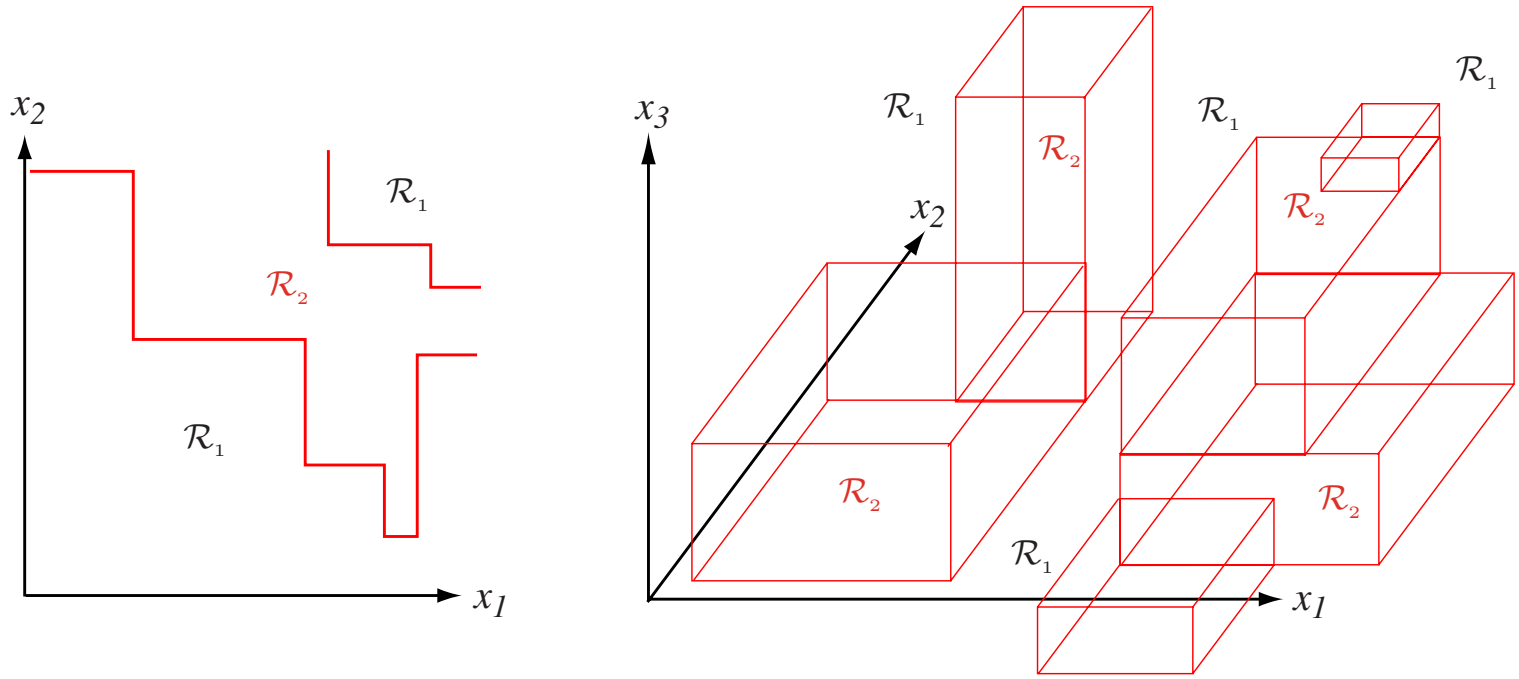
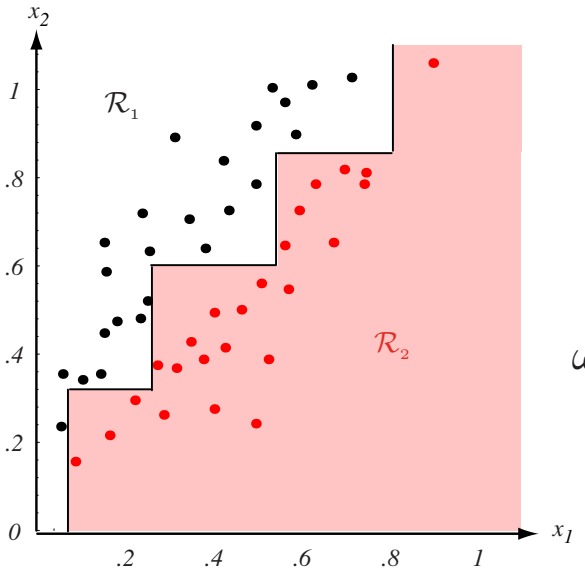
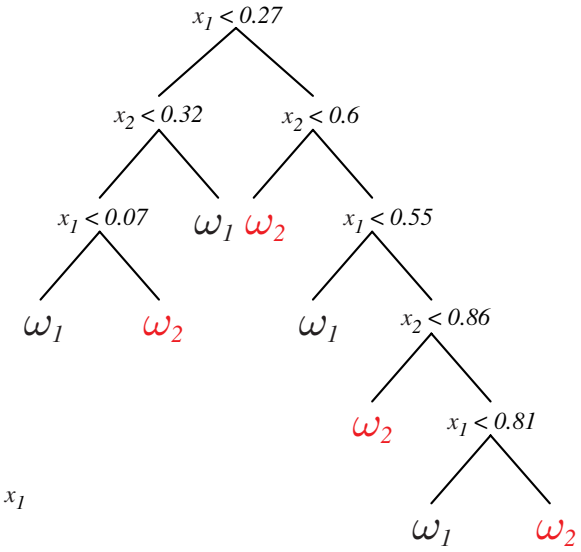


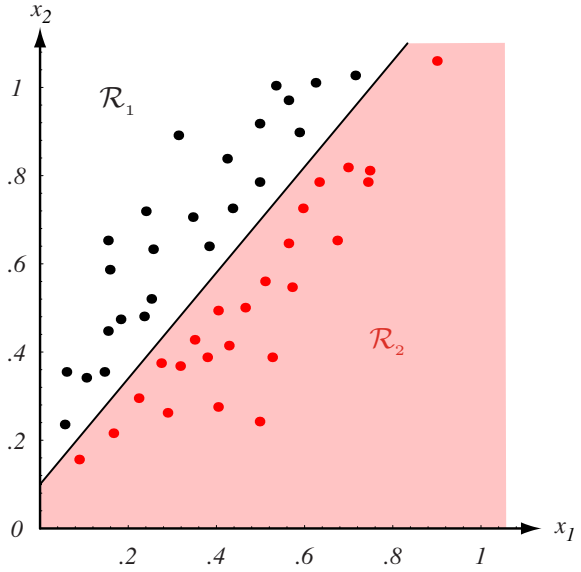
FIGURE 8.3. Monothetic decision trees create decision boundaries with portions perpendicular to the feature axes. The decision regions are marked $\mathcal{R}_1$ and $\mathcal{R}_2$ in these two-dimensional and three-dimensional two-category examples. With a sufficiently large tree, any decision boundary can be approximated arbitrarily well in this way. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

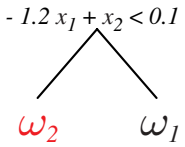
Árvore de decisão: superfície de decisão

Escolha da característica usada na pergunta pode não ser adequada









Hiperplanos arbitrários (calculados usando a técnica de gradiente descendente, por exemplo)

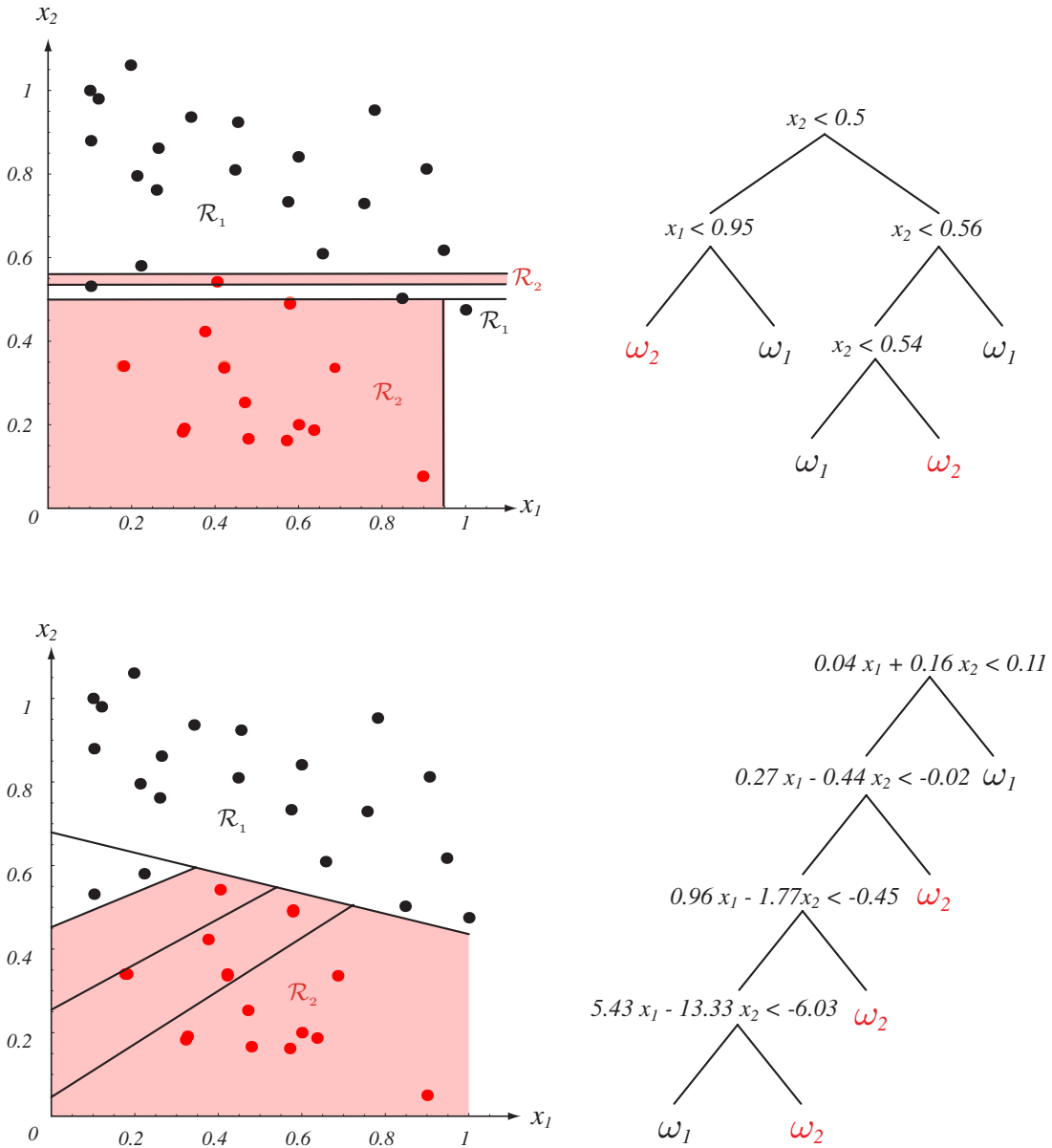
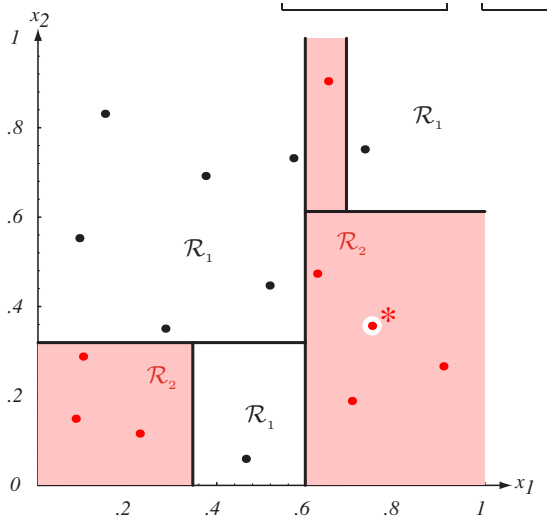


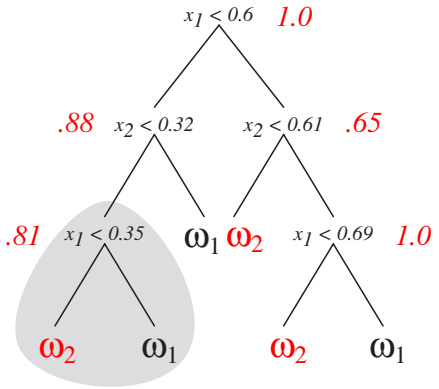
FIGURE 8.6. One form of multivariate tree employs general linear decisions at each

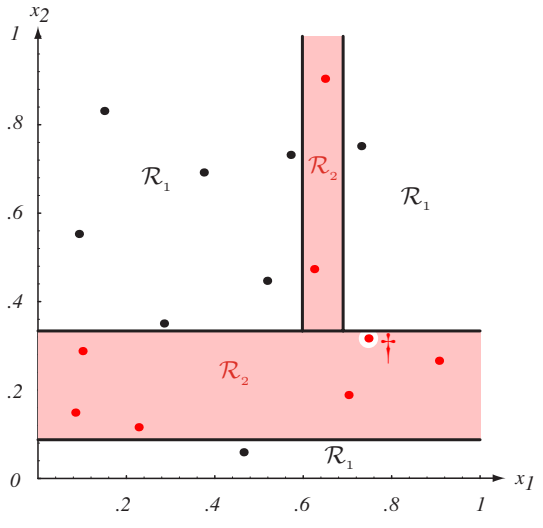
Árvore de decisão: sensível a ruídos

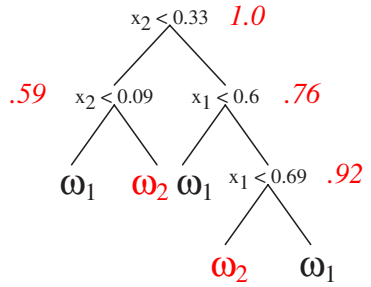
ω_1 (black)	
x_1	x_2
.15	.83
.09	.55
.29	.35
.38	.70
.52	.48
.57	.73
.73	.75
.47	.06

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









Árvore de decisão: missing values

Tratamento de missing values no treinamento: Descartar itens com valores faltantes

Se a pergunta envolve a característica x_i, para calcular a impureza pode-se desconsiderar apenas as amostras com valores faltantes em x_i. Amostras com valores faltantes em outras características não afetam o cálculo.

Tratamento de missing values na classificação: para classificar itens com valores faltantes, pode-se associar a cada nó "perguntas suplentes" que são usadas quando o valor da característica associada à pergunta principal está ausente.

Por outro lado, um valor faltante pode por si só representar uma informação relevante. Nesses casos, pode-se atribuir um valor artificial para os dados faltantes.
(note que isso não funcionaria para abordagens baseadas em métricas, pois o valor artificial usado pode afetar o resultado de forma indesejada)

Árvore de decisão: resumo

- pode lidar tanto com características numéricas ou categóricas (é uma técnica que não depende de uma métrica definida no espaço de características)
- idéia simples (embora a construção não seja necessariamente)
- em alguns casos, as decisões que levam a uma classificação é passível de interpretação (regras bem definidas)

  Cada nó folha corresponde a uma conjunção de perguntas. Uma classe é uma disjunção de conjunções.
- vários programs bem conhecidos: ID3, CART, C4.5, ASSISTANT
- possibilita tratamento de dados faltantes (*missing values*)