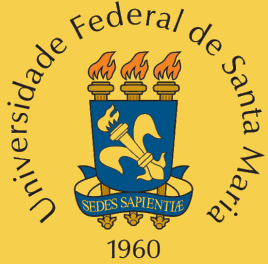

Universidade Federal de Santa Maria
SEDES SAPIENTIÆ
1960


UFVJM
1953
SAPIENTIA ET LIBERTAS

Reconhecimento de padrões e aprendizagem computacional

Árvores de decisão

Árvores de decisão

Definição

Estrutura hierárquica onde cada nó interno representa um teste em um atributo, cada ramo representa uma saída do teste, e cada nó folha (ou nó terminal) representa o rótulo de uma classe.

Criação empírica

O processo de criação de uma árvore de decisão é composto por três etapas:

1. Adotar um critério para a criação de um nó;
2. Classificação de um nó como terminal ou não terminal;
3. Geração de um conjunto de árvores podadas.

Critério para avaliação

$$Ganho = info(T) - \sum_{t=1}^{m} \frac{T_t}{T} \tag{1}$$

em que:

$$Info(T) = -\sum_{j=1}^{k} \frac{freq(C_j, T)}{T} * log_2(\frac{freq(C_j, T)}{T} \tag{2}$$

Siglas

- $freq(C_j, T)$ é o número de amostras T subdivididas no subespaço C_j.
- T é o número total de amostras.
- k é o número total de classes existentes.
- m é o número de subespaços criados na divisão de T.

Exemplo 1

Atributos categóricos
(algoritmo Itemized Dichotomizer 3)

Observações empíricas

Obs	Céu	Temp	Umidade	Vento	Decisão
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
3	nublado	alta	alta	não	joga
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
7	nublado	baixa	normal	sim	joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
10	chuva	suave	normal	não	joga
11	sol	suave	normal	sim	joga
12	nublado	suave	alta	sim	joga
13	nublado	alta	normal	não	joga
14	chuva	suave	alta	sim	não joga

Entropia do sistema

$$info(T) = -p_1 log_2 p_1 - p_2 log_2 p_2 - ... - p_k log_2 p_k \quad (3)$$

Calculando a entropia para o espaço atual $T = [9+, 5-]$:

$$-(\frac{9}{14})log_2(\frac{9}{14}) - (\frac{5}{14})log_2(\frac{5}{14}) = 0,940 \quad (4)$$

Subespaços

- Céu
- Temperatura
- Umidade
- Vento

Obs	Céu	Temp	Umidade	Vento	Decisão
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
3	nublado	alta	alta	não	joga
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
7	nublado	baixa	normal	sim	joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
10	chuva	suave	normal	não	joga
11	sol	suave	normal	sim	joga
12	nublado	suave	alta	sim	joga
13	nublado	alta	normal	não	joga
14	chuva	suave	alta	sim	não joga

Divisão pelo Céu

$$T_{sol} = [2+, 3-]; T_{nublado} = [4+, 0-]; T_{chuva} = [3+, 2-] \tag{5}$$

$$info(sol) = -(\frac{2}{5})log_2(\frac{2}{5}) - (\frac{3}{5})log_2(\frac{3}{5}) = 0,97094 \tag{6}$$

$$info(nublado) = -(\frac{4}{4})log_2(\frac{4}{4}) = 0 \tag{7}$$

$$info(chuva) = -(\frac{3}{5})log_2(\frac{3}{5}) - (\frac{2}{5})log_2(\frac{2}{5}) = 0,97094 \tag{8}$$

$$ganho = 0,940 - \frac{5}{14}info(sol) - \frac{4}{14}info(nublado) - \frac{5}{14}info(chuva) \tag{9}$$

$$ganho = 0,940 - \frac{5}{14}0,97094 - \frac{4}{14}0 - \frac{5}{14}0,97094 = 0,2464 \tag{10}$$

Divisão pela Temperatura

$$T_{alta} = [3+, 2-]; T_{suave} = [3+, 1-]; T_{baixa} = [3+, 2-] \tag{11}$$

$$info(alta) = -(\frac{3}{5})log_2(\frac{3}{5}) - (\frac{2}{5})log_2(\frac{2}{5}) = 0,97094 \tag{12}$$

$$info(suave) = -(\frac{3}{4})log_2(\frac{3}{4}) - (\frac{1}{4})log_2(\frac{1}{4}) = 0,811 \tag{13}$$

$$info(baixa) = -(\frac{3}{5})log_2(\frac{3}{5}) - (\frac{2}{5})log_2(\frac{2}{5}) = 0,97094 \tag{14}$$

$$ganho = 0,940 - \frac{5}{14}info(alta) - \frac{4}{14}info(suave) - \frac{5}{14}info(baixa) \tag{15}$$

$$ganho = 0,940 - \frac{5}{14}0,97094 - \frac{4}{14}0,811 - \frac{5}{14}0,97094 = 0,015 \tag{16}$$

Divisão pela umidade

$$T_{alta} = [3+, 4-]; T_{baixa} = [6+, 1-] \tag{17}$$

$$info(alta) = -(\frac{3}{7})log_2(\frac{3}{7}) - (\frac{4}{7})log_2(\frac{4}{7}) = 0,985228 \tag{18}$$

$$info(baixa) = -(\frac{6}{7})log_2(\frac{6}{7}) - (\frac{1}{7})log_2(\frac{1}{7}) = 0,591672 \tag{19}$$

$$ganho = 0,940 - \frac{7}{14}info(alta) - \frac{7}{14}info(baixa) \tag{20}$$

$$ganho = 0,940 - \frac{7}{14}0,985228 - \frac{7}{14}0,591672 = 0,151 \tag{21}$$

Divisão pelo vento

$$T_{sim} = [3+, 3-]; T_{nao} = [6+, 2-] \tag{22}$$

$$info(sim) = -(\frac{3}{6})log_2(\frac{3}{6}) - (\frac{3}{6})log_2(\frac{3}{6}) = 1 \tag{23}$$

$$info(nao) = -(\frac{6}{8})log_2(\frac{6}{8}) - (\frac{1}{8})log_2(\frac{1}{8}) = 0,811278 \tag{24}$$

$$ganho = 0,940 - \frac{8}{14}info(sim) - \frac{6}{14}info(nao) \tag{25}$$

$$ganho = 0,940 - \frac{8}{14}0,811278 - \frac{6}{14}1 = 0,047841 \tag{26}$$

Divisão do espaço

O atributo Céu apresenta o maior ganho de informação e, portanto será o primeiro nó da árvore de decisão.

O subespaço nublado tem entropia 0, indicando não haver dúvida na decisão de 'jogar'. Os subespaços sol e chuva apresentam entropia. E por isto serão analisados adiante.

Subespaço Céu = sol

Obs	Céu	Temp	Umidade	Vento	Decisão
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
11	sol	suave	normal	sim	joga

Ganho de informação no subespaço Céu = sol

Divisão pela Temperatura

$$T_{alta} = [2+, 0-]; T_{suave} = [1+, 1-]; T_{baixa} = [0+, 1-] \tag{27}$$

$$info(alta) = -(\frac{2}{2})log_2(\frac{2}{2}) = 0 \tag{28}$$

$$info(suave) = -(\frac{1}{2})log_2(\frac{1}{2}) - (\frac{1}{2})log_2(\frac{1}{2}) = 1 \tag{29}$$

$$info(baixa) = -(\frac{1}{1})log_2(\frac{1}{1}) = 0 \tag{30}$$

$$ganho = 0,97094 - \frac{2}{5}info(alta) - \frac{2}{5}info(suave) - \frac{1}{5}info(baixa) \tag{31}$$

$$ganho = 0,97094 - \frac{2}{5}0 - \frac{2}{5}1 - \frac{1}{5}0 = 0,57094 \tag{32}$$

Ganho de informação no subespaço Céu = sol

Divisão pela Umidade

$$T_{alta} = [3+, 0-]; T_{baixa} = [0+, 2-] \tag{33}$$

$$info(alta) = -(\frac{3}{3})log_2(\frac{3}{3}) = 0 \tag{34}$$

$$info(baixa) = -(\frac{2}{2})log_2(\frac{2}{2}) = 0 \tag{35}$$

$$ganho = 0,97094 - \frac{3}{5}info(alta) - \frac{2}{5}info(baixa) \tag{36}$$

$$ganho = 0,97094 - \frac{3}{5}0 - \frac{2}{5}0 = 0,97094 \tag{37}$$

Ganho de informação no subespaço Céu = sol

Divisão pelo Vento

$$T_{sim} = [1+, 1-]; T_{nao} = [2+, 1-] \tag{38}$$

$$info(sim) = -(\frac{1}{2})log_2(\frac{1}{2}) - (\frac{1}{2})log_2(\frac{1}{2}) = 1 \tag{39}$$

$$info(nao) = -(\frac{2}{3})log_2(\frac{2}{3}) - (\frac{1}{3})log_2(\frac{1}{3}) = 0,918295 \tag{40}$$

$$ganho = 0,97094 - \frac{2}{5}info(sim) - \frac{3}{5}info(nao) \tag{41}$$

$$ganho = 0,97094 - \frac{2}{5}1 - \frac{3}{5}0,918295 = 0,019963 \tag{42}$$

Divisão do subespaço Céu = sol

O atributo umidade apresenta o maior ganho de informação e, portanto será o segundo nó da árvore de decisão, no subespaço Céu = sol.

Os subespaços alta e normal tem entropia 0, indicando não haver dúvida na decisão de ’jogar’ ou ’não jogar’.

Subespaço Céu = chuva

Obs	Céu	Temp	Umidade	Vento	Decisão
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
10	chuva	suave	normal	não	joga
14	chuva	suave	alta	sim	não joga

Ganho de informação no subespaço Céu = chuva

Divisão pela Temperatura

$$T_{alta} = [0+, 1-]; T_{suave} = [1+, 1-]; T_{baixa} = [1+, 1-] \tag{43}$$

$$info(alta) = -(\frac{1}{1})log_2(\frac{1}{1}) = 0 \tag{44}$$

$$info(suave) = -(\frac{1}{2})log_2(\frac{1}{2}) - (\frac{1}{2})log_2(\frac{1}{2}) = 1 \tag{45}$$

$$info(baixa) = -(\frac{1}{2})log_2(\frac{1}{2}) - (\frac{1}{2})log_2(\frac{1}{2}) = 1 \tag{46}$$

$$ganho = 0,97094 - \frac{1}{5}info(alta) - \frac{2}{5}info(suave) - \frac{2}{5}info(baixa) \tag{47}$$

$$ganho = 0,97094 - \frac{1}{5}0 - \frac{2}{5}1 - \frac{2}{5}1 = 0,17090 \tag{48}$$

Ganho de informação no subespaço Céu = chuva

Divisão pela Umidade

$$T_{alta} = [1+, 1-]; T_{normal} = [2+, 1-] \tag{49}$$

$$info(alta) = -(\frac{1}{2})log_2(\frac{1}{2}) - (\frac{1}{2})log_2(\frac{1}{2}) = 1 \tag{50}$$

$$info(normal) = -(\frac{2}{3})log_2(\frac{2}{3}) - (\frac{1}{3})log_2(\frac{1}{3}) = 0,9182958 \tag{51}$$

$$ganho = 0,97094 - \frac{2}{5}info(alta) - \frac{3}{5}info(normal) \tag{52}$$

$$ganho = 0,97094 - \frac{3}{5}1 - \frac{2}{5}0,9182958 = 0,019962 \tag{53}$$

Ganho de informação no subespaço Céu = chuva

Divisão pelo Vento

$$T_{sim} = [2+, 0-]; T_{nao} = [0+, 3-] \tag{54}$$

$$info(sim) = -(\frac{2}{2})log_2(\frac{2}{2}) = 0 \tag{55}$$

$$info(nao) = -(\frac{3}{3})log_2(\frac{3}{3}) = 0 \tag{56}$$

$$ganho = 0,97094 - \frac{2}{5}info(sim) - \frac{3}{5}info(nao) \tag{57}$$

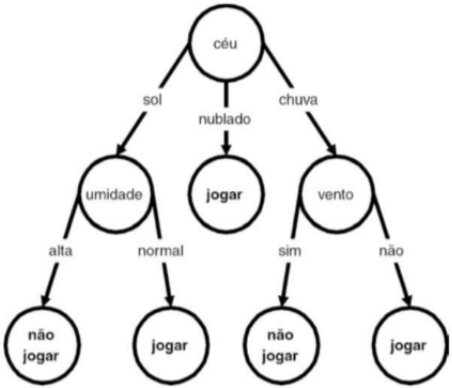
$$ganho = 0,97094 - \frac{2}{5}0 - \frac{3}{5}0 = 0,97094 \tag{58}$$

Divisão do subespaço Céu = chuva

O atributo vento apresenta o maior ganho de informação e, portanto será o segundo nó da árvore de decisão, no subespaço Céu = chuva.

Os subespaços sim e não tem entropia 0, indicando não haver dúvida na decisão de ’jogar’ ou ’não jogar’.

Árvore de decisão



Exemplo 2

Atributos categóricos e contínuos
(algoritmo C4.5)

Observações empíricas

Obs	Céu	Temp	Umidade	Vento	Decisão
1	sol	85	85	não	não joga
2	sol	80	90	sim	não joga
3	nublado	83	78	não	joga
4	chuva	70	96	não	joga
5	chuva	68	80	não	joga
6	chuva	65	70	sim	não joga
7	nublado	64	65	sim	joga
8	sol	72	95	não	não joga
9	sol	69	70	não	joga
10	chuva	75	80	não	joga
11	sol	75	70	sim	joga
12	nublado	72	90	sim	joga
13	nublado	81	75	não	joga
14	chuva	71	80	sim	não joga

Contexto

Para simplificar o exemplo, vamos assumir que a primeira divisão do espaço será feito pelo atributo Céu = *sol*, *nublado*, *chuva*.
Agora iremos analisar o ganho de informação de usar o atributo Umidade para subdividir o espaço de decisão Céu = sol.

Subespaço Céu = sol

Obs	Céu	Temp	Umidade	Vento	Decisão
1	sol	85	85	não	não joga
2	sol	80	90	sim	não joga
8	sol	72	95	não	não joga
9	sol	69	70	não	joga
11	sol	75	70	sim	joga

Pontos de partição do atributo Umidade

Ordena-se o atributo numérico:

Umidade
70
70
85
90
95

Três partições são possíveis:

- Entre 70 e 85, $v_p1 = 77,5$
- Entre 85 e 90, $v_p2 = 87,5$
- Entre 90 e 95, $v_p3 = 92,5$

Ganho de informação na partição 1

$$info(umidade < 77,5) = -\frac{2}{2}log_2(\frac{2}{2}) - \frac{0}{2}log_2(\frac{0}{2}) = 0 \quad (59)$$

$$info(umidade > 77,5) = -\frac{0}{3}log_2(\frac{0}{3}) - \frac{3}{3}log_2(\frac{3}{3}) = 0 \quad (60)$$

$$ganho = 0,97094 - \frac{2}{5} * info(umidade < 77,5) - \frac{3}{5} * info(umidade > 77,5) = 0,97094 \quad (61)$$

Ganho de informação na partição 2

$$info(umidade < 87,5) = -\frac{2}{3} log_2(\frac{2}{3}) - \frac{1}{3} log_2(\frac{1}{3}) = 0,918 \quad (62)$$

$$info(umidade > 87,5) = -\frac{0}{2} log_2(\frac{0}{2}) - \frac{2}{2} log_2(\frac{2}{2}) = 0 \quad (63)$$

$$ganho = 0,97094 - \frac{3}{5} * info(umidade < 87,5) - \frac{2}{5} * info(umidade > 87,5) = 0,420 \quad (64)$$

Ganho de informação na partição 3

$$info(umidade < 92,5) = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1 \quad (65)$$

$$info(umidade > 92,5) = -\frac{0}{1}log_2(\frac{0}{1}) - \frac{1}{1}log_2(\frac{1}{1}) = 0 \quad (66)$$

$$ganho = 0,97094 - \frac{4}{5} * info(umidade < 92,5) - \frac{1}{5} * info(umidade > 92,5) = 0,170 \quad (67)$$

Divisão do subespaço Céu = sol

A partição 1 da umidade apresentou o maior ganho de informação e, portanto será o segundo nó da árvore de decisão, no subespaço Céu = sol.

Overfitting

Quando árvores de decisão são construídas, muitos ramos ou subárvores podem conter ruídos ou erros.

Por isso, prevenir o sobreajuste é fundamental ao se modelar uma árvore de decisão. Isso pode ser feito de duas maneiras:

- Definindo restrições no tamanho da árvore
- Podando a árvore

Restrições no tamanho da árvore

- Número mínimo de amostras para uma divisão de nó: valores mais elevados impedem o modelo de aprender sobre relações que podem ser muito específicas da amostra selecionada para a árvore,
- Número mínimo de amostras para um nó de término: valores mais baixos para problemas de classe mais desbalanceados.
- Profundidade máxima da árvore: uma maior profundidade permitirá ao modelo aprender relações cada vez mais específicas.
- Número máximo de nós de término.
- Número máximo de atributos a considerar para uma divisão: valores mais altos podem levar a um sobreajuste.

Pré-poda

Regras de parada que previnem a construção daqueles ramos que não parecem melhorar a precisão preditiva da árvore.

- Todas as observações alcançando um nó pertencem à mesma classe.
- Todas as observações alcançando um nó têm o mesmo vetor de características (mas não necessariamente pertencem à mesma classe).
- O número de observações no nó é menor que um certo limiar.
- O mérito atribuído a todos os possíveis testes que particionam o conjunto de observações no nó é muito baixo.

Pós-poda

Um dos métodos mais simples é baseado em duas medidas (Bratko, 1984): o erro estático e o erro de backed-up:

- O erro estático é o número de classificações incorretas considerando que todos os exemplos que chegam nesse nó são classificados usando a classe majoritária da distribuição de classes desse nó.
- O erro de backed-up é a soma das classificações incorretas de todas as subárvores do nó corrente. Se o erro de backed-up é maior ou igual ao erro estático, então o nó é trocado por uma folha com a classe majoritária do nó.

Pseudo-algoritmo de poda

- Para cada nó interno da árvore, é calculada a taxa de erro caso esse nó vire folha (e tudo abaixo dele seja eliminado).
- Em seguida, é calculada a taxa de erro caso não haja a poda.
- Se a diferença entre essas duas taxas de erro for menor que um valor pré-estabelecido, a árvore é podada; caso contrário, não ocorre a poda.
- Esse processo se repete progressivamente, gerando um conjunto de árvores podadas.
- Por fim, para cada uma delas é calculado erro na classificação de um conjunto de dados teste, e a árvore que obtiver o menor erro será a escolhida.

Vantagens

Flexibilidade

Árvores de decisão não assumem nenhuma distribuição para os dados. Elas são métodos não paramétricos. O espaço de objetos é dividido em subespaços, e cada subespaço é ajustado com diferentes modelos. Uma árvore de decisão fornece uma cobertura exaustiva do espaço de instâncias. Havendo exemplos suficientes, pode aproximar o erro de Bayes de qualquer função.

Robustez

Árvores univaridas são invariantes a transformações (estritamente) monótonas de variáveis de entrada. Por exemplo, usar x, $log(x)$, ou e^x; como a j-ésima variável de entrada produz árvores com a mesma estrutura. Como uma consequência dessa invariância, a sensibilidade a distribuições com grande cauda e outliers é reduzida.

Seleção de atributos

O processo de construção de uma árvore de decisão seleciona os atributos a usar no modelo de decisão. Essa seleção de atributos produz modelos que tendem a ser bastante robustos contra a adição de atributos irrelevantes e redundantes.

Interpretabilidade

Decisões complexas e globais podem ser aproximadas por uma série de decisões mais simples e locais. Todas as decisões são baseadas nos valores dos atributos usados para descrever o problema. Ambos os aspectos contribuem para a popularidade das árvores de decisão.

Eficiência

O algoritmo para aprendizado de árvore de decisão é um algoritmo guloso construído de cima para baixo (top-down), usando uma estratégia dividir para conquistar sem backtracking. Sua complexidade de tempo é linear com o número de exemplos.

Desvantagens

Replicação

O termo refere-se à duplicação de uma sequência de testes em diferentes ramos de uma árvore de decisão, levando a uma representação não concisa, que também tende a ter baixa precisão preditiva. A replicação é inerente à representação da árvore de decisão.

Valores ausentes

Uma árvore de decisão é uma hierarquia de testes. Se o valor de um atributo é desconhecido, isso causa problemas em decidir que ramo seguir. Algoritmos devem empregar mecanismos especiais para abordar falta de valores.

Atributos contínuos

O gargalo do algoritmo é a presença de atributos contínuos. Nesse caso, uma operação de ordenação é solicitada para cada atributo contínuo de cada nó de decisão.

Instabilidade

pequenas variações no conjunto de treinamento podem produzir grandes variações na árvore final. A cada nó, o critério de mérito de divisão classifica os atributos, e o melhor atributo é escolhido para dividir os dados. Se dois ou mais atributos são classificados similarmente, pequenas variações da classificação dos dados podem alterar a classificação. Todas as subárvores abaixo desse nó mudam. A estratégia da partição recursiva implica que a cada divisão que é feita o dado é dividido com base no atributo de teste. Depois de algumas divisões, há usualmente muitos poucos dados nos quais a decisão se baseia. Há uma forte tendência a inferências feitas próximo das folhas serem menos confiáveis que aquelas feitas próximas da raiz.