Sistemas de Apoio à Decisão

Modelagem Preditiva: Correlação e Métodos de Segmentação

Vahid Nikoofard

Universidade do Estado do Rio de Janeiro - Faculdade de Tecnologia
Laboratório Interdisciplinar Virtual de Inteligência Artificial (LIVIA)

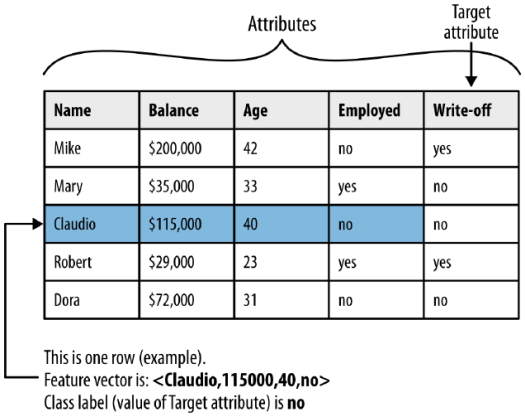
Novembro 2022

Modelos, Indução e Previsão

- Um modelo é uma representação simplificada da realidade criada para servir um propósito. Ele é simplificado com base em alguns pressupostos sobre o que é e o que não é importante para a finalidade específica ou, às vezes, com base nas limitações de informações ou tratabilidade.
- Em Data Science, um modelo preditivo é uma fórmula para estimar o valor desconhecido de interesse: o alvo. (precisão)
- Em um modelo descritivo, o principal objetivo do modelo não é estimar um valor, mas obter informações sobre o fenômeno ou processo subjacente. (inteligibilidade)

Aprendizagem Supervisionado

Aprendizagem supervisionada é a criação de um modelo que descreve uma relação entre um conjunto de variáveis selecionadas (atributos ou características) e uma variável predefinida chamada variável alvo.

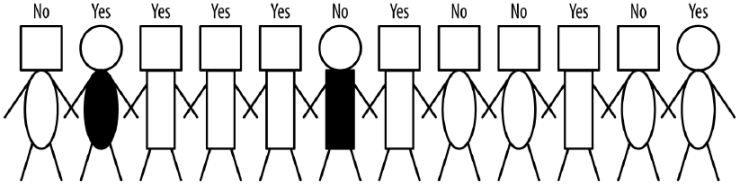


Segmentação Supervisionada

- Segmentar a população em subgrupos que possuem diferentes valores para a variável alvo (e dentro do subgrupo os exemplos possuem valores semelhantes para a variável alvo).
- A segmentação pode ser usada para previsão.
- Profissionais de meia-idade que residem na cidade de Nova York, em média, têm uma taxa de rotatividade de 5

Seleção de Atributos Informativos

- Como podemos julgar se uma variável contém informações importantes sobre a variável alvo?
- Dado um grande conjunto de exemplos, como selecionamos um atributo para dividi-los de maneira informativa?



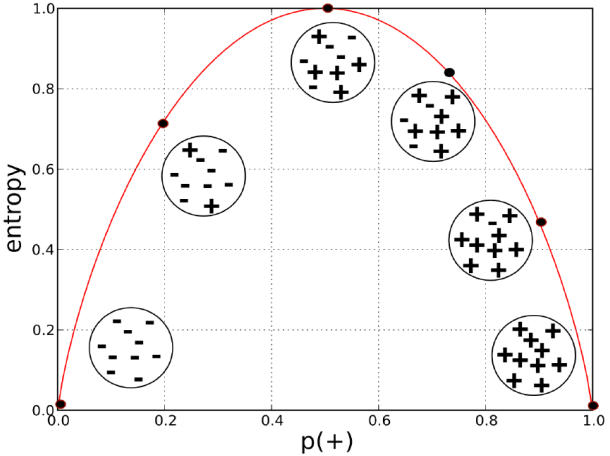
Entropia

- Medida de Pureza: entropia
- Critério de divisão: Ganho de informação.
- entropia é uma medida de desordem que pode ser aplicada a um conjunto.

$$\text{Entropia} = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$$

Onde cada p_i é a probabilidade (porcentagem relativa) da propriedade i dentro do conjunto, que varia de $p_i = 1$, quando todos os membros do conjunto têm a propriedade i, e $p_i = 0$ quando nenhum membro do conjunto tem propriedade i.

Entropia: cont.



Entropia: exemplo

considere um conjunto S de 10 pessoas, com sete sendo da classe positiva de crédito e três da classe negativa de crédito. Assim:

$$p(\text{positiva}) = \frac{7}{10} = 0.7$$
$$p(\text{negativa}) = \frac{3}{10} = 0.3$$

Então a entropia fica

$$\begin{aligned} \text{entropia}(S) &= -[0.7 \times \log_2(0.7) + 0.3 \times \log_2(0.3)] \\ &\approx -[0.7 \times -0.51 + 0.3 \times -1.74] \\ &\approx 0.88 \end{aligned}$$

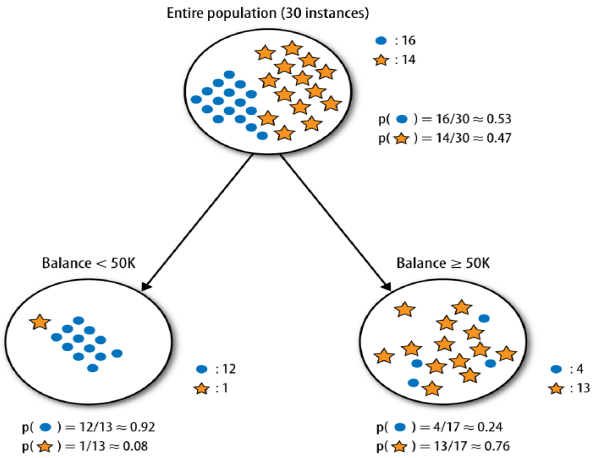
Ganho de Informação (GI)

- A entropia apenas nos diz o quanto um subconjunto individual é impuro.
- Gostaríamos de medir quão *informativo* é um atributo com relação ao nosso alvo: quanto ganho de informação isso nos dá sobre o valor da variável alvo.

$$\text{GI(pai, filhos)} = \text{entropia(pai)} - [p(c_1) \times \text{entropia}(c_1) + p(c_2) \times \text{entropia}(c_2) + ...]$$

Notavelmente, a entropia de cada filho (c_i) é ponderada pela proporção dos exemplos pertencentes a esse filho, $p(c_i)$.

Ganho de Informação (GI): exemplo1



Ganho de Informação (GI): exemplo1

considere a divisão na figura da pagina anterior. Este é um problema de duas classes ($\circ$ e $\star$). Analisando a figura, o conjunto de filhos, sem dúvida, parece "mais puro" do que o do pai. O conjunto pai tem 30 exemplos consistindo de 16 pontos e 14 estrelas, assim:

$$\begin{aligned}\text{entropia(pai)} &= -[p(\circ) \times \log_2 p(\circ) + p(\star) \times \log_2 p(\star)] \\ &\quad - [0.53 \times -0.9 + 0.47 \times -0.11] = 0.99(muitoimpuro)\end{aligned}$$

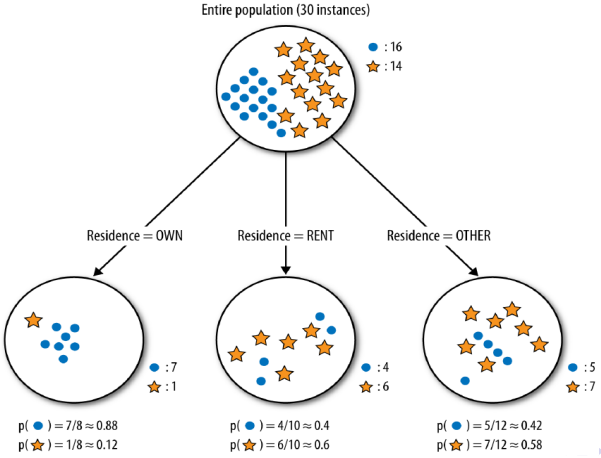
$$\begin{aligned}\text{entropia(saldo} < \text{50mil)} &= -[p(\circ) \times \log_2 p(\circ) + p(\star) \times \log_2 p(\star)] \approx 0.39 \\ \text{entropia(saldo} > \text{50mil)} &= -[p(\circ) \times \log_2 p(\circ) + p(\star) \times \log_2 p(\star)] \approx 0.79\end{aligned}$$

Então o GI é calculado

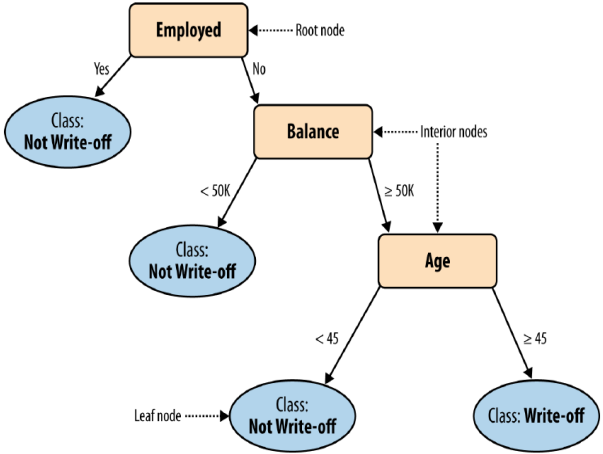
$$\begin{aligned}\text{GI} = entropia(pai) &- [p(saldo < 50mil \times entropia(saldo < 50mil) \\ &+ p(saldo > 50mil \times entropia(saldo > 50mil] \\ &\approx 0.99 - [0.43 \times 0.39 + 0.57 \times 0.79] \approx 0.37\end{aligned}$$

Ganho de Informação (GI): exemplo2

entropia(pai) ≈0.99; entropia(Residence=OWN) ≈0.54; entropia(Residence=RENT) ≈0.97; entropia(Residence=OTHER) ≈0.98; GI ≈0.13



Segmentação Supervisionada com Modelos com Estrutura de Árvore



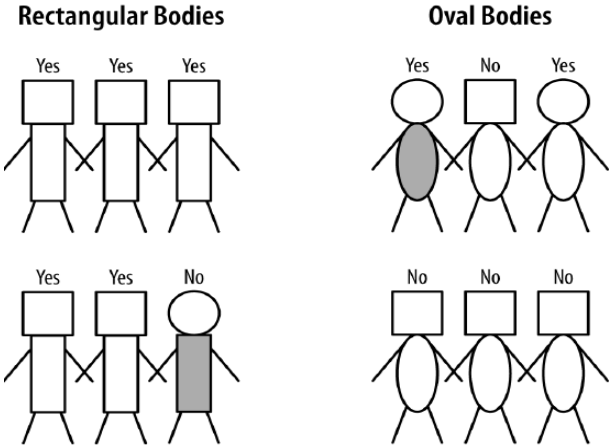
Árvores de decisão

- Árvores de classificação são um tipo de modelo estruturado em árvore.
- Árvore de estimativa de probabilidade
- Árvore de regressão.
- São intuitivas e fáceis de analisar
- Indução de árvore de decisão: segmentação supervisionada

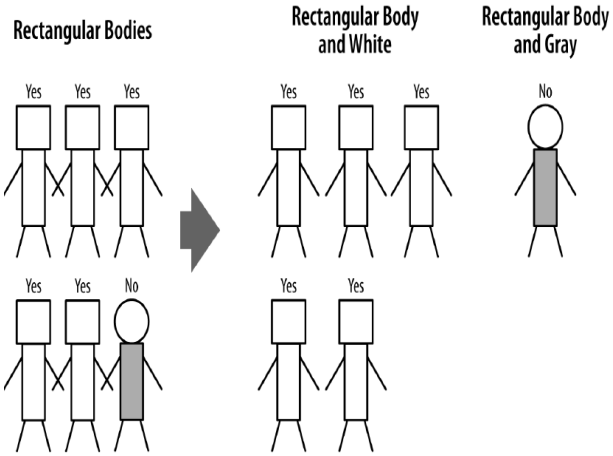
Indução de árvore de decisão

Em resumo, o procedimento de indução de árvore de classificação é um processo recursivo de dividir e conquistar, onde a meta em cada etapa é selecionar um atributo para dividir o grupo atual em subgrupos que sejam os mais puros possíveis, no que diz respeito à variável alvo. Realizamos essa divisão de forma recursiva, dividindo repetidas vezes até o fim. Escolhemos os atributos para divisão testando todos e selecionando aqueles que produzem os subgrupos mais puros. Quando terminamos? (Em outras palavras: quando paramos de dividir?) Deve ficar claro que o procedimento estará finalizado quando os nós forem puros ou quando não tivermos mais variáveis para dividir. Mas é melhor parar antes; retomaremos essa questão mais pela frente.

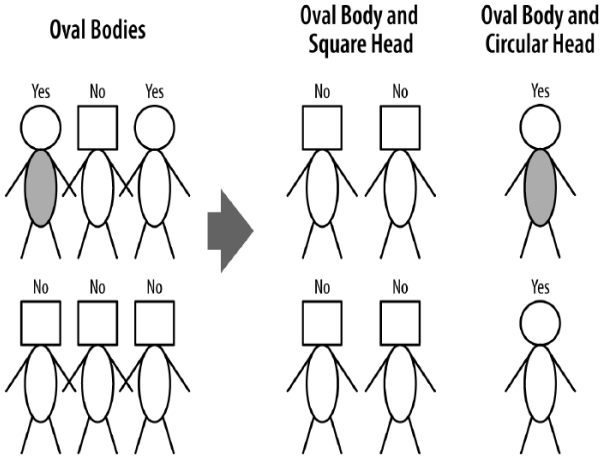
Indução de árvore de decisão: exemplo



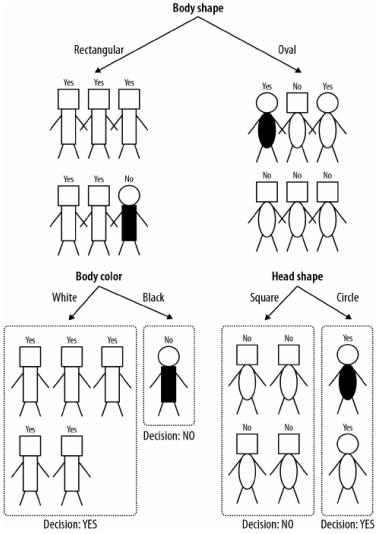
Indução de árvore de decisão: exemplo



Indução de árvore de decisão: exemplo



Indução de árvore de decisão: exemplo



Estimativa de Probabilidade

- Uma previsão mais informativa do que apenas uma classificação.
- Resolver problema de casos raros
- Um método: Estimativa baseada em frequência de probabilidade de pertencer à classe.

$$p(c) = \frac{n}{n+m}$$

- Correção de Laplace para suavizar

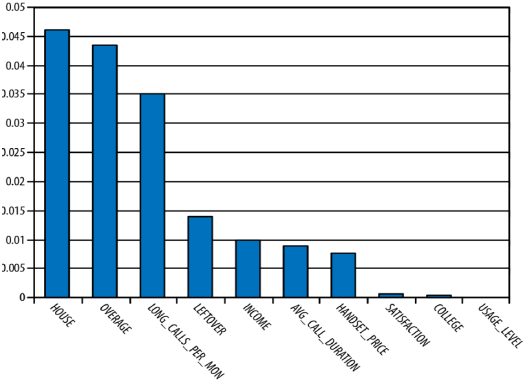
$$p(c) = \frac{n+1}{n+m+2}$$

Indução de Árvore de Decisão: exemplo de rotatividade

Variável	Explicação
FACULDADE	O cliente possui ensino superior?
RENDA	Rendimento anual
EXCESSO	Média de cobranças em excesso por mês
RESTANTE	Média de minutos sobrando por mês
CASA	Valor estimado da habitação (do censo)
PREÇO_APARELHO	Custo do telefone
LIGAÇÕES_LONGAS_POR_MÊS	Quantidade média de ligações longas (15 minutos ou mais) por mês
DURAÇÃO_MÉDIA_LIGAÇÃO	A duração média de uma ligação
SATISFAÇÃO_INFORMADA	Nível de satisfação informado
NÍVEL_USO_INFORMADO	Nível de utilização autorrelatado
ABANDONAR O SERVIÇO (*Variável alvo*)	O cliente permaneceu ou abandonou o serviço (rotatividade)?

Indução de Árvore de Decisão: exemplo de rotatividade

Ganho de informação



Indução de Árvore de Decisão: exemplo de rotatividade