IA – Eduardo

Parte 2) Classificação

Árvores de Classificação

- Atributos únicos (identificadores) terão entropia 0
- Atributos contínuos → estabelecer pontos de corte

1) Qual atributo discrimina melhor: SEXO ou PAÍS?

Se SEXO=M então Classe=Não; Senão Classe=Sim. → Acurácia (A) = 50%.

A = 50%, erra 2 Classe= Sim (M) e erra 3 Classe= Não (F) / 10

- Mas regra default (atribuir sempre CLASSE=Não) retorna → A=60%!

Se PAÍS=Inglaterra então Sim; Senão Não → A = 80%.

A = 80%, erra 2 Classe= Sim (França) / 10

Atributo meta → "Comprar": **Entropia**: E = 0,97, sendoP(+)=0,4 e P(-)=0,6 (Symbolab)

Parte 3) Regressão

Problemas de **classificação**:

- Modelo mais simples → Chutar na **moda**

Problemas de **regressão**:

Modelo mais simples → Chutar na **média**

Arvores de Regressão – exemplo

- Modelo da média: u = 3,7

Gastos Família:

MSE(Sim) = $[(1-2)^2 + \ldots + (3-2)^2] / 4 =$ **1** - u = 2

MSE(Não) = $[(7-6)^2 + (6-6)^2 + (5-6)^2] / 3 = 2/3 =$ **0,67** - u = 6

MSE = 4/7 * **1** + 3/7 * **0,67** = 0,86

Cuidado com superajuste

- Parar de aumentar a arvore quando o erro de validação aumentar

K-NN para regressão → calcular **média** dos valores dos vizinhos

---

K-means

Sensibilidade em relação a inicialização:

- k-means pode "ficar preso" em **ótimos locais**
- solução ótima local

-> Solução: Iniciar algoritmo várias vezes

Premissa: se selecionar um de cada grupo, converge p/ solução ótima

- Chance pequena: P(k!/K^k)

menor J -> melhor solução (para k fixo), se aumentar k, diminui J

Silhueta simplificada

b(i) - distancia entre objeto ao centroide do cluster vizinho mais próximo

a(i) - distancia entre objeto ao centroide ao cluster

Coeficiente de silhueta bem próximo de 1

Pode ser utilizado para selecionar o "**melhor**" número de clusters

- Selecionar o valor de **k** dando a **maior média** de s(i)