Métodos básicos de aprendizado supervisionado

Prof. Rodrigo Pedrosa

12 de dezembro de 2023

1 Leitura

1. Wikipedia - Método do mínimos quadrados
2. Wikipedia - Regressão linear
3. Wikipedia - Regressão polinomial
4. ChatGPT (Depois de estudar todas as opções anteriores)

2 Questões teóricas

Regressão linear

1. O que é a regressão linear e em que contexto é usada?
2. Dado um conjunto de dados de salários e anos de experiência em uma empresa, como seria um modelo de regressão linear simples para prever o salário com base na experiência? Como podemos interpretar os coeficientes de regressão? Como podemos avaliar a qualidade do modelo?
3. Em um conjunto de dados de vendas mensais de uma empresa, como seria um modelo de regressão linear múltipla para prever as vendas com base em variáveis como preço do produto, gastos com publicidade e desemprego na região? Caso notemos *underfitting* como podemos tentar resolver?
4. Considere a base de dados abaixo:

x_1	x_2	y
4	5	12
3	8	17
1	3	5

   (a) Escreva a expressão genérica de um modelo linear para as variáveis deste problema.

   (b) Escreva a expressão da soma do erro quadrado médio em função do pesos do modelo para a base de dados apresentada.

   (c) Apresente o gradiente da função de erro apresentada na questão anterior.

   (d) Dado o vetor de pesos $\mathbf{w} = [1, 2, 3]^t$. Qual a previsão do modelo para a entrada $\mathbf{x} = [1, 1, 1]^t$? Qual o erro absoluto total deste modelo para a base de dados apresentada.

5. Como podemos gerar modelos de regressão não-linear utilizando a técnica dos mínimos método dos mínimos quadrados

Regressão Logística

1. O que é a regressão logística e em que contexto é usada?
2. Como a função sigmoide é usada na regressão logística?
3. Como avaliar a qualidade de um modelo de regressão logística?
4. Em um conjunto de dados que descreve estudantes que se candidataram a programas de pós-graduação, cada registro contém informações sobre a pontuação do teste GRE, a pontuação no TOEFL, o GPA da graduação e o departamento escolhido. Apresente um modelo de regressão logística para prever se um candidato será admitido em um programa de pós-graduação com base em suas informações. Como podemos interpretar os coeficientes de regressão para identificar as características mais importantes para a admissão.

Árvores de decisão

1. O que é uma árvore de decisão e como ela é usada na aprendizagem de máquina?
2. Qual é a diferença entre os algoritmos de construção de árvore de decisão para classificação e para regressão?
3. O que podemos fazer para evitar o *overfitting* em árvores de decisão?
4. Considere um conjunto de dados com 100 exemplos, dos quais 60 pertencem à classe A e 40 pertencem à classe B. Calcule o índice de Gini desse conjunto de dados.
5. Em um conjunto de dados com 80 exemplos, dos quais 45 pertencem à classe X e 35 pertencem à classe Y. Calcule o índice de Gini desse conjunto de dados.
6. Suponha que um conjunto de dados seja dividido em dois subconjuntos, onde o subconjunto A contém 30 exemplos, dos quais 20 pertencem à classe P e 10 pertencem à classe Q, e o subconjunto B contém 70 exemplos, dos quais 40 pertencem à classe P e 30 pertencem à classe Q. Calcule o ganho de Gini para essa divisão com base no índice de Gini inicial do conjunto de dados.
7. Considere um conjunto de dados com duas características, "Altura"(com valores "Alto"e "Baixo") e "Idade"(com valores "Jovem"e "Adulto"), e uma classe "Classe"(com valores "A"e "B").

   Considere a seguinte tabela de dados:

Altura	Idade	Classe
Alto	Jovem	A
Alto	Adulto	A
Baixo	Jovem	B
Baixo	Adulto	B
Alto	Jovem	B
Baixo	Adulto	A

   Construa uma árvore de decisão para classificar os exemplos com base nessas características, usando o critério de Gini.
8. Dado um conjunto de dados com 100 exemplos, onde 70 pertencem à classe X e 30 pertencem à classe Y, avalie duas divisões possíveis com base no índice de Gini, e determine qual delas é mais preferível em termos de impureza.
9. Considere a seguinte base de dados:

Example	*Author*	*Thread*	*Length*	*Where_read*	*User_action*
e_1	*known*	*new*	*long*	*home*	*skips*
e_2	*unknown*	*new*	*short*	*work*	*reads*
e_3	*unknown*	*followup*	*long*	*work*	*skips*
e_4	*known*	*followup*	*long*	*home*	*skips*
e_5	*known*	*new*	*short*	*home*	*reads*
e_6	*known*	*followup*	*long*	*work*	*skips*
e_7	*unknown*	*followup*	*short*	*work*	*skips*
e_8	*unknown*	*new*	*short*	*work*	*reads*
e_9	*known*	*followup*	*long*	*home*	*skips*
e_{10}	*known*	*new*	*long*	*work*	*skips*
e_{11}	*unknown*	*followup*	*short*	*home*	*skips*
e_{12}	*known*	*new*	*long*	*work*	*skips*
e_{13}	*known*	*followup*	*short*	*home*	*reads*
e_{14}	*known*	*new*	*short*	*work*	*reads*
e_{15}	*known*	*new*	*short*	*home*	*reads*
e_{16}	*known*	*followup*	*short*	*work*	*reads*
e_{17}	*known*	*new*	*short*	*home*	*reads*
e_{18}	*unknown*	*new*	*short*	*work*	*reads*
e_{19}	*unknown*	*new*	*long*	*work*	?
e_{20}	*unknown*	*followup*	*short*	*home*	?

(a) Apresente uma árvore de decisão para a classificação das *User-actions* e calcule o grau de impureza (I_G) médio do nó raiz da sua árvore. (Obs: $I_G(p) = 1 - \sum_{i=1}^{J} p_i^2$)

(b) De acordo com a árvore apresentada, qual a classificação dos exemplos e_{19} e e_{20}?