Inteligência Artificial - Lista de Aprendizagem de Máquina

Roberto Sérgio Ribeiro de Meneses - 520403

17/02/2025

1. Deseja-se obter o modelo de regressão linear para os seguintes dados

x_1	x_2	y
0	1	3
1	2	6
2	2	7
3	1	8
4	2	11

Calcule os coeficientes da regressão.
A equação da regressão linear múltipla é dada por:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1}$$

Para encontrar os coeficientes, usamos o método dos mínimos quadrados:

1. Construção da matriz X e do vetor y

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \\ 1 & 3 & 1 \\ 1 & 4 & 2 \end{bmatrix}, \quad y = \begin{bmatrix} 3 \\ 6 \\ 7 \\ 8 \\ 11 \end{bmatrix} \tag{2}$$

2. Cálculo de $X^T X$

$$X^T X = \begin{bmatrix} 5 & 10 & 8 \\ 10 & 30 & 17 \\ 8 & 17 & 11 \end{bmatrix} \tag{3}$$

4. Cálculo da inversa de $X^T X$

$$(X^T X)^{-1} X^T = \begin{bmatrix} \frac{61}{55} & -\frac{13}{55} & -\frac{17}{55} & \frac{49}{55} & -\frac{5}{11} \\ -\frac{9}{55} & \frac{3}{55} & -\frac{3}{55} & \frac{9}{55} & \frac{2}{11} \\ -\frac{4}{11} & \frac{5}{11} & \frac{4}{11} & -\frac{2}{11} & \frac{2}{11} \end{bmatrix} \tag{4}$$

5. Cálculo dos coeficientes

Multiplicamos a inversa por $X^T y$:

$$B = (X^T X)^{-1} X^T y = \begin{bmatrix} 1,87 \\ 1.69 \\ 1.09 \end{bmatrix} \tag{5}$$

Portanto, os coeficientes da regressão são:

$$\beta_0 = 1.87, \quad \beta_1 = 1.69, \quad \beta_2 = 1.09 \tag{6}$$

A equação da regressão resultante é:

$$\hat{y} = 1.87 + 1.69x_1 + 1.09x_2 \tag{7}$$

2.Explique a necessidade de se utilizar conjuntos de dados separados para treinamento e para testes em algoritmos de aprendizado de máquina.

A separação dos conjuntos de treinamento e teste é fundamental para avaliar corretamente o desempenho de um modelo de aprendizado de máquina.

O conjunto de treino é utilizado para ajustar os parâmetros do modelo, enquanto o conjunto de teste permite medir sua capacidade de generalização para novos dados.

Essa separação evita o *overfitting*, garantindo que o modelo não apenas memorize os exemplos vistos, mas também consiga fazer previsões precisas em dados desconhecidos.

3. Papel das Épocas de Treinamento em um Algoritmo de Aprendizado de Máquina

As épocas de treinamento desempenham um papel fundamental no ajuste dos modelos de aprendizado de máquina. Durante cada época, o modelo processa todo o conjunto de dados de treinamento e ajusta seus parâmetros para minimizar o erro. Esse processo pode ser descrito pelos seguintes aspectos:

- **Ajuste progressivo**: A cada época, os parâmetros do modelo são refinados com base no erro cometido, permitindo um aprendizado gradual e mais preciso.
- **Redução de erro**: Com o passar das épocas, o modelo melhora sua capacidade de generalização, reduzindo a diferença entre os valores previstos e os reais.
- **Convergência**: O monitoramento da perda ao longo das épocas permite verificar se o modelo está convergindo para uma solução estável ou se há sinais de overfitting ou underfitting.

4. Os seguintes dados (círculos claros) formam um conjunto de dados onde deseja-se descobrir dois agrupamentos. Para esta tarefa foi utilizado o algoritmo k-médias. Inicializando o algoritmo nos centróides representados pelos círculos escuros, esboce a posição final destes centróides, determine os dados pertencentes a cada cluster e explique se o resultado foi satisfatório.

Considere os pontos representados pelos círculos claros e os centróides iniciais (círculos escuros) no diagrama abaixo. O objetivo é encontrar dois agrupamentos ($k = 2$). A seguir, descrevemos o processo e o resultado:

Esboço da posição final dos centróides

- Após a primeira atribuição de rótulos, cada ponto é associado ao centróide mais próximo em termos de distância euclidiana.
- Em seguida, recalculam-se as posições dos centróides como a média dos pontos que lhe foram atribuídos.
- Este processo é repetido até a convergência, isto é, até que as posições dos centróides não se alterem significativamente.

No esboço final, cada círculo claro (ponto de dado) estará colorido ou marcado de acordo com o cluster ao qual pertence, e os centróides (círculos escuros) se deslocam para as regiões mais densas de cada grupo.

Dados pertencentes a cada cluster

- **Cluster 1**: Agrupa os pontos localizados mais à esquerda (ou na região superior, dependendo da distribuição exata dos dados). O centróide deste cluster estará aproximadamente no meio destes pontos.
- **Cluster 2**: Agrupa os pontos localizados mais à direita (ou na região inferior). O centróide deste cluster estará no meio destes pontos, após as iterações do algoritmo.

Discussão do resultado

Embora o algoritmo k-médias tenha separado os dados em dois grupos, é possível observar que o **Cluster 1** engloba, na verdade, dois subconjuntos de pontos que acabam sendo forçados a pertencer a um único cluster. Isso indica que o resultado não foi totalmente satisfatório, pois há evidências de que o conjunto de dados poderia ser melhor representado por mais de dois agrupamentos, ou que a escolha dos centróides iniciais e parâmetros não foi adequada para capturar a estrutura real dos dados. Assim, mesmo que se observe uma separação básica entre dois grandes conjuntos, a presença desses dois subconjuntos em um mesmo cluster sugere que:

- É possível que o número de clusters (k) tenha sido subestimado, não refletindo a real distribuição dos pontos.
- A inicialização dos centróides possa ter direcionado o algoritmo para uma solução subótima, fazendo com que subgrupos distintos fossem agrupados indevidamente.
- Poderiam ser necessárias técnicas de pré-processamento ou métricas de similaridade diferentes para melhorar a qualidade do agrupamento.

6. Esboçar o dendrograma utilizando a distância euclidiana para os seguintes dados:

x_1	x_2
-3	0
-3	2
-3	-2
1	4
2	2
2	-2
2	-3
3	4
3	2
4	-1
4	-3

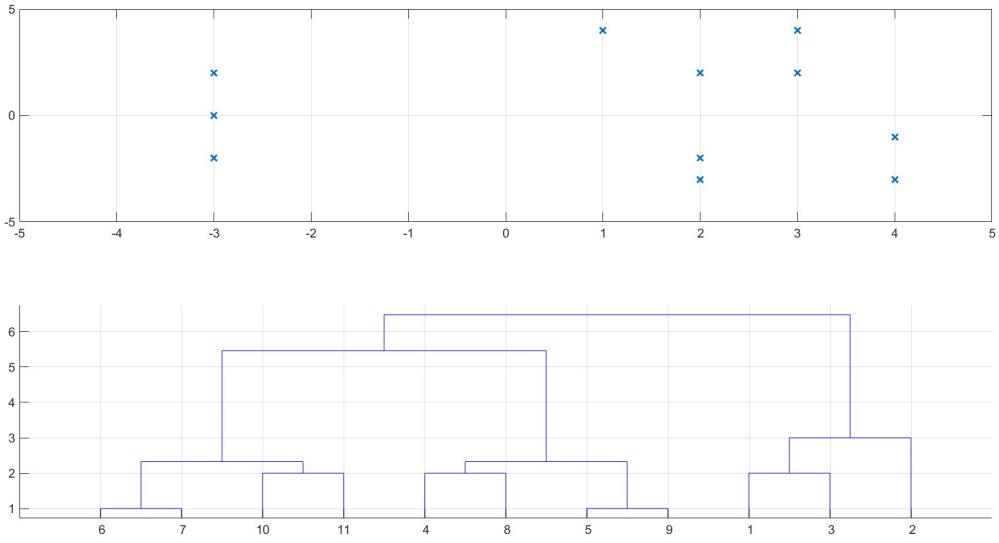

Figura 1: Rede Bayesiana com a solução.