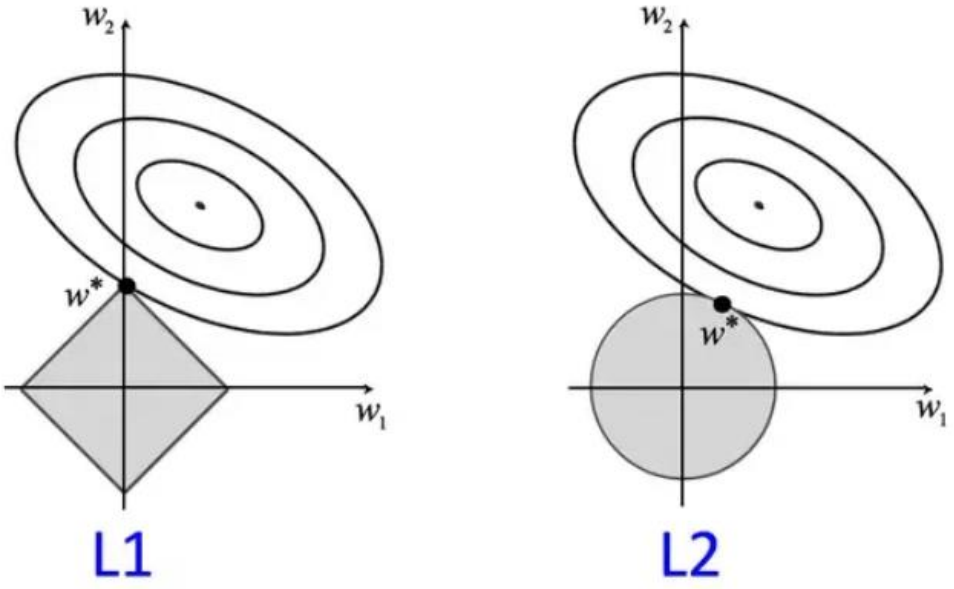


Regularización

Aprendizaje esperado.

El alumno aprenderá que la regularización es un método que permite evitar un posible sobreajuste del modelo (overfitting). Por ello, su importancia para la validación del modelo a fin de lograr un correcto análisis predictivo.

Contenido

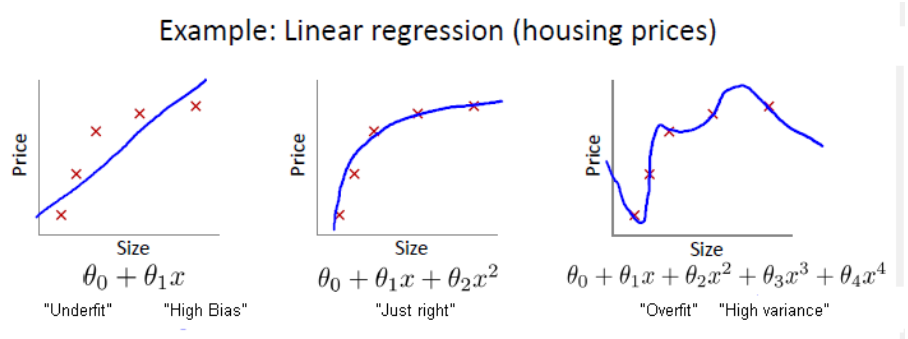
Regularización

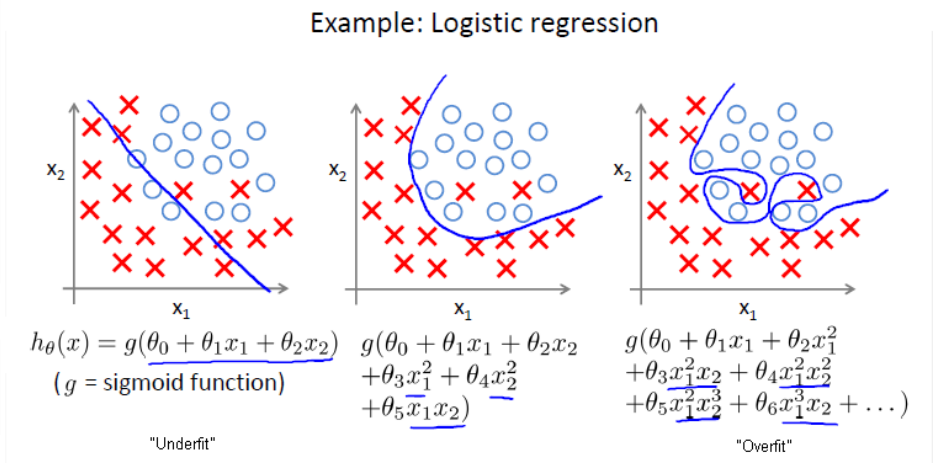
1. El problema del sobreajuste
2. Tratamiento del sobreajuste
3. Idea intuitiva
4. Regresión lineal con regularización L2 (Ridge)
5. Gradiente descendiente para regresión lineal con regularización L2
6. Ejemplos ilustrativos 1, 2 y 3
7. Regresión lineal con regularización L1 (Lasso)
8. Ejemplos ilustrativos 4 y 5
9. Regularización con Elastic-Net
10. Regresión logística con regularización L2

El problema del sobreajuste

Sobreajuste: Se produce cuando la hipótesis o modelo aprendido se ajusta bien a los datos de entrenamiento, pero falla en la generalización a nuevos ejemplos.

Una de las causas del sobreajuste es la utilizacion de **demasiadas características.**



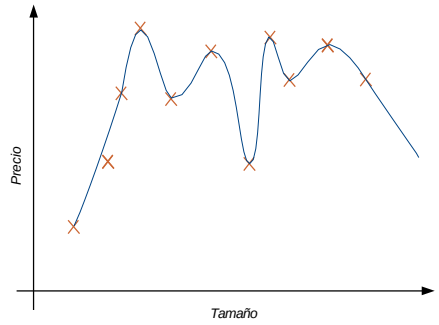


Tratamiento del sobreajuste

1. Reducción del número de características
   - ➜ Seleccionar manualmente las características
   - ➜ Algoritmo de selección de modelos

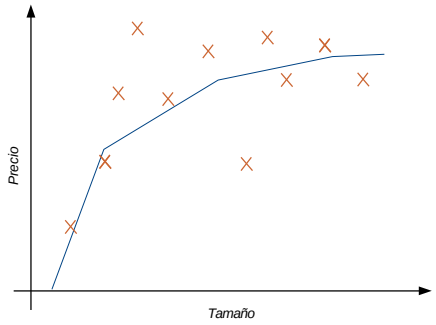
2. Regularización
   - ➜ Mantener todas las características, pero reducir la magnitud de los valores de los parámetros θ_j
   - ➜ Funciona bien si se dispone de muchas características, y todas ellas contribuyen en cierto grado a predecir el valor de salida y.

Idea intuitiva



$\theta_3 \approx 0, \theta_4 \approx 0$:

$\Longrightarrow$



$\theta_0+\theta_1 x+\theta_2 x^2+\theta_3 x^3+\theta_4 x^4$

$\theta_0 + \theta_1 x + \theta_2 x^2$

¿Cómo introducir esta idea en el proceso de aprendizaje?

Si se penalizan θ_3 y θ_4, estos parámetros tenderán a valores pequeños, salvo que ésto suponga un aumento considerable en la función de coste:

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + 1000\theta_3 + 1000\theta_4$$

Regresión lineal con regularización L2 (Ridge)

- Valores más pequeños para los parámetros:
  - Hipótesis más simples
  - Menos tendencia al sobreajuste

- Para conseguirlo, se modifica la **función de coste** introduciendo un término que **penaliza la complejidad**. Obtenemos:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right],\ \lambda > 0$$

- El objetivo es, por tanto:

$$\min_\theta J(\theta) = \min_\theta \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

¿Qué ocurriría si se fija un valor extremadamente alto para λ?

Gradiente descendiente para regresión lineal con regularización L2

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{m}\left(\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) \quad \text{para } j = 0$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\left[\left(\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) + \lambda\theta_j\right] \quad \text{para } j \geq 1$$

inicializar aleatoriamente θ

repetir hasta convergencia {

$$\theta_0 := \theta_0 - \alpha\frac{1}{m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_0^{(i)}\right]$$

$$\theta_j := \theta_j - \alpha\frac{1}{m}\left[\left(\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) + \lambda\theta_j\right]$$

}

Regresión lineal con regularización L2 (Ridge)

Regresión lineal con regularización L2

$$\min_{\theta} J(\theta) = \min_{\theta} \frac{1}{2m} \left[\sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right]$$

Ecuación normal

$$X = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix} \qquad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

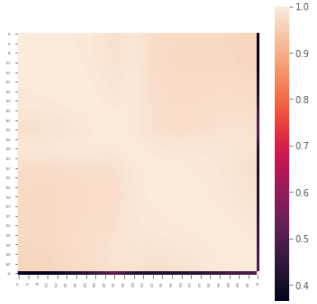
$$\min_{\theta} J(\theta) \Rightarrow \frac{\partial}{\partial \theta_j} J(\theta) = 0$$

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \right)^{-1} X^T Y$$

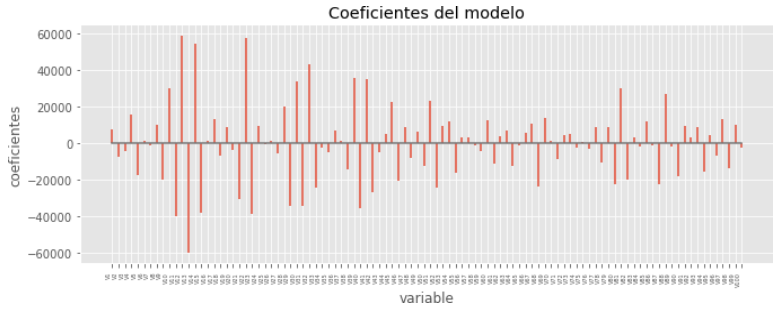
El tamaño de la matriz diagonal es $(n+1) \times (n+1)$

Ejemplo ilustrativo 1

- Problema con 100 atributos altamente correlacionados.



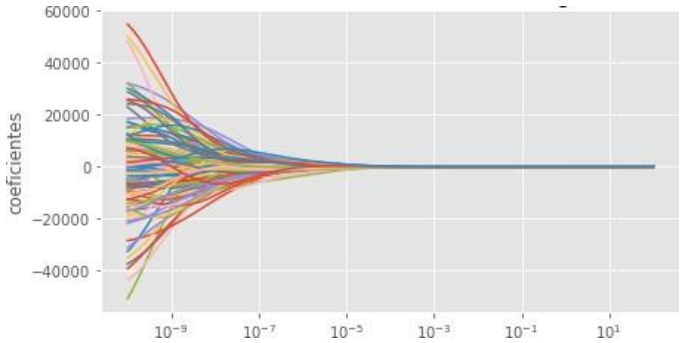
- Coeficientes estimados por mínimos cuadrados (OLS):



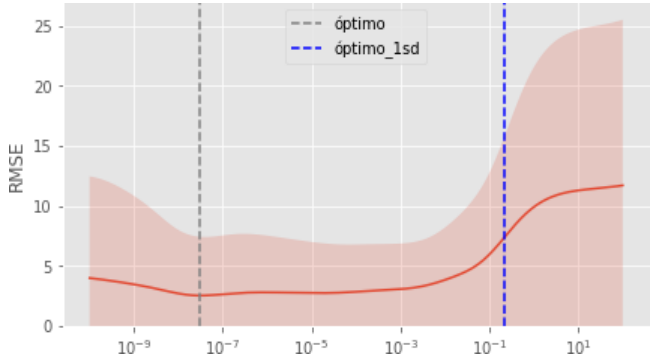
[1]Figuras tomadas de https://www.cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python.html

Ejemplo ilustrativo 2

➜ Evolución de los coeficientes en función de λ:

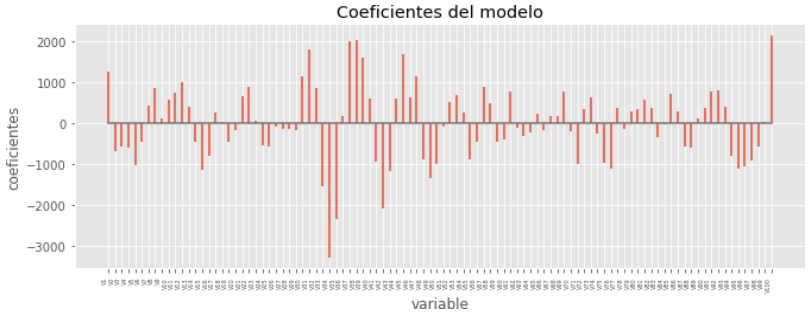


➜ Evolución del error en función de λ:

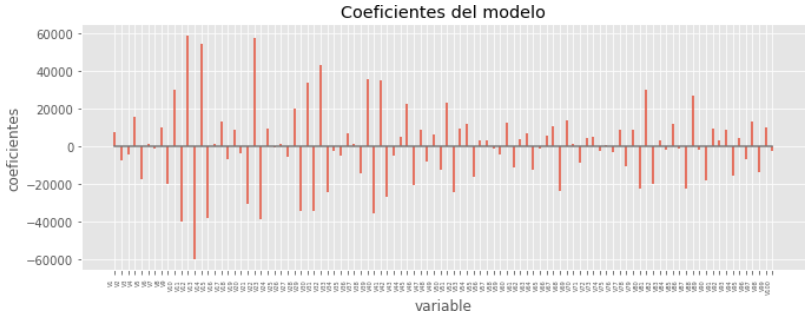


Ejemplo ilustrativo 3

- Valor óptimo para λ: 2.9673e-08



- El orden de magnitud de los coeficientes es mucho menor comparando con regresión sin regularización:



Regularización L2 (Ridge): resumen

- Necesario estandarizar las variables.
- Muy útil cuando algunas de las variables predictoras están correlacionadas.
  - Disminuir los coeficientes tiene el efecto de disminuir la correlación entre las variables.
- Funciona mejor cuando la mayoría de los atributos son relevantes.
- El parámetro λ tiene mucha influencia $\Rightarrow$ ajustar.
- Con un valor de λ adecuado, el método de Ridge es capaz de reducir varianza sin apenas aumentar el sesgo (bias), consiguiendo así un menor error total.
- Ridge no supone ganar en interpretabilidad, puesto que (casi) todos los atributos son incluidos en el modelo.

Regresión lineal con regularización L1 (Lasso)

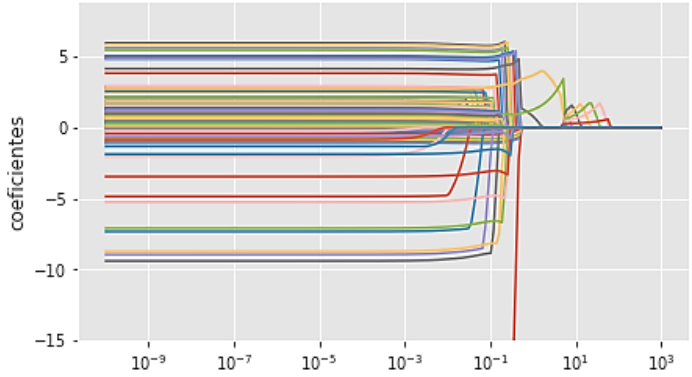
- La idea es similar a Ridge, cambia la norma L2 por L1.
- En la regularización Lasso o L1, la complejidad *C* se mide como la media del valor absoluto de los coeficientes del modelo.
- Ahora la función de coste a minimizar es:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}|\theta_j|\right],\ \lambda > 0$$

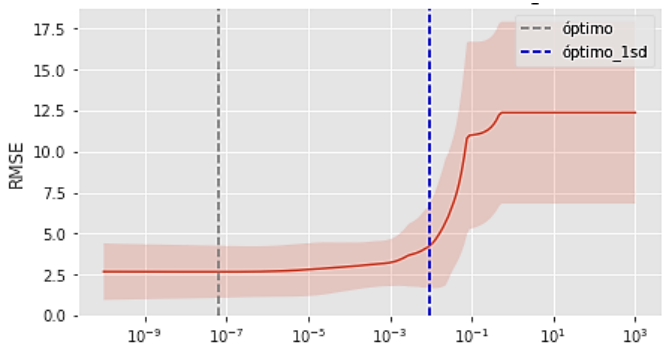
- Lasso es muy útil cuando sospechamos que varios de los atributos son irrelevantes.
- Algunos de los coeficientes acabarán valiendo 0, filtrando así atributos irrelevantes.
- Se obtiene un modelo menos denso, más comprensible, que generaliza mejor.
- Lasso funciona mejor cuando hay poca correlación entre los atributos.

Ejemplo ilustrativo 4

- Evolución de los coeficientes en función de λ:

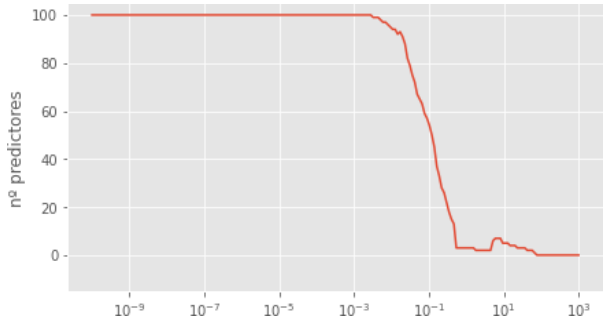


- Evolución del error en función de λ:

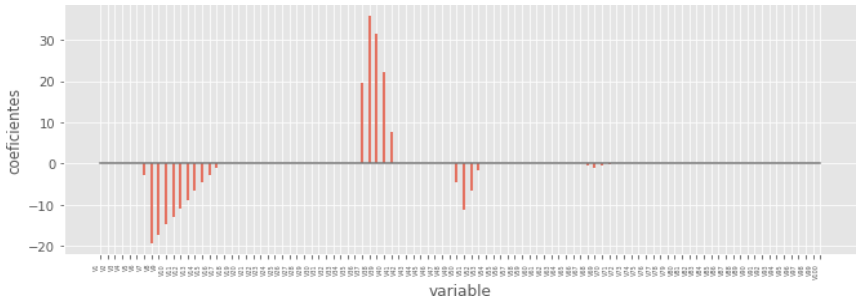


Ejemplo ilustrativo 5

- Evolución del número de variables incluidas en función de λ:



- Valor óptimo para λ: 6.4423e-08. (+1sd = 0.0093293)
- Solo 24 variables son incluidas en el modelo final.

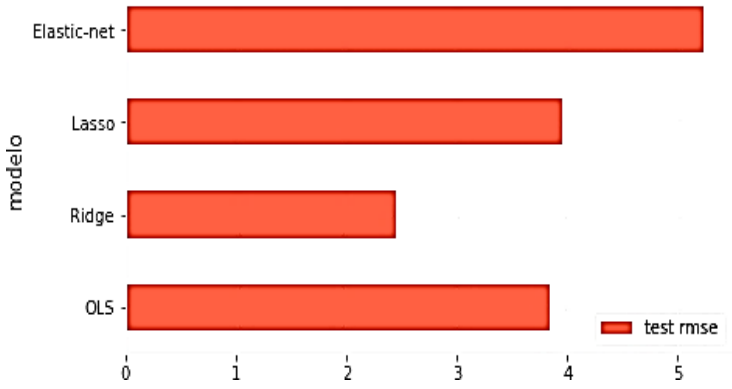


Regularización con Elastic-Net

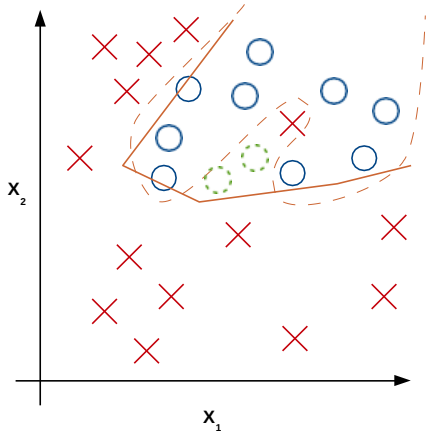
- **Elastic-Net** puede verse como una combinación convexa de L1 y L2.
- El término de penalización es ahora:

$$C = \frac{1}{2m}\left(r \cdot (\lambda \sum_{j=1}^{n} |\theta_j|) + (1-r)(\lambda \sum_{j=1}^{n} \theta_j^2)\right)$$

- Es efectiva cuando tengamos un gran número de atributos, algunos de ellos irrelevantes y otros redundantes (correlacionados entre ellos).
- En el ejemplo:



Regresión logística con regularización L2



$$h_\vartheta(x) = \frac{1}{1 + e^{-(\vartheta_0 + \vartheta_1 x + \vartheta_2 x^2 + \vartheta_3 x^3 + \vartheta_4 x^4)}}$$

Función de coste:

$$J(\theta) = \frac{1}{m}\left[\sum_{i=1}^{m} -y^{(i)} \log h_\theta(x^{(i)}) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta_j^2$$

Gradiente descendiente para regresión logística con regularización L2

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{m}\left(\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) \quad \text{para } j = 0$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\left[\left(\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) + \lambda\theta_j\right] \quad \text{para } j \geq 1$$

inicializar aleatoriamente θ

repetir hasta convergencia {

$$\theta_0 := \theta_0 - \alpha\frac{1}{m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_0^{(i)}\right]$$

$$\theta_j := \theta_j - \alpha\frac{1}{m}\left[\left(\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) + \lambda\theta_j\right]$$

}

Logros.

Al término de la sesión el alumno tendrá una visión global y clara de la regularización como un método que le permite restringir el proceso de estimación, se usa para evitar un posible sobre ajuste del modelo (overfitting).

.

Conclusiones.

Las técnicas de regularización nos ayudan a minimizar los problemas que hemos descrito realizando restricciones en los coeficientes de regresión del modelo, lo que ayuda a controlar su complejidad y a evitar que los coeficientes tomen valores extremos..

Referencias.

1. Ian h: Witten, Eibe Frank, Mark A. Hall (2011). Data Mining. Third Edition. Chapter 6.
2. Gorunescu Florin (2017). Data Mining: Concepts, Models and Terchniques. Intellligent Systems Reference Library Volume12. Chapter 5.
3. Aurélien Géron (2019). Hands-On Machine Learning with Scikit-Learn and TensorFlow. 2nd Edition. Chapter 9.
4. Aggarwal & Reddy (2014). Data Clustering , Algorithms and Applications. CRC Press.