Aprendizaje de máquina

Regularización y entrenamiento

Leonardo Flórez-Valencia

Facultad de ingeniería

Regularización

1. Cresta (*ridge*, o norma L^2)

$$J_R(\mathbf{w}, b) = J(\mathbf{w}, b) + \frac{\lambda}{[m]} \left(\sum_{j=1}^{n} \left[w^j \right]^2 + b^2 \right)$$

2. *Least Absolute Shrinkage and Selection Operator* (LASSO, o norma L^1)

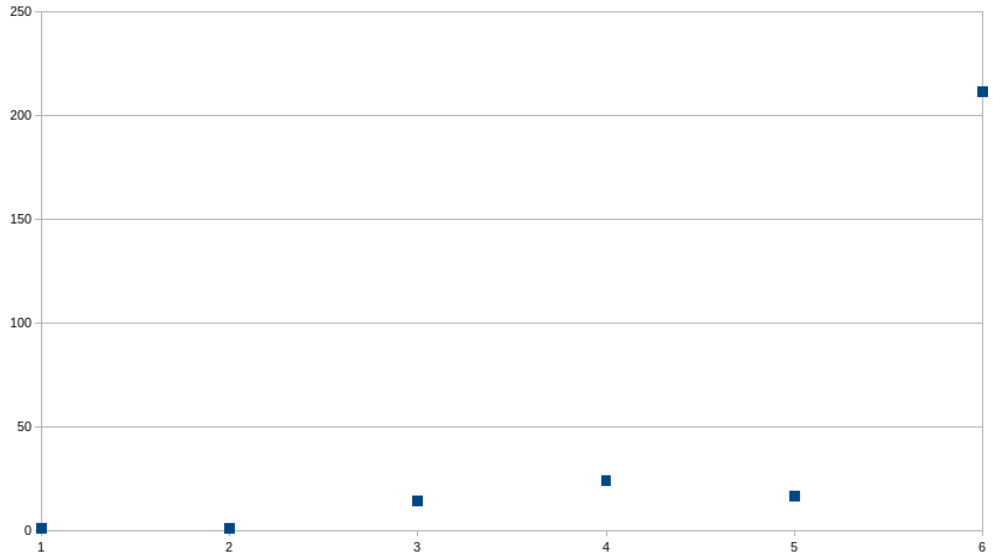
$$J_R(\mathbf{w}, b) = J(\mathbf{w}, b) + \frac{\lambda}{[m]} \left(\sum_{j=1}^{n} \left| w^j \right| + |b| \right)$$

¿Qué es un buen entrenamiento?

- Hay 3 hiperparámetros reales: $\epsilon \approx 0$, $\alpha \in (0, 1)$ y $\lambda \geq 0$.
- ¿Cómo configurar cada valor?
- Tres configuraciones algorítmicas:
  - ***Función de activación***.
  - Tipo de regularización.
  - Tipo de normalización.
- ¿Cuál escoger?
- Los resultados pueden ser:
  - Perfectos.
  - Imperfectos.
  - ***Satisfactorios***.

Entrenamiento imperfecto

- Más ejemplos $\mathbf{X}$, $\mathbf{y}$.
- Cambiar la cantidad de variables.
- Adicionar características polinomiales.
- Cambiar λ: Resultados de alto sesgo o alta varianza.



Matriz de confusión

		Realidad	
		0	1
Predicción	0	Tz	Fz
	1	Fo	To

1. Sensibilidad (clase 0):

$$\frac{Tz}{Tz + Fo}$$

2. Especificidad (clase 1):

$$\frac{To}{To + Fz}$$

3. Exactitud:

$$\frac{Tz + To}{Tz + To + Fz + Fo}$$

4. Calificación F1 (*F1-score*):

$$\frac{2To}{2To + Fz + Fo}$$

Grupos de entrenamiento

De los ejemplos de entrenamiento:

$$\mathbf{X} = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^n \\ x_2^1 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \vdots \\ x_m^1 & x_m^2 & \cdots & x_m^n \end{bmatrix} \in \mathbb{R}^{m \times n}$$

extraer ***aleatoriamente*** los ejemplos $\mathbf{X}_{tra} \in \mathbb{R}^{m_{tra} \times n}$ (entrenamiento), $\mathbf{X}_{val} \in \mathbb{R}^{m_{val} \times n}$ (validación) y $\mathbf{X}_{tes} \in \mathbb{R}^{m_{tes} \times n}$ (prueba), donde $m_{tra} \approx 0.6m$, $m_{val} \approx 0.2m$ y $m_{tes} \approx 0.2m$

$$m = m_{tra} + m_{val} + m_{tes}$$

Selección del modelo

Aumentar con características polinomiales

1. Estimar $[\mathbf{w}, b]$ con $\mathbf{X}_{tra}$, medir $J_{val}(\mathbf{w}, b)$
2. Estimar $\left[\mathbf{w} \bigcup w_1^2, b\right]$ con $\mathbf{X}_{tra}$, medir $J_{val}\left(\mathbf{w} \bigcup w_1^2, b\right)$
3. Estimar $\left[\mathbf{w} \bigcup w_1^2 \bigcup w_2^2, b\right]$ con $\mathbf{X}_{tra}$, medir $J_{val}\left(\mathbf{w} \bigcup w_1^2 \bigcup w_2^2, b\right)$
4. Quedarse con el J_{val} mínimo.

Sesgo vs. varianza

- Sesgo: bajo ajuste (*fit*) de los datos.
  - J_{tra} alto, $J_{val} \approx J_{tra}$
- Varianza: algo ajuste (*fit*) de los datos.
  - J_{tra} bajo, $J_{val} \gg J_{tra}$
- "Jugar" con $\lambda = [0.1, 1, 10, 100, 1000, \cdots]$
  - Estimar $[\mathbf{w}, b]$ con regularización a partir de $\mathbf{X}_{tra}$.
  - Medir $J_{val}\,(\mathbf{w}, b)$ sin regularización ($\lambda = 0$).
  - Quedarse con el λ que minimiza J_{val}.

Taller 1: identificar dígitos

1. Recuperar la base de datos MNIST.
2. Proponer un sistema, que use únicamente la regresión logística, para identificar los dígitos del 0 al 9:
    1. Diseñar el proceso de entrenamiento.
    2. Ejecutar el proceso de entrenamiento.
    3. Implementar un sistema simple para usuarios finales.
3. Escribir un documento corto con sus observaciones, resultados, conclusiones y perspectivas.