Modelos Estadísticos para Clasificación y Regresión

Práctico 2 - Regresión lineal

IMERL - FIng

30 de agosto de 2023

Aprendizaje supervisado

Se tiene un conjunto de datos **etiquetados**.
Es decir que el conjunto de datos consiste en pares $(\vec{X}_i, y_i)$, donde $\vec{X}$ contiene los atributos del dato i e y su etiqueta.

Ejemplo: conjunto de datos *Boston Housing*.

- CRIM - per capita crime rate by town
- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per $10,000
- PTRATIO - pupil-teacher ratio by town
- B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in $1000's

Objetivo

Generalizar

Objetivo

Generalizar

Aprender o inferir una relación funcional $f : \vec{X} \rightarrow y$ entre el conjunto de atributos $\vec{X}$ y sus etiquetas y correspondientes.

Con un modelo que generalice bien podremos **predecir** la etiqueta de cualquier otro conjunto de datos no etiquetados.

Preguntas

- ¿Cómo se obtiene un modelo que generalice?
- ¿Qué quiere decir que un modelo generalice **bien**?

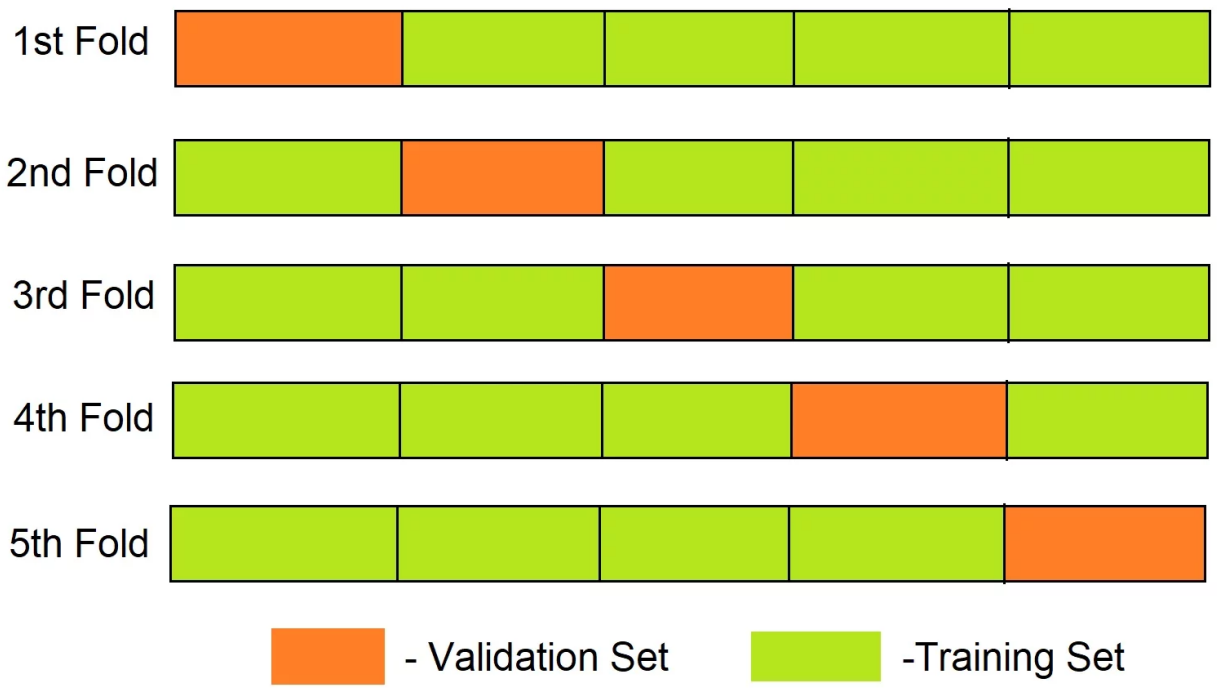
Obtención de un modelo

1. Elegir un modelo.
   Para esto nos podemos basar en ciertas hipótesis sobre la relación funcional. Por ejemplo, podemos asumir que existe una relación **lineal** entre el año y el número promedio de transistores en un microprocesador.
2. Definir una métrica o función de costo, que nos permita decidir qué tan **bueno** es un modelo.
3. Separar en conjuntos de entrenamiento, validación y test.
4. Entrenar: ajustar los parámetros del modelo a los datos.
5. Validar: observar el desempeño del modelo en un conjunto de datos con los que **no** haya entrenado. Esto permite comparar distintos modelos y seleccionar **hiperparámetros**.
6. Evaluar en el conjunto de test.

Validación cruzada

Si nuestro conjunto de datos es muy pequeño, no es conveniente apartar datos para validación que no podremos usar para entrenamiento.

Solución: utilizar un esquema de validación cruzada.



Sobreajuste y regularización

Modelos muy complejos pueden llevar a **sobreajuste**, que impide la generalización.
Solución: regularizar, penalizando la complejidad de los modelos.
Ejemplo: regularización de Ridge.

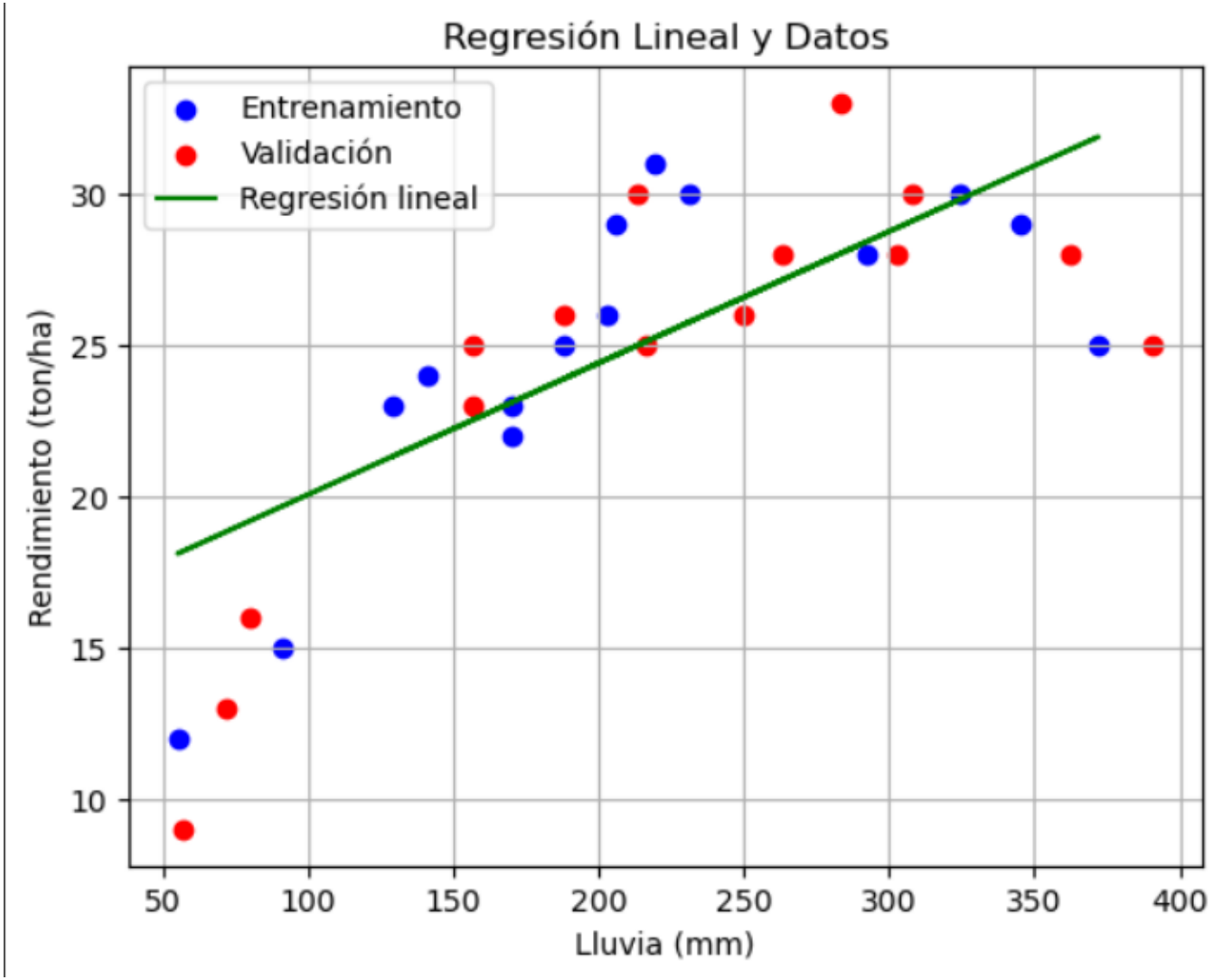
$$\mathcal{L}' = \mathcal{L} + \lambda \mathbf{w}^\top \mathbf{w}.$$

Práctico 2

- Modelo: lineal (en los parámetros!). Atributos polinomiales.
- Métrica de evaluación: *MSE*
- Regularización de Ridge
- Validación cruzada 5-folds
- Uso de Pipelines de SkLearn para definir series de transformaciones de los datos, seguidas de un estimador.

Preguntas?

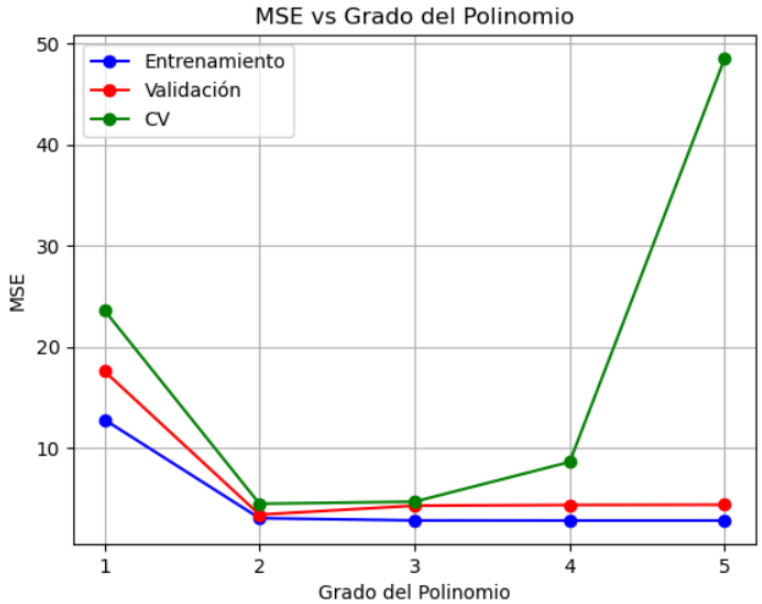
1. Regresión lineal

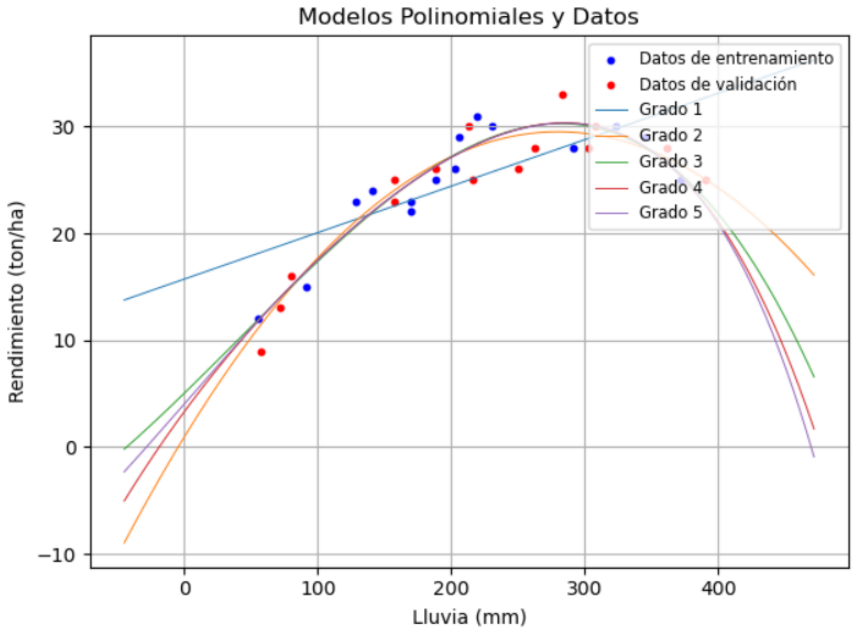


2. MSE en entrenamiento, validación y CV

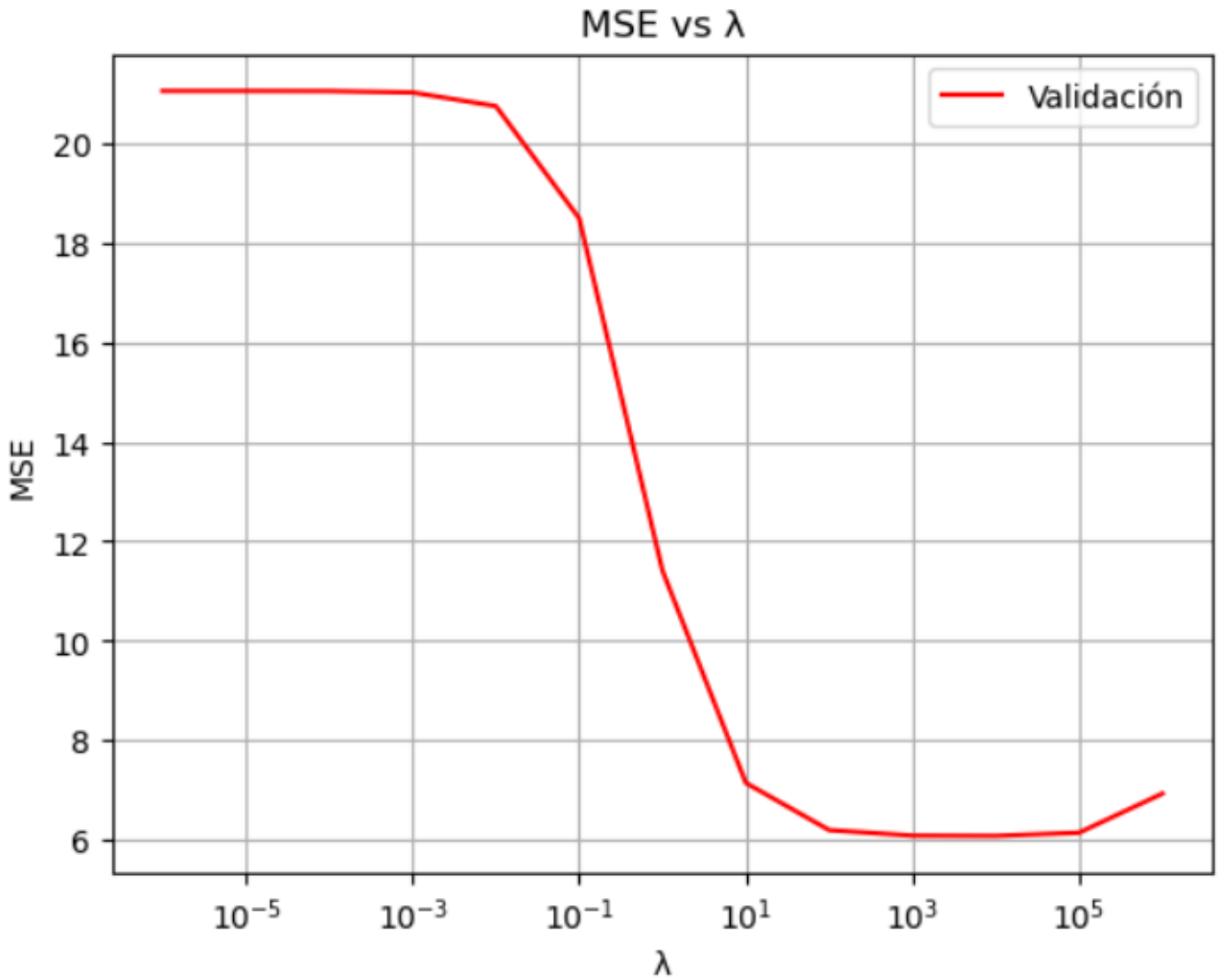
```
MSE Entrenamiento: 12.780502233244631
MSE Validación: 17.578253618682638
MSE CV: 23.60932500511161
```

3. Determinar el grado óptimo





4. λ óptimo para regresión polinomial de grado 5 con regularización



4. λ óptimo para regresión polinomial de grado 5 con regularización