Prácticas de Aprendizaje Automático

Práctica 3

Ajuste de Modelos Lineales

Pablo Mesejo
Universidad de Granada
Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

Ajuste de Modelos Lineales

- Ajuste y selección del mejor modelo lineal, y estimación del error E_{out} del modelo final

- **Casuística (relativamente) real**: te llega un problema y… ¿cómo lo resuelves?
  - **Análisis** del Problema, **Exploración** de los Datos, Formulación de **Hipótesis**, **Entrenamiento**, **Validación**, y **Discusión** de Resultados

Ajuste de Modelos Lineales

– Problema de clasificación

https://archive.ics.uci.edu/ml/datasets/Dataset+for+Sensorless+Drive+Diagnosis

Dataset for Sensorless Drive Diagnosis Data Set

Download: Data Folder, Data Set Description

Abstract: Features are extracted from motor current. The motor has intact and defective components. This results in 11 different classes with different condit

Data Set Characteristics:	Multivariate	**Number of Instances:**	58509	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	49	**Date Donated**	2015-02-24
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	81488

– Problema de regresión

https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data

Superconductivty Data Data Set

Download: Data Folder, Data Set Description

Abstract: Two file s contain data on 21263 superconductors and their relevant features.

Data Set Characteristics:	Multivariate	**Number of Instances:**	21263	**Area:**	Physical
Attribute Characteristics:	Real	**Number of Attributes:**	81	**Date Donated**	2018-10-12
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	53055

1. Analizar y **comprender el problema**

   a) ¿Qué es X? ¿Qué es Y? ¿En qué consiste el problema que tengo que resolver (f: X → Y)?

   b) ¡Visualizar datos! t-SNE sería una opción…

   c) ¿Cómo se relacionan los datos? ¿Hay correlaciones?

2. Selección de clases de **funciones a usar**

   a) sabemos que vamos a usar modelos lineales, pero ¿qué combinaciones/transformaciones lineales/no-lineales de los valores observados vamos a emplear? ¿Por qué?

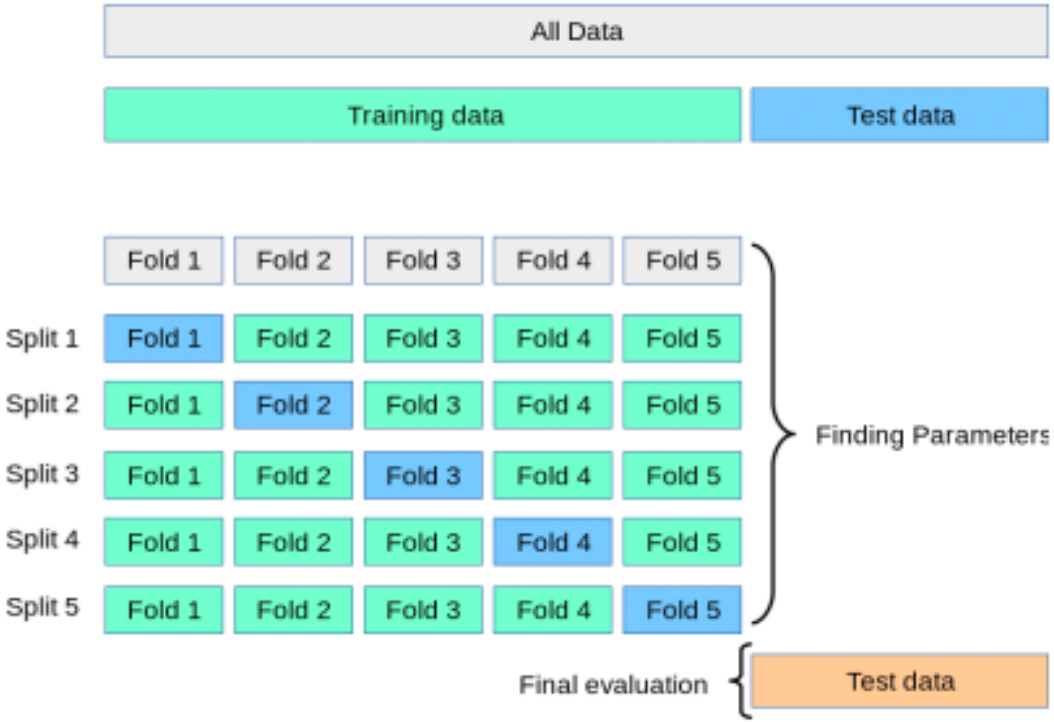
   b) ¿qué modelos concretos planteamos usar? → **Identificar qué modelos vamos a emplear**!

3. Definición de los conjuntos de **training, validación y test**

   a) Si la base de datos ya define conjuntos de training y test → unirlos y definir conjuntos de training y test propios. Más adelante, ¿comparar los resultados con los nuevos conjuntos de datos que con los conjuntos predefinidos?

   b) ¿Uso de cross-validation? ¿Por qué?

5-fold cross-validation



https://scikit-learn.org/stable/modules/cross_validation.html

4. **Preprocesado de los datos**: todas las manipulaciones sobre los datos iniciales que nos permitan fijar el conjunto de vectores de características que se usarán en el entrenamiento.
   a) eliminar datos sin variabilidad,
   b) reducción/aumento de dimensionalidad,
   c) normalización o codificación de datos, datos faltantes, datos extremos,…
5. Fijar la **métrica de error** a usar. Discutir su idoneidad para el problema
   a) MSE, MAE, Accuracy,…
6. Discutir todos los **parámetros** y el tipo de **regularización** a usar.
   a) Discutir la idoneidad de los valores de los parámetros de la técnica de ajuste. No podéis emplear los métodos por defecto incluidos en Scikit-learn sin saber qué hacen.
      - learning rate, tamaño de minibatch, criterio de parada, etc.
   b) L2 regularization /weight decay /ridge regression, L1 regularization

7. **Estimación de hiperparámetros** y selección de la mejor hipótesis.
    a) Entrenamiento (Ein) y Estimación de Eout
    b) Posibilidades de análisis:
        a) Comparar el Eout de la selección de modelos (validación cruzada) con el Eout de la mejor hipótesis
        b) Comparar el Eout obtenido con distintos porcentajes de training y test
        c) Emplear baselines con los que comparar.
            a) Por ejemplo, si tengo un 3% de Etest no sé si es mucho, porque a lo mejor un estimador *naive* (la media en regresión, o un clasificador aleatorio en clasificación) ya me da un 4% de error.

8. **Caso real en donde no se distingue entre training y test** (es decir, no hay conjunto de test definido).
    a) ¿Cómo escoger el mejor modelo y qué error Eout tiene?
    b) Posibilidades de análisis:
        a) Analizar el optimismo de un conjunto de test pequeño vs test grande y pocos datos de ajuste
        b) Analizar el compromiso que representa el uso de validación cruzada
        c) Analizar qué pasaría si validásemos y entrenásemos con los mismos datos. ¿Eout sería optimista o pesimista?

Ajuste de Modelos Lineales

- En la práctica:
  - Solamente se pide emplear **modelos lineales** (regresión lineal, regresión logística y perceptrón+pocket), junto con las **transformaciones en los datos**, técnicas de **regularización** y **preprocesado** que consideréis más conveniente
  - Si alguien quiere probar a mayores SVM, MLP, RF. ¡Perfecto! Que compare con los modelos lineales y justifique su uso. ¡Pero hay que usar modelos lineales!

Ajuste de Modelos Lineales

.zip = Códigos (.py) + Informe (.pdf)

1 fichero para regresión
1 fichero para clasificación

Fecha de entrega: 30 de Mayo

SE VALORARÁ ENORMEMENTE LA **JUSTIFICACIÓN** DE LAS DECISIONES TOMADAS Y LA **DISCUSIÓN** DE LOS RESULTADOS OBTENIDOS.