ADPI, 2º MIT

Práctica 3: Regresión Lineal

Curso académico 17/18

---

1. Introducción

En esta práctica nos enfrentaremos al problema de predecir la producción de energía de una central eólica a partir de variables meteorológicas. Los principales objetivos de la práctica son los siguientes:

- Comparar las prestaciones de un modelo basado en el conocimiento de la relación física entre las variables de entrada y la variable a predecir, con las prestaciones de modelos estadísticos genéricos.
- Estudiar la sensibilidad de un modelo de regresión lineal con respecto a la presencia de *outliers*.

  En principio, la relación entre las variables de entrada y la de salida está caracterizada desde el punto de vista físico, dado que una central eólica extrae una fracción de la energía cinética de la columna de aire que atraviesa el molino. Dicha energía cinética está bien definida en función de la velocidad del viento:

  $$E_c \propto v^3, \tag{1}$$

  donde v es la velocidad del viento en la dirección paralela al eje del rotor del molino. Sin embargo, los datos pueden estar afectados por un elevado nivel de ruido procedente de distintos fenómenos, como la parada accidental o intencionada de los rotores, la existencia de un efecto umbral (por debajo de una determinada velocidad de viento, la energía generada es nula), etc. Por estas razones puede ser conveniente recurrir a un método estadístico o de aprendizaje máquina, en el que se trata el sistema que relaciona entradas y salidas como una caja negra.

2. Datos

Usaremos una sencilla base de datos proporcionados por una central eólica experimental. Los datos consisten en la evolución temporal de tres variables: velocidad de viento, dirección y potencia generada. La tarea consistirá en predecir la energía generada a partir de las otras dos variables. La variable temporal no será tenida en cuenta, ya que el sistema se puede considerar sin memoria (la potencia generada en cada instante sólo depende de las variables meteorológicas de ese instante, y no de las pasadas).

El aspecto de cada muestra de datos es el siguiente:

Velocidad, v (m/s)	Dirección, ϕ (Âº)	Energía (kWh)
10.67	194	11.655,73

Los datos vienen directamente separados en dos conjuntos: $(\mathbf{X}, \mathbf{Y})$ para el entrenamiento y $(\mathbf{X}_{\text{tt}}, \mathbf{Y}_{\text{tt}})$ para la validación de los resultados.

3. Preprocesado

Por otra parte, es aconsejable preprocesar la variable correspondiente a la dirección del viento, que viene expresada en grados. Se deberá llevar algún tipo de reparametrización de dicha variable (transformarla en otra/s variable/s) para que evitar el efecto de discontinuidad (1Âº y 359Âº son ángulos muy similares, a pesar de su valor escalar tan dispar). Una opción puede consistir en basar el modelo de regresión en las coordenadas cartesianas del viento:

$$v_x = v\cos(\phi), v_y = v\sin(\phi). \tag{2}$$

Si se considera necesario, se podrá llevar a cabo cualquier otro preprocesdo que se considere oportuno, como una normalización o escalado de los datos.

4. Medida de la Calidad del Regresor

La medida de calidad más extendida para un regresor es el error cuadrático medio (MSE):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})^2.$$

Sin embargo dicho valor puede no sernos de utilidad si no tenemos en cuenta la varianza de la variable a predecir. Por tanto, ambas magnitudes deberán compararse para poder saber en qué medida estamos reduciendo la incertidumbre sobre y. Para ello, utilizaremos como medida de calidad la varianza explicada (explained variance, EV), que viene dada por la expresión

$$\text{EV} = 1 - \frac{\text{MSE}}{\sigma_y^2}, \tag{3}$$

donde σ_y^2 es la varianza empírica de la magnitud que queremos predecir. Nótese que, a diferencia del MSE, la EV es una magnitud a maximizar. Su valor está comprendido entre cero (peor predicción posible) y uno (cuando el error de predicción es nulo.

5. Métodos de regresión

Se deberán utilizar los siguientes métodos de regresión:

- Un modelo basado en el conocimiento a priori del problema. De acuerdo a (1), podemos estimar la energía como $y = C\ v^3$, donde podemos estimar C como

$$C = \frac{1}{N}\sum_{i=1}^{N}\frac{y^{(i)}}{v^3} \tag{4}$$

- Un regresor lineal regularizado basado en aquellas características que considere oportunas v, v_x, v_y ... Se aconseja probar distintos modelos y hacer una comparativa razonada.

6. Análisis a realizar

Se proporciona un conjunto de datos de entrenamiento y otro de test en el fichero. Se deberán entrenar distintos tipos de modelos a partir de los datos de entrenamiento, y obtener las prestaciones (en términos de la medida de calidad que se ha mencionado anteriormente) sobre el conjunto de test. Se debe tener en cuenta en cada momento que el conjunto de test se utiliza para medir las prestaciones del predictor, y no para usarlo para validación de parámetros. En caso de que se necesiten optimizar los parámetros de un modelo a partir de datos de validación, éstos deben proceder del conjunto de entrenamiento. Se debe hacer uso, por tanto, del procedimiento de validación cruzada (CV) sobre los datos de entrenamiento.
Los experimentos llevados a cabo sobre los datos deberán analizar la diferencia de prestaciones entre el modelo basado en la física del problema y los modelos basados en aprendizaje estadístico, justificando las diferencias que se aprecien.

7. Prestaciones frente a *outliers*

A la hora de valorar las prestaciones de nuestro modelo, un punto importante será valorar si la presencia de outliers está condicionando nuestra estima. Al trabajar con datos de dimensionalidad baja, la presencia de outliers puede realizarse por mera visualización.
Una forma más precisa para analizar la presencia de outliers es estudiar el efecto de usar subconjuntos de datos de distinto tamaño en el entrenamiento (curvas de aprendizaje), dibujando

tanto el error promedio en un cierto número de simulaciones (medido en el conjunto de test y entrenamiento) **como la desviación típica**.
Se deja a criterio del alumno el análisis de la presencia y efecto de *outliers*, así como la aplicación de técnias sencillas para entrenar un regresor que sea robusto frente a la presencia de los mismos.