Práctica II

Regresión lineal

Especificaciones

- El dataset a utilizar en esta práctica es el de cal_housing.csv
  - Las columnas *longitude, latitude, housingMedianAge, totalRooms, totalBedrooms, population, households, medianIncome* son las características de las casas
  - La columna *medianHouseValue* es el valor a predecir (target)
- Con el dataset especificado realice lo siguiente:
  - Cargue el dataset
  - Genere un conjunto de entrenamiento con 80% de los datos y 20% para pruebas (con shuffle)
  - Con el conjunto de entrenamiento genere un conjunto de validación de 10 pliegues

Especificaciones

- Usando el conjunto de entrenamiento (80% de los datos) realice lo siguiente:
  - Grafique la relación entre cada una de las variables y el valor a predecir
  - Genere una matriz de correlación y grafique un mapa de calor
- Usando el conjunto de validación (80% de los datos en 10 pliegues) realice lo siguiente para cada pliegue
  - Usando los datos sin escalar
    - Usando sólo la variable que presenta mayor correlación
      - Utilice regresión lineal mediante gradiente descendiente estocástico para generar una línea que se ajuste lo mejor posible a los datos (modificando parámetros *max_iter, learning_rate y eta0*) y reporte el error cuadrado medio (MSE) y el coeficiente de determinación (r^2)
      - Utilice regresión polinomial mediante gradiente descendiente estocástico con polinomios de grado 2 y 3 para generar una línea curva que se ajuste lo mejor posible a los datos (modificando parámetros *max_iter*, *learning_rate y eta0*) y reporte MSE y r^2
      - Almacene los valores MSE y r^2 de cada pliegue y al terminar los 10 pliegues reporte el resultado promedio estos
    - Usando las 2 variables con mayor correlación repita los experimentos anteriores
    - Usando las 3 variables con mayor correlación repita los experimentos anteriores
    - Usando todas las variables repita los experimentos anteriores

Especificaciones

- Usando el conjunto de validación (80% de los datos en 10 pliegues) realice lo siguiente para cada pliegue
  - Usando los datos con escalado estándar repita todos los experimentos anteriores
  - Usando los datos con escalado robusto repita todos los experimentos anteriores
- Usando el conjunto de entrenamiento (80% de los datos) entrene un modelo utilizando el tipo de regresión (lineal o polinomial) con la(s) variable(s) y método de escalado (sin escalado, estándar o robusto) que mejor resultado obtuvo en promedio en el conjunto de validación
- Usando el modelo anterior realice la predicción de los datos del conjunto de prueba y reporte MSE y r^2

Evidencias

- El código fuente
- Un documento que contenga lo siguiente
  - Una introducción donde se explique el problema a resolver y una descripción del método de regresión lineal
  - Gráficas de la relación de las variables con el valor a predecir y algunas conclusiones que se pueden obtener de estas
  - Gráfico del mapa de color con la matriz de correlación de las variables y algunas conclusiones que se pueden obtener de estas
  - Tabla que reporte los resultados en el conjunto de validación y algunas conclusiones que se pueden obtener de estas, indicado las caracteríticas (tipo de regresión, escalamiento, etc.) que tendrá el modelo que usará en el conjunto de prueba

Tipo de regresión	Tipo de escalamiento	Variables seleccionadas	Learning rate utilizado	Número de iteraciones utilizadas	MSE promedio	R^2 promedio

  - Resultados del MSE y R^2 en el conjunto de prueba obtenidos por el modelo entrenado
  - Conclusiones generales de la práctica realizada, mencionando las dificultades encontradas y alguna idea para mejorar los resultados obtenidos