Lab 2: Nearest Neighbor (`sklearn Version`)

1. Introducción

Vecinos más cercanos (Neareast neighbor) es uno de los enfoques más populares y sencillos para la clasificación y la predicción. En este laboratorio estudiará una implementación particular de este enfoque en `Python`. Se le pedirá que añada varios métodos a la implementación y luego que pruebe estos métodos para problemas de clasificación y regresión.

2. **Tareas de laboratorio**

A. Estudie la clase `kNN` de la aproximación del vecino más próximo proporcionada en el `Jupiter notebook Lab-02` que acompaña a este laboratorio. Tenga en cuenta que esta clase está diseñada para conjuntos de datos con atributos de entrada numéricos.

B. Añade a la clase `kNN` el método `normalize` que normaliza los atributos de entrada de los datos de entrenamiento `X_train` y los datos de prueba `X_test.` La normalización de atributos es importante, ya que todos los atributos reciben el mismo peso cuando se calculan las distancias entre instancias.

Para implementar el método `normalize` se podría utilizar el método `max` de `pandas.DataFrame` ya que tanto `X_train` como `X_test` son `pandas.DataFrame.`

Pruebe el clasificador `kNN` en los conjuntos de datos de clasificación de `diabetes` y `glass` (véase el apéndice A) para el caso en que los datos no están normalizados y el caso en que los datos están normalizados. Indique si los índices de precisión de entrenamiento y de retención mejoran con la normalización.

Para probar el clasificador `kNN` puede utilizar el script proporcionado en el cuaderno `Jupiter.` El script proporciona un gráfico con las tasas de precisión de entrenamiento y de retención en función del parámetro `k` del clasificador `kNN`.

Pruebe el clasificador `kNN` en los conjuntos de datos de clasificación de vidrio los datos se normalizan para diferentes valores del parámetro `exp` de la distancia de Minkowski . Indique si los índices de precisión de entrenamiento y de retención cambian debido a `exp.` Para esta tarea puede utilizar el segundo script de prueba proporcionado en la nota de Jupiter.

C. Añadir a la clase `kNN` el método `getClassProbs` que calcula para todas las instancias de prueba en `X_test` las probabilidades de clase posteriores. Esto significa que el método calcula para cada fila (instancia) de `X_test` una fila con

probabilidad de clase 1, probabilidad de clase 2 y probabilidad de clase N. Combine las filas de las probabilidades de clase posteriores en el objeto `pandas.DataFrame` que será la salida del método `getClassProbs`.

D. Añadir a la clase `kNN` el método `getPrediction` que calcula para todas las instancias de prueba en `X_test` valores de regresión para el atributo de salida Esto significa que el método calcula para cada instancia (fila) en `X_test` un valor de regresión igual a la media de los valores y en `Y_train` de los `k` vecinos más cercanos de la instancia en `X_train`. Combine los valores de regresión calculados para todas las instancias de `X_test` en un objeto `pandas.DataFrame` que será la salida del método `getPrediction`.

Pruebe el método `getPrediction` en el conjunto de datos `autoprice`, que es un conjunto de datos de regresión (véase el Apéndice A). Para ello, puede adaptar el script de prueba que ya ha utilizado para la Tarea B. Utilice el error absoluto medio como métrica principal para estimar el rendimiento de la regresión1 en lugar de la tasa de precisión. Para calcular el error medio absoluto puede utilizar el método `mean_absolute_error` de `sklearn.metrics`.

Informe: Prepare un archivo pdf del cuaderno Jupiter junto con su código para las tareas B, C y D. Es decir entregue tanto el pdf como el archivo del código original. No olvide anexar introducción, conclusiones y referencias.

Proporciona respuestas en campos "markdown" para las preguntas de la tarea B.

Error absoluto medio (MAE) es la diferencia absoluta media entre lo que el valor predicho ��̂�� y el valor verdadero ���� para todas las instancias de prueba (���� , ����) en los datos de entrenamiento D ; es decir, ������ = ∑(����,����)∈�� |���� - ��̂�� |. **Apéndice A**: Conjuntos de datos

I. Datos sobre la diabetes (clasificación)

1. Título: Pima Indians Diabetes Database

2. Fuentes:

A. Propietarios originales: National Institute of Diabetes and Digestive and Kidney Diseases

B. Donante de la base de datos: Vincent Sigillito (vgs@aplcen.apl.jhu.edu)
Research Center, RMI Group Leader
Applied Physics Laboratory
The Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20707
(301) 953-6231

3. La selección de estos casos a partir de una base de datos más amplia está sujeta a varias restricciones. En particular, todos los pacientes son mujeres con de al menos 21 años y de ascendencia india pima. ADAP es una rutina de aprendizaje adaptativo que genera y ejecuta análogos digitales de dispositivos similares al perceptrón. Se trata de un algoritmo único; véase el artículo para más detalles.

4. Número de instancias: 768

5. Número de atributos: 8 más clase

6. Para cada atributo: (todos con valor numérico)

    1. Número de embarazos
    2. Concentración plasmática de glucosa a las 2 horas en una prueba oral de tolerancia a la glucosa
    3. Presión arterial diastólica (mm Hg)
    4. Grosor del pliegue cutáneo del tríceps (mm)
    5. Insulina sérica a las 2 horas (mu U/ml)
    6. Índice de masa corporal (peso en kg/(altura en m)^2)
    7. Función pedigrí de la diabetes
    8. Edad (años)
    9. Variable de clase (0 o 1)

7. Faltan valores de atributos: Ninguno

8. Distribución de clases: (el valor de clase 1 se interpreta como "dio positivo en diabetes")

    Valor clase Número de instancias
    0 500
    1 268

9. Breve análisis estadístico:

    Número de atributo: Media: Desviación estándar:
    1. 3.8 3.4
    2. 120.9 32.0
    3. 69.1 19.4
    4. 20.5 16.0
    5. 79.8 115.2
    6. 32.0 7.9
    7. 0.5 0.3
    8. 33.2 11.8

    Valores reetiquetados en el atributo 'class'
    Desde: 0 a: tested_negative
    Desde: 1 a: tested_positive

II. Datos del cristal (clasificación)

1. Titulo: Glass Identification Database

2. Fuentes:

    (a) Creador: B. German
    -- Central Research Establishment

Home Office Forensic Science Service
Aldermaston, Reading, Berkshire RG7 4PN

(b) Donor: Vina Spiehler, Ph.D., DABFT
Diagnostic Products Corporation
(213) 776-0180 (ext 3014)

3. Información pertinente:n
Vina realizó una prueba comparativa de su sistema basado en reglas, BEAGLE, el algoritmo del vecino más próximo y el análisis discriminante. BEAGLE es un producto disponible a través de VRS Consulti ng, Inc.; 4676 Admiralty Way, Suite 206; Marina Del Ray, CA 90292 (213) 827 -7890 y FAX: -3189. Para determinar si el vidrio era un tipo de vidrio "flotado" o no, se obtuvieron los siguientes resultados (# respuestas incorrectas):

Tipo de muestra Beagle NN DA Ventanas que se procesaron por flotación (87) 10 12 21 Ventanas que no lo eran: (76) 19 16 22

El estudio de la clasificación de tipo de vidrio fue motivado por la investigación criminológica. En la escena del crimen, el vidrio que queda puede utilizarse como prueba... ¡si se identifica correctamente!

4. Número de instancias: 214

5. Número de atributos: 10 (incluido un Id#) más el atributo de clase
-- todos los atributos se valoran de forma continua

6. Información sobre atributos:
1. RI: índice de refracción
2. Na: Sodio (unidad de medida: porcentaje en peso en el óxido correspondiente, al igual que los atributos 4-10)
3. Mg: Magnesio
4. Al: Aluminio
5. Si: Silicio
6. K: Potasio
7. Ca: Calcio
8. Ba: Bario
9. Fe: Hierro
10. clase: tipo de vidrio
-- 1 building_windows_float_processed
-- 2 building_windows_non_float_processed
-- 3 vehicle_windows_float_processed
-- 4 vehicle_windows_non_float_processed (ninguna en esta base de datos) -- 5 contenedores
-- 6 vajilla
-- 7 faros

7. Faltan valores de atributos: Ninguno
Resumen estadístico:

Atributo: Mín Máx Media DE Correlación con la clase 2. RI: 1.5112 1.5339 1.5184 0.0030 -0.1642 3. Na: 10.73 17.38 13.4079 0.8166 0.5030 4. Mg: 0 4.49 2.6845 1.4424 -0.7447 5. Al: 0.29 3.5 1.4449 0.4993 0.5988 6. Si: 69.81 75.41 72.6509 0.7745 0.1515 7. K: 0 6.21 0.4971 0.6522 -0.0100 8. Ca: 5,43 16,19 8.9570 1.4232 0.0007 9. Ba: 0 3.15 0.1750 0.4972 0.5751 10. Fe: 0 0.51 0.0570 0.0974 -0.1879

8. Distribución por clases: (de un total de 214 instancias)
   -- 163 Vidrios de ventanas (ventanas de edificios y vehículos)
   -- 87 flotadores procesados
   -- 70 ventanas de edificios
   -- 17 ventanas del vehículo
   -- 76 no flotantes procesados
   -- 76 ventanas de edificios
   -- 0 ventanas del vehículo
   -- 51 Vidrio sin ventana
   -- 13 contenedores
   -- 9 vajilla
   -- 29 faros
   Valores reetiquetados en la clase de atributos
   De: '1' a: 'build wind float'
   De: '2' a: 'build wind non-float'
   De: '3' a: 'vehic wind float'
   De: '4' a: 'vehic wind non-float'
   De: '5' a: containers
   De: '6' a: tableware
   De: '7' a: headlamps

III. Datos del precio del automóvil (R egresión)

Este conjunto de datos consta de tres tipos de entidades:
(a) la especificación de un automóvil en función de diversas características; (b) su calificación de riesgo de seguro asignada,;
(c) sus pérdidas normalizadas de uso en comparación con otros coches.
La segunda calificación corresponde al grado en que el coche es más arriesgado de lo que indica su precio. Los coches son inicialmente se le asigna un símbolo de factor de riesgo asociado a su precio e. A continuación, si es más arriesgado (o menos), este símbolo se ajusta por moviéndolo hacia arriba (o hacia abajo) en la escala. Los actuarios llaman a este proceso "simbolización". Un valor de +3 indica que el auto es
El tercer factor es la siniestralidad media relativa por año asegurado.
Este valor se normaliza para todos los automóviles de una determinada clasificación de tamaño (pequeños de dos puertas, familiares...), deportivo/especialidad, etc...), y representa la pérdida media por coche y año. -
Nota: Varios de los atributos de la base de datos podrían utilizarse como atributo de "clase". Los datos originales (del repositorio de la UCI) (http://www.ics.uci.edu/~mlearn/MLSummary.html) tienen 205 instancias descritos por 26 atributos :
- 15 continua
- 1 entero

- 10 nominal

A continuación se ofrece más información sobre estos atributos:
1. simbolización: -3, -2, -1, 0, 1, 2, 3.
2. pérdidas-normalizadas: continua de 65 a 256.
3. marca: alfa-romero, audi, bmw, chevrolet, dodge, honda,
isuzu, jaguar, mazda, mercedes-benz, mercury,
mitsubishi, nissan, peugot, plymouth, porsche,
renault, saab, subaru, toyota, volkswagen, volvo
4. tipo de combustible: diesel, gas.
5. aspiración: std, turbo.
6. número de puertas: cuatro, dos.
7. Tipo de carrocería: techo duro, wagon, berlina, portón trasero, descapotable.
8. ruedas motrices: 4x4, fwd, rwd.
9. ubicación del motor: delantero, trasero.
10. distancia entre ejes: continua de 86,6 a 120,9.
11. longitud: continua de 141,1 a 208,1.
12. anchura: continua de 60,3 a 72,3.
13. altura: continua de 47,8 a 59,8.
14. peso en vacío: continuo de 1488 a 4066.
15. tipo de motor: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. número de cilindros: ocho, cinco, cuatro, seis, tres, doce, dos.
17. cilindrada: continua de 61 a 326.
18. sistema de combustible: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. calibre: continuo de 2,54 a 3,94.
20. carrera: continua de 2,07 a 4,17.
21. relación de compresión: continua de 7 a 23.
22. caballos de fuerza: continua de 48 a 288.
23. pico-rpm: continua de 4150 a 6600.
24. mpg en ciudad: continua de 13 a 49.
25. en carretera: continua de 16 a 54.
26. precio: continuo de 5118 a 45400.

En los datos originales también faltan algunos valores de atributos
indicados con " ?" : Atributo #: Número de casos en los que falta un valor:

2. 41
6. 2
19. 4
20. 4
22. 2
23. 2
26. 4
He cambiado los datos originales de la siguiente manera :
- Se eliminaron todas las instancias con incógnitas, con lo que se obtuvieron 159 instancias. - La variable objetivo es "precio"
- Se eliminaron todos los atributos nominales (10).

Fuente original: UCI machine learning repository.
(http://www.ics.uci.edu/~mlearn/MLSummary.html).
Source: collection of regression datasets by Luis Torgo (ltorgo@ncc.up.pt)
at http://www.ncc.up.pt/~ltorgo/Regression/DataSets.html
Characteristics: 159 cases; 14 continuous variables; 1 nominal vars..