Centro de investigación en Computación

Clasificación inteligente de Patrones

Empates, k-NN y Leave-one-out

Autor:
Diego Noguez Ruiz

Profesor:
Dr. Cornelio Yáñez Márquez

24 de abril de 2023

Índice general

Índice de tablas

1. Introducción

Previamente habíamos estudiado un clasificador de la familia de los k-NN, específicamente el 1-NN. Sin embargo, durante nuestra última sesión analizamos en general los clasificadores k-NN.

Por otro lado, hemos estado analizando a profundidad algunos de los métodos de validación más reconocidos en el estado del arte, como la partición fija y el Hold-Out; sin embargo, es importante destacar que estos métodos no son de validación cruzada, los cuales se destacan por ser los métodos favoritos actualmente por la comunidad. Es por ello que hemos añadido un nuevo método a nuestro *toolkit* llamado *Leave-One-Out Cross Validation* (LOOCV), el cual presenta mejoras en comparación con los anteriores.

En esta tarea se realizará un ejemplo práctico para reforzar el entendimiento de estos nuevos conocimientos.

2. Desarrollo

2.1. Propósito de la tarea

Ejemplificar y comparar algunos algoritmos de la familia k-NN.

2.2. Parte 1

En relación con el contenido del presente RD 10:

- 1.a) Para cada una de las iteraciones 3 a 28 del Leave-one-out al aplicar el algoritmo 3-NN en el Nutt [1] dataset, crear una tabla de distancias ordenadas como se hizo para el patrón 2.

  Para esta parte se considero la clase positiva igual a la clase 0, entonces los primeros 14 patrones son de la clase positiva, y los restantes de la negativa. Una vez teniendo en cuenta se realizaron las respectivas tablas.

  - Patrón 3. Ver tabla 1

Distancia	Patrón	Clase del patrón
5664.3553	16	1
5753.5756	9	0
5849.4868	22	1

Tabla 1: Tabla para el patrón 3 (NUTT Dataset)

  - Patrón 4. Ver tabla 2
  - Patrón 5. Ver tabla 3

Distancia	Patrón	Clase del patrón
4315.7133	28	1
4740.8204	25	1
4783.1153	12	0

Tabla 2: Tabla para el patrón 4 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
5245.4231	17	1
5279.4483	2	0
8673.2522	28	1

Tabla 3: Tabla para el patrón 5 (NUTT Dataset)

- Patrón 6. Ver tabla 4
- Patrón 7. Ver tabla 5
- Patrón 8. Ver tabla 6
- Patrón 9. Ver tabla 7
- Patrón 10. Ver tabla 8
- Patrón 11. Ver tabla 9
- Patrón 12. Ver tabla 10
- Patrón 13. Ver tabla 11
- Patrón 14. Ver tabla 12
- Patrón 15. Ver tabla 13
- Patrón 16. Ver tabla 14
- Patrón 17. Ver tabla 15
- Patrón 18. Ver tabla 16
- Patrón 19. Ver tabla 17
- Patrón 20. Ver tabla 18
- Patrón 21. Ver tabla 19
- Patrón 22. Ver tabla 20
- Patrón 23. Ver tabla 21

Distancia	Patrón	Clase del patrón
7806.3448	8	0
8285.9729	3	0
8623.1005	16	1

Tabla 4: Tabla para el patrón 6 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
5700.9404	11	0
5880.8870	8	0
5906.5303	9	0

Tabla 5: Tabla para el patrón 7 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
5880.8870	7	0
6078.3792	16	1
6093.7694	3	0

Tabla 6: Tabla para el patrón 8 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
4483.9667	24	1
4712.4760	23	1
4750.0025	22	1

Tabla 7: Tabla para el patrón 9 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
5960.6671	26	1
6004.9222	25	1
6101.1409	23	1

Tabla 8: Tabla para el patrón 10 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
3989.7087	24	1
4252.9691	26	1
4379.3388	23	1

Tabla 9: Tabla para el patrón 11 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
4081.7533	28	1
4457.7164	25	1
4630.2699	21	1

Tabla 10: Tabla para el patrón 12 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
6190.2404	7	0
6841.5697	9	0
7660.9439	24	1

Tabla 11: Tabla para el patrón 13 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
8965.2156	17	1
9242.5488	27	1
9555.3553	5	0

Tabla 12: Tabla para el patrón 14 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
1782.9013	21	1
2623.5162	20	1
2647.3289	28	1

Tabla 13: Tabla para el patrón 15 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
3092.4857	22	1
3281.5750	23	1
3283.9665	24	1

Tabla 14: Tabla para el patrón 16 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
4774.0011	28	1
5245.4231	5	0
5419.6050	21	1

Tabla 15: Tabla para el patrón 17 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
3400.0162	15	1
3411.9351	20	1
3433.5380	23	1

Tabla 16: Tabla para el patrón 18 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
3890.2023	15	1
4144.9885	23	1
4180.5001	25	1

Tabla 17: Tabla para el patrón 19 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
2623.5162	15	1
3266.5461	23	1
3309.0057	21	1

Tabla 18: Tabla para el patrón 20 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
1782.9013	15	1
2274.4975	28	1
3289.9307	25	1

Tabla 19: Tabla para el patrón 21 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
2305.6310	23	1
3092.4857	16	1
3146.5852	26	1

Tabla 20: Tabla para el patrón 22 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
2037.6318	26	1
2305.6310	22	1
2553.3917	25	1

Tabla 21: Tabla para el patrón 23 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
2635.4879	26	1
2639.8445	23	1
3147.7924	25	1

Tabla 22: Tabla para el patrón 24 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
2514.3428	26	1
2553.3917	23	1
2864.8115	15	1

Tabla 23: Tabla para el patrón 25 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
2037.6318	23	1
2514.3428	25	1
2635.4879	24	1

Tabla 24: Tabla para el patrón 26 (NUTT Dataset)

Distancia	Patrón	Clase del patrón
2957.6804	25	1
3004.2389	23	1
3190.0915	26	1

Tabla 25: Tabla para el patrón 27 (NUTT Dataset)

- Patrón 24. Ver tabla 22
- Patrón 25. Ver tabla 23
- Patrón 26. Ver tabla 24
- Patrón 27. Ver tabla 25
- Patrón 28. Ver tabla 26

- 1.b) Generar la matriz de confusión y reportar los valores de accuracy, sensitivity, specificity y balanced accuracy

  Recordemos que los patrones 1 y 2 se calcularon en clase, considerándolos tenemos que el $IR = 1$, entonces el dataset es balanceado, por lo tanto, se pueden calcular todas las medidas de desempeño conocidas.
  A continuación la matriz de confusión.

$$CM = \begin{bmatrix} 5 & 9 \\ 0 & 14 \end{bmatrix}$$

  - Accuracy:0.68
  - Recall:0.36
  - Specificity:1.0
  - Balanced Accuracy:0.68

Distancia	Patrón	Clase del patrón
2274.4975	21	1
2647.3289	15	1
3240.2942	25	1

Tabla 26: Tabla para el patrón 28 (NUTT Dataset)

- 1.c) Generar la matriz de confusión que resulta de aplicar el algoritmo 5-NN en el Nutt dataset con el método de validación Leave-one-out. Reportar los valores de accuracy, sensitivity, specificity y balanced accuracy y comparar con los valores del inciso 1.b). ¿Es válida esta comparación? Justificar

  De igual manera considerando el mismo data set obtenemos que al aplicar el 5-NN se tiene la siguiente matriz de confusión.

$$CM = \begin{bmatrix} 5 & 9 \\ 0 & 14 \end{bmatrix}$$

  - Accuracy:0.68
  - Recall:0.36
  - Specificity:1.0
  - Balanced Accuracy:0.68

  Recordemos que se puede comparar el desempeño de dos clasificadores sii se satisfacen las 3 siguientes condiciones:

  - Es el mismo dataset
  - Es la misma medida de desempeño
  - Es el mismo método de validación

  En particular, si es válido hacer la comparación ya que se cumplen los requisitos, y como tenemos la misma matriz de confusión concluimos que aumentar el valor de k no implica una mejora en la clasificación.

2.3. Parte 2

En esta parte se usará el dataset Electricity (lo recibirán por correo)

- 2.a) Aplicar el algoritmo 1-NN en el dataset Electricity con el método de validación Leave-one-out. Generar y reportar la matriz de confusión, así como los valores de accuracy, sensitivity, specificity y balanced accuracy.

  Primero se realizó la limpieza del dataset, ya que había columnas redundantes. A continuación la matriz de confusión definiendo a la clase 1 como la clase positiva.

$$CM = \begin{bmatrix} 1219 & 211 \\ 210 & 760 \end{bmatrix}$$

  Por otro lado, el dataset es balanceado ya que $IR = 1430/970 = 1.47$, donde la clase mayoritaria es la clase positiva, calculando las respectivas medidas de desempeño se obtiene lo siguiente.

- Accuracy: 0.8246
- Recall: 0.8524
- Specificity: 0.7835
- Balanced Accuracy: 0.818

- 2.b) Aplicar el algoritmo 5-NN en el dataset Electricity con el método de validación Leave-one-out. Generar y reportar la matriz de confusión, así como los valores de accuracy, sensitivity, specificity y balanced accuracy.

$$CM = \begin{bmatrix} 1260 & 170 \\ 232 & 738 \end{bmatrix}$$

Medidas de desempeño:

- Accuracy: 0.8325
- Recall: 0.8811
- Specificity: 0.7608
- Balanced Accuracy: 0.821

- 2.c) Aplicar el algoritmo 11-NN en el dataset Electricity con el método de validación Leave-one-out. Generar y reportar la matriz de confusión, así como los valores de accuracy, sensitivity, specificity y balanced accuracy

$$CM = \begin{bmatrix} 1257 & 173 \\ 279 & 691 \end{bmatrix}$$

Medidas de desempeño:

- Accuracy: 0.8108
- Recall: 0.8776
- Specificity: 0.7124
- Balanced Accuracy: 0.795

- 2.d) Aplicar el Clasificador Euclidiano en el dataset Electricity con el método de validación Leave-one-out. Generar y reportar la matriz de confusión, así como los valores de accuracy, sensitivity, specificity y balanced accuracy

$$CM = \begin{bmatrix} 729 & 701 \\ 480 & 490 \end{bmatrix}$$

- Accuracy: 0.5079
- Recall: 0.5098
- Specificity: 0.5052
- Balanced Accuracy: 0.5075

3. Conclusiones y trabajo futuro

Durante la sesión en la que se estudió el clasificador K-NN, surgió la pregunta natural de si un mayor valor de k mejoraría la clasificación. Gracias a esta tarea, pudimos despejar esa duda durante la parte 2, y la respuesta es que **no necesariamente**.

Por primera vez pudimos ver en acción el clasificador K-NN. Siendo honesto, pensé que tardaría demasiado en correr para el data set electricity, dado que contiene 2400 patrones. Como sabemos, la complejidad de estos algoritmos es $O(n * n)$, sin embargo, en mi computadora en particular, cuando realicé el 5-NN y 11-NN, tardó aproximadamente 5 y 10 segundos, respectivamente. Creo que estos tiempos no son demasiado largos, pero es evidente que a medida que aumente el número de patrones en el dataset, los tiempos de ejecución se incrementarán.

Posiblemente en un futuro cercano, estudiaremos los K-NN en conjunto con otras herramientas, como los *Fuzzy sets*.

Referencias

[1] B. Haznedar, M. T. Arslan, and A. Kalinli, "Microarray Gene Expression Cancer Data," vol. 4, Apr. 2017. Publisher: Mendeley Data.