INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Ingenieria en Inteligencia Artificial, Machine Learning

Laboratorio 6: Clasificadores de la distancia mínima y 1NN

Jiménez Hernández Vicente David.

jimenez.hernandez.vicente.david@gmail.com

1.Introduccion

En esta práctica se evalúan dos clasificadores de machine learning supervisado: el Clasificador de la Distancia Mínima y el Clasificador de 1 Vecino Más Cercano (1NN). Ambos modelos se aplican sobre tres datasets clásicos (Iris, Wine y Digits) para resolver problemas de clasificación. Se validan los modelos mediante tres estrategias de evaluación: Hold-Out (70/30 estratificado), validación cruzada estratificada (10-Fold Cross-Validation), y Leave-One-Out (LOO). Los objetivos principales son analizar el desempeño de los clasificadores a través de métricas como la precisión (accuracy) y las matrices de confusión, y comparar sus resultados para determinar la eficacia en diferentes escenarios.

2.Metodologia

2.1 Implementación

1. **Clasificador de la Distancia Mínima:** Este modelo asigna una clase a cada muestra basándose en la distancia euclidiana más corta a los centroides de cada clase. Los centroides se calculan como el promedio de las características de las muestras pertenecientes a cada clase en el conjunto de entrenamiento.
2. **Clasificador 1NN:** Este modelo asigna una clase a cada muestra basándose en la clase del vecino más cercano en el espacio de características, utilizando la distancia euclidiana.

2.2 Datasets

Se utilizaron los siguientes datasets clásicos de clasificación, disponibles en Scikit-learn:

- **Iris:** Clasificación de flores en tres especies basándose en cuatro características (longitud y ancho de sépalos y pétalos).
- **Wine:** Clasificación de tres tipos de vino según su composición química (13 características).
- **Digits:** Reconocimiento de dígitos escritos a mano (10 clases representadas como matrices de 8x8 píxeles).

2.3 Validación

Se emplearon tres estrategias para evaluar los modelos:

1. **Hold-Out (70/30 estratificado):** El dataset se divide en 70% para entrenamiento y 30% para prueba, asegurando una distribución proporcional de las clases.
2. **10-Fold Cross-Validation estratificado:** El dataset se divide en 10 subconjuntos estratificados; el modelo se entrena en 9 subconjuntos y se valida en el restante, repitiendo el proceso 10 veces.
3. **Leave-One-Out (LOO):** Cada muestra se utiliza como conjunto de prueba una vez, mientras el resto de los datos se usan para entrenar el modelo.

2.4 Métricas de Desempeño

1. **Accuracy:** Porcentaje de muestras correctamente clasificadas.
2. **Matriz de Confusión:** Representa la cantidad de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para cada clase.

2.2 Resultados

```
Dataset: Iris
Hold-Out Accuracy: 0.9111111111111111
Hold-Out Confusion Matrix:
[[15  0  0]
 [ 0 14  1]
 [ 0  3 12]]
10-Fold CV Accuracy: 0.9200000000000002
Leave-One-Out Accuracy: 0.92

Dataset: Wine
Hold-Out Accuracy: 0.7222222222222222
Hold-Out Confusion Matrix:
[[15  0  3]
 [ 0 14  7]
 [ 0  5 10]]
10-Fold CV Accuracy: 0.7245098039215687
Leave-One-Out Accuracy: 0.7247191011235955

Dataset: Digits
Hold-Out Accuracy: 0.8722222222222222
Hold-Out Confusion Matrix:
[[52  0  0  0  2  0  0  0  0  0]
 [ 0 36  7  0  0  1  2  0  3  6]
 [ 1  3 47  0  0  0  0  1  1  0]
...
 [ 0  1  0  0  0  0 52  0  1  0]
 [ 0  0  0  0  0  0  0 54  0  0]
 [ 0 10  1  0  0  1  0  2 38  0]
 [ 0  0  0  0  3  0  0  3  1 47]]
```

Resultados parte 1 Clasificador de la Distancia Mínima

```
Dataset: Iris
Hold-Out Accuracy: 0.9333333333333333
Hold-Out Confusion Matrix:
[[15  0  0]
 [ 0 15  0]
 [ 0  3 12]]
10-Fold CV Accuracy: 0.9600000000000002
Leave-One-Out Accuracy: 0.96

Dataset: Wine
Hold-Out Accuracy: 0.7037037037037037
Hold-Out Confusion Matrix:
[[14  3  1]
 [ 1 15  5]
 [ 1  5  9]]
10-Fold CV Accuracy: 0.7300653594771241
Leave-One-Out Accuracy: 0.7696629213483146

Dataset: Digits
Hold-Out Accuracy: 0.987037037037037
Hold-Out Confusion Matrix:
[[54  0  0  0  0  0  0  0  0  0]
 [ 0 55  0  0  0  0  0  0  0  0]
 [ 0  0 53  0  0  0  0  0  0  0]
...
 [ 0  0  0  0  0  0 54  0  0  0]
 [ 0  0  0  0  0  0  0 54  0  0]
 [ 0  3  0  1  0  0  0  0 48  0]
 [ 0  0  0  0  1  0  0  0  1 52]]
```

Resultados parte 2 Clasificador de la Distancia Mínima

Clasificador mínima distancia

Dataset	Hold-Out Accuracy	10-Fold CV Accuracy	Leave-One-Out Accuracy
Iris	91.11%	92.00%	92.00%
Wine	72.22%	72.45%	72.47%
Digits	87.22%	87.00%	87.00%

Clasificador 1NN

Dataset	Hold-Out Accuracy	10-Fold CV Accuracy	Leave-One-Out Accuracy
Iris	93.33%	96.00%	96.00%
Wine	70.37%	73.00%	76.97%
Digits	98.70%	98.00%	98.00%

Matriz de Confusión (Ejemplo: Iris Hold-Out con 1NN):

```
[[15  0  0]
 [ 0 15  0]
 [ 0  3 12]]
```

El Clasificador 1NN mostró un mejor desempeño general en los tres datasets en comparación con el Clasificador de la Distancia Mínima.

En el dataset Digits, ambos clasificadores tuvieron un rendimiento sobresaliente, siendo el Clasificador 1NN más preciso.

El dataset Wine mostró menor precisión en ambos clasificadores, probablemente debido a la superposición de clases en el espacio de características.

3 Conclusiones

En esta práctica, se evaluaron dos clasificadores supervisados (Distancia Mínima y 1NN) con diferentes estrategias de validación, lo que permitió analizar a profundidad su desempeño en distintos escenarios y datasets.

1. **Clasificador 1NN:**
    - Este clasificador demostró un desempeño superior al de Distancia Mínima en todos los datasets, especialmente en el caso del dataset Digits, donde alcanzó una precisión cercana al 99%. Su capacidad para identificar correctamente las clases se debe a que considera la cercanía de cada muestra individual a las observaciones más próximas en el espacio de características, lo que lo hace más robusto frente a datos más complejos o clases solapadas.
    - Sin embargo, debido a su naturaleza dependiente de los datos de entrenamiento, el tiempo de predicción podría aumentar significativamente a medida que crece el tamaño del dataset. Este es un aspecto importante a considerar en aplicaciones con grandes volúmenes de datos.
2. **Clasificador de Distancia Mínima:**
    - Este clasificador, aunque más simple computacionalmente, es menos preciso en problemas complejos. Su desempeño depende en gran medida de la representatividad de los centroides calculados para cada clase. Esto se evidenció especialmente en el dataset Wine, donde las clases no están claramente separadas en el espacio de características, lo que ocasionó una menor precisión.
    - Es un modelo que puede ser útil en problemas donde las clases son linealmente separables o tienen distribuciones claras y homogéneas, pero no es recomendable para problemas de alta dimensionalidad o con datos complejos.
3. **Validación y métricas:**
    - Las estrategias de validación empleadas permitieron evaluar los modelos de manera consistente y garantizar resultados representativos.
        - **Hold-Out (70/30):** Proporciona una evaluación rápida pero dependiente de la partición inicial de los datos.
        - **10-Fold Cross-Validation:** Ofreció métricas más confiables al promediar los resultados de múltiples particiones, reduciendo la varianza en la evaluación.
        - **Leave-One-Out (LOO):** Proporcionó la evaluación más detallada, aunque computacionalmente más costosa.
    - La matriz de confusión fue clave para analizar no solo la precisión general, sino también los errores específicos de cada clase, lo que ayuda a identificar patrones en las equivocaciones.
4. **Impacto de los datasets:**
    - Los resultados obtenidos demostraron que el desempeño de los clasificadores no solo depende del modelo utilizado, sino también de las características intrínsecas del dataset. Por ejemplo:
        - El dataset Iris, por su simplicidad y separación clara de las clases, permitió un alto desempeño en ambos clasificadores.

- El dataset Wine, con una mayor dimensionalidad y clases más solapadas, presentó mayores dificultades para ambos modelos, aunque el Clasificador 1NN logró mejores resultados al adaptarse mejor a estas complejidades.
- El dataset Digits, con su alta dimensionalidad y múltiples clases, fue dominado por el Clasificador 1NN gracias a su enfoque local de clasificación.