TC3006C.102 | MOMENTO DE RETROALIMENTACIÓN: MÓDULO 2

Análisis y Reporte sobre el Desempeño del Modelo SVM Multiclase sobre el Set de Datos Iris

Armendáriz Torres, David Fernando[1]

[1] A01570813 | Ingeniería en Ciencia de Datos y Matemáticas

1 Introducción

A lo largo de este momento de retroalimentación se implementa un modelo de apredizaje supervisado, maquinas de vectores de soporte, con capacidad multiclase, mediante la metodología *One vs One*, para clasificar los datos del dataset iris.

1.1 Conceptos y Herramientas

1.1.1 SVM

Una Maquina de Vectores de Soporte, mejor conocida como *SVM* por sus siglas en inglés, genera un hiperplano de $n-1$ dimensiones en un espacio n para segmentar los datos y a su vez clasificarlos. SVM es un técnica de clasificación binaria en donde n es la cantidad de features empleados. El hiperplano generado busca optimizar el margen entre los puntos mas cercanos de ambas clases, en caso de los datos no ser linealmente separables, se proyectan los datos en una dimensión mayor a través de una función de kenrel y se segmentan en dicha dimensión (IBM, s.f.).

El modelo de *SVM* emplea la clase sklearn.svm.SVC() de scikit-learn (Pedregosa et al., 2011).

1.1.2 One Vs One

One vs One es una técnica de clasificación multiclase que permite a un modelo de clasificación binaria hacer clasificaciones multiclase. Esta técnica funciona entrenando un modelo para cada par de clases presentes en los datos los cuales utiliza para generar predicciones y la moda de las predicciones generadas se considera como la clasificación final.

Está tecnica fue empleada para el modelo y se encuentra integrada como la metodología multiclase por defecto de la clase sklearn.svm.SVC() de scikit-learn (Pedregosa et al., 2011).

Fecha: 07 de septiembre de 2024

1.1.3 Validación Cruzada Anidada

La validación cruzada es una técnica empleada para generar una estimación mas robusta del desempeño de un modelo. Está técnica, a grandes razgos, consiste en segmentar el set de datos y rotar los segmentos que se utilizan para entrenar y evaluar a traves de multiples iteraciones.

La validación cruzada anidada es un técnica que emplea dos ciclos de validación cruzada anidados para evaluar y ajustar los hiperparámetros del un modelo reduciendo el posible overfit de un modelo al que este puede ser sujeto dada la técinca de validación cruzda y generando una evaluación robusta de su desempeño. La validación cruzada anidad utiliza emplea un ciclo interno de validación cruzda para optimizar los hiperparámetros y un ciclo externo para evaluar los resultados. Esta técnica asegura que los datos utilizados para la evaluación en cada iteración del ciclo externo no son utilizados en el ciclo interno para el ajuste de lo hiperparámetros (Ploomber, 2022). La figura 1 muestra el funcionamiento de la validación cruzada anidada.

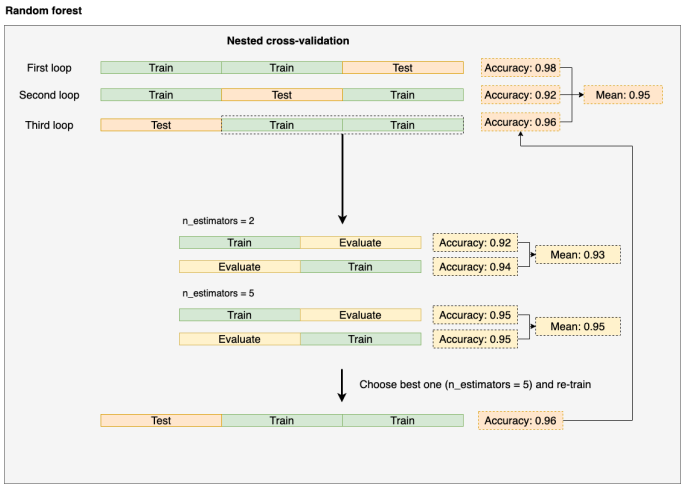

Figura 1. Diagrama de Validación Cruzada Anidada (Ploomber, 2022).

Esta técnica fue empleada utilizando las clases sklearn.model_selection.GridSearchCV() y sklearn.model_selection.cross_val_score de

scikit-learn (Pedregosa et al., 2011).

2 Exploración de los Datos: iris

El set de datos trabajado es el famosamente conocido conjunto de datos iris de Fisher (Fisher, R., 1936). Los datos de iris se comoponen de la siguiente forma:

- Características: 4 caracterísicas físicas referentes a la flor iris.
  - *sepal-length*: numérico
  - *sepal-width*: numérico
  - *petal-length*: numérico
  - *petal-width*: numérico
- Etiquetas de Clase (*class*): 3 clases correspondientes a 3 especies de la flor iris.
  - *setosa*
  - *versicolor*
  - *virginica*
- Registros: 150 registros, 50 de cada clase
- Datos Ausentes: El set de datos carece de datos ausentes.

2.1 Separabilidad Lineal

Dado que se busca emplear un modelo SVM multiclase, es necesario comprobar si el set de datos es linealmente separable para definir el tipo de margen a emplear.

Para evaluar la separabilidad lineal, se entrenó, sobre todos los datos, un modelo SVM con kernel lineal y un coeficiente de regularización alto, $C = 10,000,000,000$, que aproximara el comportamiento de un margen duro. Al evaluar el desempeño del modelo mediante exactitud, se puede identificar la separabilidad; el modelo generó un resultado de exactitud del $98\,\%$, lo que indica que el modelo no se ajustó perfectamente a los datos y, por ende, no es linealmente separable. La figura 2 presenta gráficos de dispersión de cada característica; dicha figura se utiliza para ejemplificar la no separabilidad lineal de los datos.

3 Preprocesamiento de Datos

3.1 Estandarización

Los datos característica se estandarizaron con la intención de mejorar el desempeño del modelo. Esto asegura que todas las caracterísitcas contribuyan equitativamente al modelado y facilitan la interpretación de estas.

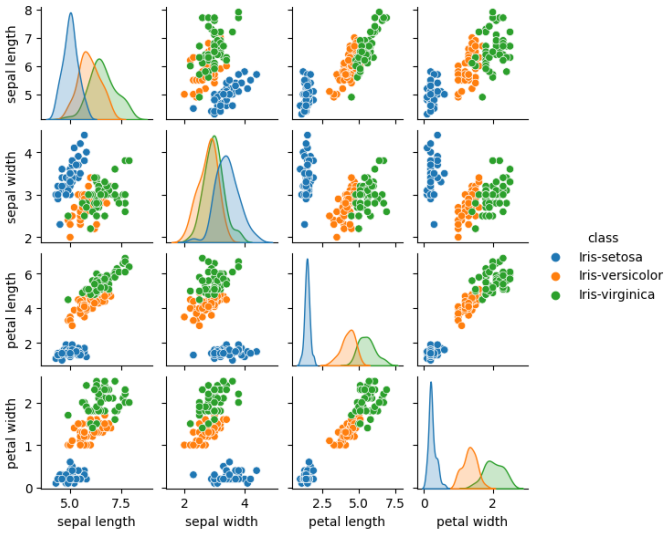

Figura 2. Gráficos de dispersión: características de irir.

3.2 Validación Cruzada Anidada

Como parte del modelo y la evaluación se emplea del método de validación cruzada anidada para determinar los hiperparámetro de mejor desempeño y obtener una evaluación robusta. Para esto de designaron dos ciclos de validación cruzada una interna para el ajuste de hiperparamétros y una externa para la evaluación. Ambos ciclos cuentan con la misma estrategia: Un enfoque de *K-fold* de 5 segmentos y mezcla de datos.

Es relevante recalcar que dada la validación cruzada anidada, no se generó una segmentación inicial y global de entrenamiento y prueba sobre el set de datos. En cambio se utilizan diversas configuraciones de los 5 segmentos 30 registros cada uno para entrenar (4 de 5) y validar/evaluar (1 de 5).

4 Modelado: *SVM One vs One*

Como se mencionó ya previamente, el modelo empleado es una máquina de vectores de soporte multiclase a través de una metodología *One vs One, SVM OvO*. A continuación se presenta el proceso de modelado.

4.1 Validación Cruzada Anidada: Ciclo Interno

Retomando lo previamente mencionado, se utilizó validación cruzda anidad para ajustar los hiperparametros. El ciclo interno varió el kernel y el coeficiente de regularización del modelo, C.

Cuadrícula de validación cruzada:

- kernel: lineal (*linear*), polinomial (*poly*), Función de núcleo de base radial (rbf) y sigmoide (*sigmoid*).
- C: 0,1, 1, 10, 100

La métrica empleada en el proceso de validación es exactitud, el razonamiento de esta selección se detalla en la sección de evaluación.

El cuadro 1 presenta los resultados de la validación cruzada bajo la métrica mencioanda. Delinado de color verde se denota la configuración seleccionada, $C = 10$, kernel: lineal. Es de interés mencionar que la configuración $C = 100$, kernel: lineal, tiene el mismo desempeño, no obstante al tener un mayor valor de C, representa un mayor costo computacional.

C	Kernel	Promedio	Desviación Estándar
0.1	linear	0.96	0.03887
0.1	poly	0.84	0.08794
0.1	rbf	0.89333	0.08
0.1	sigmoid	0.87333	0.08
1	linear	0.95333	0.03887
1	poly	0.92	0.08
1	rbf	0.95333	0.08
1	sigmoid	0.9	0.08
10	linear	0.96667	0.03887
10	poly	0.94667	0.08
10	rbf	0.94	0.08
10	sigmoid	0.86	0.08
100	linear	0.96667	0.03887
100	poly	0.92	0.08
100	rbf	0.92	0.08
100	sigmoid	0.82	0.08

Cuadro 1. Validación cruzada anidada: ciclo interno, resultados, métrica exactitud.

4.2 Modelado

El modelo se entrenó utilizando los siguientes hiperparámetros, aquellos no presentados emplearon los valores por defecto de la función sklearn.svm.SVC() (Pedregosa et al., 2011).

Hiperparámetros:

- Estado Aleatorio: 666; garantiza la reproducbilidad
- Metodología Multiclase: *One vs One*
- Kernel: lineal
- Coeficiente de Regularización (C): 10

5 Evaluación

La métrica de evaluación utilizada es la exactitud, la cual evalúa el porcentaje de predicciones correctas generadas por el modelo. La selección de dicha métrica se debe a el balance perfecto de las clases objetivo lo que hace de la exactitud una métrica robusta.

5.1 Validación Cruzada Anidada: Ciclo Externo

El ciclo externo de validación cruzada anidada genera el siguiente set de exactitudes por segmento: $100{,}00\,\%$, $100{,}00\,\%$, $93{,}33\,\%$, $90{,}00\,\%$, $93{,}33\,\%$. De esto se obtiene un promedio de $95{,}33\,\%$ con una desviación estándar de $0{,}0400$.

5.2 Curvas de Validación

Para generar un análisis mas detallado de sobre el desempeño del modelo se genraron las siguientes curvas de validación.

Las figuras 2 y 3 denotan el desempeño del modelo conforme se varían los híperparametros de la cuadrilla de validación (kernel, C). La figura 3 varía el kernel mientras mantiene el valor de $C = 10$. La figura 4 en cambio varía el coeficiente de regularización C mientras mantiene un kernel lineal.

Ambas figuras comparten ciertas características, el modelo tiene en promedio un mejor desempeño sobre los sets de entrenamiento que sobre los de validación, adicionalmente los sets de validación cuentan con una mayor varainza. Exceptuando el modelo de la función sigmoide, todos cuentan con un desempeño, exactitud, mayor al $90\,\%$ y una desviación estandar menor a $0{,}10$. El valor alto de exactitud denota poco sesgo en el modelo, es decir no presenta subajuste. La diferncia entre los desempeños de los sets de entrenamiento y validación, siendo el primero mas alto, no es suficientemente significativa como para ser signo de overfit. Adicionalmente, la baja desviación estandar ($< 0{,}10$) se puede interpretar como resultado de una buen balance entre el el sesgo y la varaianza lo que denota un buen ajuste y generalización, exceptuando el caso de la función sigmoide.

De las diversas configuraciones de kernel se puede observar claramente que el linear cuenta con el mejor desempeño promedio y la menor desviación estandar sobre los sets de validación. A su vez, el valor de $C = 10$ y $C = 100$ tienen el mismo desempeño en el set de validación, sin embargo la diferencia entre desempeño de set es mayor para $C = 100$. Debido a esto, la configuración ideal, como ya se mencionó, es la de un kernel lineal con $C = 10$.

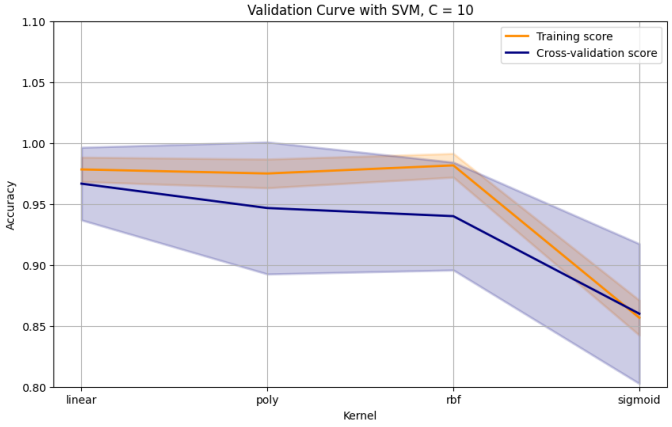

Figura 3. Curva de Validación: Kernel Variable

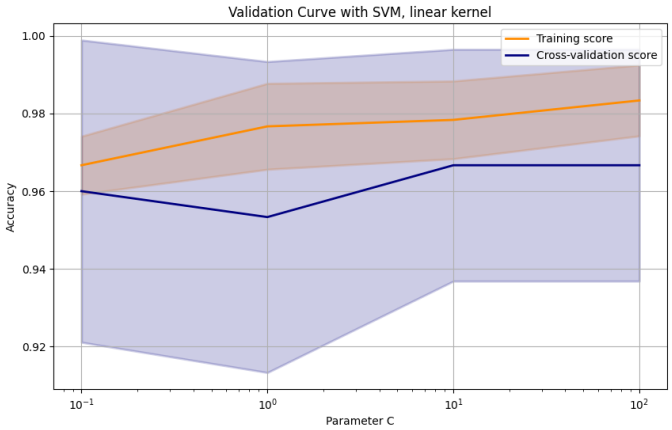

Figura 4. Curva de Validación: C Variable

La figura 5 muestra la exactitud del modelo final a lo largo de los diversos ciclos externos de la validación cruzada. Las medias de entrenamiento y validación son de 98,17 % y 95,33 % respectivamente, las desviaciones estándar son de 0,0062 y de 0,0400. La diferencia entre medias de desemempeño y los valores de desviación estándar son suficientemnte bajos como para no denotar un sobreajuste.

6 Conclusión

El desempeño del modelo es satisfactorio y las técinas empleadas muestran una clara mejora los resultados obtenidos, en especial tras el ajusto de los hiperparámetros utilizanod validción cruzada anidada como se puede observar en el cuadro 1. El modelo no muestra signos de subajuste y aunque tampoco recea en el sobreajuste, no obstante el modelo podría aún ser mejorado para distanciarse mas de un sobreajuste. Como futuras recomendaciones se plantea lo siguiente:

- Emplear un set de datos mayor. El iris dataset cuenta una cantidad módica de 150 registros. El incrementar los datos empleados podría mejorar el desempeño.
- Variar mas hiperparámetros. Solo se variarion 2 hiperparámetros en este modelelado, el hacer pruebas con otros hiperparamétros, incluso aquellos exclusivos de ciertos kernels podría generar mejores resultados.
- Emplear diferentes técincas de validación cruzaca como *K-folds* estratificada para mantener una cantidad proporcional de clases a lo largo de los segmentos y asi generalizar mejor.

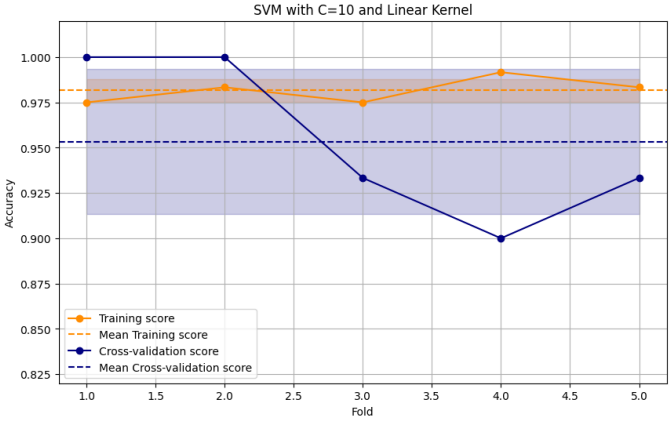

Figura 5. Curva de Validación por segmentos, modelo final

Referencias

[1] Fisher, R. (1936). Iris [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56C76.

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[3] IBM. (s.f.). Support Vector Machine. IBM. Retrieved from https://www.ibm.com/topics/support-vector-machine#:~:text=What%20are%20SVMs%3F,in%20an%20N%2Ddimensional%20space.

[4] Ploomber. (2022, abril). Model selection done right: A gentle introduction to nested cross-validation. Ploomber. Recuperado de https://ploomber.io/blog/nested-cv/.