Challenge 7

Valerie Parra Cortés

27 de abril de 2020

Preprocesamiento

Una vez se descargaron los datos, se eliminaron los registros asociados a las clases que no nos interesaban para el problema de clasificación binario. Posteriormente codificamos las clases. En este caso codificamos la clase *Jazz&Blues* con 0, mientras que la clase *Soul & Reagge* representado por el número 1.

Dependencias

1. Keras Versión 2.3.1
2. Numpy Versión: 1.18.1
3. Sklearn Versión 0.22.2

División de los datos

Se usó el 10 % de los datos totales para la etapa final de prueba. Del 90 % restante se tomó el 10 % para validación entre modelos. La división final de los datos se muestra a continuación:

	Clase 0	**Clase 1**
Entrenamiento	3509	3254
Validación	425	410
Prueba	400	352
Total	4334	4016

Cuadro 1: División de los datos para el entrenamiento, la validación y la prueba

Selección de modelo

Para resolver el SVM se escogió un kernel polinomial, por lo que uno de los hiperparámetros a ajustar era el orden de dicho polinomio, sumado a esto escoger un correcto valor de la constante C también era importante, por lo que se uso validación cruzada para ajustar estos parámetros. Los valores de los hiperparámetros utilizados en la validación cruzada se lista a continuación:

1. **Constante C** 1,0.1,0.001,0.001,0.0001,0.00001
2. **Grado polinomio** 1,2,3,4,5,6
3. **Gamma** Los valores que gamma podía tomar según la documentación de *Keras* puedes escoger dos diferentes valores para el coeficiente del kernel $1/(|features| * \sigma^2_{features})$ o bien $1/|features|$. Se iteró sobre estos dos valores pero no hubo diferencia en el desempeño final del modelo sobre este parámetro.

Resultados

Se graficaron los resultados obtenidos. En la Figura 1 vemos que hay gran variación los *accuracy*, hay en general buenos resultados (encima del 60 %) para la gran mayoría de modelos excepto para valores muy pequeños de la constante C. Adicionalmente se miro la gráfica desde el en dos diferentes cortes para poder ver como influía de manera independiente tanto C como el grado del polinomio en el desempeño final del modelo. En la Figura 2 vemos el comportamiento de la SVC para varios valores de C. En el gráfico se muestra que para valores de C muy pequeños, (menores a 1) el desempeño del modelo cae notoriamente, por otro lado para los valores de C entre 1 y 25 se tienen mejores resultados, estando todos por encima del 70 % en datos de validación. Paralelamente vemos en la Figura 3 un desempeño bajo en cada grado probado que corresponde a los valores de C pequeños, a excepción es ta línea hay resultados muy variados en el SVM implementado, en general hay muy buenos desempeños, pero el mejor desempeño se logra con C=10 con un polinomio de grado 3, con una exactitud del 85 %. La tabla completa con los resultados de todos los modelos realizados puede encontrarse en la sección de Anexos

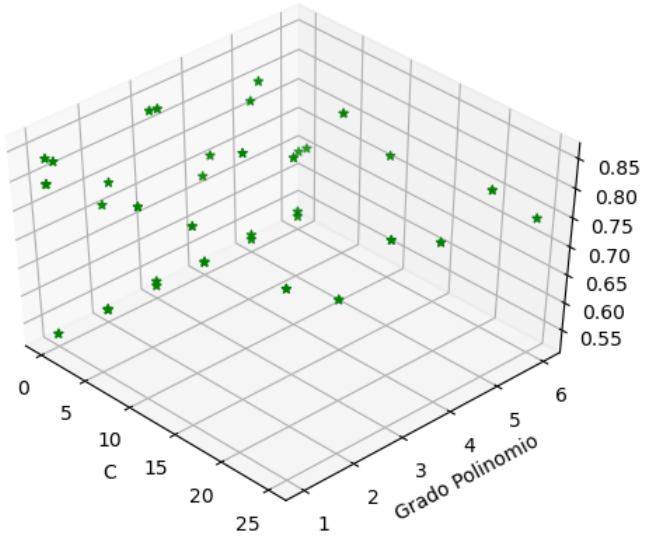

Figura 1: Vista lateral de los *Accuracy* logrados. En el eje x se encuentra el valor de C, en el eje y se encuentra el grado del polinomio y en el eje z el *accuracy* logrado

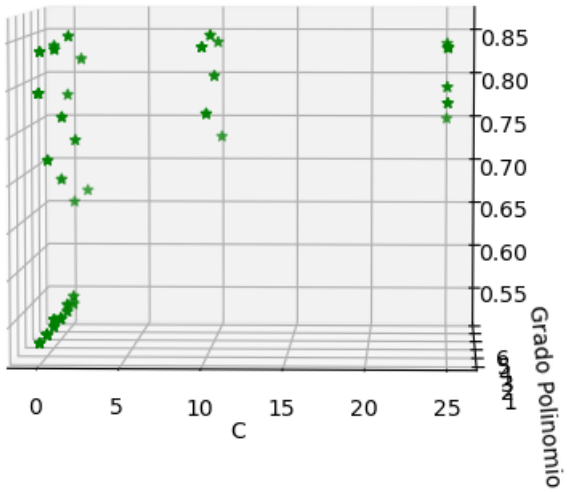

Figura 2: Vista Accuracy-Grado, En el eje x se encuentra el valor de C, en el eje y se encuentra el grado del polinomio y en el eje z el accuracy logrado

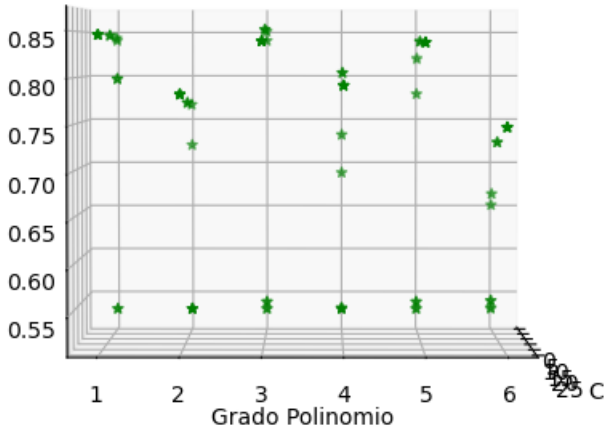

Figura 3: Vista Accuracy-C. En el eje x se encuentra el valor de C, en el eje y se encuentra el grado del polinomio y en el eje z el *accuracy* logrado

Entrenamiento y validación del modelo final

		Predicho	
		0	1
Real	0	352	73
	1	61	349

Cuadro 2: Matriz de confusión del modelo final con los datos de prueba

Una vez seleccionado el mejor modelo se utilizaron estos parámetros para entrenar el modelo final. En los datos finales de prueba la exactitud fue del 0.839 %, muy similar a lo obtenido en los datos de validación. La matriz de confusión en los datos final de prueba puede verse en la Tabla 2, esta matriz nos muestra que discriminando por clase tenemos una exactitud de 82 % para la clase 0 y de 86 % para la clase 1, por lo que nuestro modelo comete menos errores con esta última clase. El desempeño final del modelo está alrededor del 2 % por debajo del desempeño en validación, lo cual es un cambio menor y nos indica que nuestro modelo es capaz de hacer predicciones correctas cuando se alimenta con datos nuevos.

Anexos

C	Degree	Gamma	Accuracy	C	Degree	Gamma	Accuracy
25	1	scale	0.8457447	1	1	scale	0.8430851
25	1	auto	0.8457447	1	1	auto	0.8430851
25	2	scale	0.7832447	1	2	scale	0.7672872
25	2	auto	0.7832447	1	2	auto	0.7672872
25	3	scale	0.8390957	1	3	scale	0.8510638
25	3	auto	0.8390957	1	3	auto	0.8510638
25	4	scale	0.7925532	1	4	scale	0.7327128
25	4	auto	0.7925532	1	4	auto	0.7327128
25	5	scale	0.837766	1	5	scale	0.8204787
25	5	auto	0.837766	1	5	auto	0.8191489
25	6	scale	0.7486702	1	6	scale	0.6648936
25	6	auto	0.7486702	1	6	auto	0.6648936
10	1	scale	0.8457447	0.1	1	scale	0.8404255
10	1	auto	0.8457447	0.1	1	auto	0.8404255
10	2	scale	0.7712766	0.1	2	scale	0.7207447
10	2	auto	0.7712766	0.1	2	auto	0.7207447
10	3	scale	0.8523936	0.1	3	scale	0.8404255
10	3	auto	0.8523936	0.1	3	auto	0.8404255
10	4	scale	0.8045213	0.1	4	scale	0.6888298
10	4	auto	0.8045213	0.1	4	auto	0.6888298
10	5	scale	0.8390957	0.1	5	scale	0.7792553
10	5	auto	0.8390957	0.1	5	auto	0.7792553
10	6	scale	0.7273936	0.1	6	scale	0.6515957
10	6	auto	0.7273936	0.1	6	auto	0.6515957

C	Degree	Gamma	Accuracy	C	Degree	Gamma	Accuracy
0.001	1	scale	0.7965426	0.0001	1	scale	0.5319149
0.001	1	auto	0.7965426	0.0001	1	auto	0.5319149
0.001	2	scale	0.5319149	0.0001	2	scale	0.5319149
0.001	2	auto	0.5319149	0.0001	2	auto	0.5319149
0.001	3	scale	0.5398936	0.0001	3	scale	0.5319149
0.001	3	auto	0.5398936	0.0001	3	auto	0.5319149
0.001	4	scale	0.5319149	0.0001	4	scale	0.5332447
0.001	4	auto	0.5319149	0.0001	4	auto	0.5332447
0.001	5	scale	0.5398936	0.0001	5	scale	0.5319149
0.001	5	auto	0.5398936	0.0001	5	auto	0.5319149
0.001	6	scale	0.5412234	0.0001	6	scale	0.5319149
0.001	6	auto	0.5412234	0.0001	6	auto	0.5319149
0.001	1	scale	0.7965426	1.00E-05	1	scale	0.5319149
0.001	1	auto	0.7965426	1.00E-05	1	auto	0.5319149
0.001	2	scale	0.5319149	1.00E-05	2	scale	0.5319149
0.001	2	auto	0.5319149	1.00E-05	2	auto	0.5319149
0.001	3	scale	0.5398936	1.00E-05	3	scale	0.5319149
0.001	3	auto	0.5398936	1.00E-05	3	auto	0.5319149
0.001	4	scale	0.5319149	1.00E-05	4	scale	0.5319149
0.001	4	auto	0.5319149	1.00E-05	4	auto	0.5319149
0.001	5	scale	0.5398936	1.00E-05	5	scale	0.5319149
0.001	5	auto	0.5398936	1.00E-05	5	auto	0.5319149
0.001	6	scale	0.5412234	1.00E-05	6	scale	0.5332447
0.001	6	auto	0.5412234	1.00E-05	6	auto	0.5332447