Fundamentos de Machine y Deep Learning

Ejercicio N° 2 (Mejoras y Comparación de Modelos de Clasificación)
11 / 09 / 2023

INTEGRANTES

- Camila Constanza Aguilera Bustamante
- Wladimir Richard Parada Rebolledo
- Néstor Patricio Rojas Ríos
- Ramiro Andrés Uribe Garrido

Pregunta 1

¿Cuál es la proporción entrenamiento/test que logra mejor desempeño y con cuál de los modelos entre RandomForest, SVM, NB?

Al desarrollar los 3 modelos con las 4 proporciones establecidas en el enunciado para dividir los datasets de entrenamiento y validación, aquella que presentó mejor desempeño fue ***80/20***, tanto en ***exactitud***, como ***especificidad*** y ***sensibilidad***. Ahora bien, de los 3 modelos evaluados, el que mejor rendimiento mostró fue ***Random forest***. La tabla 1 resume el desempeño de los modelos, siendo destacado aquel que obtuvo el mejor resultado.

Pregunta 2.1

Viendo que un balance de clases de 1.2 (sólo un 20% más de ejemplos de la clase negativa sobre la positiva), donde reduce notoriamente la cantidad de ejemplos de la clase negativa, ¿por qué considera que se logra esa mejoría, a pesar de eliminar de entrenamiento y evaluación esa cantidad de ejemplos originales? (Justifique con claridad, según lo que se conoce sobre la forma en que se entrenan los modelos).

Al balancear los datasets para obtener una proporción más cercana entre las clases de la variable endógena (en este caso ***1,2 a 1*** entre registros OK ***NO*** y ***YES***) estamos forzando a que el modelo encuentre aquellas diferencias en las variables exógenas que se asocian a dichas clases, según las técnicas propias a cada algoritmo. De esta forma disminuimos el sesgo que se pueda generar al entrenar un modelo con una importante desproporción de las clases de la variable a explicar.

Pregunta 2.2

Habiendo determinado en el ejercicio 1 cuál es el modelo que tiene mejor desempeño entre todos, con una mejor proporción de entrenamiento/test ¿qué características del modelo apoyan su mejor desempeño sobre los otros modelos, aunque la diferencia haya sido menor? (Justifique con claridad, según lo que se conoce sobre las características particulares de los modelos y por qué ese modelo muestra mejor desempeño que los otros).

Las características que apoyan el modelo elegido en el ejercicio 1 (***Random forest***), está directamente relacionado a los valores obtenidos en la aplicación de este modelo utilizando las distintas proporciones de datos de entrenamiento y datos de validación. En 3 de las 4 proporciones evaluadas (90/10, 80/20 y 70/30), el modelo de ***Random forest*** presenta una mayor presición, sensibilidad y especificidad, siendo solo superada por la precisión del modelo ***Support vector machine*** en la proporción 60/40. ***Random forest*** es mejor para este ejemplo al considerar su naturaleza de *ensemble learning*: el modelo es resultado de desarrollar múltiples algoritmos de árboles de decisión cuyos resultados se promedian o se deciden por moda; cada árbol es entrenado con distintos subconjuntos de datos aleatorios por lo que sus estructuras son diferentes, evitando así el sobreajuste con los datos de entrenamiento.

Pregunta 3.1

¿Cuáles son los parámetros de ejecución del modelo que dan el mejor desempeño de la Red Neuronal?

De entre los atributos se probaron más de 15 combinaciones, siendo finalmente seleccionados:

Atributos 1: Edad, Ocupación, EstadoCivil, Educación, Duración, NumContactos, EmpTasaVar y NumEmpleados.
Atributos 2: Edad, Ocupación, EstadoCivil, Educación, Duración, NumContactos y EmpTasaVar.
Atributos 3: Edad, Ocupación, EstadoCivil, Educación, Duración, EmpTasaVar y NumEmpleados.
Atributos 4: Edad, Ocupación, EstadoCivil, Duración, NumContactos, EmpTasaVar y NumEmpleados.

Para los nodos se probaron valores del 1 al 100, siendo finalmente seleccionados: ***25***, ***58***, ***85*** y ***100***.

De entre las iteraciones se probaron más de 100 candidatos entre 5 y 5.000, siendo seleccionados: ***600***, ***1.150***, ***3.000*** y ***3.400***. El no cambio en los indicadores desde las 1.500 iteraciones indica que no fueron necesarias más para lograr la convergencia en el ajuste de los parámetros.

Finalmente, se buscó el mejor modelo combinando las 4 opciones candidatas para cada uno de los 3 parámetros, siendo seleccionado aquel con ***Atributos 3***, ***58*** nodos y ***1.150*** iteraciones. La tabla 2 resume el desempeño de los modelos, siendo destacado aquel que obtuvo el mejor resultado.

Pregunta 3.2

¿Logra superar al mejor modelo de los primeros 3 modelos? ¿Por qué considera que si o no y qué característica distinta entre estos 2 modelos hace la diferencia? (En cualquier caso, se pide una posible y teórica explicación de por qué es mejor/peor que ese otro modelo).

Al comparar el mejor candidato de los modelos generados en la pregunta 1 (***Random forest***) con el mejor modelo de redes neuronales (***Red neuronal con 1 capa intermedia de 58 nodos***) se observa que el primero presentó un mejor desempeño en exactitud (***88,79%*** contra ***86,34%***) y en sensibilidad (***90,83%*** versus ***85,63%***), mostrando sólo 0,09 puntos porcentuales menos en especificidad (***87,10%*** contra ***87,19%*** de la red neuronal), por lo que consideramos un mejor desempeño del modelo ***Random forest***.

La diferencia podría explicarse por la naturaleza de *ensemble learning* que tiene ***Random forest***: al generarse múltiples árboles de decisión se gana en complejidad del modelo, por lo que es capaz de captar mejor la complejidad de los datos; además, al decidir el resultado mediante promedio o moda (según la naturaleza de los datos a predecir), el modelo es más robusto frente a ruidos en los datos o a valores *outliers*. Quizás si se hubiese probado con un mayor número de capas intermedias o aumentaran significativamente la cantidad de registros podría obtenerse un mejor desempeño con la ***Red neuronal***.

Anexos

Modelo	*Proporción*	*Exactitud*	*Sensibilidad*	*Especificidad*
Random forest	90/10	0,8815	0,8950	0,8697
Naive Bayes	90/10	0,7930	0,8311	0,7595
Support Vector Machine	90/10	0,8207	0,8174	0,8236
Random forest	**80/20**	**0,8879**	**0,9083**	**0,8710**
Naive Bayes	80/20	0,7999	0,8284	0,7761
Support Vector Machine	80/20	0,8250	0,8308	0,8201
Random forest	70/30	0,8826	0,9036	0,8655
Naive Bayes	70/30	0,7997	0,8292	0,7755
Support Vector Machine	70/30	0,8182	0,8419	0,7988
Random forest	60/40	0,7674	0,9059	0,8665
Naive Bayes	60/40	0,6068	0,8410	0,7727
Support Vector Machine	60/40	0,8199	0,8469	0,7979

Tabla 1. Indicadores de los modelos no redes neuronales.

Atributos	*Nodos*	*Iteraciones*	*Exactitud*	*Sensibilidad*	*Especificidad*
Atributos 1	25	3.000	0,8527	0,8240	0,8872
Atributos 2	25	3.000	0,8351	0,8074	0,8684
Atributos 3	25	3.000	0,8303	0,7498	0,9271
Atributos 4	25	3.000	0,8522	0,8280	0,8813
Atributos 1	***25***	3.000	0,8538	0,8240	0,8895
Atributos 1	***58***	3.000	0,8485	0,8172	0,8860
Atributos 1	***83***	3.000	0,8356	0,7859	0,8954
Atributos 1	***100***	3.000	0,8634	0,8514	0,8778
Atributos 1	25	***600***	0,8527	0,8231	0,8884
Atributos 1	25	***1.150***	0,8538	0,8240	0,8895
Atributos 1	25	***3.000***	0,8538	0,8240	0,8895
Atributos 1	25	***3.400***	0,8538	0,8240	0,8895
Atributos 3	**58**	**1.150**	**0,8634**	**0,8563**	**0,8719**

Tabla 2. Indicadores de los modelos de redes neuronales.