Selección de características

Diana Cordero

June 2024

1 Introducción

Dado el conjunto de datos referente a las variables que determinan el desempeño escolar de estudiantes, se tomó como base los artículos *Machine Learning to predict student performance based on well-being data: a technical and ethical discussion* y *Predicting Student Academic Performance Using Machine Learning*. Se realiza el análisis de componentes principales y, posteriormente, se entrena un modelo de Gradient Boosting usando las características reducidas por PCA, utilizando Google Colab.

2 Conjunto de datos

En la figura 1 se muestra la lectura de los datos, seguidos de la figura 2 que contiene la descripción básica del conjunto de datos

3 Estadística descriptiva

4 Metodología

4.1 Análisis de componentes principales

La técnica de *análisis de componentes principales* propone la transformación del conjunto a un nuevo conjunto sintético de variables que no están correlacionados y se encuentran ordenados de mforma que los primeros conservan la mayor parte de la variación en todas las variables originales.

	StudentID	Age	Gender	Ethnicity	ParentalEducation	StudyTimeWeekly	Absences	Tutoring	ParentalSupport	Extracurricular	Sports	Music	Volunteering	GradeClass
0	1001	17	1	0	2	19.833723	7	1	2	0	0	1	0	2
1	1002	18	0	0	1	15.408756	0	0	1	0	0	0	0	1
2	1003	15	0	2	3	4.210570	26	0	2	0	0	0	0	4
3	1004	17	1	0	3	10.028829	14	0	3	1	0	0	0	3
4	1005	17	1	0	2	4.672495	17	1	3	0	0	0	0	4

Figure 1: Lectura de datos

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2392 entries, 0 to 2391
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   StudentID          2392 non-null   int64  
 1   Age                2392 non-null   int64  
 2   Gender             2392 non-null   int64  
 3   Ethnicity          2392 non-null   int64  
 4   ParentalEducation  2392 non-null   int64  
 5   StudyTimeWeekly    2392 non-null   float64
 6   Absences           2392 non-null   int64  
 7   Tutoring           2392 non-null   int64  
 8   ParentalSupport    2392 non-null   int64  
 9   Extracurricular    2392 non-null   int64  
 10  Sports             2392 non-null   int64  
 11  Music              2392 non-null   int64  
 12  Volunteering       2392 non-null   int64  
 13  GradeClass         2392 non-null   int64  
dtypes: float64(1), int64(13)
memory usage: 261.8 KB
```

Figure 2: Descripción básica

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Figure 3: n = individuos, p = variables

Se trata de una técnica estadística multivariada de simplificación. Las nuevas variables son combinaciones linealmente independientes de las variables originales.

Consideremos un grupo de variables $x_1, x_2, ..., x_p$ cada uno con el mismo número de muestras; cada variable registra información en un vector del mismo tamaño el cual, matricialmente da el 100% de la información y se puede representar matricialmente como se muestra en la figura 3.

La comparación de dos individuos i y j es evaluada con la distancia euclidiana clásica entre:

$$d^2(i,j) = \sum_{P=1}^{p} (x_{i_P} - x_{j_P})^2 \tag{1}$$

A partir de esta matriz calculamos un nuevo conjunto $(C_1, C_2, ..., C_p)$ no correlacionados entre si, cuyas varianzas van decreciendo progresivamente. Cada elemento C_j (j=1,. . . , p) representa una combinación lineal de las variables originales $(x_1, x_2, ..., x_p)$

$$C_j = a_{j,1}x_1 + a_{j,2}x_2 + ... + a_{j,p}x_p = a'_j x \tag{2}$$

donde

$$a'_j = a_{1_j}.a_{2_j}, ..., a_{p_j} \tag{3}$$

$$X = \begin{bmatrix} x_1 \\ ... \\ x_n \end{bmatrix},$$

a'_j es una matriz de constantes y representa el peso de las variables en cada componente; las componentes explican una parte de la varianza total, pretendiendo encontrar k componentes que representen casi toda la varianza de la nube de información.

El primer componente principal representa la mayor varianza observada sujeta a cumplir la condición de ortonormalidad; el segundo componente principal se obtiene calculando los valores de a_2 de tal manera que C_1 y C_2 sean no correlacionados (ortogonales). Igualmente se eligen los siguientes componentes cada uno con menor varianza que el anterior.

```
Number of components: 5
Explained variance ratio: [0.08862982 0.08453462 0.08167878 0.08065752 0.07947324]
Cumulative explained variance: [0.08862982 0.17316444 0.25484321 0.33550073 0.41497397]
Dimensions of transformed data: (1913, 5)
```

Figure 4: Resultados

4.2 Gradient Boosting

Gradient Boosting es una técnica de aprendizaje automático que se utiliza para problemas tanto de regresión como de clasificación. Es un método basado en la idea de mejorar iterativamente modelos más débiles, comúnmente árboles de decisión, agregándolos de forma secuencial para crear un modelo final robusto y preciso. La clave de este enfoque es que cada nuevo modelo intenta corregir los errores cometidos por los modelos anteriores.

5 Resultados y conclusiones

Luego de realiar el PCA y entrenar el modelo de gradient boostong, se obtuvieron los resultados mostrados en la figura 4.

- Se han retenido 5 componentes principales, de los 14 originales. Estos 5 componentes, en conjunto, explican aproximadamente el 41.50% de la varianza total del conjunto de datos original, lo cual indica que na cantidad significativa de la información original no se captura en estos 5 componentes. Esto puede afectar el rendimiento del modelo predictivo.
- La precisión global del modelo es del 54.07%, lo que sugiere que el modelo tiene un rendimiento moderado.
- Las métricas por clase son:
  - Clase 0: Muy baja precisión, recall y f1-score, lo que indica un rendimiento muy pobre en la predicción de esta clase.
  - Clase 1: Precisión, recall y f1-score bajos, indicando un rendimiento deficiente.
  - Clase 2 y 3: Rendimiento moderado con precisión, recall y f1-score alrededor del 30%-35%.
  - Clase 4: Mejor rendimiento con una precisión del 70%, recall del 86% y f1-score del 77%.
- El modelo de Gradient Boosting muestra un rendimiento aceptable en la clase 4, pero es deficiente en las otras clases. La precisión global del 54.07% es moderada y sugiere que el modelo no está capturando bien las complejidades de los datos.

Como pasos siguientes, se considerarán otros métodos de reducción de dimensionalidad o modelos de aprendizaje supervisado que puedan manejar mejor la estructura de los datos.

6 Bibliografía

- Andrade Saltos, V. A., & Flores M., P. (2018). Comparativa entre classification trees, random forest y gradient boosting; en la predicción de la satisfacción laboral en Ecuador. Ciencia Digital, 2(4.1.), 42-54. https://doi.org/10.33262/cienciadigital.v2i4.1.
- Blanco-Murillo, D. M., García-Domínguez, A., Galván-Tejada, C. E., & Celaya-Padilla, J. M. (2018). Comparación del nivel de precisión de los clasificadores Support Vector Machines, k Nearest Neighbors, Random Forests, Extra Trees y Gradient Boosting en el reconocimiento de actividades infantiles utilizando sonido ambiental. Res. Comput. Sci., 147(5), 281-290.
- Gradient Boosting con python. (s.f.)