

Clase 4: Datos desbalanceados

Pamela E. Pairo

Posgrado Digital Accounting



En la clase de hoy...

En la primera parte:

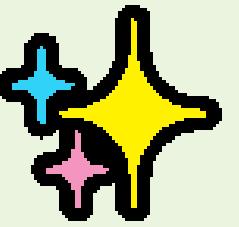
- Avisos
- Técnicas para el manejo de datos desbalanceados

Recreo ☕ 🍃

En la segunda parte:

- Integración

Avisos



LatinR 2022 (virtual)

Tutoriales (10 y 11 de octubre)



- Comunicando seus resultados: Criando apresentações com Quarto.
- Introducción a RMarkdown, creando reportes con R y Python en RStudio.
- Herramientas y atajos para programar eficientemente con RStudio.
- Introduction to Text Analysis with R.
- Conquistando errores en R.
- Creating Shiny Apps with Rhino: the new framework Shiny apps.

Visitar [la página del evento](#) para mayor información

GitHub



GitHub es un servicio basado en la nube que aloja un sistema de control de versiones (VCS) llamado Git.

¿Qué es el control de versiones?

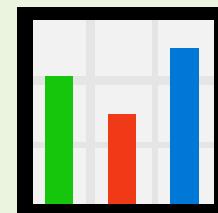
El control de versiones es un sistema que ayuda a rastrear y gestionar los cambios realizados en un archivo o conjunto de archivos.

Todo el material de las clases de aprendizaje automático, se encuentra en el siguiente repositorio

https://github.com/PamelaPairo/clases_pda

Rstudio Cloud

Datos desbalanceados



¿Qué es el desbalance de clases?



Es un problema recurrente del aprendizaje supervisado en el que una clase supera en gran medida a otra clase. Este problema se enfrenta con más frecuencia en problemas de clasificación binaria que en problemas de clasificación multiclase

El término **desbalance se refiere a la disparidad encontrada en la variable dependiente (respuesta).**



Algunos Ejemplos

- Las transacciones fraudulentas en un banco.
- Detección de un tipo de cáncer en los residentes de un área elegida.
- Predecir si los mails son spams o no.





Algunas formas de lidiar con el desbalance

- ✓ Cambiar la métrica de evaluación
- ✓ Cambiar el algoritmo
- ✓ Sobremuestrear la clase minoritaria
- ✓ Submuestrear la clase mayoritaria
- ✓ Generación de muestras sintéticas



Cambiar la métrica de evaluación

Accuracy no es la métrica adecuada cuando se tiene un dataset desbalanceado.

En su lugar, es mejor utilizar métricas que tengan más en cuenta los datos de las clases minoritarias como son la f1, la sensitividad o la precisión.

- **Matriz de confusión**
- **Precision**: el número de verdaderos positivos (TP) dividido por todas las predicciones positivas (TP+ FP). La baja precisión indica un alto número de falsos positivos.
- **Recall o Sensibilidad**: el número de verdaderos positivos (TP) dividido por el número de valores positivos en los datos de la prueba (TP+ FN). Se la denomina también Tasa de verdaderos positivos. Valores bajos indican una gran cantidad de falsos negativos.
- **F1**: el promedio ponderado de **Precision** y **Recall**



Cambiar el algoritmo

Es una buena práctica probar varios algoritmos en nuestro problema de clasificación con desbalanceo de datos.

Los vistos en clase:

- Árboles de decisión
- Ensamblés (Bagging y Boosting)
- SVM
- Naive Bayes

Existen muchísimos más!!!

Sobremuestrear la clase minoritaria



El sobremuestreo se puede definir como agregar más copias de la clase minoritaria. Puede ser una buena opción cuando no se tiene una gran cantidad de datos con los que trabajar.



Submuestrear la clase mayoritaria

El submuestreo se puede definir como eliminar algunas observaciones de la clase mayoritaria. Puede ser una buena opción cuando se tiene una cantidad grande de datos (ej. millones de datos).

Desventaja : se está eliminando información que puede ser valiosa. Esto podría dar lugar a un ajuste inadecuado y una mala generalización del conjunto de testeo.



Generación de muestras sintéticas

SMOTE or Synthetic Minority Oversampling Technique

SMOTE usa un algoritmo de vecinos más cercanos (KNN) para generar datos nuevos y sintéticos que podemos usar para entrenar nuestro modelo.

Se genera un conjunto aleatorio de observaciones de la clases minoritaria para cambiar el sesgo de aprendizaje del clasificador hacia la clase minoritaria.



Tener en cuenta

- En **R**, necesitamos instalar el siguiente paquete

```
1 install.packages("themis")
2 library(tidymodels)
3 library(themis)
```

- Hacer el split de los datos **ANTES** de probar técnicas de sobremuestreo submuestreo.

Si se hace lo contrario, puede ocurrir que los mismos datos estén presentes tanto en el conjunto de entrenamiento como en el conjunto de testeo (*data leakage*), causando el *overfitting* y la baja generalización del modelo.



Referencias

- Artículo de Towards to Data Science
- Tratamiento de clases desbalanceadas
- Practical Guide to deal with Imbalanced Classification Problems in R
- Themis