Subject: Probability and Statistics

Class XVII_1: Imbalanced Data

UdeA

Bioengineering

Francisco José Campuzano Cardona

Bioengineerer, MSc in Engineering

Evaluación de modelos de Clasificación

UdeA

Como mencionamos antes, un caso con datos desbalanceados, constituye un problema del caso raro. Y hay varias estrategias para enfrentarse a este.

Lo primero es modificar el umbral de clasificación.

Se puede hacer usando la gráfica PR buscando el umbral que acorte la distancia de la curva a la esquina superior izquierda, por inspección.

O tanteando hasta mejorar completamente la clasificación de 1s, si estos son muy costosos y los falsos positivos no son costosos.

Existen otras estrategias.

UdeA

Submuestreo

Si se cuenta con muchos datos, se puede submuestrear la clase dominante, dado que se supone que al ser tan prevalente, hay muchos *records* redundantes.

Entonces se hace un muestreo aleatorio simple, de la clase dominante para reducir la cantidad de *records* de esta clase, y al ajustar un modelo con menos datos pero más balanceados, puede mejorar el desempeño del modelo. Pero cuánto puedo reducir los datos, es un pregunta que depende del contexto y de la separación de las clases.

UdeA

Sobremuestro

El submuestreo tiene el problema de que no usa todo los datos que tiene a la mano, lo cual es una crítica razonable, máxime si se tiene un set de datos no muy grande, este enfoque puede ser perjudicial. Entonces se tiene el enfoque de sobremuestreo de la clase rara, para esto se repiten *records* de la clase rara.

Luego se evalua el modelo nuevamente y se determina si mejora la clasificación.

UdeA

Modificar los pesos

Otra alternativa, puede ser ajustar manualmente los pesos que se asigna a cada *record*, haciendo que los pesos de la clase rara sean mayores, de modo que la sumatoria de pesos de la clase rara, sea aproximadamente igual a la sumatoria de pesos de la otra clase. Para esto el método *fit( )* tiene este parámetro:

sample_weight : ***array-like of shape (n_samples,), default=None***

Weights applied to individual samples (1. for unweighted).

Nuevamente se evalúa el modelo y se determina su desempeño

Subject: Probability and Statistics

UdeA

Bioengineering

¡Thanks!

Francisco José Campuzano Cardona

Bioengineering. MSc in Engineering