Tarea 5

1. Detalles administrativos

ESTA TAREA ES OPCIONAL. Ver como calcular la nota con/sin entregar esta tarea en el programa del curso.

La entrega de esta tarea es el Lunes 20 de diciembre, al **mediodia**, por cuestionario en SIDING. Debe entregarse un zip que contenga los archivos necesarios para correr un solo notebook, que ya debe venir con todos los pedazos de código ya ejecutados. El archivo debe llamarse **NumeroAlumno_Apellido_Nombre**.

2. El problema de los datos desbalanceados

Un problema de clasificación con datos desbalanceados se presenta cuando la mayoría de los datos corresponden a una clase en particular, o cuando una clase solo es representada por una minoría. Los datos desbalanceados presentan un problema para la clasificación: muchos algoritmos van a tender a ignorar las predicciones hacia clases representadas por las minorías, aunque son precisamente estas predicciones las importantes!

2.1. Visualizar el problema

Trabaja con el set de datos Fashion MNIST de la tarea 4.

- Genera un conjunto de datos de entrenamiento $X_{\text{train}}, y_{\text{train}}$ que tenga 10.000 ejemplos de una prenda p, y solo 500 de otra, digamos q, y un set de test $X_{\text{test}}, y_{\text{test}}$ que contenga 1000 ejemplos de cada una de las dos prendas (y que sean distintos a los del conjunto de datos, por supuesto).

- Entrena una regresión logística o un random forest con los datos de entrenamiento. Mira la matriz de confusión, y el accuracy de predecir cada clase, cuando usamos el set de test. ¿Qué puedes decir al respecto?

2.2. Solucionar el problema

Para solucionar el problema vamos a usar un enfoque llamado Synthetic Minority Oversampling Technique, o SMOTE (`https://www.cs.cmu.edu/afs/cs/project/jair/pub/`

volume16/chawla02a-html/chawla2002.html), que consiste en una forma para samplear datos sintéticos de la clase minoritaria. Vamos a comparar los siguientes enfoques (la definición de SMOTE que vamos a usar está en la próxima página).

1. Genera un nuevo conjunto de datos de entrenamiento de la siguiente forma. Toma los 500 datos de la prenda q, y copialos 9 veces más para obtener un dataframe de 5000 filas, y únelo con todos los datos de la prenda p de X_{train} (serían en total 15.000). Entrena ahora tu clasificador con estos datos, y vuelve a probarlos con el set de test. ¿Mejora algo?

2. Genera un nuevo conjunto de datos de entrenamiento de la siguiente forma. Utiliza SMOTE para samplear otros 500 datos adicionales de la prenda q, y agregalos a X_{train}. Entrena ahora tu clasificador con estos datos, y vuelve a probarlos con el set de test. ¿Mejora algo?

3. Genera un nuevo conjunto de datos de entrenamiento de la siguiente forma. Utiliza SMOTE para samplear otros 4500 datos adicionales de la prenda q, y agregalos a X_{train}. Entrena ahora tu clasificador con estos datos, y vuelve a probarlos con el set de test. ¿Mejora algo?

4. Entrena un autoencoder para reducir X_{train} a 10 dimensiones. Utiliza SMOTE sobre el dataset encode(X_{train}) para samplear otros 500 datos adicionales de la prenda q, y luego decodifica todo junto para obtener un nuevo conjunto de entrenamiento sobre las dimensiones originales, con 11.000 filas. Entrena ahora tu clasificador con estos datos, y vuelve a probarlos con el set de test. ¿Mejora algo?

5. Entrena un autoencoder para reducir X_{train} a 10 dimensiones. Utiliza SMOTE sobre el dataset encode(X_{train}) para samplear otros 4500 datos adicionales de la prenda q, y luego decodifica todo junto para obtener un nuevo conjunto de entrenamiento sobre las dimensiones originales, con 15.000 filas. Entrena ahora tu clasificador con estos datos, y vuelve a probarlos con el set de test. ¿Mejora algo?

Finalmente, cuál de los 5 enfoques funciona mejor? Responde las siguientes tres preguntas:

- ¿Sirve de algo SMOTE?
- ¿Conviene samplear poco o harto con SMOTE?
- ¿Conviene samplear en el espacio reducido por un autoencoder, o en el espacio completo?

2.3. SMOTE

La implementación de SMOTE que haremos es simple. Dado un dataset X de elementos, con D dimensiones (features) en el que todos pertenecen a una cierta clase y, la generación de un conjunto S de N nuevos elementos sintético para la clase y es:

- $N = \emptyset$.
- Para j entre 1 y N:
    - Inicializar a como vector de D dimensiones.
    - Seleccionar un elemento al azar de X, llamemosle p
        - Sean $\ell_1, \ldots, \ell_5$ los k vecinos más cercanos a p. Seleccionar un ℓ_i al azar
        - Para cada dimensión $d \in D$:
            - Sea p_d y ℓ_{id} el valor de la dimension d de p y ℓ_i, respectivamente.
            - Hacer que a_d sea un número al azar k en $[p_d, \ell_{id}]$.
    - $N = N \cup a$