Maestría en Inteligencia Artificial Aplicada

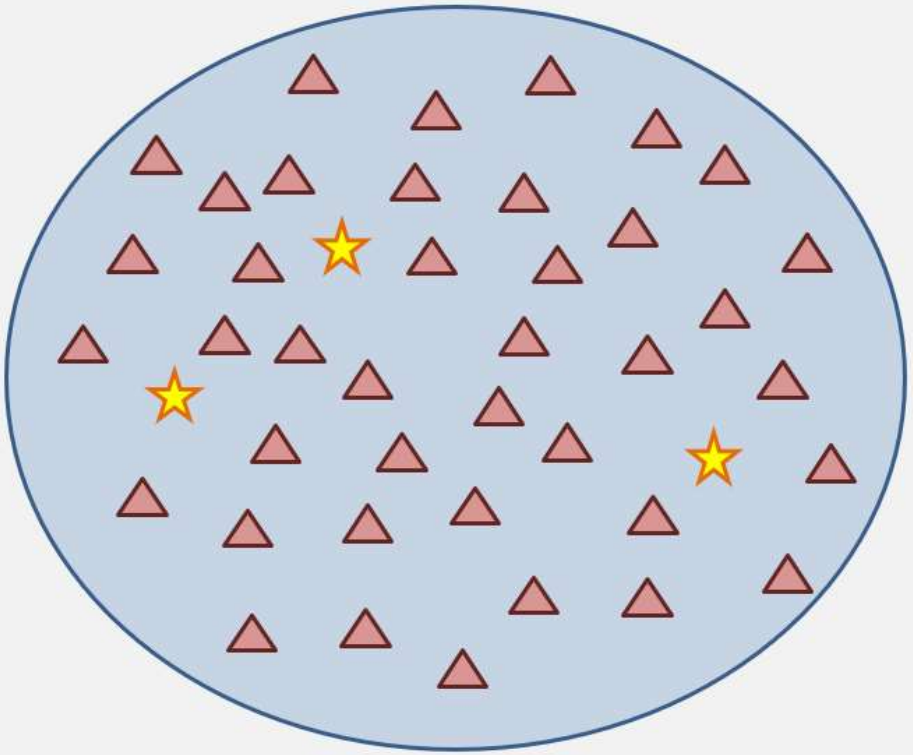
Clases no balanceadas y técnicas de Submuestreo y Sobremuestreo

Dr. Luis Eduardo Falcón Morales
ITESM

Tecnológico de Monterrey

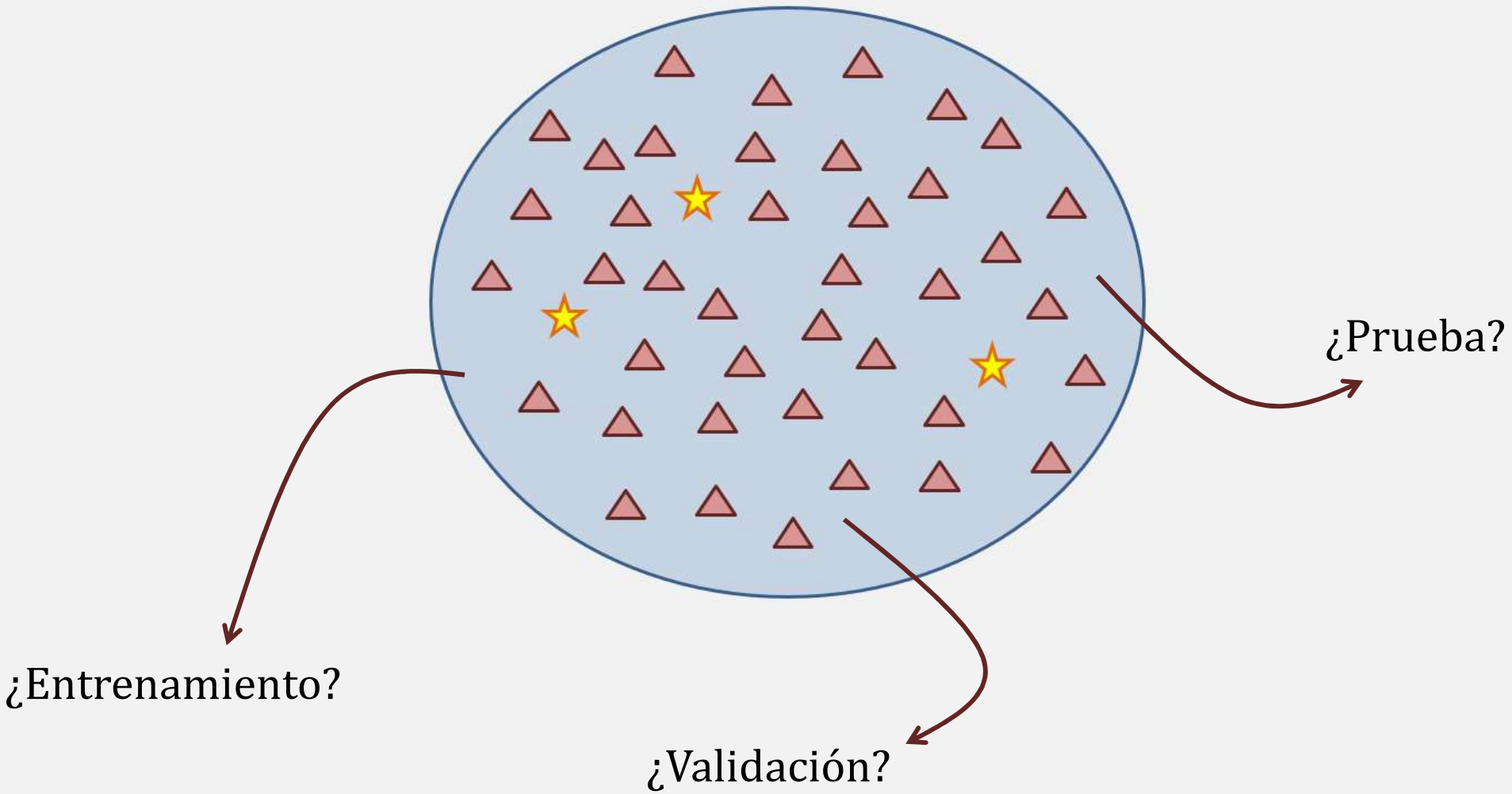
Muchos problemas de clasificación reales tienen clases muy desbalanceadas.

En general podríamos decir que un problema de clasificación bi-calse tiene clases no balanceadas, si una de las clases es menor al 20% del total de datos del universo, pero este umbral es relativo.



Es un problema que sigue actualmente abierto y sobre el cual se sigue realizando mucha investigación.

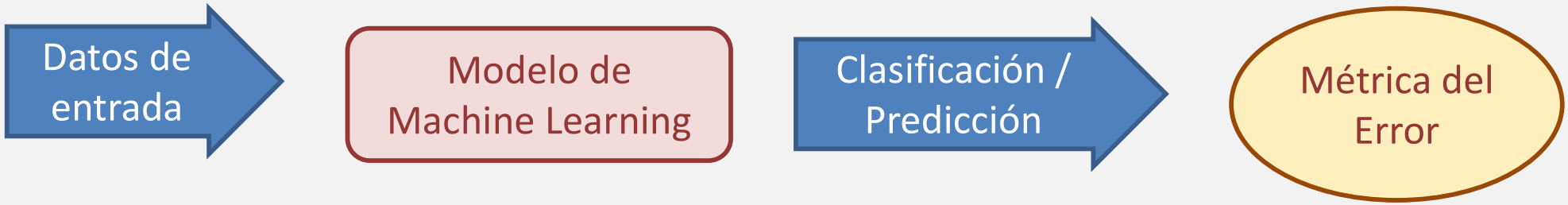
¿Cómo realizar una partición del universo de datos en el conjunto de entrenamiento, validación y prueba, respetando el porcentaje o representatividad de cada clase?



Por ejemplo:

Supongamos que tenemos una población de 1’000,000 de personas en la cual deseamos detectar si una rara enfermedad está o no presente.

Sin embargo, si dicha enfermedad se manifiesta solamente en 1 de cada millón de personas, podemos definir un modelo que siempre realice una predicción a la clase mayoritaria, de esta manera nuestro método de clasificación tendrá una exactitud del 99.9999% aunque dicho modelo no esté haciendo nada por detectar la clase minoritaria.



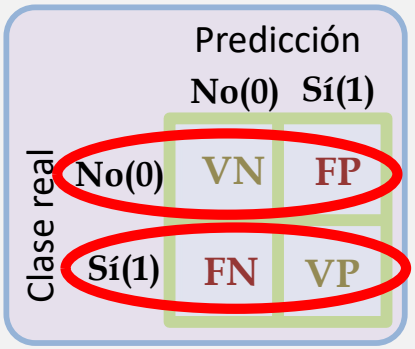
- En general se mide el desempeño de un modelo usando la métrica de la exactitud (accuracy):

Clase real	Predicción: No(0)	Predicción: Sí(1)
No(0)	VN	FP
Sí(1)	FN	VP

$$Exactidud\ (accuracy) = \frac{VP + VN}{VP + VN + FP + FN}$$

Esta métrica puede funcionar muy bien en general cuando las clases están aproximadamente igualmente distribuidas

Clases No-Balanceadas Igualmente Importantes: (métrica G-mean)



Si las clases son no balanceadas, al menos un 80%-20% aproximadamente, conviene usar la métrica *G-mean* en lugar de la exactitud (*accuracy*). Esta métrica es independiente de la distribución de los casos entre clases.

Sensibilidad (recall)

$$sensibilidad = \frac{VP}{VP + FN}$$

Tasa de verdaderos positivos

Especificidad (specificity)

$$especificidad = \frac{VN}{VN + FP}$$

Tasa de verdaderos negativos

$$G\text{-}mean = \sqrt{Sensitivity \times Specificity}$$

Media geométrica de sensibilidad y especificidad: un valor bajo en cualquier de los dos, implica un valor bajo para G-mean.

$0 \leq G\text{-}mean \leq 1$

Maestría en Inteligencia Artificial Aplicada

Clases No Balanceadas:

Sub-muestreo
y
Sobre-mustreo

Luis Eduardo Falcón Morales
ITESM – Campus Guadalajara

Tecnológico de Monterrey

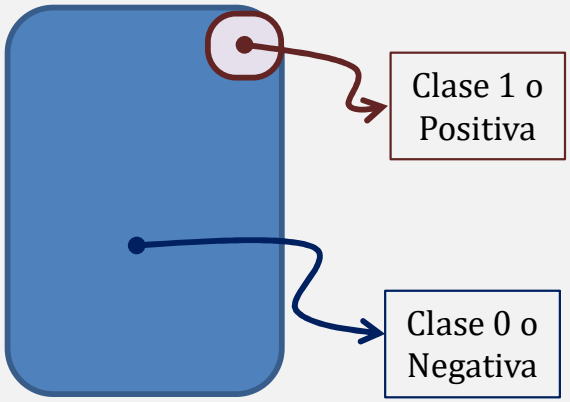
Métricas para clases No-Balanceadas

En lo sucesivo estaremos suponiendo que tenemos dos clases solamente.

Sin embargo, todo lo expuesto se puede extender a cualquier problema multiclase.

Una clase mayoritaria, que asociamos con la clase 0, "No" o negativa.

Una clase minoritaria, que asociamos con la clase 1, "Sí" o positiva.



Cuando las clases no están balanceadas conviene usar métricas que se enfocan a una sola clase:

- Sensibilidad/Especificidad (Sensitivity/Specificity)
- Precisión/Exhausitividad (Precision/Recall)

Cuando el desbalanceo empieza a ser muy marcado, se requiere algo más que el uso de otras métricas o la penalización con pesos a las clases en los métodos de optimización.

Técnicas de muestreo en problemas desbalanceados

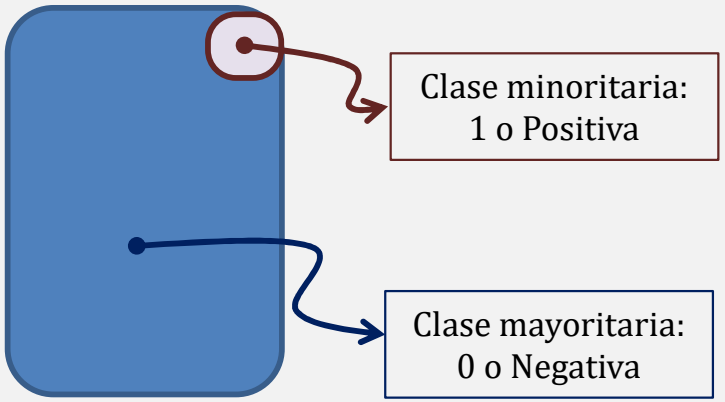
Técnicas de sobre-muestreo (over-sampling)

(Estos métodos trabajan sobre la clase minoritaria)

Técnicas de sub-muestreo (under-sampling)

(Estos métodos trabajan sobre la clase mayoritaria)

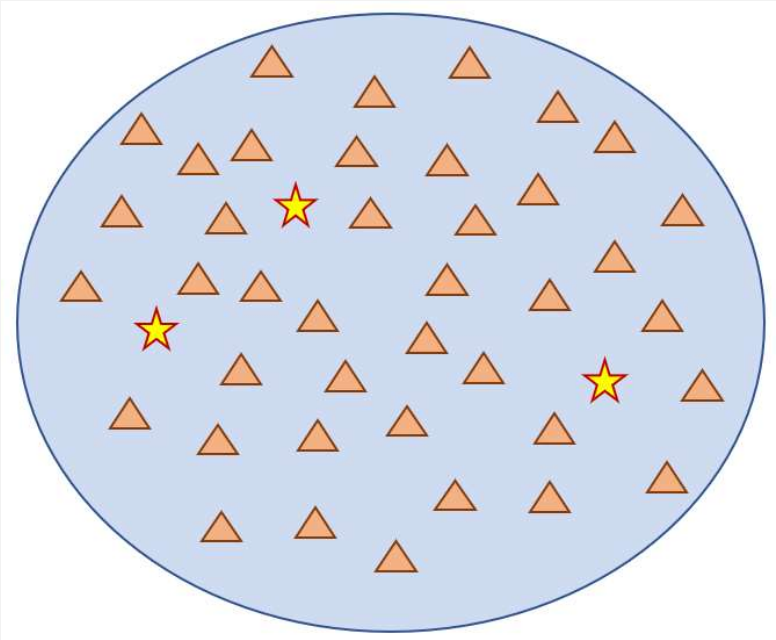
Clases no balanceadas



Técnicas de sobre-muestreo (over-sampling)

Estos métodos trabajan sobre la clase minoritaria.

Sobre-muestreo aleatorio (random over-sampling)

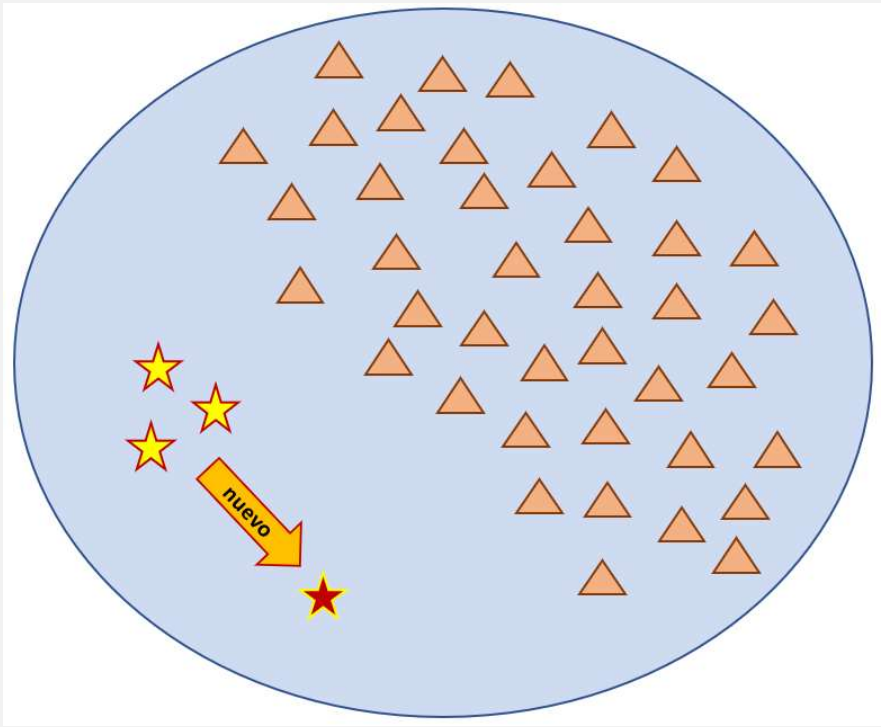


Sobre-muestreo aleatorio:

- Seleccionar aleatoriamente de la clase menor y con reemplazo (bootstrapping), un porcentaje de dichos datos.
- En general se realiza el sobre-muestreo hasta obtener dos clases del mismo porcentaje, pero pudiera ser menor.
- Repetir aleatoriamente dicho proceso y promediar los resultados, usando por ejemplo validación-cruzada.
- Tiene el inconveniente de que se están introduciendo datos *ficticios* y que son copia de algunos ya existentes.
- Solo debe aplicarse en el conjunto de entrenamiento para evitar el filtrado de información (data leakage).
- Aunque puede usarse esta técnica de sobre-muestreo de manera única, se recomienda combinarlo con sub-muestreo.

Inconvenientes de la técnica de sobremuestreo aleatorio (random oversampling):

- Tiene el inconveniente de que se están introduciendo datos *ficticios*.
- Los datos que se duplican y agregan al conjunto minoritario no están agregando valor o nueva información al modelo que se está entrenando, por ser simplemente duplicados de los datos que ya se tienen.



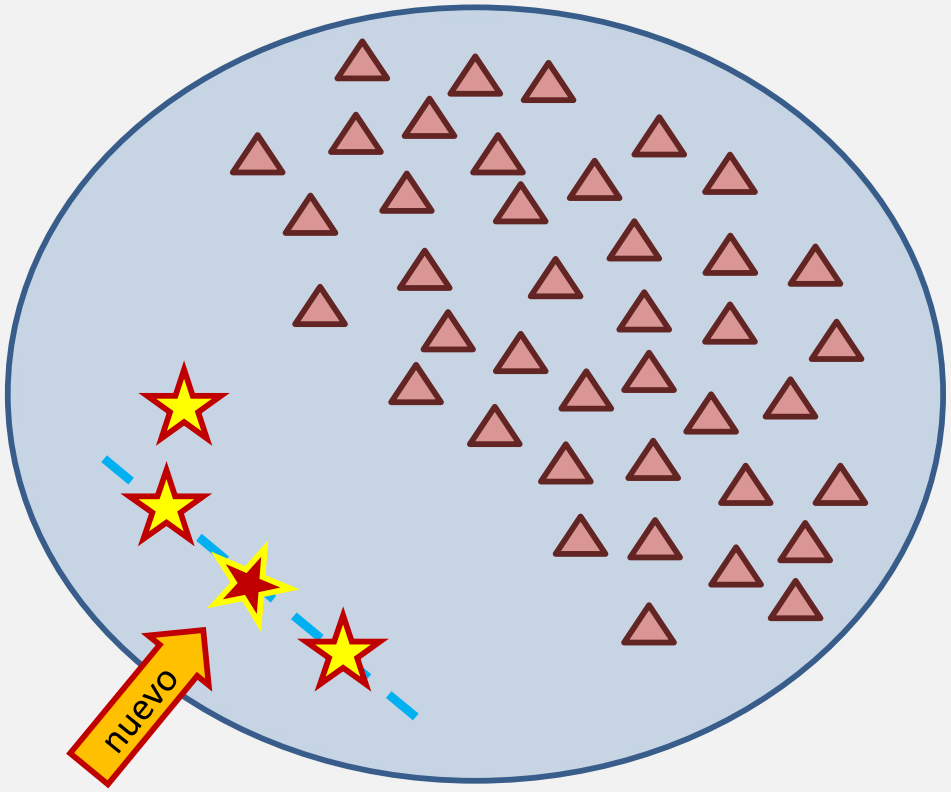
Se desearía agregar datos al conjunto minoritario que no sean una simple copia de los que ya se tienen, sino que se puedan generar datos nuevos y diferentes (*sintetizar*), a partir de los pocos datos que se tengan. Técnica que también podríamos considerar como ***data augmentation***., técnica utilizada usualmente en aprendizaje profundo (deep learning).

Synthetic Minority Oversampling Technique (SMOTE) 2002

Artículo original: https://arxiv.org/abs/1106.1813

Oversampling : SMOTE

La idea principal de SMOTE es interpolar datos nuevos a lo largo de la línea recta que une dos datos "*cercanos*" del conjunto minoritario.

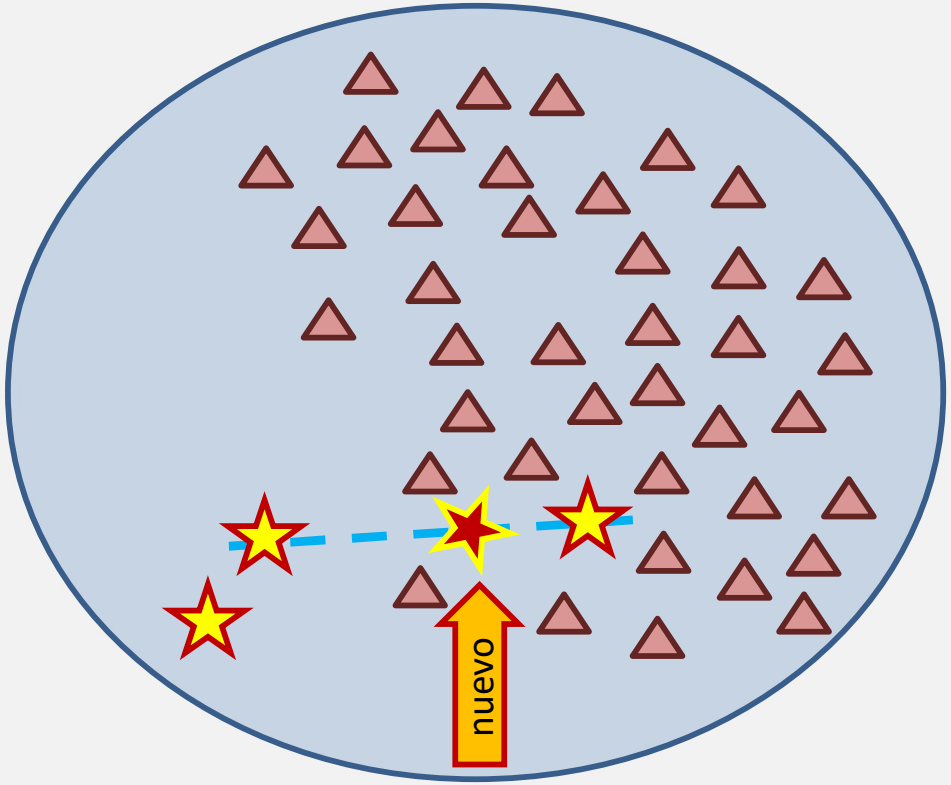


Método SMOTE:

Definir los enteros k y m:

- Seleccionar aleatoriamente un dato del conjunto minoritario, a.
- Encontrar los k vecinos más cercanos de a, y que pertenezcan a la clase minoritaria. Como valor predeterminado se toma $k = 5$.
- Seleccionar aleatoriamente un dato b de los k vecinos seleccionados. NOTA: Una variante del método es que todo el proceso se repetirá con cada uno de los k vecinos.
- Trazar la línea recta que une los datos a y b.
- Generar aleatoriamente m (1 por default) nuevos datos a lo largo de dicha línea y que estén entre a y b.

Se trabaja sobre la clase minoritaria.



Una de las desventajas de SMOTE es que se están agregando nuevos datos sin tomar en cuenta la clase mayoritaria.

En caso de existir cierto traslape entre ambas clases, los datos nuevos agregados pueden ser más parecidos a la clase mayoritaria que a la minoritaria que se desea, creando mayor ambigüedad en el modelo.

Oversampling : Borderline-SMOTE

Introduce dos mejoras a SMOTE:
Definir los enteros k y m:

i. Para cada dato p de la clase minoritaria se encuentran sus m vecinos más cercanos de ambas clases. Llamamos Ω_{pm} a cada uno de estos conjuntos.

ii. Si todos los datos de Ω_{pm} son de la clase mayoritaria, se ignora p, *i.e.*, ya no se hace nada con él pues se le considera como *ruido*. Igualmente se ingora p si la mayoría de los datos en Ω_{pm} son de la clase minoritaria y por lo tanto ya es probable que se clasifique bien y no se requiere hacerle algún ajuste.

iii. Por último, si la mayoría (pero no todos) de los datos en Ω_{pm} son de la clase mayoritaria, se agrega p a un conjunto que llamaremos DANGER, por estar formado de datos que pueden ser fácilmente mal clasificados. Este conjunto es llamado la frontera (borderline).

iv. Borderline SMOTE-1: para cada dato en DANGER seleccionar aleatoriamente cuando mucho sus k vecinos más cercanos de la clase minoritaria y aplicar SMOTE.

v. Borderline SMOTE-2 genera además datos trazando una línea con los vecinos de la clase mayoritaria en Ω_{pm}, pero a una distancia menor a 0.5, para que queden más cercanos a la clase minoritaria.

Han, H., Wang, WY., Mao, BH. (2005). **Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning**. In: Huang, DS., Zhang, XP., Huang, GB. (eds) Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, vol 3644. Springer, Berlin, Heidelberg.

Imágenes obtenidas del artículo de los autores en la liga:
https://link.springer.com/chapter/10.1007/11538059_91

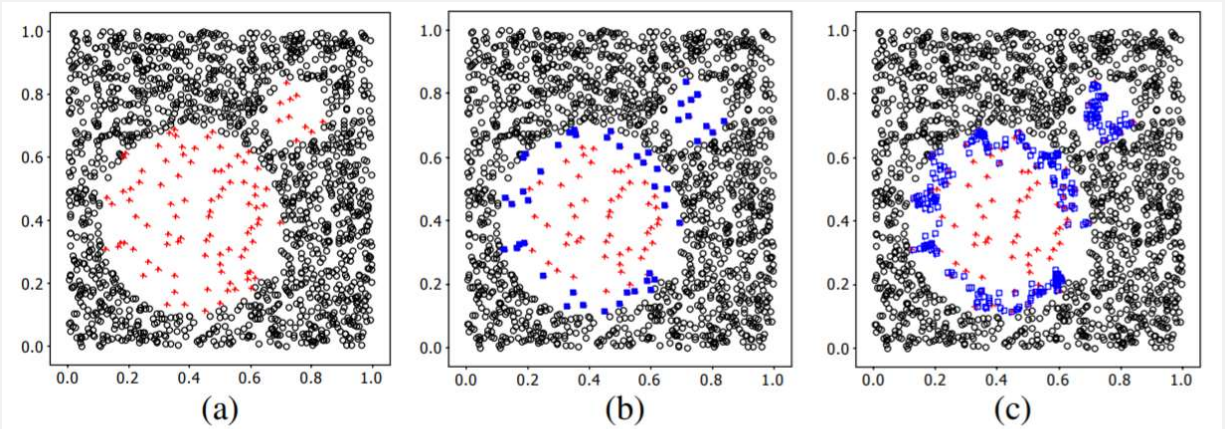


Fig. 1. (a) The original distribution of Circle data set. (b) The borderline minority examples (*solid squares*). (c) The borderline synthetic minority examples (*hollow squares*).

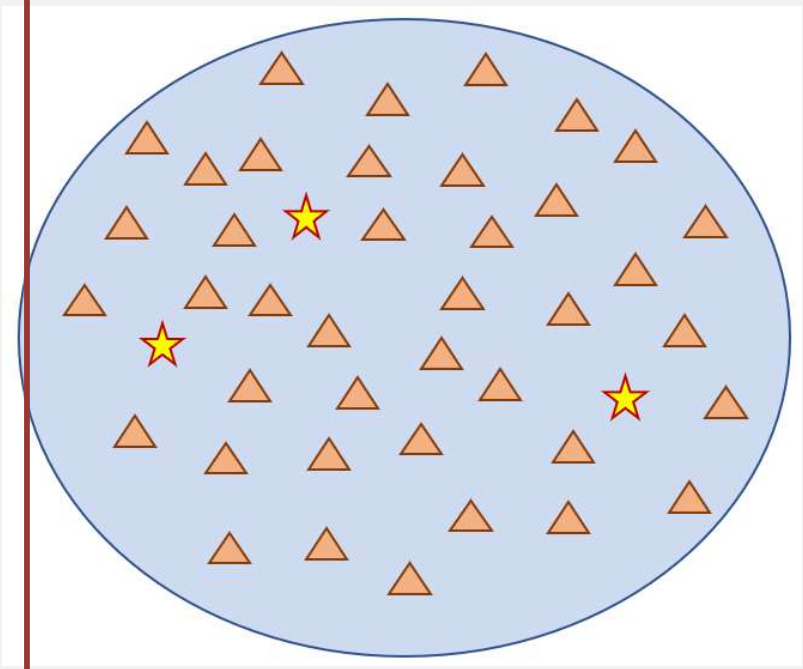
Este método trabaja solo con datos de la clase minoritaria que se encuentran en la *frontera* de ambas clases, los cuales se supone tienen mayor probabilidad de ser mal clasificados.

Se trabaja sobre la clase minoritaria.

Técnicas de sub-muestreo (under-sampling)

Estos métodos trabajan sobre la clase mayoritaria.

Sub-muestreo aleatorio (random under-sampling)



Sub-muestreo aleatorio:

- Seleccionar aleatoriamente de la clase mayor y sin reemplazo para eliminar dicho dato. En general, se realiza el sub-muestreo hasta obtener dos clases del mismo porcentaje, pero podría ser otro porcentaje.
- Existen diversas variaciones de sub-muestreo, pero en particular se basan en tres casos: seleccionar para elminar datos; seleccionar para conservarlos; el caso mixto que selcciona unos para borrar y otros para conservar.
- Repetir aleatoriamente cualquiera de dichos métodos y promediar los resultados mediante por ejemplo validación-cruzada.
- Aplicar la técnica solo en el conjunto de entrenamiento para evitar el filtrado de información (data leakage).
- **En general, es mejor usar de forma combinada sub y sobre muestreo en un conjunto de entrenamiento.**

- El tema del sub y sobre muestreo sigue siendo tema de investigación y continuamente aparecen nuevas publicaciones proponiendo nuevas formas de abordar el tema.

- En particular para las técnicas de submuestreo se buscan nuevos métodos que no borren datos de la clase mayoritaria que tienen información valiosa para el entrenamiento del modelo.

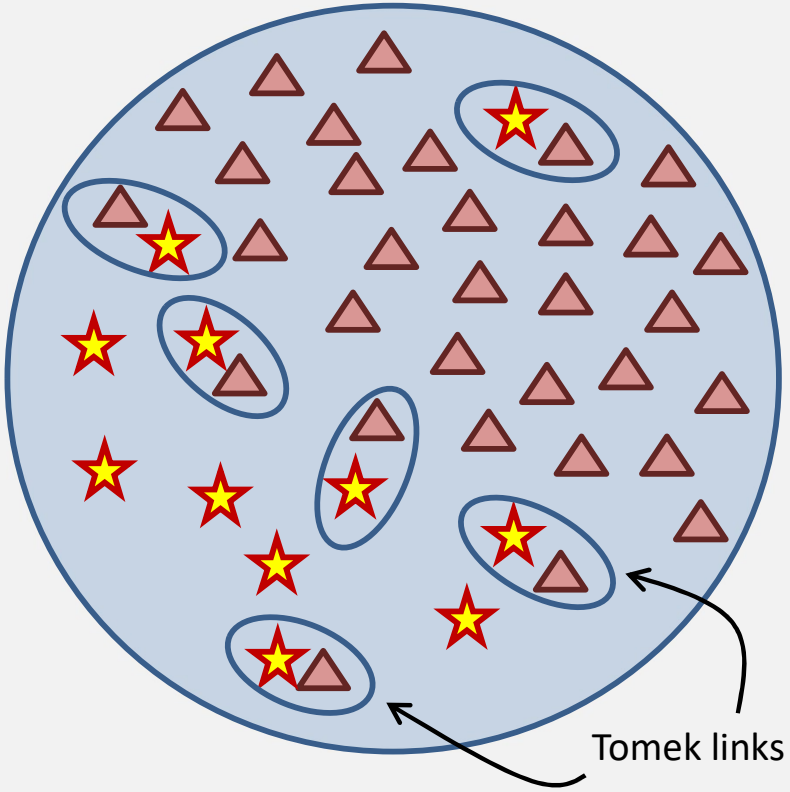
- Se proponen en general reglas heurísticas para la selección adecuada de los datos a borrar o conservar durante el submuestreo.

- Los métodos más populares de sub-muestreo son los basados en borrar los datos seleccionados y estudiaremos a continuación dos de los más populares: Tomek-links y ENN, pero existen muchos más.

Submuestreo : Undersampling : Tomek-Links (1976)

https://ieeexplore.ieee.org/document/4309452

- En técnicas de sub-muestreo los datos en la frontera generalmente son los que mayor información proporcionan.
- Se seleccionan pares de puntos, uno de la clase minoritaria y otro de la clase mayoritaria, tal que cada uno de ellos es el más cercano a cualquier otro de la clase opuesta. Usualmente se los conoce como parejas Tomek-links.
- De cada pareja Tomek-link se elimina el dato de la clase mayoritaria. En la figura, los triángulos son el conjunto de la clase mayoritaria.
- Es decir, se están eliminado los datos ambiguos, ya sea de la frontera o ruidosos.
- Como solo se eliminan los datos ambiguos, al final no necesariamente pueden quedar balanceadas las clases con algún porcentaje deseado.



- El método Tomek-link por sí solo no es de mucha ayuda, generalmente se combina con otros métodos de sub-muestreo o de sobre-muestreo.

Se trabaja sobre la clase mayoritaria.

Submuestreo : Undersampling : Edited Nearest Neighbors (ENN)

(1972)

https://www.semanticscholar.org/paper/Asymptotic-Properties-of-Nearest-Neighbor-Rules-Wilson/dea8658ee4750ec6bb408a2281cf922cbb300a0a

- Para cada dato x del conjunto de entrenamiento se encuentran sus $k = 3$ vecinos más cercanos.

- Se aplica el criterio de votación por mayoría para asignar el dato x a una clase.
  - Si x queda bien clasificado se deja igual.
  - En el caso de que el dato x quede mal clasificado:
    - Eliminar x en caso de que sea de la clase mayoritaria.
    - Si x es de la clase minoritaria, eliminar todos los datos de la clase mayoritaria que están en la vecindad de los k vecinos más cercanos.

- Nuevamente, este método no necesariamente genera al final clases balanceadas con algún porcentaje predeterminado.

- Además, ENN también por sí solo no es de mucha ayuda, generalmente se combina con otros métodos de sub-muestreo o de sobre-muestreo.

Submuestro + Sobremuestreo (Undersampling + Over sampling)

- En general se recomienda combinar ambas técnicas de submuestreo y sobremuestreo para aprovechar lo mejor de ambas técnicas.

En el siguiente artículo puedes encontrar otras muchas técnicas de submuestro y sobremuestreo:

A study of the behavior of several methods for balancing machine learning training data. Gustavo E. A. P. A. Batista and Ronaldo C. Prati and Maria Carolina Monard. SIGKDD Explor, 2004. Vol.6, pp. 20-29.

https://www.semanticscholar.org/paper/A-study-of-the-behavior-of-several-methods-for-data-Batista-Prati/6aae0dc122102693e8136856ffc8b72df7f78386

- Veamos algunos de los principales casos que combinan ambas técnicas.

Submuestreo aleatorio + Sobre muestreo aleaorio (Random undersampling + Random oversampling)

En principio uno puede proponer la combinación de los métodos anteriores en cualquier orden, siempre que te de buenos resultados. Sin embargo hay algunas combinaciones que se sabe dan en general buenos resultados. Veamos primero una de las más simples de todas.

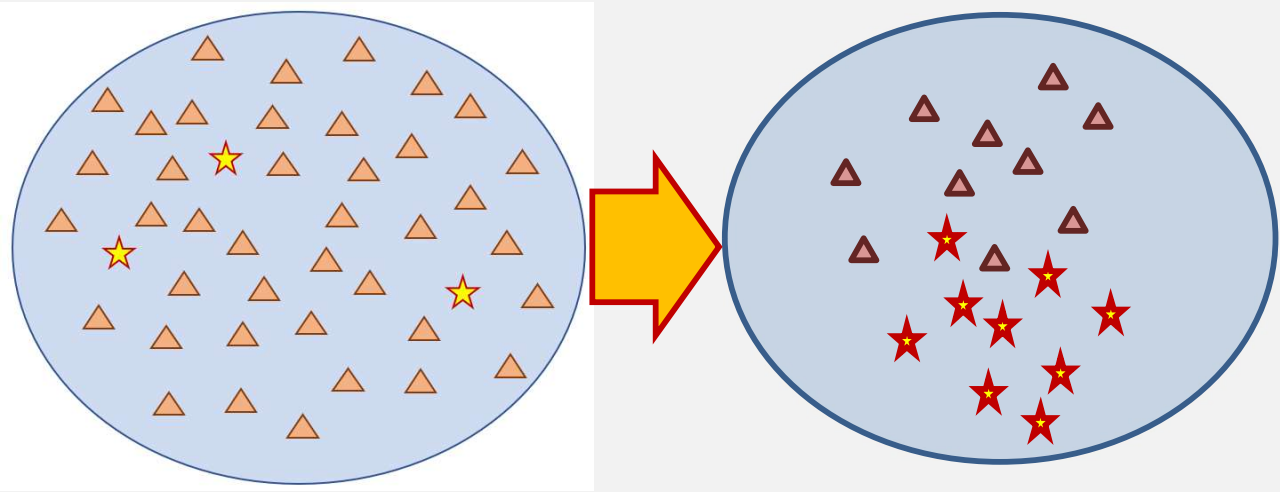
- **Se hace un muestreo aleatorio tanto con submuestreo como con sobremuestreo:**
  - Este caso de combinación de ambas técnicas tiene la ventaja de ser muy rápdo, aunque con la desventaja de que selecciona sin analizar la naturaleza de los datos.
  - Sin embargo, para los casos en lo que el desvalanceo no es muy marcado, puede tener muy buenos resultados.
  - Puede usarse también como punto de partida (baseline) para evaluar si alguna otra la mejora sutancialmente.

Sobre-muestreo aleatorio:

- Duplicar datos de la clase minoritaria con bootstrapping, indicando el porcentaje deseado.

Sub-muestreo aleatorio:

- Seleciona la cantidad de datos deseados de la clase mayoritaria, ya sea para borrarlos o para conservarlos.



En general se pueden combinar ambas técnicas como se desee, pero las siguientes son otras de las combinaciones más comunes y con las cuales puedes inciar para familiarizarte con dichas técnicas. Recuerda en todas ellas apoyarte en métricas como G-mean, recall, roc-auc, etc.

Combinación 1: SMOTE + submuestreo aleaorio (SMOTE + random undersampling)

Combinación 2: SMOTE + Tomek_Links

Combinación 3: SMOTE + ENN

⋮ ⋮

Generalmente se aplica primero la técnica de sobremuestreo y luego la de submuestreo, pero inclusive hay métodos que los van aplicando de manera simultánea o al revés.
Lo importante es experimentar y encontrar la técnica que mejor aplique en tu caso.