Temas complementarios

Contenido

- Medidas de desempeño basadas en la matriz de confusión
- Curva ROC e Interpretación del AUC
- Tratamiento de datos desbalanceados

Matriz de Confusión

Matriz de confusión

MATRIZ DE CONFUSIÓN		PRONOSTICADOS	
		Positivo	**Negativo**
REAL	**Positivo**	Verdaderos Positivos (TP)	Falsos Negativos (FN)
	Negativo	Falsos Positivos (FP)	Verdaderos Negativos (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

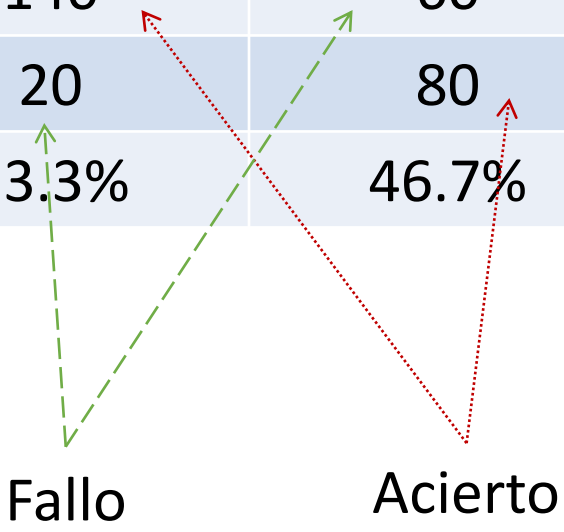
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Matriz de Confusión

Observado	Pronosticado	Pronosticado	Pronosticado
	Sí	No	% correcto
Sí	140	60	70%
No	20	80	80%
Global	53.3%	46.7%	73.3%



Matriz de Confusión

Observado	Pronosticado	Pronosticado	Pronosticado
	Sí	No	% correcto
Sí	140	60	70%
No	20	80	80%
Global	53.3%	46.7%	73.3%

Verdaderos positivos (TP)

Falso positivo (FP)

Falso negativo (FN)

Verdaderos negativos (TN)

Precisión y Exhaustividad

- **Precisión / Especificidad (precision):** Verdaderos positivos de todas las predicciones positivas.

- **Exhaustividad / Sensibilidad (recall):** Verdaderos Positivos de todos los que son realmente positivos.

Precisión y Exhaustividad

Observado	Pronosticado	Pronosticado	Pronosticado
	Sí	No	% correcto
Sí	140	60	70%
No	20	80	80%
Global	53.3%	46.7%	73.3%

Verdaderos positivos (TP)

Falso positivo (FP)

Falso negativo (FN)

Verdaderos negativos (TN)

$$Precisión = \frac{TP}{TP + FP} = \frac{140}{140 + 20} = 87.5\%$$

$$Exhaustividad = \frac{TP}{TP + FN} = \frac{140}{140 + 60} = 70\%$$

F-Score

Media Armónica de precisión y exhaustividad

$$F - Score = 2 * \frac{precisión \ * exhaustividad}{precisión + exhaustividad} = \ 77.7\%$$

$\mathbf{F_\beta}$-Score

Es una generalización del F-Score en la que un valor β pondera el peso (importancia) de la precisión y la exhaustividad.

$$F_\beta - Score = (1 + \beta^2) \times \frac{precisión \times exhaustividad}{(\beta^2 \times precisión) + exhaustividad}$$

- $\beta < 1$ le da mayor importancia a la precisión (más interés en reducir falsos positivos).
- $\beta > 1$ le da mayor peso a la exhaustividad (más interés en reducir falsos negativos).
- Si $\beta = 1$ se obtiene la media armónica de la precisión y la exhaustividad: F1-Score.

Overall Error Rate

Total de falsos positivos y falsos negativos dentro de las predicciones hechas.

$$OER = \frac{FP + FN}{FP + FN + TP + TN} = 26.7\%$$

Curva ROC y AUC

Métricas para comparar modelos

Matriz de Confusión - Precisión

Limitaciones de la precisión ("accuracy") :

Supongamos un problema con 2 clases:

- 9990 ejemplos de la clase 1
- 10 ejemplos de la clase 2

Si el modelo de clasificación siempre dice que los ejemplos son de la clase 1, su precisión es

9990/10000 = 99.9%

Totalmente engañosa, ya que nunca detectaremos ningún ejemplo de la clase 2.

Métricas para comparar modelos

Curva ROC

- Desarrolladas en los años 50 para analizar señales con ruido: caracterizar el compromiso entre aciertos y falsas alarmas.
- Permiten comparar visualmente distintos modelos de clasificación.
- El área que queda bajo la curva es una medida de la precisión (accuracy) del clasificador:
  - Cuanto más cerca estemos de la diagonal (área cercana a 0.5), menos preciso será el modelo.
  - Un modelo "perfecto" tendrá área 1.

El espacio ROC

- Se normaliza la matriz de confusión por columnas: TPR, FNR TNR, FPR.

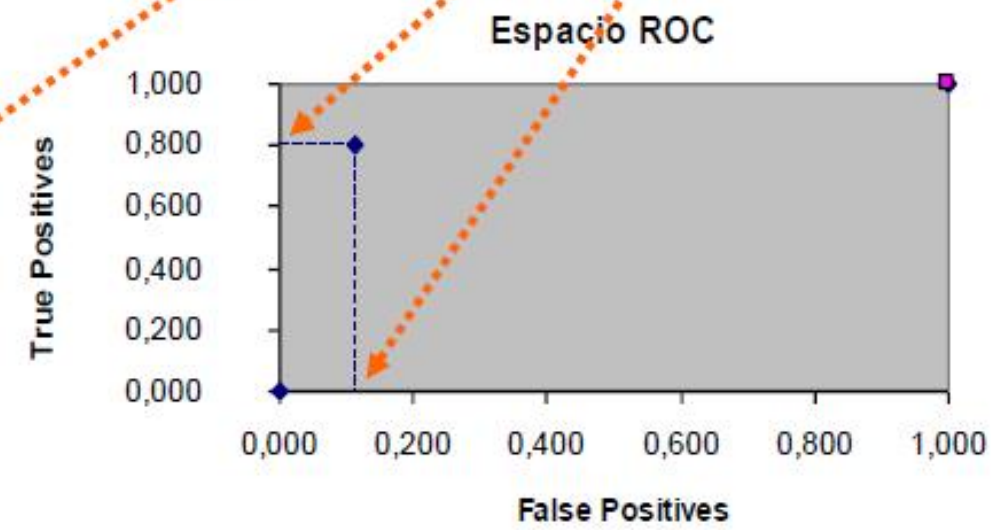
Real

Pred		abrir	cerrar
	ABRIR	400	12000
	CERRAR	100	87500

TPR= 400 / 500 = 80%
FNR= 100 / 500 = 20%
TNR= 87500 / 99500 = 87,9%
FPR= 12000 / 99500 = 12,1%

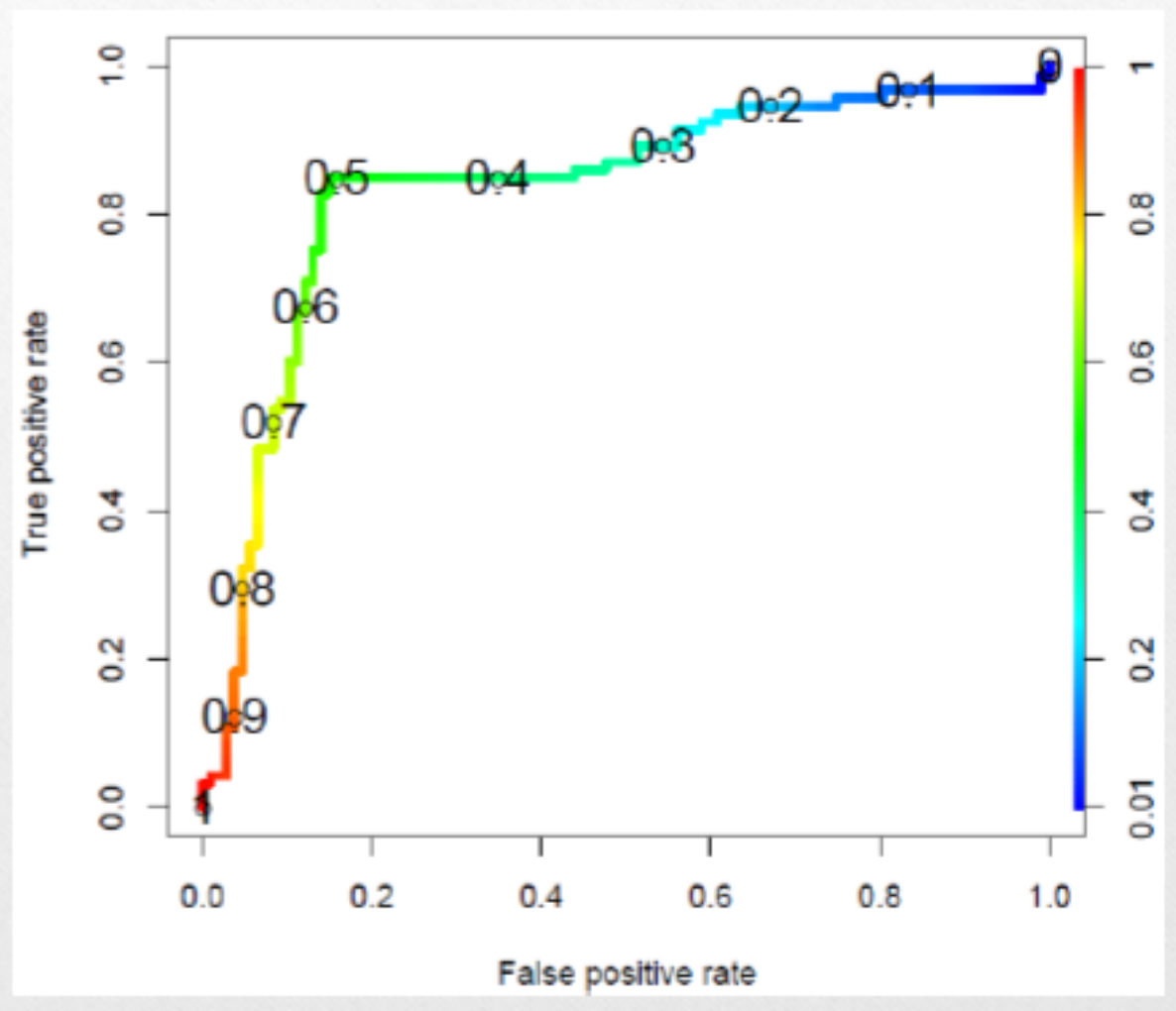
Real

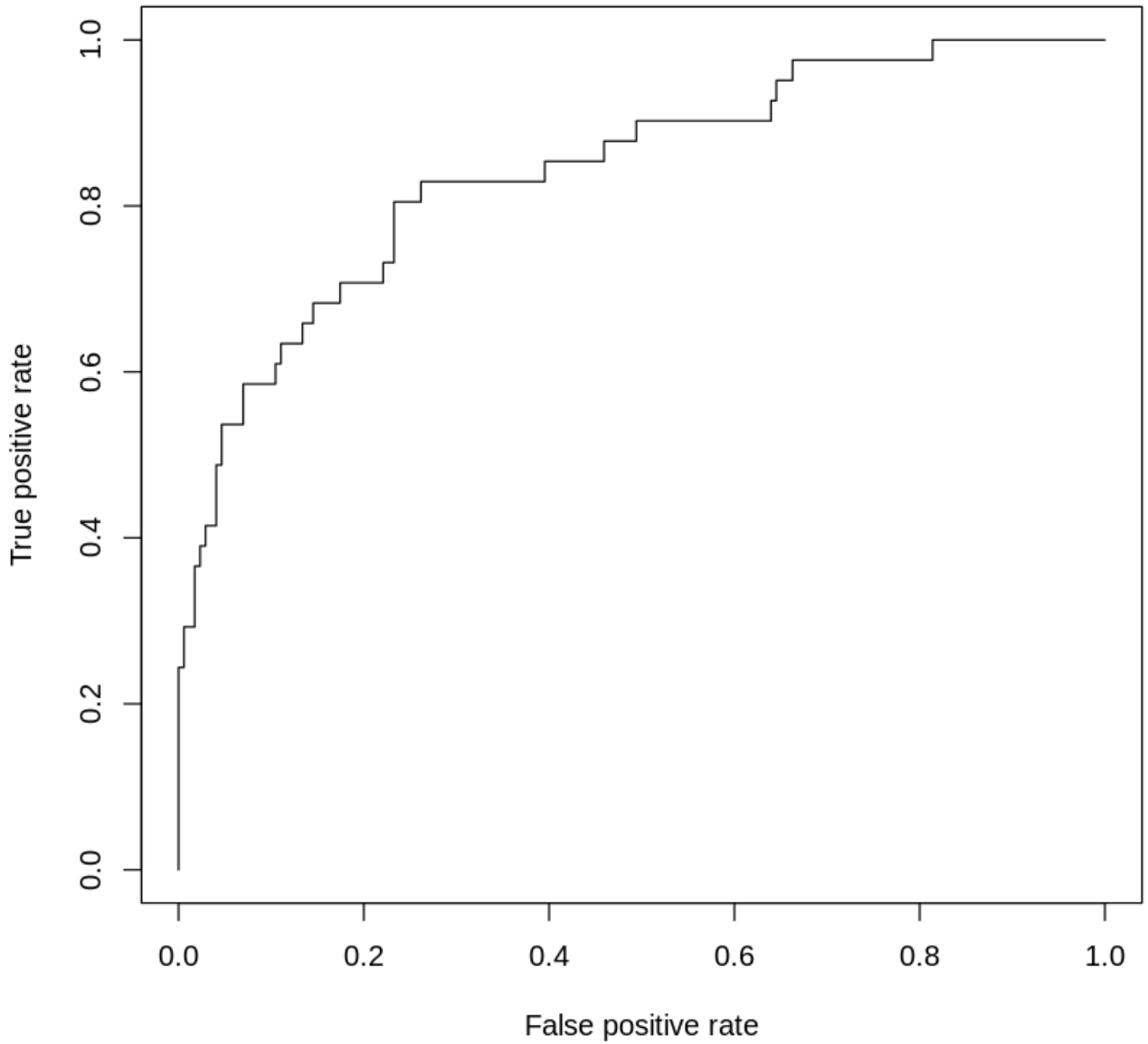
Pred		abrir	cerrar
	ABRIR	0,8	0,121
	CERRAR	0,2	0,879



Curva ROC

- **Discretización:**
  Seleccionando un valor de corte:
  $f(x) \geq c \rightarrow$ clase = 1
  $f(x) < c \rightarrow$ clase = -1




True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Intrepretación del AUC

- El area bajo la curva es de 84.3%, lo cual está bien por encima del 50% de una predicción al azar.
- La curva ROC ranquea todos las empresas desde el que tiene la probabilidad más alta de rechazo hasta la más baja (equivale a variar el umbral desde alto hasta bajo). Comenzando en el origen, mapea todos las empresas en orden descendiente de probabilidad (desde el “mejor” hasta el “peor”). Un clasificador perfecto, con exactitud 100%, primero predeciría correctamente todos los positivos, y luego predeciría correctamente todos los negativos; es decir, la curva iría recto hasta el punto (0,1), y luego cambiaría y sería horizontal hasta el punto (1,1). Esto, por supuesto, no es posible en la práctica, y los “pasos” en la curva reflejan los errores ocasionales que el modelo comete. Un buen modelo cometería pocos errores positivos para los mejores clientes y pocos eroreres negativos para los peores.
- Es importante observar que un modelo que simplemente adivina al azar, tendrá como curva ROC una línea de 45 grados. Tal modelo tendría la misma probabilidad de hacer una predicción correcta que una incorrecta, sin importar si el cliente tiene una alta o baja probabilidad predecida.
- En este caso el AUC es 84.3% el cual es un buen resultado. El AUC indica la proporción de parejas concordantes en los datos; en este caso el porcentaje de parejas concordantes es aproximadamente 84.3%. Las parejas concordantes son aquellas parejas de casos positivo y negativo en el dataset para las cuales el modelo de Regresión logística puede clasificarlos correctamente.
- En el dataset de prueba, tenemos 41 positivos (empresas rechazadas) y 172 negativos (empresas sin rechazo); el número total de parejas (positivos y negativos) es 41 x 172 = 7052, de los cuales 84.3% (== 5945) tienen unos parámetros del modelo de Regresión Logística que pueden clasificarlos correctamente.

Datos desbalanceados

Datos desbalanceados

Se dice que los datos están desbalanceados si la clase a predecir tiene un bajo porcentaje dentro de los datos (< 10%). La clase a predecir es entonces la menor, y la clase con más casos es la mayor.

¿Cómo balancear?

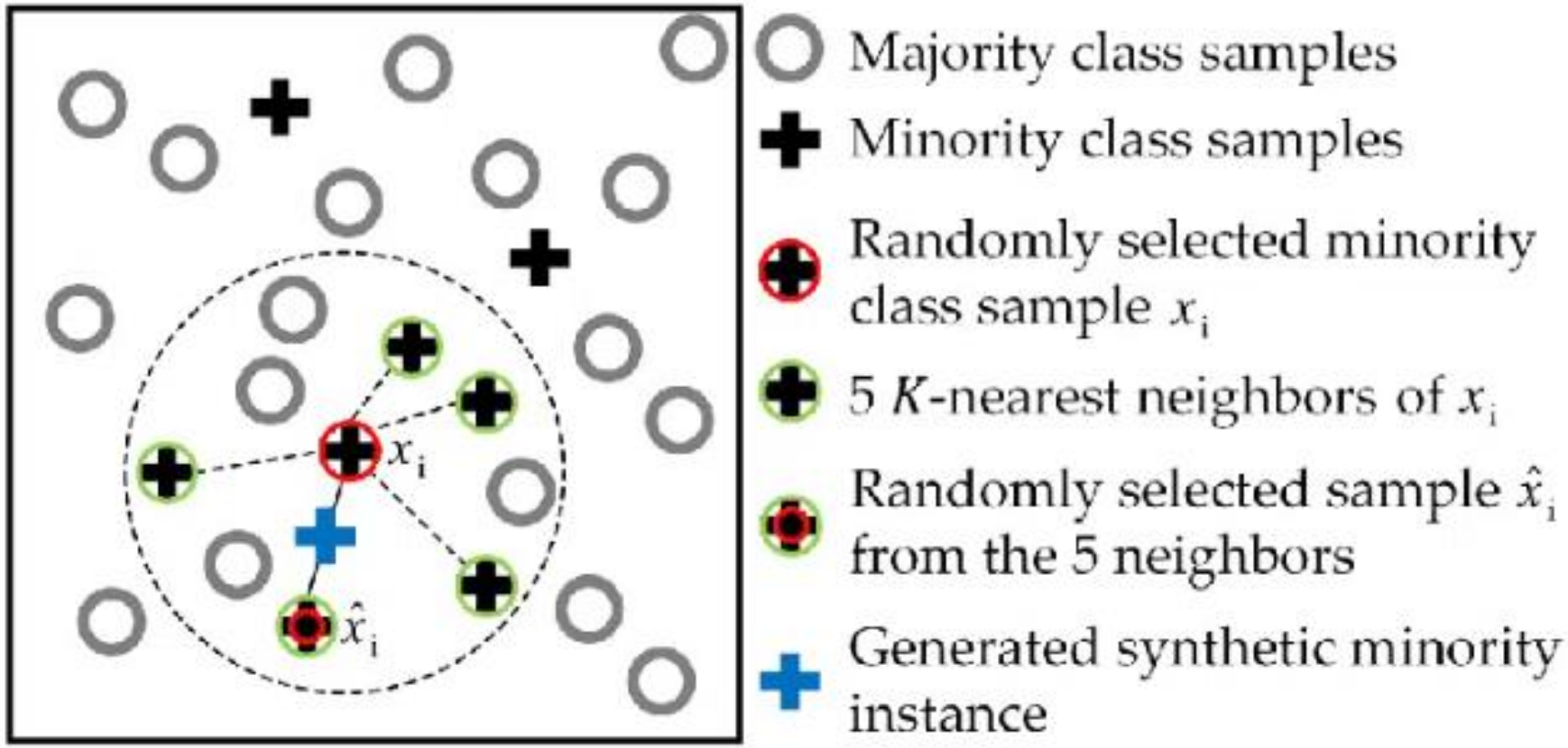
- **Undersampling:** El riesgo es la pérdida de datos.
- **Oversampling:** El riesgo es overfitting.
- **Under or Over mejorado:** Trata de superar las dificultades dadas por estos iniciales.

SMOTE

SMOTE (Synthetic Minority Over-sampling Technique): técnica de sobremuestreo utilizada para equilibrar la distribución de clases de un conjunto de datos mediante la creación de muestras sintéticas de clases minoritarias, interpolando entre las muestras de clases minoritarias existentes.

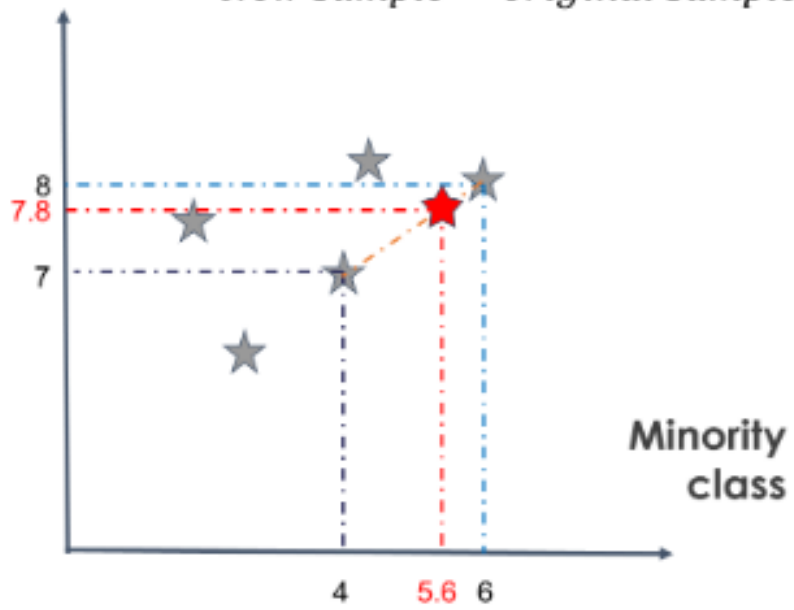
SMOTE funciona creando muestras sintéticas a lo largo de las líneas que unen a los vecinos más cercanos en el espacio de características. La idea básica de SMOTE es crear nuevas muestras de clase minoritaria dando pequeños pasos desde una de las muestras de clase minoritaria hasta uno de sus k vecinos más cercanos en el espacio de características, donde k es un parámetro del algoritmo.

SMOTE


x_i
$\hat{x}_i$
Majority class samples
Minority class samples
Randomly selected minority class sample x_i
5 K-nearest neighbors of x_i
Randomly selected sample $\hat{x}_i$ from the 5 neighbors
Generated synthetic minority instance

SMOTE: numerical example

$$New\ sample = original\ sample - factor * (original\ sample - neighbour)$$



$X_{ori} = (4,7)$

$X_{neig} = (6, 8)$

$$New\ sample = (4,7) - 0.8 * ((4,7) - (6,8))$$

$$New\ sample = (4,7) - 0.8 * ((-2,-1))$$

$$New\ sample = (4,7) - ((-1.6, -0.8))$$

$$New\ sample = (5.6, 7.8)$$

SMOTE – Algoritmo

1. Seleccione una muestra de clase minoritaria del conjunto de datos original.
2. Busque sus k vecinos de clase minoritaria más cercanos en el espacio de características.
3. Seleccione aleatoriamente uno de los k vecinos más cercanos.
4. Genere una nueva muestra sintética interpolando entre la muestra de clase minoritaria seleccionada y el vecino seleccionado aleatoriamente.
5. Repita los pasos 1-4 hasta generar el número deseado de muestras sintéticas.

Smote (Over)

Herrera (2013) Imbalanced classification: Common approaches and open problems. 2013 International School on Trends in Computing

Otras técnicas

- https://machinelearningmastery.com/data-sampling-methods-for-imbalanced-classification/