Aprendizaje Automático, conjuntos de datos desequilibrados y más métricas de clasificación

Inteligencia Artificial e Ingeniería del Conocimiento

Constantino Antonio García Martínez

Universidad San Pablo Ceu

Conjuntos de datos desequilibrados

Conjuntos de datos desequilibrados

Code Example: Clasificación de enfermedades raras

Conjuntos de Datos Desequilibrados: Definición

- **Definición:** Un conjunto de datos se considera *desequilibrado* cuando una clase supera significativamente en número a la(s) otra(s).
- **Ejemplo:** En una clasificación binaria, si el 90 % de los datos pertenece a una clase y el 10 % a la otra, el conjunto de datos está desequilibrado.
- **Problema:** Los clasificadores estándar tienden a favorecer a la clase mayoritaria, lo que lleva a predicciones sesgadas.

Matriz de Confusión y Métricas Relacionadas

Matriz de Confusión

¿Cómo detectar un clasificador que se comporta mal con datos desequilibrados?

- **Matriz de Confusión:** Una tabla que resume el rendimiento de un algoritmo de clasificación.
- **Terminología:**
- **Visualización:**

	Predicho Positivo	Predicho Negativo
Real Positivo	TP	FN
Real Negativo	FP	TN

  - **Verdaderos Positivos (TP):** Instancias positivas correctamente clasificadas.
  - **Falsos Positivos (FP):** Instancias negativas incorrectamente clasificadas como positivas.
  - **Verdaderos Negativos (TN):** Instancias negativas correctamente clasificadas.
  - **Falsos Negativos (FN):** Instancias positivas incorrectamente clasificadas como negativas.

Métricas Derivadas de la Matriz de Confusión

- **Exactitud (accuracy):** $\frac{TP+TN}{TP+FP+TN+FN}$ (Predicciones correctas totales.)

  Problema: La accuracy puede ser engañosa en conjuntos de datos desequilibrados.

Métricas Derivadas de la Matriz de Confusión

- **Exactitud (accuracy):** $\frac{TP+TN}{TP+FP+TN+FN}$ (Predicciones correctas totales.)

  Problema: La accuracy puede ser engañosa en conjuntos de datos desequilibrados.

- **Precisión:** $\frac{TP}{TP+FP}$ (Proporción de predicciones positivas que son correctas.)
- **Exhaustividad o Sensibilidad (recall):** $\frac{TP}{TP+FN}$ (Proporción de positivos reales correctamente predichos.)
- **Puntuación F1:** $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ (Media armónica de precisión y recall.)

Code Example: Matriz de Confusión

Promedio Macro, Promedio Ponderado y Clasificación Multiclase

- **Problema:** Para clasificación multiclase, calculamos precisión y recall para cada clase, pero a menudo queremos una métrica resumen del rendimiento general.
- **Promedio Macro:**
  - Promedia la métrica (ej., precisión, recall) entre todas las clases por igual, sin considerar el desequilibrio de clases.
- **Promedio Ponderado:**
  - Promedia la métrica entre clases, pero cada clase se pondera por su soporte (el número de instancias en esa clase).
- **Ejemplo:**
  - Si una clase domina el conjunto de datos, un promedio ponderado dará más importancia a esa clase, mientras que un promedio macro trata todas las clases por igual.

Promedios Macro y Ponderados en Clasificación Binaria

- **¿Podemos usarlos?**
  - Sí, tanto los promedios macro como ponderados pueden usarse técnicamente en clasificación binaria.
- **¿Por qué no se usan típicamente?**
  - La diferencia entre promedios macro y ponderados suele ser insignificante en problemas binarios, ya que solo hay dos métricas de clase para promediar.
  - Más comúnmente, se informan métricas directas como accuracy, precisión y recall para una clase (generalmente la clase positiva).
- **¿Cuándo considerarlos?**
  - En conjuntos de datos binarios altamente desequilibrados, los promedios ponderados pueden ser útiles para tener en cuenta la clase dominante.

Soluciones para la Clasificación Desequilibrada

Soluciones para la Clasificación Desequilibrada

Ajuste de Umbral y Curvas ROC

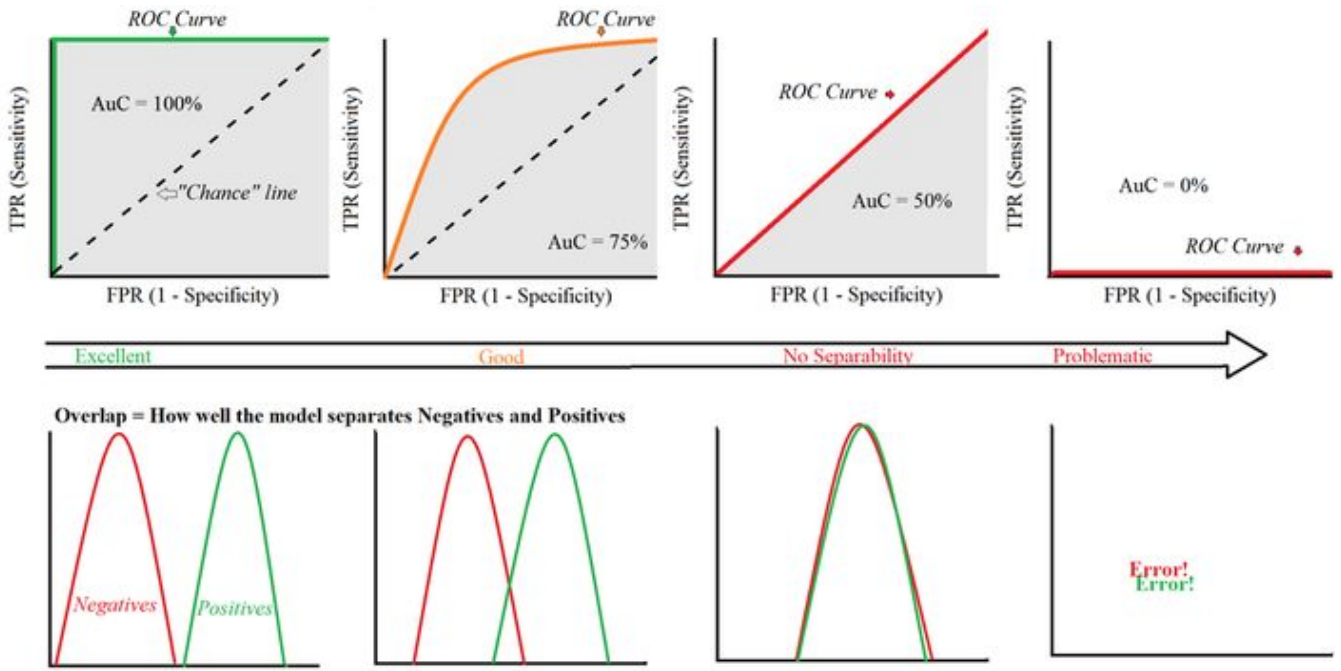
Ajustando el Umbral

- **Umbral Estándar:** El clasificador típicamente usa un umbral de 0.5 para predecir clases positivas/negativas.
- **Ajustando el Umbral:** Al bajar o subir el umbral, puedes influir en el balance entre precisión y recall.
- **Ejemplo:** Bajar el umbral aumenta la recall pero puede reducir la precisión, y viceversa.

Code Example: Umbral en Clasificación Desequilibrada

Curva ROC y AUC-ROC

- **Curva ROC:** Grafica la Tasa de Verdaderos Positivos (recall) vs. la Tasa de Falsos Positivos (TFP = $\frac{FP}{FP+TN}$) en varios umbrales.
- **AUC-ROC:** El área bajo la curva ROC, que mide la capacidad general del clasificador para distinguir entre clases.
- **Interpretación:** Un AUC más alto significa mejor rendimiento del modelo al distinguir entre clases.



Curva ROC y AUC-ROC

Code Example: Curva ROC y AUC-ROC

Soluciones para la Clasificación Desequilibrada

Ponderación de Clases (class weighting)

Ponderación de Clases (class weighting)

- **Pesos de Clase:** Asignar pesos más altos a la clase minoritaria y pesos más bajos a la clase mayoritaria durante el entrenamiento del modelo.
- **Por qué:** Esto obliga al modelo a enfocarse más en la clase minoritaria y puede ayudar a mitigar el sesgo.
- **Implementación:** Muchos clasificadores (ej., SVM, Bosques Aleatorios) permiten la ponderación de clases como parámetro.
- **Ejemplo:** Establecer pesos de clase inversamente proporcionales a las frecuencias de clase.

Code Example: Ponderación de Clases

Soluciones para la Clasificación Desequilibrada

Aprendizaje Sensible al Costo

Aprendizaje Sensible al Costo

- **Aprendizaje Sensible al Costo:**
  - Una técnica que considera el costo de los errores de clasificación, asignando diferentes penalizaciones a diferentes tipos de errores.
  - Útil en escenarios donde algunos errores (ej., falsos negativos en diagnósticos médicos) son más costosos que otros.
- **Matriz de Costos:**
  - Una matriz de costos asigna una penalización a cada resultado de clasificación.
- **Ejemplo de Matriz de Costos:**

	Predicho Positivo	Predicho Negativo
Real Positivo	0 (correcto)	5 (falso negativo)
Real Negativo	1 (falso positivo)	0 (correcto)

- **Ajuste del Clasificador:**
  - Los algoritmos pueden ajustarse para minimizar el costo total de clasificación errónea, en lugar de optimizar métricas estándar como la accuracy.
  - Muchos modelos (ej., árboles de decisión, SVM) pueden integrar matrices de costo directamente.

Research Project: Aprendizaje sensible al costo en Sklearn
Ver trabajo propuesto.

Soluciones para la Clasificación Desequilibrada

Técnicas de Muestreo

Submuestreo, Sobremuestreo, SMOTE

- **Submuestreo:**
  - Reduce el tamaño de la clase mayoritaria eliminando instancias aleatoriamente.
  - **Pros:** Reduce el sesgo hacia la clase mayoritaria.
  - **Contras:** Puede descartar información útil, llevando a subajuste.
- **Sobremuestreo:**
  - Aumenta el tamaño de la clase minoritaria replicando instancias o generando datos sintéticos.
  - **Pros:** Equilibra las distribuciones de clase sin perder datos.
  - **Contras:** Puede llevar a sobreajuste al duplicar muestras de la clase minoritaria.
- **SMOTE (Técnica de Sobremuestreo de Minorías Sintética):**
  - Genera muestras sintéticas para la clase minoritaria interpolando entre muestras existentes.
  - **Pros:** Reduce el sobreajuste introduciendo variabilidad en las muestras sintéticas.
  - **Contras:** Puede crear muestras poco realistas, potencialmente introduciendo ruido.

Research Project: Sobremuestreo, submuestreo, SMOTE, ...
Ver trabajo propuesto.

Soluciones para la Clasificación Desequilibrada

Técnicas de Ensembles (conjuntos)

Técnicas de Ensembles para Conjuntos de Datos Desequilibrados

- **Visión General:**
  - Los métodos de ensemble combinan múltiples clasificadores para mejorar el rendimiento, especialmente en conjuntos de datos desequilibrados.
  - Pueden reducir efectivamente el sesgo hacia la clase mayoritaria y mejorar la predicción para la clase minoritaria.
- **EasyEnsemble:**
  - Crea aleatoriamente múltiples subconjuntos equilibrados de la clase mayoritaria mediante submuestreo.
  - Entrena un clasificador separado en cada subconjunto equilibrado.
  - Combina las predicciones de todos los clasificadores (ej., votación por mayoría).
- **BalanceCascade:**
  - Entrena el primer clasificador en los datos originales y predice etiquetas.
  - Elimina ejemplos mal clasificados de la clase mayoritaria antes de entrenar el siguiente clasificador.
  - Continúa hasta que se construye un número específico de clasificadores.

Research Project: Métodos de Ensembles
Ver trabajo propuesto.

Para Saber Más

Para Saber Más

- **Implementación de Aprendizaje Sensible al Costo en scikit-learn**
  - **Lectura de Artículo:** Buscar al menos un artículo relevante relacionado con el aprendizaje sensible al costo con desequilibrio de clases, leerlo y resumirlo.
  - **Programación:**
    - Implementar aprendizaje sensible al costo usando matrices de costo en modelos de scikit-learn como Árboles de Decisión, SVM.
    - Comparar rendimiento con clasificadores tradicionales en conjuntos de datos desequilibrados.
- **Sobremuestreo vs. Submuestreo vs. SMOTE**
  - **Lectura de Artículo:**
    - Artículo Sugerido: "SMOTE: Synthetic Minority Over-sampling Technique"(Chawla et al., 2002).
  - **Programación:**
    - Implementar submuestreo, sobremuestreo y SMOTE usando imbalanced-learn.
    - Evaluar el impacto en el rendimiento de cada técnica usando diferentes clasificadores.
- **Métodos Avanzados de Ensembles para Datos Desequilibrados**
  - **Lectura de Artículo:**
    - Artículo Sugerido: “Exploratory Undersampling for Class-Imbalance Learning"(Liu et al., 2009).
  - **Programación:**
    - Usar EasyEnsemble y BalanceCascade.
    - Comparar su rendimiento contra métodos de conjunto tradicionales como Random Forest.