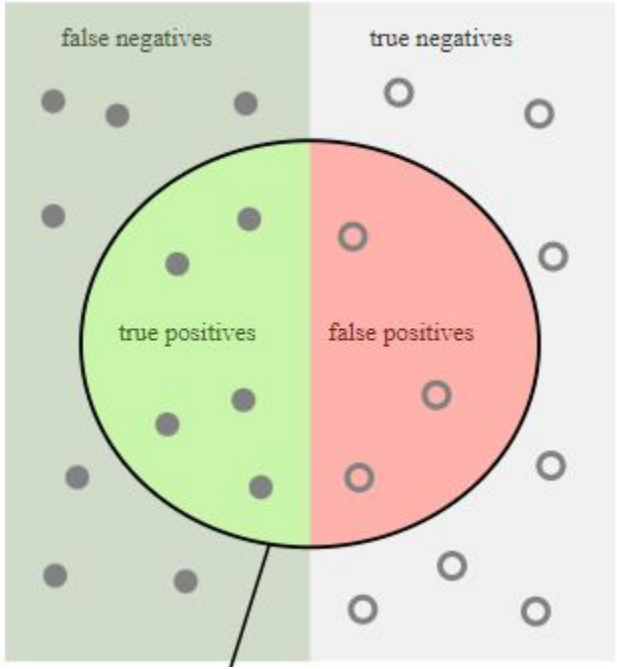
Métricas de evaluación de modelos

Curso de inteligencia artificial-Intermedio

Objetivo

Profundizar en las métricas para evaluar modelos de ML vistos en clase.



Métricas para modelos de clasificación

Exactitud

La **exactitud** se define como la proporción de las predicciones correctas (tanto positivas como negativas) sobre el total de predicciones realizadas. La fórmula es:

$$\text{Exactitud} = \frac{TP + TN}{TP + TN + FP + FN}$$

Exactitud

Interpretación de la Exactitud

- **Exactitud alta**: Indica que el modelo está haciendo un buen trabajo al clasificar correctamente tanto las instancias positivas como las negativas. Esto puede ser útil cuando las clases están bien balanceadas.

Exactitud

Interpretación de la Exactitud

- **Exactitud alta**: Indica que el modelo está haciendo un buen trabajo al clasificar correctamente tanto las instancias positivas como las negativas. Esto puede ser útil cuando las clases están bien balanceadas.

- **Exactitud baja**: Indica que el modelo está cometiendo muchos errores al clasificar las instancias. Puede ser una señal de que el modelo necesita mejorar.

Exactitud

Exactitud en Modelos Desbalanceados

- Uno de los problemas con la exactitud surge cuando tienes clases desbalanceadas. En estos casos, la exactitud puede dar una impresión falsa de que el modelo está funcionando bien, cuando en realidad podría estar ignorando la clase minoritaria.

Exactitud

Ejemplo en un conjunto de datos desbalanceado:

- Supongamos que estás entrenando un modelo para detectar fraudes en transacciones bancarias. Solo el 1% de las transacciones son fraudulentas, mientras que el 99% son legítimas. Si tu modelo simplemente predice que todas las transacciones son legítimas (es decir, nunca predice fraude), obtendrás una exactitud del 99%. Sin embargo, esto no significa que tu modelo esté haciendo un buen trabajo en detectar fraudes.

Exactitud

¿Cuándo es útil la Exactitud?

La exactitud es útil cuando las clases están aproximadamente balanceadas o cuando los errores en la clasificación de cualquiera de las clases tienen un costo similar. Algunos ejemplos incluyen:

- **Clasificación de imágenes**: Si estás clasificando imágenes en varias categorías (por ejemplo, diferentes especies de animales) y todas las categorías tienen aproximadamente la misma cantidad de ejemplos, la exactitud puede ser una métrica adecuada.
- **Reconocimiento de voz**: En aplicaciones donde el objetivo es identificar si una palabra específica ha sido pronunciada (como en asistentes de voz), la exactitud puede ser útil cuando hay un balance razonable entre las palabras positivas y negativas.

Precisión

- La **Precisión** es una métrica que indica la proporción de predicciones correctas entre todas las predicciones realizadas para la clase positiva. Es decir, de todas las veces que el modelo predijo que una instancia era positiva, ¿qué porcentaje de esas predicciones fueron correctas?

Fórmula:

$$\text{Precisión} = \frac{TP}{TP + FP}$$

Precisión

Interpretación:

- **Precisión alta**: Un modelo con alta precisión comete pocos falsos positivos. Es decir, cuando el modelo dice que algo es positivo, generalmente lo es. Esto es valioso cuando los falsos positivos son costosos o indeseables (por ejemplo, en el diagnóstico de enfermedades raras).

Precisión

Interpretación:

- **Precisión alta**: Un modelo con alta precisión comete pocos falsos positivos. Es decir, cuando el modelo dice que algo es positivo, generalmente lo es. Esto es valioso cuando los falsos positivos son costosos o indeseables (por ejemplo, en el diagnóstico de enfermedades raras).
- **Precisión baja**: Indica que el modelo genera muchos falsos positivos, prediciendo incorrectamente la clase positiva con frecuencia. Esto puede ser problemático si cada falso positivo tiene consecuencias graves (por ejemplo, etiquetar injustamente a personas como fraudulentas).

Recall (Sensibilidad)

- El **Recall** (también conocido como **Sensibilidad**) es una métrica que mide la capacidad del modelo para identificar correctamente las instancias positivas. En otras palabras, de todas las instancias que realmente son positivas, ¿cuántas veces el modelo las detectó correctamente?

Fórmula:

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

Recall (Sensibilidad)

Interpretación:

- **Recall alto**: Un modelo con alto recall detecta la mayoría de las instancias positivas, es decir, comete pocos falsos negativos. Es crucial cuando los falsos negativos son costosos o peligrosos (por ejemplo, en sistemas de seguridad o detección de enfermedades graves, donde no detectar una amenaza real es inaceptable).

Recall (Sensibilidad)

Interpretación:

- **Recall alto**: Un modelo con alto recall detecta la mayoría de las instancias positivas, es decir, comete pocos falsos negativos. Es crucial cuando los falsos negativos son costosos o peligrosos (por ejemplo, en sistemas de seguridad o detección de enfermedades graves, donde no detectar una amenaza real es inaceptable).

- **Recall bajo**: Indica que el modelo está perdiendo muchas instancias positivas, produciendo muchos falsos negativos. Esto puede ser problemático en aplicaciones donde es crítico capturar todos los casos positivos posibles (por ejemplo, en la detección de fraude financiero).

Precisión vs Sensibilidad

Relación entre Precisión y Recall

Existe una **tensión inherente entre Precisión y Recall**. Optimizar para uno generalmente significa comprometer el otro. El motivo de esta relación es que, en muchos casos, aumentar la precisión reducirá el recall y viceversa:

- **Alta Precisión, Bajo Recall**: El modelo es muy conservador en sus predicciones positivas. Solo predice positivo cuando está muy seguro, pero puede perder muchos casos positivos (alto FN).
- **Alto Recall, Baja Precisión**: El modelo es más agresivo en sus predicciones positivas. Captura la mayoría de los positivos, pero al hacerlo también comete más errores al clasificar negativos como positivos (alto FP).

Precisión vs Sensibilidad

¿Cuándo preferir Precisión sobre Recall (y viceversa)?

- **Prefieres Precisión** cuando los **falsos positivos son costosos o indeseables**.
  Ejemplos incluyen:
  - Diagnóstico médico para enfermedades tratables: No quieres alarmar a pacientes sanos innecesariamente.
  - Clasificación de correo electrónico como spam: Si la precisión es baja, podrías perder correos importantes al etiquetarlos incorrectamente como spam.
- **Prefieres Recall** cuando los **falsos negativos son costosos o peligrosos**.
  Ejemplos incluyen:
  - Detección de fraudes: Es mejor capturar todos los posibles fraudes, incluso si algunos no lo son.
  - Diagnóstico de enfermedades graves como el cáncer: No quieres perder ningún paciente que realmente esté enfermo.

F1-Score

El F1-Score equilibra estas la precisión y la sensibilidad, proporcionando un valor único que resume su rendimiento conjunto. Se define como:

$$\text{F1-Score} = 2 \times \frac{\text{Precisión} \times \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

F1-Score

Interpretación

- **F1-Score alto**: Indica que el modelo tiene un buen balance entre precisión y sensibilidad, es decir, identifica correctamente la mayoría de los casos positivos (alta sensibilidad) y también tiene pocos falsos positivos (alta precisión).

F1-Score

Interpretación

- **F1-Score alto**: Indica que el modelo tiene un buen balance entre precisión y sensibilidad, es decir, identifica correctamente la mayoría de los casos positivos (alta sensibilidad) y también tiene pocos falsos positivos (alta precisión).
- **F1-Score bajo**: Indica que hay un desequilibrio entre la precisión y la sensibilidad, lo que significa que el modelo o bien está perdiendo muchos verdaderos positivos (baja sensibilidad) o bien está cometiendo muchos falsos positivos (baja precisión).

F1-Score

Interpretación

- **F1-Score alto**: Indica que el modelo tiene un buen balance entre precisión y sensibilidad, es decir, identifica correctamente la mayoría de los casos positivos (alta sensibilidad) y también tiene pocos falsos positivos (alta precisión).
- **F1-Score bajo**: Indica que hay un desequilibrio entre la precisión y la sensibilidad, lo que significa que el modelo o bien está perdiendo muchos verdaderos positivos (baja sensibilidad) o bien está cometiendo muchos falsos positivos (baja precisión).

El F1-Score es útil cuando tienes un **desequilibrio** entre clases y quieres asegurarte de que el modelo tenga un rendimiento balanceado entre la capacidad de detectar correctamente los positivos y minimizar los falsos positivos.

F1-Score

Relación con otras métricas

- **Precisión vs Sensibilidad**: La precisión se enfoca en minimizar los falsos positivos, mientras que la sensibilidad se enfoca en minimizar los falsos negativos. El F1-Score equilibra estas dos métricas y es útil cuando ambos son igualmente importantes.

F1-Score

Relación con otras métricas

- **Precisión vs Sensibilidad**: La precisión se enfoca en minimizar los falsos positivos, mientras que la sensibilidad se enfoca en minimizar los falsos negativos. El F1-Score equilibra estas dos métricas y es útil cuando ambos son igualmente importantes.
- **Exactitud (Accuracy)**: La exactitud mide la proporción total de predicciones correctas, pero en conjuntos de datos desbalanceados (donde una clase es mucho más común que la otra), puede dar una impresión engañosa del rendimiento del modelo. El F1-Score es más adecuado en estos casos.

F1-Score

Aplicaciones prácticas

1. **Detección de fraudes o anomalías**: En estos casos, las clases positivas (por ejemplo, transacciones fraudulentas) suelen ser raras. El F1-Score es más útil que la exactitud porque enfatiza el equilibrio entre detectar fraudes y evitar etiquetar transacciones legítimas como fraudulentas.

F1-Score

Aplicaciones prácticas

1. **Detección de fraudes o anomalías**: En estos casos, las clases positivas (por ejemplo, transacciones fraudulentas) suelen ser raras. El F1-Score es más útil que la exactitud porque enfatiza el equilibrio entre detectar fraudes y evitar etiquetar transacciones legítimas como fraudulentas.
2. **Clasificación de textos (spam, opiniones, etc.)**: En problemas como la clasificación de correos electrónicos como spam o la clasificación de reseñas como positivas o negativas, el F1-Score puede proporcionar una mejor evaluación del rendimiento del modelo que simplemente la exactitud, ya que las clases suelen estar desbalanceadas.

F1-Score

Aplicaciones prácticas

1. **Detección de fraudes o anomalías**: En estos casos, las clases positivas (por ejemplo, transacciones fraudulentas) suelen ser raras. El F1-Score es más útil que la exactitud porque enfatiza el equilibrio entre detectar fraudes y evitar etiquetar transacciones legítimas como fraudulentas.
2. **Clasificación de textos (spam, opiniones, etc.)**: En problemas como la clasificación de correos electrónicos como spam o la clasificación de reseñas como positivas o negativas, el F1-Score puede proporcionar una mejor evaluación del rendimiento del modelo que simplemente la exactitud, ya que las clases suelen estar desbalanceadas.
3. **Diagnóstico médico**: En tareas de diagnóstico, el F1-Score ayuda a equilibrar la detección correcta de casos positivos (sensibilidad) con la precisión de esos diagnósticos. Un buen F1-Score es esencial cuando tanto los falsos negativos como los falsos positivos tienen un costo alto.

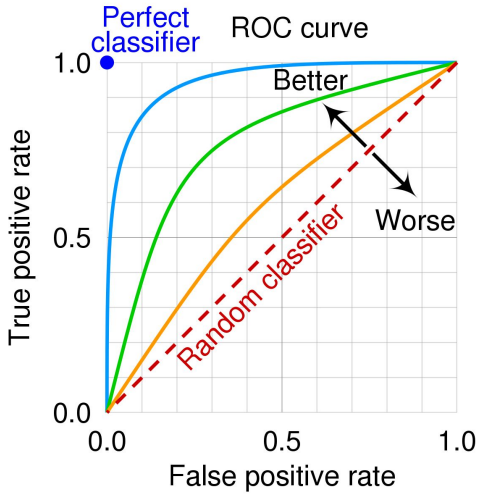
F1-Score

Aplicaciones prácticas

1. **Detección de fraudes o anomalías**: En estos casos, las clases positivas (por ejemplo, transacciones fraudulentas) suelen ser raras. El F1-Score es más útil que la exactitud porque enfatiza el equilibrio entre detectar fraudes y evitar etiquetar transacciones legítimas como fraudulentas.
2. **Clasificación de textos (spam, opiniones, etc.)**: En problemas como la clasificación de correos electrónicos como spam o la clasificación de reseñas como positivas o negativas, el F1-Score puede proporcionar una mejor evaluación del rendimiento del modelo que simplemente la exactitud, ya que las clases suelen estar desbalanceadas.
3. **Diagnóstico médico**: En tareas de diagnóstico, el F1-Score ayuda a equilibrar la detección correcta de casos positivos (sensibilidad) con la precisión de esos diagnósticos. Un buen F1-Score es esencial cuando tanto los falsos negativos como los falsos positivos tienen un costo alto.

AUC-ROC

El **Área bajo la curva ROC (AUC-ROC)** es una métrica utilizada para evaluar el rendimiento de los modelos de clasificación binaria, aunque también puede aplicarse a problemas multiclase. Es una métrica que mide la capacidad de un modelo para discriminar entre clases, sin importar el umbral de decisión elegido.



AUC-ROC

- **Curva ROC (Receiver Operating Characteristic)**: La curva ROC es un gráfico que muestra la relación entre la **tasa de verdaderos positivos (True Positive Rate, TPR)**, que es la **sensibilidad**, y la **tasa de falsos positivos (False Positive Rate, FPR)**, en diferentes umbrales de decisión.
  - **TPR (Sensibilidad)**: Es la proporción de casos positivos correctamente identificados.
  - **FPR**: Es la proporción de casos negativos incorrectamente clasificados como positivos.

El eje **y** de la curva ROC muestra la TPR (sensibilidad), mientras que el eje **x** muestra la FPR.

AUC-ROC

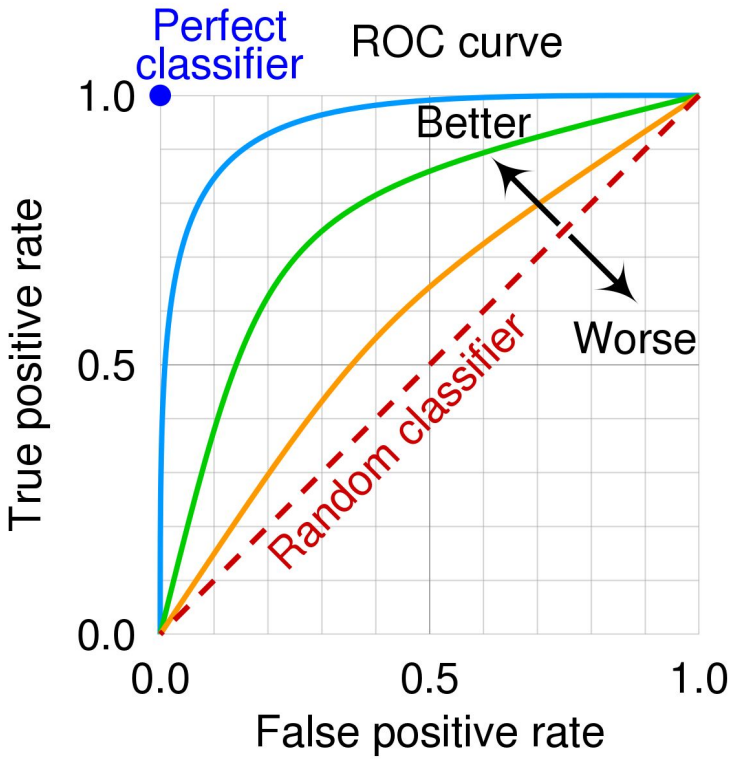
El **AUC (Area Under the Curve)** es el área bajo la curva ROC. Es un valor numérico que oscila entre **0** y **1**, y representa la capacidad del modelo para separar las clases positivas y negativas.

- **AUC cercano a 1**: Indica que el modelo tiene una muy buena capacidad para discriminar entre las clases. En este caso, la curva ROC estará más cerca del rincón superior izquierdo del gráfico.

- **AUC cercano a 0.5**: Indica que el modelo tiene un desempeño similar al de hacer predicciones al azar. En este caso, la curva ROC será cercana a la diagonal.

- **AUC menor a 0.5**: Significa que el modelo tiene un peor desempeño que adivinar al azar, lo que indica que podría estar invertido (clasifica sistemáticamente mal).

AUC-ROC


Perfect classifier
ROC curve
1.0
Better
0.5
Worse
Random classifier
True positive rate
0.0
0.0
0.5
1.0
False positive rate

AUC-ROC

Relación con otras métricas

- **Precisión y Sensibilidad**: A diferencia de la precisión y la sensibilidad, que se calculan para un umbral específico, el AUC-ROC tiene en cuenta todos los posibles umbrales, proporcionando una visión más general del rendimiento del modelo.

- **Exactitud**: La exactitud mide el porcentaje de predicciones correctas en general, pero puede ser engañosa en conjuntos de datos desbalanceados. El AUC-ROC es mucho más robusto en estos casos, ya que evalúa la capacidad del modelo para diferenciar entre clases sin depender de un umbral específico.

AUC-ROC

Aplicaciones prácticas

1. **Medicina**: En modelos de diagnóstico, el AUC-ROC es valioso para evaluar la capacidad de un modelo para identificar correctamente a los pacientes enfermos frente a los sanos a lo largo de diferentes umbrales. En estos casos, una curva ROC puede ayudar a decidir qué umbral es más apropiado según las necesidades del sistema (por ejemplo, minimizar falsos negativos en un diagnóstico grave).
2. **Finanzas y detección de fraudes**: En la detección de fraudes, es importante tener una visión global del rendimiento del modelo en diferentes niveles de umbral, para garantizar que tanto los fraudes como las transacciones legítimas estén siendo adecuadamente diferenciados.
3. **Tareas de clasificación**: En problemas como la clasificación de textos, imágenes o reconocimiento de patrones, el AUC-ROC permite evaluar la calidad general del modelo sin centrarse en un único umbral.

AUC-ROC

Ventajas del AUC-ROC

- **Independiente del umbral**: A diferencia de métricas como la precisión, la sensibilidad o la exactitud, el AUC-ROC resume el rendimiento del modelo en todos los posibles umbrales.

- **Adecuado para datos desbalanceados**: El AUC-ROC es útil en conjuntos de datos donde una clase es mucho más prevalente que la otra, ya que no se ve tan afectado por este desbalance como la exactitud.

AUC-ROC

Consideraciones

- **AUC-ROC no siempre es la mejor métrica**: En problemas con datos extremadamente desbalanceados, incluso un buen AUC puede no ser representativo del rendimiento real del modelo, ya que la FPR (tasa de falsos positivos) puede parecer pequeña debido al bajo número de negativos. En estos casos, otras métricas como la precisión, la sensibilidad, o incluso el **AUC-PR (Precision-Recall AUC)** pueden ser más informativas.

- **Elección del umbral**: Aunque el AUC-ROC proporciona una visión global del rendimiento, la elección de un umbral específico sigue siendo importante según los objetivos del problema. El AUC-ROC no te dice cuál es el mejor umbral, sino que muestra el rendimiento general del modelo en todos ellos.

Métricas para modelos de regresión

MSE

El **Error Cuadrático Medio** es una métrica común que mide el promedio de los cuadrados de los errores o desviaciones entre los valores reales y los valores predichos por el modelo. Es una métrica sensible a grandes errores, ya que eleva al cuadrado las diferencias.

Fórmula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Donde:

- y_i son los valores reales.
- $\hat{y}_i$ son los valores predichos.
- n es el número total de observaciones.

MSE

Interpretación:

- **MSE bajo** indica que el modelo predice bien, ya que los errores son pequeños.
- **MSE alto** significa que hay una alta variabilidad entre las predicciones y los valores reales, lo que indica que el modelo no está ajustando bien los datos.

MSE

Características:

- **Escalado**: El MSE está en las mismas unidades que el cuadrado de la variable dependiente. Si la variable es "precio de casas" en dólares, el MSE estará en "dólares al cuadrado", lo que puede ser difícil de interpretar.

- **Sensibilidad a valores atípicos**: Como eleva los errores al cuadrado, el MSE es altamente sensible a outliers (valores atípicos), amplificando su impacto en la métrica.

RMSE

El **RMSE** es simplemente la raíz cuadrada del MSE. Corrige el problema de la escala de MSE, devolviendo los errores a las mismas unidades que la variable objetivo.

Fórmula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Interpretación:

- **RMSE bajo** indica que las predicciones están cercanas a los valores reales.
- **RMSE alto** sugiere que las predicciones están lejos de los valores reales.

RMSE

Interpretación:

- **RMSE bajo** indica que las predicciones están cercanas a los valores reales.

- **RMSE alto** sugiere que las predicciones están lejos de los valores reales.

RMSE

Características:

- **Escalado**: El RMSE está en las mismas unidades que la variable dependiente, lo que facilita la interpretación directa.

- **Sensibilidad a valores atípicos**: Al igual que el MSE, el RMSE es sensible a los outliers debido al término cuadrático.

MAE

El **Error Absoluto Medio** mide la media de las diferencias absolutas entre los valores reales y los predichos. A diferencia del MSE, no eleva los errores al cuadrado, por lo que es más robusto ante outliers.

Fórmula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MAE

Interpretación:

- **MAE bajo** indica que las predicciones están muy cercanas a los valores reales.

- **MAE alto** sugiere que las predicciones tienen grandes desviaciones con respecto a los valores reales.

MAE

Características:

- **Escalado**: El MAE está en las mismas unidades que la variable dependiente.

- **Robustez a outliers**: A diferencia de MSE y RMSE, el MAE no amplifica grandes errores, lo que lo hace menos sensible a valores atípicos.

Coeficiente de determinación R^2

- El **coeficiente de determinación (R^2)** mide la proporción de la varianza de la variable dependiente que es explicada por las variables independientes en el modelo. Evalúa la bondad del ajuste del modelo, comparando el rendimiento del modelo con el de una línea horizontal que representa la media de los valores reales.

Fórmula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Donde:

- $\bar{y}$ es la media de los valores reales.

Coeficiente de determinación R^2

Interpretación:

- **$R^2 = 1$**: El modelo predice perfectamente los valores.

- **$R^2 = 0$**: El modelo no predice mejor que la media de los valores reales.

- **R^2 negativo**: Significa que el modelo es peor que simplemente predecir la media de los valores reales.

Coeficiente de determinación R^2

Características:

- **Escalado**: R^2 es adimensional, con un rango entre $-\infty$ y 1.

- **No sensible a la escala de las variables**: Como es un cociente, no depende de las unidades de las variables.

Coeficiente de determinación R^2

Características:

- **Escalado**: R^2 es adimensional, con un rango entre $-\infty$ y 1.

- **No sensible a la escala de las variables**: Como es un cociente, no depende de las unidades de las variables.