8. Métricas

Ejercicio 8.1. Matriz de Confusión

(a) En un problema de clasificación binaria, ¿a qué se denomina clase positiva y a qué clase negativa? Si nuestro problema consiste en clasificar spam vs. no-spam, ¿cuál es la clase positiva? Si nuestro problema es clasificar imágenes de perros vs. gatos, ¿cuál es la clase positiva?

(b) Explicar con tus palabras la definición de *verdadero positivo*, *verdadero negativo*, *falso positivo* y *falso negativo*.

(c) Completar la Primera Parte del notebook `notebook_metricas.ipynb`. El Test 1 debería pasar.

(d) ¿Por qué podría un falso positivo ser considerado más (o menos) importante que un falso negativo? Dar un ejemplo en donde es más grave tener falsos negativos que falsos positivos.

Ejercicio 8.2. Métricas de umbral fijo

(a) Explicar con tus palabras la definición de *accuracy*, *precision* y *recall*.

(b) Completar la Segunda Parte del notebook `notebook_metricas.ipynb`. El Test 2 debería pasar.

(c) ¿Por qué es un problema medir *accuracy* de un clasificador para compararlo con otro? Dar un ejemplo en donde sería engañoso utilizar esta comparación.

(d) Demuestre que F_β puede ser reescrito como $F_\beta = \dfrac{(1+\beta^2)\cdot \mathrm{TP}}{(1+\beta^2)\cdot \mathrm{TP} + \beta^2 \cdot \mathrm{FN} + \mathrm{FP}}$

(e) Demuestre que la métrica recall vista como función del umbral de decisión, es una función monótona decreciente ($recall_{M,D}(\mu_1) \leq recall_{M,D}(\mu_2)$ si $\mu_1 > \mu_2$)

(f) Muestre cómo la métrica precision vista como función del umbral de decisión, **no necesariamente** es una función monótona creciente ($precision_{M,D}(\mu_1) \not\leq precision_{M,D}(\mu_2)$ si $\mu_1 < \mu_2$)

Ejercicio 8.3. Métricas de regresión. Consideremos un problema de regresión en el cual entrenamos un modelo f sobre un conjunto de datos $\mathcal{D}$, obteniendo prediciones para cada instancia $f(\mathbf{x}^{(i)}) = \hat{y}^{(i)}$, responder

(a) Supongamos que queremos aplicar una transformación lineal de escala de las etiquetas y el modelo, es $y^{escalado} = a \cdot y + b$, ¿cómo afecta esta transformación al MSE y el MAE del modelo en $\mathcal{D}$?

(b) En este mismo escenario, ¿cómo se vería afectado el R^2 de nuestro modelo en $\mathcal{D}$?

(c) Ahora supongamos que cada etiqueta $\mathbf{y}^{(i)}$ tiene asociada una constante de normalización $m^{(i)}$ que le queremos aplicar que es siempre conocida, y queremos ver que tan bien predice nuestro modelo a este nuevo target sin reentrenarlo, es decir, vamos a tomar la predicción que ya teniamos y aplicar esa misma transformación:

$$y^{(i)}_{(norm)} = \frac{y^{(i)}}{m^{(i)}}, \hat{y}^{(i)}_{(norm)} = \frac{\hat{y}^{(i)}}{m^{(i)}}$$

¿el R^2 será el mismo que en el modelo anterior? Escribirlo en función del R^2 original si es posible.

Ejercicio 8.4. Considerar la Figura Umbral de clasificación. En esta figura se ven instancias ordenadas según la probabilidad detectada por un clasificador (entre 0 y 1). Además, se encuentran marcados cuatro umbrales de decisión.

(a) Calcular las tablas de confusión resultantes para cada uno de los cuatro umbrales de decisión. Recordar que si la probabilidad está por debajo del umbral, la instancia será clasificada como perteneciente a la clase negativa; si está por encima, como clase positiva.

(b) ¿Cuál es el mejor umbral?

(c) Con dichos umbrales, graficar la curva ROC para dicha clasificación.

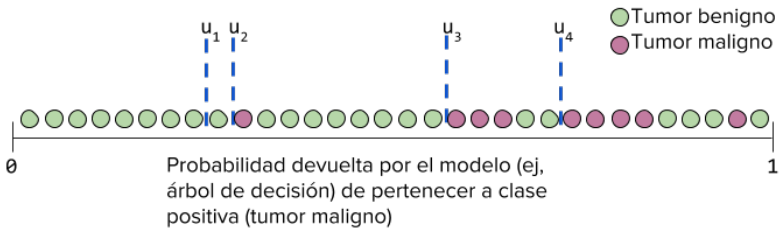

Figura 5: Umbral de clasificación

Ejercicio 8.5.
Escribir el pseudocódigo de la siguiente función:
CURVA-ROC(LABELS : LIST⟨BOOL⟩, PROBAS: LIST⟨FLOAT⟩) : LIST⟨TUPLE⟨UMBRAL, VALORREC, VALORTPR⟩⟩
que devuelve los valores de Recall y TPR junto a cada umbral explorado.

Ejercicio 8.6. Verdadero o Falso (justificar)

(a) AUC-ROC es invariante de umbral de clasificación. Mide la calidad de las predicciones del modelo independientemente del umbral de clasificación elegido.

(b) El AUC-ROC es invariante de escala (invariante a transformaciones lineales en las predicciones). Mide qué tan bien se clasifican las predicciones, en lugar de sus valores absolutos.

(c) En caso que el ordenamiento de las instancias sea el mismo y suponiendo que no hay scores repetidos, AUC-ROC va poder distinguir un clasificador muy confiado (instancias clasificadas como positiva tienen scores muy altos, instancias clasificadas como negativas scores muy bajos) de uno poco confiado (scores menos extremos).

Ejercicio 8.7. En los casos en los que existen grandes disparidades en el costo de los falsos negativos frente a los falsos positivos, puede ser fundamental minimizar un tipo de error de clasificación. Por ejemplo, al detectar spam, es probable que desee priorizar la minimización de los falsos positivos (incluso si eso resulta en un aumento significativo de los falsos negativos). ¿Cuál de las siguientes métricas sería la más adecuada y por qué?

- Precision
- F_b con algún valor de b que favorezca precision.
- AUC-ROC

Ejercicio 8.8. (*) Dada la siguiente matriz de costos:

Costos =

Clases\Decisiones	$Ignorar$	$ReChequear$	$Operar$
$Maligno$	100	5	0
$Benigno$	0	15	50

Suponiendo que $\hat{P}$ es un modelo entrenado y calibrado que produce las siguientes probabilidades:

(I) $\hat{P}(Y = Maligno|X = \mathbf{x}^{(1)}) = 0.3$

(II) $\hat{P}(Y = Maligno|X = \mathbf{x}^{(2)}) = 0.99$

(III) $\hat{P}(Y = Maligno|X = \mathbf{x}^{(3)}) = 0.01$

Indicar la decisión óptima para cada instancia.

Ejercicio 8.9. Sea A un clasificador que tiene un F_1 de 0.80 (con un umbral de clasificación de 0.5), y sea B un clasificador que tiene un F_1 0.70 (también con un umbral de clasificación de 0.5). Sin embargo al cambiar el umbral a 0.4 obtenemos F_1 de 0.76 y 0.80 respectivamente.

(a) Explicar por qué puede suceder este fenómeno dando un ejemplo aproximado.

(b) ¿Podemos concluir algo sobre el AUC Prec-Recall de estos dos modelos?

Ejercicio 8.10. Verdadero o Falso (justificar)

(a) Tanto *recall* como *precision* no toman en cuenta qué tan bien el modelo maneja los casos negativos.

(b) Un modelo que no produce falsos positivos tiene *precision* = 1.0.

(c) Un modelo que no produce falsos negativos tiene *recall* = 1.0.

(d) Si un clasificador devuelve probabilidades, la matriz de confusión se construye de manera ponderada según la probabilidad de cada clase.

(e) Si un clasificador devuelve probabilidades, hay muchas matrices de confusión asociadas dependiendo del umbral de clasificación.

(f) Si un clasificador devuelve probabilidades, hay infinitas matrices de confusión asociadas dependiendo del umbral de clasificación.

(g) Aumentar el umbral de clasificación produce que la *precision* siempre suba.

(h) Aumentar el umbral de clasificación produce que el *recall* baje o se mantenga igual.

(i) La métrica *precision* es parte fundamental del cálculo de la *curva ROC*.

Ejercicio 8.11. ¿Binaria o 2 clases?

Sean A y B clasificadores que distinguen entre imágenes de perros e imágenes de gatos. Al medir la performance del clasificador (utilizando F_1 para evitar los problemas de utilizar *accuracy*) y "gato" como clase positiva, obtenemos $F_1(A) = 0.9$, $F_1(B) = 0.8$. ¿Podemos concluir que el clasificador A es mejor que el clasificador B para este problema?
Resolver los siguientes ítems para poder responder a la pregunta:

(a) Al calcular F_1 utilizando "gato" como clase positiva, ¿importa qué ocurre con los perros que fueron clasificados correctamente? Revisar la Tercera Parte del notebook `notebook_metricas.ipynb` y decidir cuál clasificador funciona mejor, basándose en las métricas obtenidas. Observar el cambio que ocurre al intercambiar cuál es la clase positiva.

(b) ¿Para qué sirve el parámetro average en la función `f1_score` de la librería sklearn?

(c) ¿Qué sucede si la cantidad de instancias sobre las que fueron testeados es distinta? ¿Cómo se ve afectada la métrica F_1 al cambiar los True Negatives? Correr la Cuarta Parte del notebook `notebook_metricas.ipynb`. El gráfico muestra cómo varía la métrica F_1 al aumentar la cantidad de True Negatives (observar que estamos cambiando la cantidad de instancias sobre las que testeamos). ¿Qué se puede concluir de este experimento?

Ejercicio 8.12. Recordemos la métrica de `accuracy` para un problema de clasificación, para simplificar solo consideraremos el caso binario, sea y el vector de etiquetas verdadera, $\hat{y}$ las respectivas predicciones, y N la cantidad total de muestras ($y, \hat{y} \in \{0,1\}^N$):

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} 1(\hat{y}^{(i)} = y^{(i)})$$

Donde $1(x)$ es la función indicadora (vale 1 si el argumento es verdadero, 0 caso contrario). Vamos a introducir una pequeña modificación en como contamos los aciertos para cada clase, supongamos que p es la proporción de la clase minoritaria (positiva):

$$\texttt{balanced\_accuracy}(y, \hat{y}) = \frac{1}{N} \left(\frac{1}{2p} \sum_{i=1}^{N} 1(\hat{y}^{(i)} = y^{(i)} = 1) + \frac{1}{2(1-p)} \sum_{i=1}^{N} 1(\hat{y}^{(i)} = y^{(i)} = 0) \right)$$

(a) ¿Qué rango de valores posibles tiene esta métrica? ¿Cuánto vale esta métrica si tenemos un clasificador constante que predice siempre la clase mayoritaria $\hat{y} = 0$? ¿Y si predice siempre la clase minoritaria?

(b) Escribir `balanced_accuracy` en términos de TP, FP, TN y FN.

(c) Mostrar que en el caso binario, esta métrica es equivalente al promedio aritmético entre *sensitividad* (true positive rate) y la *especificidad* (true negative rate).