Pontificia Universidad JAVERIANA Cali

Aprendizaje Automático y Análisis de Datos

Diseño de experimentos y métricas de desempeño

Julián Gil González

julian.gil@javerianacali.edu.co (Periodo 2023-I)

10 de febrero de 2023

Agenda

Temas:

- Introducción al aprendizaje supervisado.
- Evaluación del aprendizaje.

Objetivos del aprendizaje: Al final de esta clase los estudiantes estarán en la capacidad de:

- Proponer una tarea de aprendizaje supervisado.
- Evaluar la calidad del aprendizaje de una tarea de aprendizaje supervisado.

Table of Contents

Definición de los problemas de clasificación y regresión

- Se tiene un conjunto de datos de **atributos** de entrada, representados a partir de la matriz

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix},$$

  donde cada fila $\mathbf{x}_i$ es una instancia.

- También se tiene un conjunto de **salidas** (o etiquetas) representados a partir del vector

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}.$$

Definición de los problemas de clasificación y regresión

- Si las etiquetas $y_i \in \mathbb{Z}$ pertenecen a los números **enteros**, se dice que el problema es de clasificación.
- Si las etiquetas $y_i \in \mathbb{R}$ pertenecen a los números **reales**, se tiene un problema de regresión.
- Consideremos que hemos entrenado un modelo a partir del cual realizamos una predicción $\hat{y}_i$ correspondiente a la instancia $\mathbf{x}_i$ y cuya etiqueta es y_i.
- Necesitamos **evaluar** el modelo a partir de la comparación entre la etiqueta y_i y la predicción del modelo $\hat{y}_i$.

Función de costo vs métrica de evaluación

- La función usada para la comparación entre las etiquetas y las predicciones depende, entre otras cosas, de la naturaleza de las salidas (regresión o clasificación).
- Es importante diferenciar dos conceptos claves: Función objetivo (Función de costo) y Métrica de Evaluación.
- La función de costo u objetivo, se usa para medir el rendimiento durante la etapa de entrenamiento del modelo. Debe cumplir con algunas propiedades: Ser continua y diferenciable (Gradiente descendente).
- Por el contrario, la métrica de evaluación se usa para analizar el rendimiento del modelo una vez el entrenamiento ha sido entrenado.

Pontificia Universidad JAVERIANA Cali

Función de costo vs métrica de evaluación

- En algunas ocasiones una función puede ser usada como función de costo y como métrica de evaluación. Esto ocurre principalmente en tareas de regresión.
- Por ahora nos enfocaremos en analizar las métricas de evaluación. Las funciones de costo las estudiaremos en la medida que vamos analizando los diferentes modelos de aprendizaje estadístico y supervisado.
- Estudiaremos en las métricas más usadas en tareas de regresión, clasificación binaria y clasificación de múltiples clases.

Table of Contents

Error cuadrático medio (MSE)

- Matemáticamente, el error cuadrático medio se define como:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2.$$ ¿Por qué se eleva al cuadrado?

- El MSE es un número real cuyo valor mínimo es 0 (el mejor de los casos) e ∞ (peor escenario).

Pontificia Universidad JAVERIANA Cali

Valor R^2

- Es una extensión del error cuadrático medio. Se define como:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{y}_i - y_i)^2}{(y_i - \bar{y})^2}.$$

- El R^2 corresponde al cuadrado del coeficiente de correlación de Pearson.
- Es un número real entre 0 y 1. ¿Cuál es el mejor y peor escenario?

Error absoluto medio (MAE)

- Similar al MSE, se cambia el operador cuadrático por el operador de valor absoluto:

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

- Es un número real entre 0 e ∞.
- Comparado con el MSE y el R^2, el MAE tiene menor sensibilidad ante los datos atípicos.

Pontificia Universidad JAVERIANA Cali

Table of Contents

Exactitud (Accuracy)

- Se define como:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i - y_i),$$

donde

$$\mathbb{I}(a) = \begin{cases} 1 & \text{Si } a = 0 \\ 0 & \text{En otro caso} \end{cases}.$$

- La exactitud puede tomar valores entre 0 y 1.

Exactitud (Accuracy)

- El accuracy presenta dificultades en casos donde la base de datos está fuertemente desbalanceada.
- Como ejemplo, suponer que tenemos un modelo para predecir los accidentes de aéreos. El modelo alcanza un accuracy del $99{,}9\,\%$.
- El modelo usado para obtener estos resultados es el siguiente.

```
def model(x_i):
    return False
```

- ¿El modelo obtenido es útil?

Matriz de confusión

- Es claro que el accuracy no representa apropiadamente el comportamiento del modelo en casos donde los datos están desbalanceados.
- Es necesario recurrir a otro tipo de artefactos que nos permitan cuantificar, de forma apropiada, el comportamiento de los modelos.
- Una alternativa es usar la matriz de confusión.

		Predicción		
		Positivo	Negativo	Total
Clase verdadera	Positivo	TP	FN	$TP + FN$
	Negativo	FP	TN	$FP + TN$
	Total	$TP + FP$	$FN + TN$	N

Matriz de confusión

- Ejemplo: calcule la matriz de confusión de los siguientes datos.

Muestra #	Etiqueta	Predicción
1	1	1
2	1	0
3	0	1
4	1	0
5	1	1
6	1	0
7	0	1
8	1	0
9	1	1
10	1	1
11	0	0
12	1	0
13	1	1
14	1	0
15	1	1
16	1	0
17	0	1
18	1	0
19	1	1
20	1	1

Pontificia Universidad JAVERIANA Cali

Precision, recall y el F_1-score

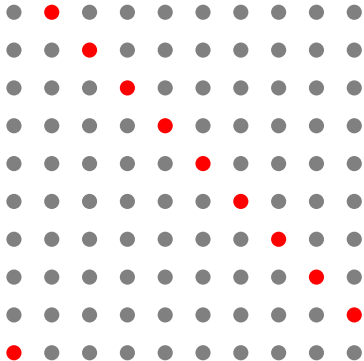
- La matriz de confusión representa la mejor descripción del comportamiento de un algoritmo.
- Sin embargo, no es una métrica de evaluación, es una matriz. En diversos escenarios es preferible representar el rendimiento del clasificador con un único valor.
- Las alternativas son, el Precision, recall y el F_1-score.

Precision, recall y el F_1-score

- Recordemos que una de las mayores problemáticas con el accuracy es su incapacidad de representar el rendimiento del modelo en presencia de datos desbalanceados.
- Los datos desbalanceados son comunes en la práctica. Casos de enfermedades raras, detección de transacciones bancarias, entre otras.

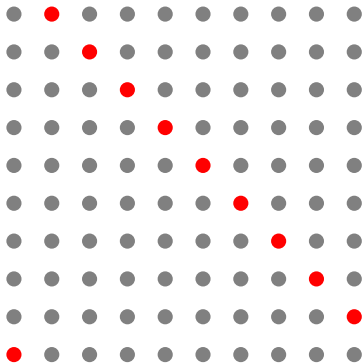
Recall

- Consideremos el ejemplo de un sistema de detección de transacciones fraudulentas. La clase positiva es la existencia de fraude y a su vez la clase minoritaria. ¿por qué?



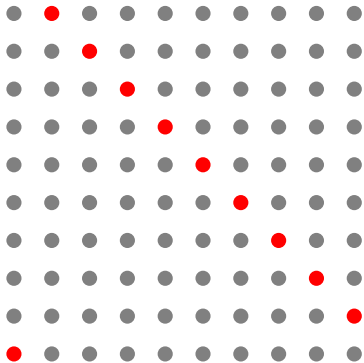
Recall

- Vamos a suponer que en nuestra base de datos hay 10 transacciones fraudulentas. Además, consideremos que el modelo es capaz de identificar 4 de esas transacciones fraudulentas.



Recall

- En este caso, el recall equivale a $4/100$, es decir el $40\,\%$. ¿Cuál es entonces la ecuación del Recall?



Recall

- Según lo analizado, ¿debemos maximizar el Recall del modelo?
- ¿Y si todo lo clasificamos como positivo?

```
def model(x_i):
    return True
```

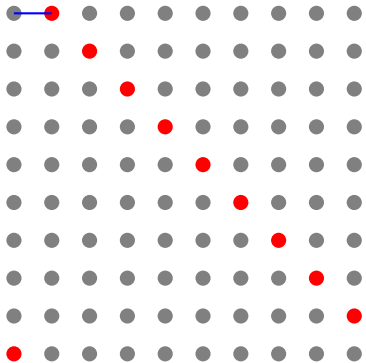
- Si todo lo clasificamos como positivo, ¿Cuál es el valor del Recall?

Precision

- Es evidente que en casos de desbalance el accuracy no es recomendable; además, maximizar el recall no es suficiente.
- En otras palabras, un recall cercano al $100\,\%$ no necesariamente representa un buen rendimiento.
- Debemos recurrir a otra métrica que nos ayude a cuantificar el rendimiento del modelo de forma correcta. Esta métrica se denomina **Precision**.

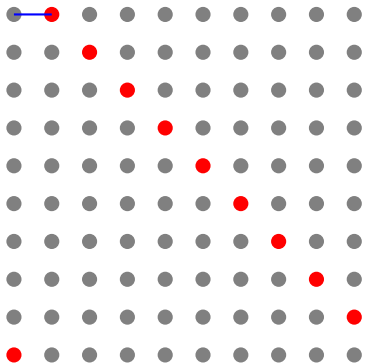
Precision

- Volvamos a nuestro ejemplo, Suponer que el modelo realizó dos predicciones, ambas positivas. Una de ellas es positiva, sin embargo la otra es negativa.



Precision

- El Recall del modelo sería del $10\,\%$, mientras que el Precision del $50\,\%$. ¿Cómo determino el Precision?



Recall vs Precision

- Si predecimos todas las muestras como negativas, nuestro Recall será cero. Lo cual evidencia un mal modelo.
- Por el contrario, si predecimos todas las muestras como fraudulentas, nuestro Recall será del $100\,\%$, mientras que el Precision puede ser bajo. ¿Por qué?
- El caso ideal es en el que tengamos valores altos para ambas métricas. Sin embargo, hay un **trade-off** entre ellas. Si se maximiza el Recall, el Precision se ve afectado.
- Se hace necesario encontrar un balance. Además, una única métrica que combine Recall y Precision.
- Calcular el Recall y Precision de los datos mostrados en la diapositva 16.

F_1-score

- El F_1-score permite combinar la información del Recall y Precision en una única métrica. Matemáticamente se describe como:

$$F_1 = 2\frac{(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}}$$

- Un valor alto para la métrica F_1-score indica que se ha mejorado el Recall, precision o ambos.

- El F_1-score da igual peso al Recall y Precision. Sin embargo, puede darse el caso donde requerimos darle más peso a alguna de las dos. En ese caso usamos la genralización del F_1-score, denominado F_β-score:

$$F_\beta = (\beta^2 + 1)\frac{(\text{Precision})(\text{Recall})}{\beta^2\,\text{Precision} + \text{Recall}}$$

Curvas ROC y su área bajo la curva (AUC)

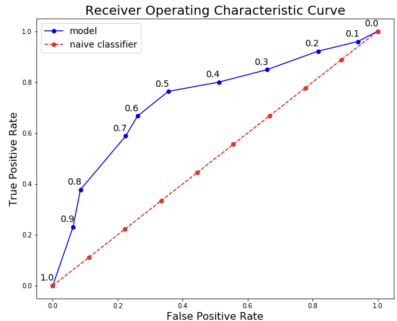
- Diferentes modelos entregan la probabilidad de que una muestra pertenezca a la clase positiva más que una predicción estática.
- En este sentido la predicción $\hat{y}_i$ es un número entre 0 y 1.
- Así, es necesario definir un umbral δ para definir si la muestra pertenece a la clase 0 o a la clase positiva.
- Por ejemplo, si la predicción fue $\hat{y}_i = 0{,}4$ y $\delta = 0{,}7$. Luego, la muestra se asigna a la clase 0.
- En otro caso si modificamos $\delta = 0{,}3$, luego la muestra corresponde a la clase positiva.
- Típicamente, se usa $\delta = 0{,}5$.

Curvas ROC y su área bajo la curva (AUC)

- Las curvas ROC, son gráficas que permiten comparar el rendimiento de un clasificador utilizando diferentes valores del umbral δ.
- Las curvas ROC son gráficas donde el eje de las abscisas representa la proporción de Falsos positivos (FPR), mientras que en el eje de las ordenadas se encuentra el la proporción de verdaderos positivos (TPR).

$$\mathrm{FPR} = \frac{FP}{FP + TN} \qquad \mathrm{TPR} = \frac{TP}{TP + FN}$$

Curvas ROC y su área bajo la curva (AUC)[1]



- En lugar de una gráfica, se requiere una métrica de desempeño. Para este caso se calcula el área bajo la curva ROC, denominado AUC.

[1]https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

Curvas ROC y su área bajo la curva (AUC)

- El AUC de un clasificador aleatorio tenderá a $0{,}5$.
- En condiciones de poco desbalance en las clases, un AUC cercano a uno evidenciará un clasificador con buen rendimiento.
- Otra aplicación interesante de las curvas ROC, es que permiten encontrar un umbral óptimo.
- ¿Cómo determinar el umbral óptimo?

Actividad

- Calcular la curva ROC a partir de los siguientes datos. ¿Cuál es el umbral recomendado?

Muestra #	Etiqueta	Predicción
1	1	0.4
2	0	0.9
3	1	0.85
4	1	0.3
5	1	0.87
6	0	0,6
7	0	0.2
8	0	0.96
9	1	0.78
10	0	0.5
11	0	0.55
12	1	0.62
13	1	0.84
14	0	0.91
15	1	0.25
16	0	0.99
17	1	0.72
18	1	0.38
19	1	0.66
20	0	0.75

Table of Contents

Definir la validación

- El aprendizaje de máquina es un proceso altamente empírico, en donde no hay una única respuesta para un problema particular.
- Son muchas las consideraciones que se deben tener en cuenta al momento de diseñar un sistema de aprendizaje automático.
  - ¿Cómo procesar los datos?
  - ¿Qué modelo debo elegir?
  - ¿Como calcular los hiper-parámetros del modelo?
- En este sentido, es fundamental que tengamos un adecuado diseño de experimentos.

Conjunto de entrenamiento y prueba

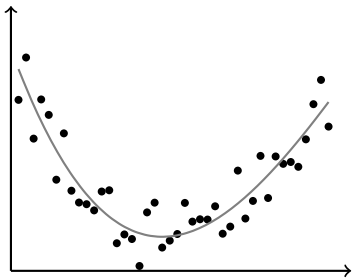
- Una de las claves al momento de realizar experimentos, es elegir de forma adecuada los conjuntos de entrenamiento y prueba.
- Esta división es necesaria con el fin de evaluar el modelo en un conjunto distinto con el que fue entrenado. ¿Por qué?
- De esta forma las bases de datos deben dividirse en al menos dos conjuntos: Entrenamiento y prueba.
- El conjunto de entrenamiento se usará exclusivamente para entrenar el modelo de aprendizaje.
- El conjunto de prueba nunca ha sido visto por el modelo y se usa con fines de validación.

Conjunto de entrenamiento y prueba

- La métricas de evaluación se calculan el conjunto de entrenamiento y prueba con el fin de analizar el comportamiento del modelo. Sin embargo, el valor más significativo es el calculado sobre el conjunto de prueba.
- Los valores de las métricas de evaluación en el conjunto de entrenamiento y prueba son útiles para diagnosticar diferentes aspectos del modelo.

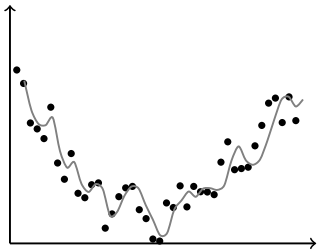
Ajuste adecuado

- Si el rendimiento es alto en el conjunto de entrenamiento y prueba, se concluye que el modelo funciona de forma adecuada.



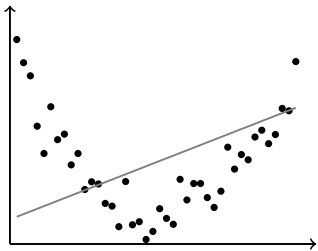
Sobre-ajuste

- Si el rendimiento en el conjunto de entrenamiento es alto y decae en el conjunto de prueba, se concluye que el modelo está sobre ajustado.
- Se asocia con modelos que tienen muchos parámetros y pocos datos de entrenamiento.
- Solución: Regularización, Adquirir más datos, reducir la complejidad del modelo (cantidad de parámetros).



Sub-ajuste

- Si el rendimiento en tanto el conjunto de entrenamiento y prueba no son los esperados, se dice que el sistema está en sub-ajuste.
- ¿Cómo saber que un rendimiento es el adecuado? Error de Bayes$\rightarrow$Rendimiento del humano$\rightarrow$Conocimiento de un experto.
- Solución: Mejorar características, aumentar la complejidad del modelo (no lineal).

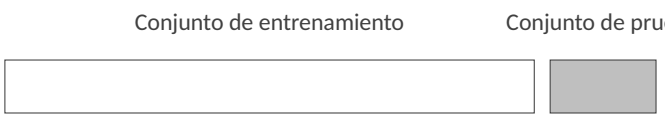


Esquemas de validación

- Se nota que el conjunto de entrenamiento y prueba no son tareas triviales y deben elegirse con cuidado.
- Por ejemplo si diseñamos un conjunto de entrenamiento con muy pocas muestras, el modelo puede ser propenso al sobre ajuste.
- En general, el tamaño del conjunto de entrenamiento debe ser mayor a la cantidad de muestras del conjunto de prueba.
- Existen distintas estrategias de validación. Hold-out validation, K-fold validation.

Hold-out validation

- Consiste en dividir, de forma aleatoria el conjunto en dos partes, entrenamiento y prueba.
- Se repite varias veces el procedimiento (Usualmente 15). En cada iteración se elige, de forma aleatoria, los conjuntos de validación y prueba. Al final se reporta el promedio de las métricas de evaluación.
- Los porcentajes de datos en el entrenamiento y prueba se eligen de forma empírica. Para bases pequeñas, se suele usar $70/30$. Para bases de datos grandes $90/10$, $95/5$.



K-fold

- Divide la base de datos en k grupos. k se define empíricamente (Típicamente un valor entre 5 y 10).
- De los k grupos, se elige uno como conjunto de prueba. Los $k-1$ restantes conforman el grupo de entrenamiento.
- El proceso se repite k veces. En cada iteración se calculan las métricas de evaluación. Al finalizar Se reporta el promedio de las métricas.