

MÉTRICAS Y VALIDACIÓN CRUZADA

ALAN REYES-FIGUEROA

INTRODUCCIÓN A LA CIENCIA DE DATOS

(AULA 35) 24.MAYO.2021

Desarrollamos métricas de desempeño para los modelos de predicción.
Regresión: Típicamente medimos el error total de predicción.

- MSE = **error medio cuadrático** o error norma L^2

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \frac{1}{n} \|\boldsymbol{\varepsilon}\|_2^2.$$

- MAE = **error medio absoluto** o error norma L^1

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| = \frac{1}{n} \|\mathbf{y} - \hat{\mathbf{y}}\|_1 = \frac{1}{n} \|\boldsymbol{\varepsilon}\|_1.$$

- log cosh

$$E = \frac{1}{n} \sum_{i=1}^n \log \cosh(y_i - \hat{y}_i).$$

Métricas para Clasificación

Clasificación: Típicamente medimos el acierto (número de clasificaciones correctas), o el error total (número de clasificaciones erróneas). Aquí puede ocurrir que los errores pesen diferente para cada clase.

- Acierto (*accuracy*)

$$accuracy = \frac{1}{n}(\text{número de datos clasificados correctamente}).$$

- Error de clasificación

$$error = \frac{1}{n}(\text{número de datos clasificados incorrectamente}).$$

- Error ponderado

$$error = \frac{1}{n} \sum_{\text{clase } k} w_k (\text{número de datos mal clasificados en la clase } k).$$

Métricas para Clasificación

- Matriz de confusión: arreglo que cuenta las buenas y malas clasificaciones, compara etiquetas reales (*ground-truth*) contra las etiquetas estimadas por el modelo (*predicted*).

		PREDICTED	
		Positive	Negative
ACTUAL	Positive	6	2
	Negative	1	1

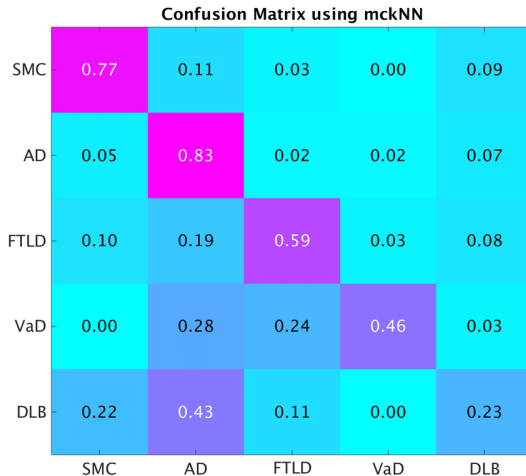
No se utiliza mucho en el caso de clasificación binaria, es más útil en casos donde hay mayor cantidad de clases o etiquetas.

Métricas para Clasificación

Confusion Matrix

Output Class	BRCA	KIRC	LUAD	LUSC	UCEC	
BRCA	342 41.0%	2 0.2%	3 0.4%	4 0.5%	1 0.1%	97.2% 2.8%
KIRC	3 0.4%	211 25.3%	0 0.0%	0 0.0%	0 0.0%	98.6% 1.4%
LUAD	4 0.5%	1 0.1%	54 6.5%	13 1.6%	3 0.4%	72.0% 28.0%
LUSC	2 0.2%	1 0.1%	8 1.0%	79 9.5%	0 0.0%	87.8% 12.2%
UCEC	0 0.0%	0 0.0%	0 0.0%	0 0.0%	104 12.5%	100% 0.0%
	97.4% 2.6%	98.1% 1.9%	83.1% 16.9%	82.3% 17.7%	96.3% 3.7%	94.6% 5.4%
	BRCA	KIRC	LUAD	LUSC	UCEC	

Target Class



Métricas para Clasificación

En el caso de clasificación binaria, se han desarrollado otras métricas útiles. Denotamos

VALORES PREDICCIÓN	Verdaderos positivos (TP)	Falsos Positivos (FP)
	Falsos Negativos (FN)	Verdaderos Negativos (TN)
VALORES REALES		

$$TP = \#\{i : \hat{y}_i = 1, y_i = 1\},$$

$$FN = \#\{i : \hat{y}_i = 0, y_i = 1\},$$

$$P = TP + FN = \#\{i : y_i = 1\},$$

$$FP = \#\{i : \hat{y}_i = 1, y_i = 0\}$$

$$TN = \#\{i : \hat{y}_i = 0, y_i = 0\},$$

$$N = FP + TN = \#\{i : y_i = 0\}.$$

Métricas para Clasificación

- **Sensitivity, Recall, hit rate, true positive rate (TPR)**

$$\text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}} = 1 - \text{FNR}$$

- **Specificity, selectivity, true negative rate (TNR)**

$$\text{TNR} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN} + \text{FP}} = 1 - \text{FPR}$$

- **Precision, positive predictive value (PPV)**

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} = 1 - \text{FDR}$$

- **Negative predictive value (NPV)**

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} = 1 - \text{FOR}$$

Métricas para Clasificación

		Fact (observation)	
		+	-
Prediction	+	True positive	False positive (Type I error)
	-	False negative (Type II error)	True negative

Confusion matrix

✓	✗
✗	✓

Specificity,
Selectivity,
True negative rate (TNR)

✓	✗
✗	✓

Accuracy (ACC)

✓	✗
✗	✓

Sensitivity, Recall,
True positive rate (TPR),
Probability of detection

✓	✗
✗	✓

Precision,
Positive predictive value

Métricas para Clasificación

- **Accuracy** (ACC)

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Balanced accuracy** (BA)

$$BA = \frac{TPR + TNR}{2}$$

- **F₁ score** Es la media armónica de la precisión y la sensibilidad

$$F_1 = 2 \frac{PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

- **Matthews correlation coefficient** (MCC)

$$MCC = \frac{TP \cdot TN + FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Métricas para Clasificación

- *Miss rate or false negative rate (FNR):* $FNR = \frac{FN}{P} = \frac{FN}{FN + TP} = 1 - TPR$
- *Fall-out or false positive rate (FPR):* $FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR$
- *False discovery rate (FDR):* $FDR = \frac{FP}{FP + TP} = 1 - PPV$
- *False omission rate (FOR):* $FOR = \frac{FN}{FN + TN} = 1 - NPV$
- *Prevalence threshold (PT):* $PT = \frac{\sqrt{TPR(1 - TNR)} + TNR - 1}{TPR + TNR - 1}$
- *Threat score (TS) or Critical success index (CSI):* $TS = \frac{TP}{TP + FN + FP}$
- *Informedness or bookmaker informedness (BM):* $BM = TPR + TNR - 1$
- *Markedness (MK) or delta P (Δp):* $MK = PPV + NPV - 1.$

Métricas para Clasificación

También son importantes métricas que se puedan derivar. Por ejemplo

- **Binary Cross-entropy:**

$$\text{Cross-entropy} = - \sum_{i=1}^n (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)).$$

- **Multiclass Cross-entropy or Categorical Cross-entropy:**

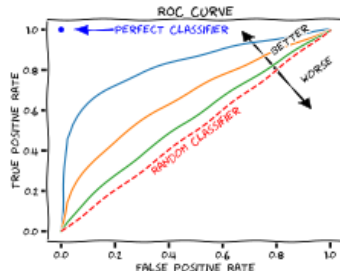
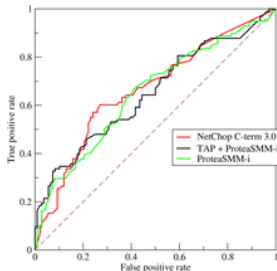
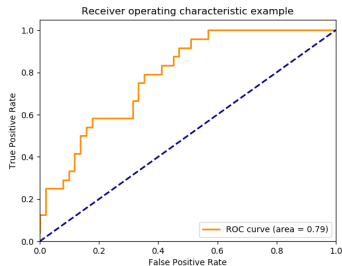
$$\text{Cross-entropy} = - \sum_{i=1}^n \sum_{k=1}^K \mathbb{P}(y_i = k) \log \mathbb{P}(\hat{y}_i = k).$$

Existen muchas otras métricas de tipo distribucional, como información mutua, divergencia de Kullback-Leibler, entre otros.

Otras métricas muy usadas, en el caso de clasificadores binarios con respuesta continua ($p \in (0, 1)$), son las siguientes:

Métricas para Clasificación

- **Curva ROC** (*Receiver Operating Characteristic curve*), es un diagrama que ilustra la capacidad de diagnóstico de un sistema clasificador binario a medida que se varía su umbral de decisión.
- **ROC AUC** área bajo la curva ROC.



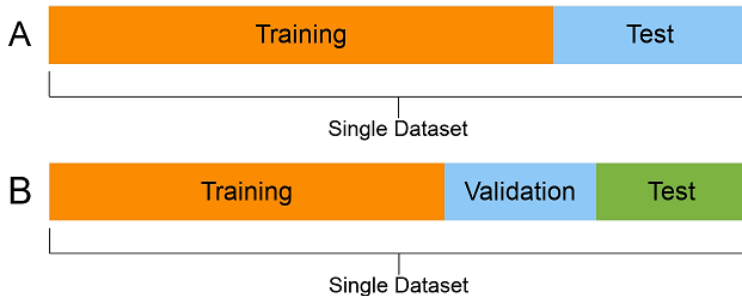
Métricas para Clasificación

Model	RMSE	CV (RMSE)	MAE
Linear Regression	847.62	1.55	630.76
Ridge	877.35	1.60	655.70
k-Nearest Neighbor	1655.70	3.02	1239.35
Random Forest	539.08	0.98	370.09
Gradient Boosting	1021.55	1.86	746.24
Neural network	2741.91	5.01	2180.89
Extra Trees	466.88	0.85	322.04

Machine learning algorithm	Descriptor type	Cross-validation accuracy (%)	Accuracy (%)	AUC	Balanced accuracy (%)	Training error	Generalization error	Training error	Generalization error
Artificial neural network	2D	86.20	85.71	0.50	50.00	MSE 0.21	RMSE 0.35	MSE 0.20	RMSE 0.35
	3D	82.75	85.71	0.91	70.50	0.16	0.37	0.16	0.37
	MD	82.75	85.71	0.66	50.00	0.27	0.39	0.22	0.34
	2D+3D	89.65	85.71	0.91	70.50	0.10	0.28	0.16	0.38
	2D+MD	75.86	85.71	0.58	50.00	0.29	0.46	0.20	0.35
	3D+MD	89.65	78.57	0.87	66.50	0.10	0.25	0.20	0.42
	2D+3D+MD	89.65	78.57	0.87	66.50	0.13	0.32	0.20	0.42
Random forest	2D	86.20	85.71	0.50	50.00	0.21	0.35	0.21	0.35
	3D	89.65	85.71	0.50	62.00	0.15	0.26	0.18	0.34
	MD	82.75	85.71	0.68	50.00	0.26	0.38	0.19	0.34
	2D+3D	86.20	85.71	0.52	50.00	0.15	0.27	0.18	0.33
	2D+MD	82.75	85.71	0.62	50.00	0.26	0.40	0.19	0.35
	3D+MD	86.20	85.71	0.89	50.00	0.16	0.26	0.17	0.31
	2D+3D+MD	86.20	85.71	0.87	50.00	0.17	0.28	0.16	0.30

Validación Cruzada

Validación cruzada se refiere a diversas técnicas de validación de modelos, para evaluar cómo los resultados de un análisis estadístico se generalizarán a un conjunto de datos independiente. Se utiliza principalmente en entornos de predicción, y se desea estimar la precisión con la que un modelo predictivo funcionará en la práctica.



Validación Cruzada

Hold-Out Cross-Validation: Literalmente, esta consiste en fijar un conjunto de entrenamiento y uno de prueba (como en la figura anterior). Se entrena o diseña el modelo usando el conjunto de entrenamiento, y se evalúa una sola vez con el de prueba.

Repeated Hold-Out Cross-Validation: Se subdivide el conjunto de datos en tres componentes: entrenamiento, validación y prueba. Esto permite que, durante el proceso de entrenamiento, evaluar parcialmente el modelo con los datos de validación, y volver a hacer ajustes necesarios siempre que se considere conveniente. Esto puede repetirse varias veces hasta obtener un modelo con un desempeño adecuado (en validación).

Una vez definido el modelo, se entrena una última vez con (entrenamiento + validación), y se evalúa con los datos de prueba.

Validación Cruzada

k-fold Cross-Validation: Es un procedimiento de remuestreo. En la validación cruzada *k*-fold, la muestra original de datos se subdivide al azar en *k* grupos de igual tamaño. De los *k* grupos, un grupo se eliminaría como un conjunto de reserva y los grupos restantes serían los datos de entrenamiento. Luego, el modelo predictivo se ajusta a los datos de entrenamiento y se evalúa en el conjunto de reserva.

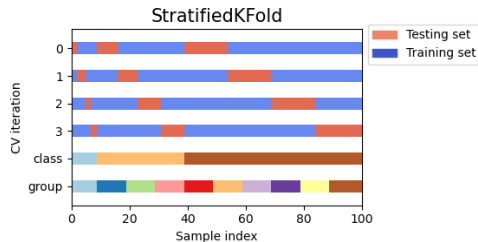
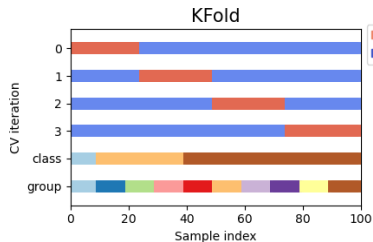
Este procedimiento se repite *k*-veces para que todos los grupos hayan servido exactamente una vez como grupo de reserva. En cada experimento se obtiene una estimación parcial del desempeño del modelo (por ejemplo un estimador parcial del error $E_i, i = 1, 2, \dots, k$).

Finalmente, el estimador del desempeño es

$$\hat{E} = \frac{1}{k}(E_1 + E_2 + \dots + E_k).$$

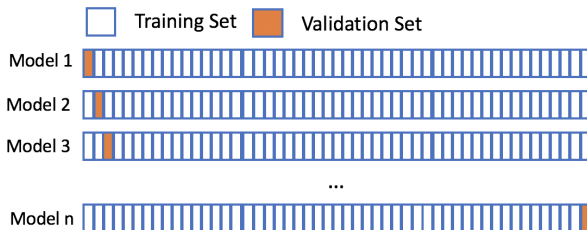
Validación Cruzada

Stratified k -fold Cross-Validation: Es una variante del anterior. Se subdivide la muestra de forma aleatoria en k grupos, asegurando que en cada grupo se mantienen la misma distribución de clases (proporciones balanceadas a lo largo de los grupos).



Validación Cruzada

Leave-One-Out Cross-Validation: es un caso particular del método de k -fold. Consiste en aplicar validación cruzada k -fold, con $k = n$ grupos, y n es el tamaño de la muestra. Así, Se hacen n grupos que consisten exactamente de un dato cada uno. Los datos se entrenarán en $n - 1$ muestras y se usarán para predecir la muestra que quedó fuera. Esto se repite n veces para que cada muestra sirva una vez como la muestra omitida.



Validación Cruzada

Group k -fold Cross-Validation: Es otra variante de la validación cruzada k -fold, que garantiza que el mismo grupo no esté representado en el conjunto de entrenamiento y en el conjunto de prueba. Por ejemplo, si quisiéramos construir un modelo predictivo que clasifique diagnósticos de pacientes (+, -), es probable que tengamos varios datos del mismo paciente. Aquí, los pacientes se considerarían como grupos (para que no queden dispersos).

