UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

# Métricas y Validación Cruzada

Alan Reyes-Figueroa

Aprendizaje Estadístico

(Aula 27) 13.mayo.2024

# Métricas

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

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

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

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

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

- log cosh

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

# 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

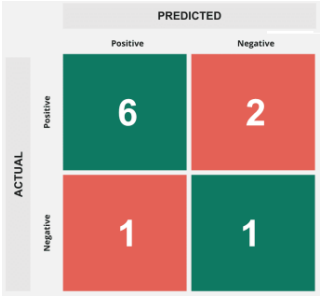
$$error = \frac{1}{n}(\text{número de datos clasificados incorrectamente}).$$

- Error ponderado

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

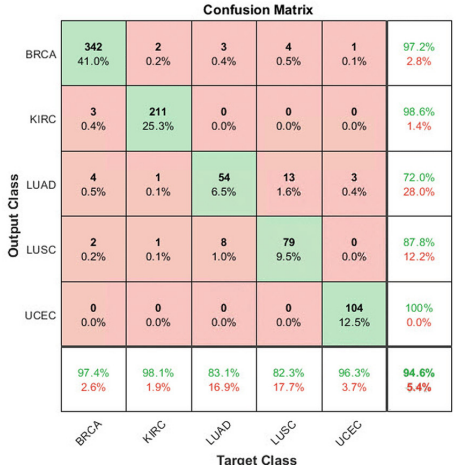
# Métricas para Clasificación

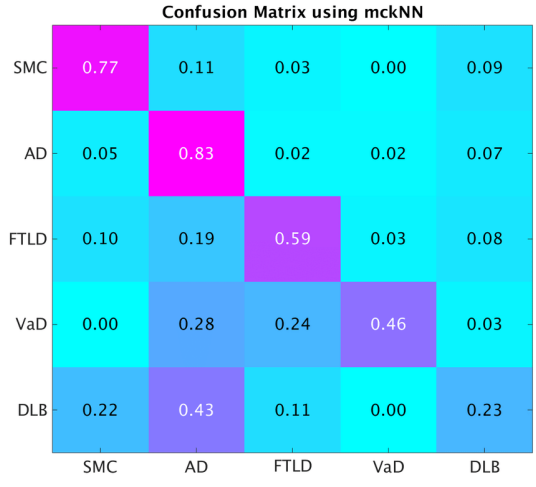
- Matriz de confusión: arreglo que cuenta las buenas y malas clasificaciones, compara etiquetas reales (*ground-truth*) contra las etiquetas estimadas por el modelo (*predicted*).




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

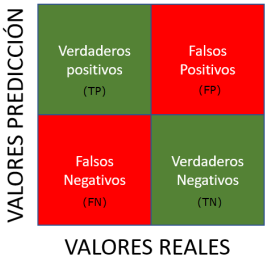
# Métricas para Clasificación



# Métricas para Clasificación

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




$$\mathrm{TP} = \#\{i : \widehat{y}_i = 1, y_i = 1\}, \qquad \mathrm{FP} = \#\{i : \widehat{y}_i = 1, y_i = 0\}$$
$$\mathrm{FN} = \#\{i : \widehat{y}_i = 0, y_i = 1\}, \qquad \mathrm{TN} = \#\{i : \widehat{y}_i = 0, y_i = 0\},$$
$$\mathrm{P} = \mathrm{TP} + \mathrm{FN} = \#\{i : y_i = 1\}, \qquad \mathrm{N} = \mathrm{FP} + \mathrm{TN} = \#\{i : y_i = 0\}.$$

# Métricas para Clasificación

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

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

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

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

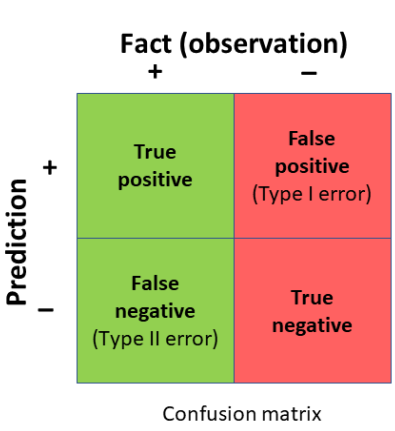
- ***Precision***, *positive predictive value* (PPV)

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

- ***Negative predictive value*** (NPV)

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

# Métricas para Clasificación




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)

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

- *Balanced accuracy* (BA)

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

- **$F_1$ *score*** Es la media armónica de la precisión y la sensitividad

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

- *Matthews correlation coefficient* (MCC)

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

- *Fowlkes-Mallows index* (FM)

$$\mathrm{FM} = \sqrt{\frac{TP}{TP + FP}\frac{TP}{TP + FN}} = \sqrt{PPV \cdot TPR}$$

# Métricas para Clasificación

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

# Métricas para Clasificación

También son importantes métricas que se puedan derivar. Por ejemplo

- ***Binary Cross-entropy***:

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

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

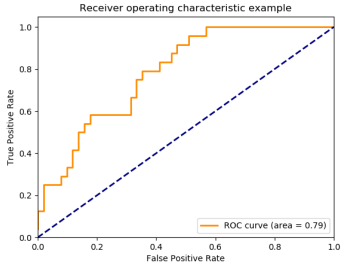
$$\text{Cross-entropy} = -\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{P}(y_i = k) \log \mathbb{P}(\widehat{y}_i = k).$$

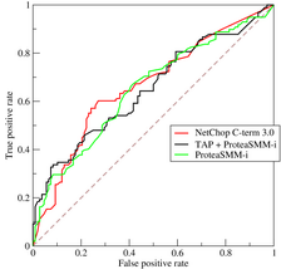
Existen muchas otras métricas de tipo distribucional, como información mutual, divergencia de Kullback-Leibler, entre otros.

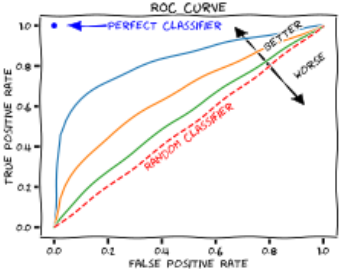
Otras métricas muy usadas, en el caso de clasificadores binarios con respuesta continua ($p \in (0, 1)$), son las siguientes:

# 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.



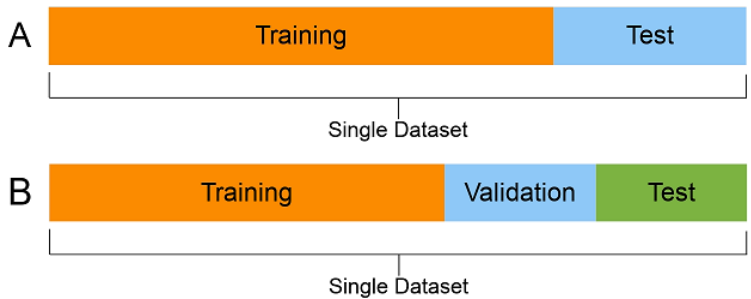
# Métricas para Clasificación

| 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 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | MSE | RMSE | MSE | RMSE |
| Artificial neural network | 2D | 86.20 | 85.71 | 0.50 | 50.00 | 0.21 | 0.35 | 0.20 | 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

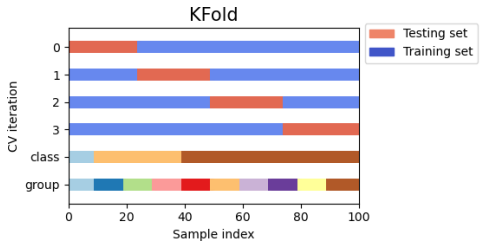
$k$-fold Cross-Validation: Es un procedimiento de remuestreo. En la validación cruzada $k$-fold, la muestra original de datos se subdivide al azar en $k$ grupos de igual tamaño. De los $k$ grupos, un grupo se eliminaría como un conjunto de reserva y los grupos restantes serían los datos de entrenamiento. 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 $E_i$, $i = 1, 2, \ldots, k$.

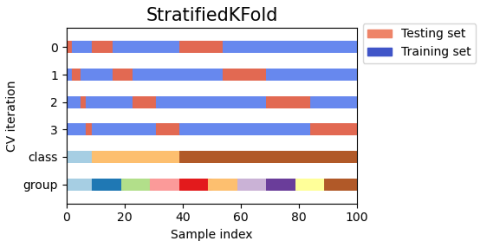
Finalmente, el estimador del desempeño es

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

# Validación Cruzada

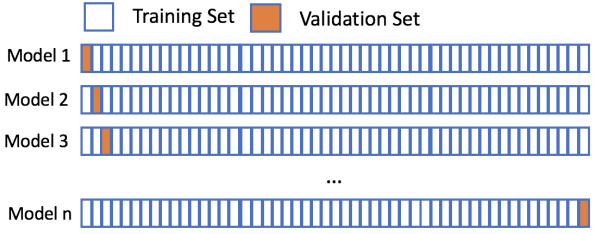
Stratified *k*-fold Cross-Validation: Es una variante del anterior. Se subdivide la muestra de forma aleatoria en *k* grupos, asegurando que en cada grupo se mantienen la misma distribución de clases (proporciones balanceadas a lo largo de los grupos).



# 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).