Universidad de Buenos Aires
Departamento de Computación

Evaluación de modelos

Temas de Procesamiento del Lenguaje Natural - NLP Aplicado

Dra. Viviana Cotik

2do bimestre 2025

Evaluación de modelos

- ¿Cómo sabemos cuán bueno es nuestro modelo?
- ¿Sobre qué conjunto de datos lo medimos?

Primera idea:

- Accuracy (eficacia) sobre cjto. entrenamiento: Porcentaje de datos de entrenamiento clasificados correctamente.
- Pero:
  - El modelo puede **memorizar** los datos de entrenamiento y tener **accuracy de 100%**. Medir **performance sobre los datos de entrenamiento** tiende a **sobreestimar los resultados**.

Conjunto Held-Out (o Control o Test)

- Al comenzar hay que **separar un conjunto de datos** (Held-Out, Control o **Test**) y **NO TOCARLOS** hasta el final
- Todas las pruebas y ajustes se realizan sobre el conjunto de **Desarrollo**
- Al terminar todas las pruebas, se evalúa el modelo obtenido con el conjunto Held-out

Estimación de performance

Se prueban distintas configuraciones.

Los datos de desarrollo se separan en:

- (100-q) % para entrenamiento
- q % para validación del modelo

Una vez definido cuál es el mejor modelo se entrena con cjto. de **desarrollo (**también llamado de **entrenamiento)**

Los datos se pueden separar al **azar** o con otro criterio (*), para evitar cualquier orden o estructura subyacente en los datos.

(*) El azar puede no ser el mejor criterio. Los datos de entrenamiento y validación deben ser **independientes** entre sí; los datos pueden estar desbalanceados, tener orden temporal, etc.

Validación cruzada (o cross-validation)

¿Y si tenemos mala suerte al separar los datos para entrenamiento/validación?

- La estimación de performance del modelo podría no ser realista.

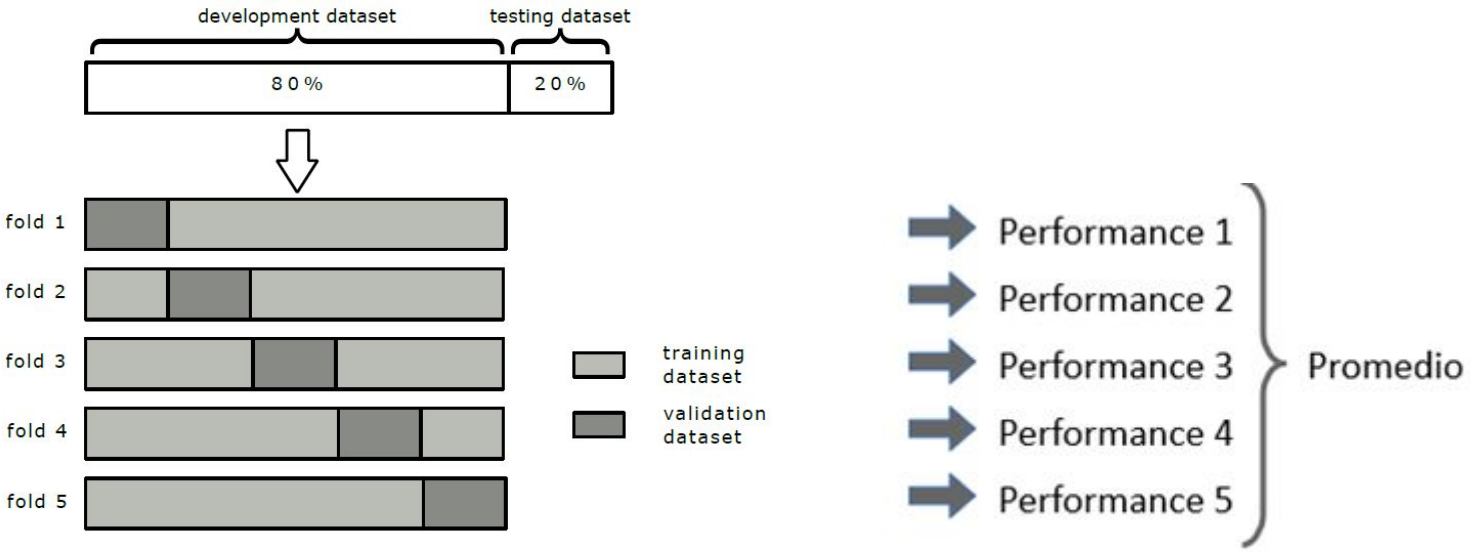
Para disminuir este riesgo: k-Fold Cross Validation

1. Desordenar los datos.
2. Separar en k folds disjuntos del mismo tamaño.
3. Para i = 1 .. k: entrenar sobre todos menos i; evaluar sobre i.

Ejemplo para k =5.

Valores usuales de D/T: 80-20, 70-30, ⅔-⅓

Validación cruzada



Selección de modelos

¿Por qué tendríamos distintos modelos para comparar?

- Distintos **atributos** (selección y transformación de atributos)
- Distintos **algoritmos** (árboles, LDA, NB, KNN, SVM, ...)
- Distintos **hiperparámetros** de cada algoritmo.

Ejemplo: hiperparámetros de los árboles de decisión

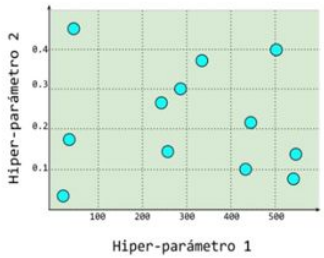
- Criterio de elección de atributos en cada nodo (Information Gain, Gini Gain...)
- Criterio de parada (ej: máxima profundidad)
- Estrategia de poda

Selección de modelos

¿Cómo buscar la mejor combinación de atributos + algoritmos + hiperparámetros?

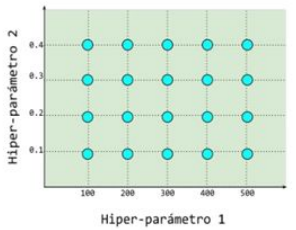
- Exploramos un espacio de búsqueda, **usando k-fold CV** para medir el desempeño de cada combinación.

Random search (best guess, 1 factor at a time) Explorar opciones y combinaciones al azar



Grid search

Plantear opciones y explorar todas las combinaciones



Al terminar, nos quedamos con la combinación con **mejor desempeño**, y **entrenamos un único modelo usando todos los datos**

Medidas de performance

Matriz de confusión
(clasificación binaria)

		valores reales	
		positivo	negativo
predicción	positivo	TP	FP
	negativo	FN	TN

TP: true positives

FP: false positives

TN: true negatives

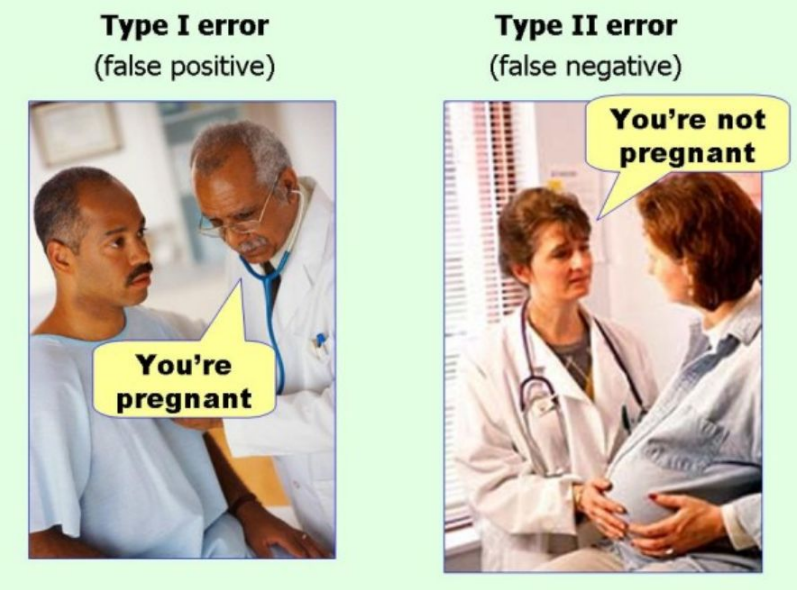
FN: false negatives

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

No dice nada sobre los tipos de aciertos y de errores que tiene el modelo.
Ej: autenticación en aplicación por voz.

- FP: autentica a un impostor
- FN: no autentica a un usuario válido

Medidas de performance



Tomado de Towards Data Science

Medidas de performance

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

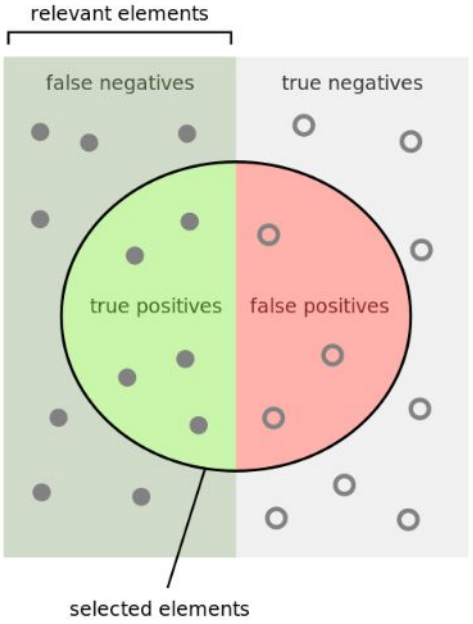
de las instancias clasificadas como positivas, cuántas lo son

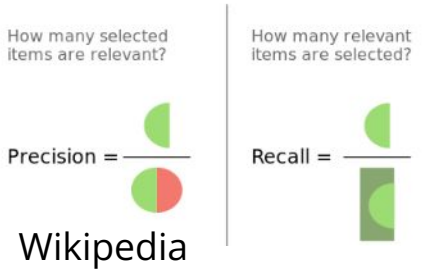
(cuán útiles son los resultados de búsqueda)

$$\text{Recall} = \frac{TP}{TP + FN}$$ (cubrimiento)

de las instancias positivas, cuántas fueron clasificadas como positivas

(cuán completos son los resultados)





Wikipedia

Medidas de performance

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

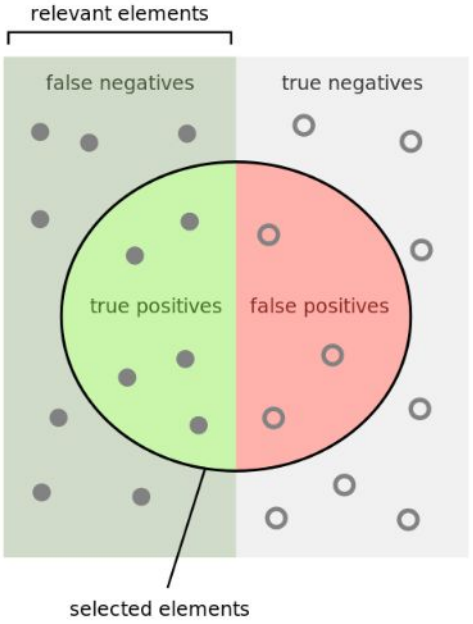
(cuán útiles son los resultados de búsqueda)

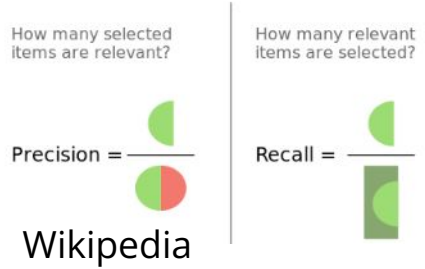
$$\text{Recall} = \frac{TP}{TP + FN}$$

(cuán completos son los resultados)

Se clasifican 4 como gatos (el primer y los últimos tres animales)

- TP: 3
- FP: 1
- P = 3/4, R = 3/3,





Wikipedia

Medidas de performance

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{Precisión} = \frac{TP}{TP + FP}$$

¿Cuál medida de performance debería priorizar cada uno de estos sistemas?

- enfermedad contagiosa
- test de embarazo

Media armónica:

$$F\text{-}measure = 2 \cdot \frac{Precisión \cdot Recall}{Precisión + Recall}$$

También llamada **F_1 score.**

Fórmula general:

$$F_\beta = (1 + \beta^2) \cdot \frac{Precisión \cdot Recall}{(\beta^2 \cdot Precisión) + Recall}$$

F_2 da más peso a Recall
$F_{0.5}$ da más peso a Precisión

Medidas de performance

$$\text{Recall} = \frac{TP}{TP + FN} = \textit{Sensitivity o bien True Positive Rate}$$

$$\frac{TN}{TN + FP} = \textit{Specificity o bien True Negative Rate}$$

Sensitivity/TPR: Porcentaje de pacientes enfermos correctamente diagnosticados.
Proporción de usuarios válidos autenticados

Specificity: Porcentaje de pacientes sanos correctamente diagnosticados.

$$\text{FPR} = \frac{FP}{FP + TN}$$

Ej. FPR: Proporción de impostores que aceptamos erróneamente.

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

¿Qué hacemos con un resultado de un estudio médico que nos da mal, pero que tiene bajo PPV?

Medidas de performance

CURVA ROC (Receiver operating characteristic)

- **Gráfico TPR (Recall) vs. FPR**

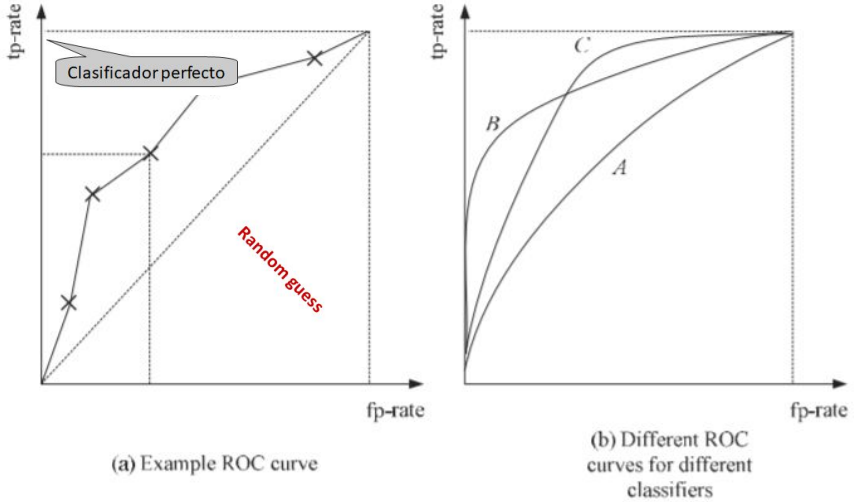
$$\text{Recall} = TPR = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{FP}{FP + TN}$$

Construcción: Variar el umbral de detección entre 0 y 100%. Para cada valor, calcular TPR y FPR (un punto en la curva).

Área bajo la curva (AUC)

- Un valor numérico. entre 0 y 1. Azar=0.5



Fuente: Introduction to ML, Alpaydin

Matriz de confusión n-aria

	Manzana (predicho)	Naranja (predicho)	Oliva (predicho)	Pera (predicho)
Manzana (real)	**MM**	**MN**	**MO**	**MP**
Naranja (real)	**NM**	**NN**	**NO**	**NP**
Oliva (real)	**OM**	**ON**	**OO**	**OP**
Pera (real)	**PM**	**PN**	**PO**	**PP**

Las medidas **precisión, recall, etc. sólo pueden formularse en forma binaria:** **cada clase contra el resto**.

$$\text{Precisión}(Manzana) = \frac{MM}{MM + NM + OM + PM}$$

$$\text{Recall}(Manzana) = \frac{MM}{MM + MN + MO + MP}$$

Matriz de confusión - Interiorizándonos

- Problema de clasificación:
  - en función de cómo clasifica
- Detección de entidades nombradas:
  - **Match exacto /total / exact match:** mismo comienzo, mismo fin, mismo tipo de entidad
  - **Match parcial:** mismo comienzo, o mismo fin, o incluido, o..

BIO format

Estándar para representar anotación de entidades en NLP
BIO/IOB Tagging/etiquetado (inside, outside, beginning)

- Formato habitual para etiquetar tokens en una tarea de chunking (ej. reconocimiento de entidades nombradas).
    - United(B_Country) States(I_Country) of(I_Country) America(E_Country)

Significado de prefijos de etiquetas

B-: indica que ese es el comienzo de la entidad.
I-: indica que el token forma parte de una entidad (no al comienzo)
O: indica que el token no pertenece a ninguna entidad.

Existen IOB y IOB 2. En IOB los chunks no siempre empiezan con B- (sólo se usa cuando una etiqueta está seguida por otra etiqueta del mismo tipo sin Os en el medio). Usamos IOB2, pero lo llamamos IOB.

En NER BERT usa etiquetas BIO para entrenamiento y devuelve los resultados también en ese formato.

Referencias:

- nlp - what is BIO Tags for creating custom NER Named entity recognization? - Data Science Stack Exchange
- BIO / IOB Tagged Text to Original Text | by Jeril Kuriakose | Analytics Vidhya | Medium

Matriz de confusión - Interiorizándonos

token	word	TAG	annotation tag according to GS	predicted tag according to algorithm
1	Ambos	TAG	O	O
2	rinones	TAG	B−AE	B-AE
3	de	TAG	O	O
4	ecoestructura	TAG	O	O
5	normal	TAG	O	B-AE
6	.	TAG	O	O
7	Derrame	TAG	B-FI	B-FI
8	pleural	TAG	I-FI	I-FI
9	derecho	TAG	I-FI	O

Según el *gold standard* hay dos entidades. Según el algoritmo hay tres.

exact match	**partial match**
1 TP (riñones), 2 FP (normal y derrame pleural) y 1 FN (derrame pleural derecho).	1 TP (rinones), 1 match parcial (derrame pleural y derrame pleural derecho), y 1 FP (normal).

Match parcial

Es necesario definir:

- ¿Qué cuenta como match parcial (partial/lenient match)?
  - right-match
  - left-match
  - el sustantivo y no los adjetivos
  - ...
- ¿Cuánto valor darle a un match parcial?
  - 0.5 de un total
  - (cant de tokens matcheados)/(cant. tokens de la entidad del Gold Standard)
  - ...

Verificar que distintos matchs parciales con una misma entidad del Gold Standard no sumen más que un match exacto.

Referencias:

- métricas de MUC y ACE (Automatic Content Extraction Program)

Tipos de error

En el cuadro vemos distintos tipos de error*,+

FI: hallazgo clínico AE: entidad anatómica

id	correct solution	system output	description
1	observed	<FI>observed</FI>	The algorithm discovers a non-existent entity (FP)
2	<FI>Edema</FI>	Edema	The algorithm misses an existent entity (FN)
3	<AE>lungs</AE>	<FI>lungs</FI>	The algorithm detects an entity, but assigns it the wrong label (*text* according to MUC, *labeling error* according to Manning)
4	<FI>[enlarged esophagus]</FI>	<FI>[enlarged]</FI>	The algorithm detects an entity, with the right label but with wrong boundaries (*type* according MUC, *boundary error* according to Manning)
5	<AE>[upper part of the liver]</AE>	<AE>liver</AE>	The algorithm detects an entity with wrong labels and wrong boundaries (label-boundary error according to Manning)
6	<FI>cyst</FI>	<FI>cyst</FI>	The algorithm detects an entity, with right label and right boundaries (TP)

*MUC: Message Understanding conference

+Manning: Doing Named Entity Recognition? Don't optimize for F1

Posible cálculo de match parcial

Según MUC-3. A modo de ejemplo.

abbrev.	category	criteria	meaning in confusion matrix
COR	correct	algorithm guess is positive and gold standard value is positive	TP
PAR	partial	algorithm and gold standard values have a partial coincidence (both with positive values)	accounted as a fraction of TP
INC	incorrect	corresponds to two incorrect results in the confusion matrix, one counted as FP for a spurious answer and one counted as FN for not getting the correct positive answer	
SPU	spurious	algorithm guess is positive and gold standard value is negative	FP
MIS	missing	algorithm guess is negative and gold standard value is positive	FN
NON	noncommittal	algorithm and GS values are negative	TN

Scoring criteria for evaluating MUC-3 results.

POS: positives

$$POS = COR + PAR + INC + MIS$$

ACT: predicted as positives

$$ACT = COR + PAR + INC + SPU$$

$$precision = \frac{COR + 0.5 * PAR}{ACT}$$

$$recall = \frac{COR + 0.5 * PAR}{POS}$$

Factores a considerar en la elección de modelos

- **Tasa de error**
- **Velocidad de entrenamiento y velocidad de test**
- **Interpretabilidad** (¿el conocimiento extraído del modelo puede ser validado por expertos?)
- **Facilidad de desarrollo**

Experimentos de aprendizaje automático

1. **Establecer objetivo de estudio** (error de un algoritmo, comparación de dos algoritmos, etc.)
2. **Seleccionar la métrica para evaluar performance**
3. Seleccionar **factores importantes** (dependen del punto 1): (hiperparámetros de un algoritmo, comparación de algoritmos: algoritmos a comparar)
4. **Elegir diseño experimental** (división de conjunto en entrenamiento y test, cross validation, hiperparámetros: modificaciones aleatorias vs. grid search)
5. **Realización de experimento** (uso de código testeado, reproducibilidad de resultados)
6. Realizar **análisis estadísticos** de los datos
7. **Conclusiones**: son sobre los datos utilizados. Realizar análisis de errores.

Resumen

Armado de conjunto de datos:

- Dejar **apartado un conjunto de test** o held-out
- Seleccionar modelos con **conjunto de desarrollo** (datos de entrenamiento y validación):
  - k-fold cross validation
  - grid search y random search
- Medidas de performance:
  - Accuracy
  - Precision, Recall, F1
  - Sensibilidad, especificidad
  - TPR, FPR, PPV, curva ROC, AUC (área bajo la curva)
  - Exact y partial match

Bibliografía

Capítulos de libros:

.ISLR, Cap. 2 (2.2)
.Alpaydin, Cap. 19 (hasta 19.7 inclusive)

Páginas web

. Doing Named Entity Recognition? Don’t optimize for F1, Chris Manning.