Universidad Autónoma de Baja California

Maestría y Doctorado en Ciencias e Ingeniería
Ingeniería en Computación

5. Clasificación

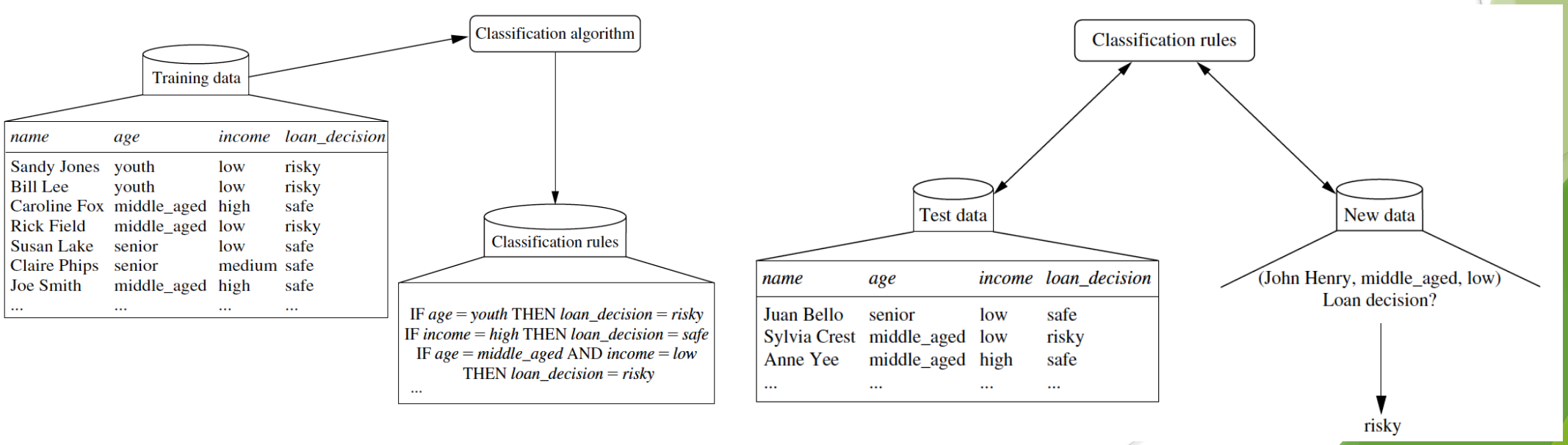
Minería de Datos

¿Qué es clasificación?

- Es una forma de análisis de datos que extreae modelos que describen clases importantes a partir de datos. Tales modelos, llamado clasificadores, predicen etiquetas de clases categóricas (discretas y no ordenadas)

Conceptos básicos

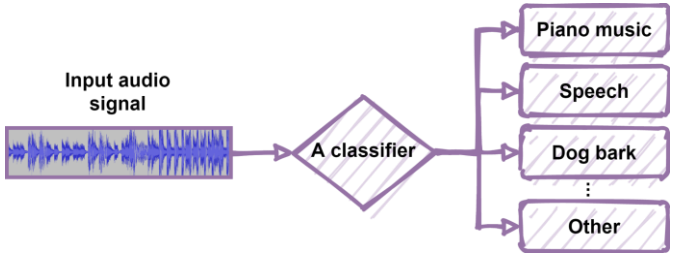
- Ejemplos de clasificadores y sus usos…
- ¿Cómo funciona la clasificación? Se realiza en dos pasos, que consisten en un paso de entrenamiento (donde se construye el modelo de clasificación) y uno de clasificación (donde el modelo es utilizado para predecir clases etiquetadas para un conjunto dado de datos).

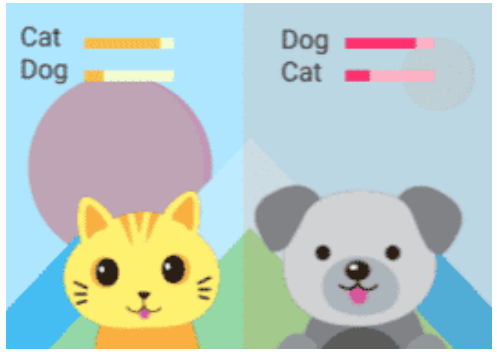


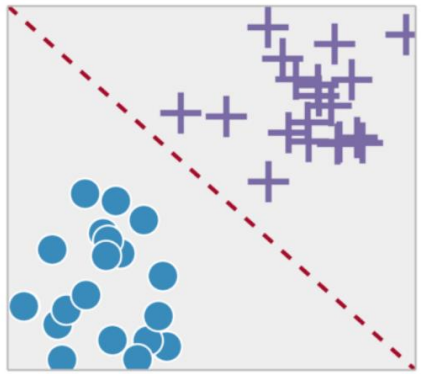
name	*age*	*income*	*loan_decision*
Sandy Jones	youth	low	risky
Bill Lee	youth	low	risky
Caroline Fox	middle_aged	high	safe
Rick Field	middle_aged	low	risky
Susan Lake	senior	low	safe
Claire Phips	senior	medium	safe
Joe Smith	middle_aged	high	safe
...	...	...	...

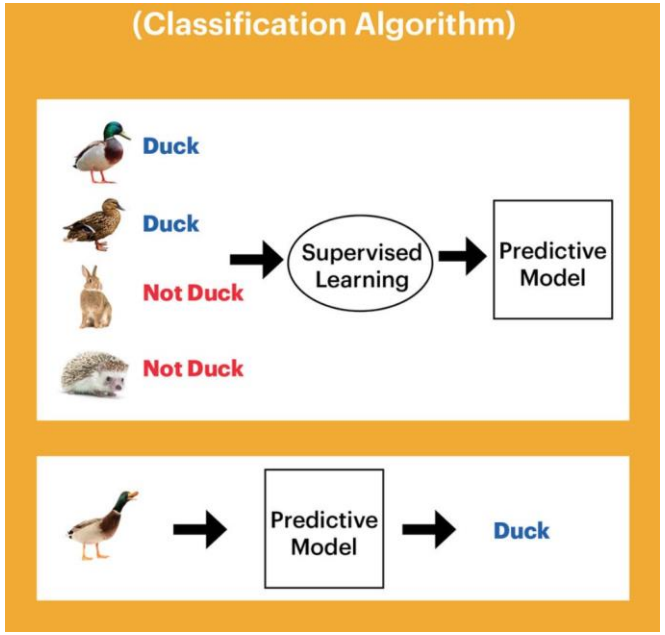
name	*age*	*income*	*loan_decision*
Juan Bello	senior	low	safe
Sylvia Crest	middle_aged	low	risky
Anne Yee	middle_aged	high	safe
...	...	...	...

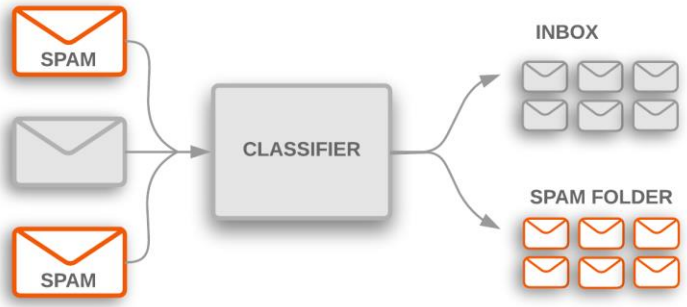
Ejemplos de clasificadores





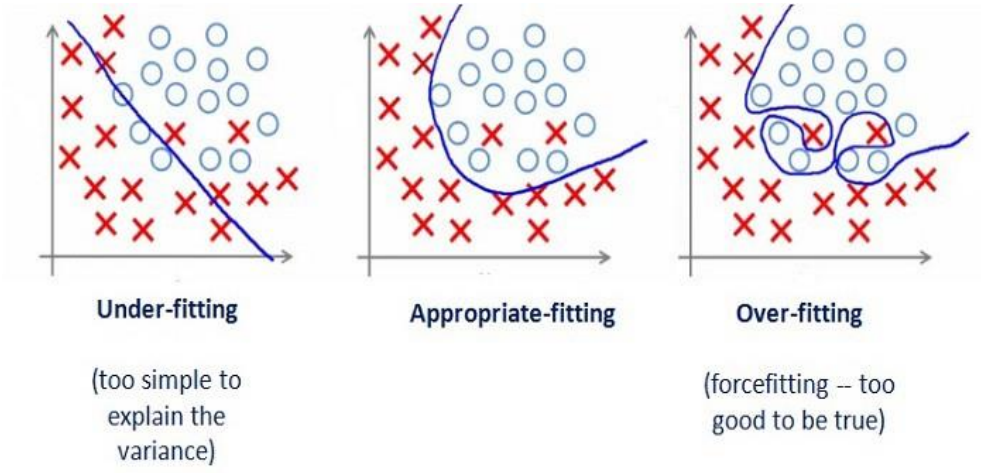






Conceptos básicos

- Aprendizaje supervisado, clasificadores y regresión
- Aprendizaje no supervisado, agrupación
- Hay que separar el conjunto de datos de prueba del de entrenamiento para evitar el *overfitting*



- La precisión de un clasificador se da a través del porcentaje de clasificaciones correcta que realiza con el conjunto de datos de prueba

Ejemplos de clasificadores

- Árboles de decisión
- Redes de Bayes
- A base de reglas
- Reglas de asociación
- Neuronales
- Deep learning
- Difusos
- KNN
- SVM
- Etc…

Validación

Métricas para evaluar el desempeño de clasificadores

- Aunque *exactitud* es una medida específica, la palabra también puede ser usada como un término que refiere a las habilidades predictivas de un clasificador.
- Términos básicos para medir exactitud en clasificadores:
  - **True Positives (TP):** Tuplas positivas que fueron etiquetadas correctamente por el clasificador.
  - **True Negatives (TN):** Tuplas negativas que fueron etiquetadas correctamente por el clasificador.
  - **False Positives (FP):** Tuplas negativas que fueron incorrectamente etiquetadas por el clasificador como positivas.
  - **False negatives (FN):** Tuplas positivas que fueron incorrectamente etiquetadas por el clasificador como negativas.
  - **Positives (P):** Tuplas positivas que refieren a la clase de interés.
  - **Negatives (N):** Tuplas negativas que refieren al resto de las clases.

Matriz de confusión

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P + N$

Ejemplo de matriz de confusión:

Classes	*buys_computer = yes*	*buys_computer = no*	*Total*	*Recognition (%)*
buys_computer = yes	**6954**	**46**	7000	99.34
buys_computer = no	**412**	**2588**	3000	86.27
Total	7366	2634	10,000	95.42

Exactitud

- La exactitud de un clasificador en una dada prueba es el porcentaje de las tuplas de conjunto de prueba que fueron correctamente clasificados por el clasificador.

$$accuracy = \frac{TP + TN}{P + N}$$

Tasa de error

- La tasa de error o tasa de mala clasificación de un clasificador, M, donde 1-precisión(M), donde precisión(M) es la precisión del clasificador M.

$$error\ rate = \frac{FP + FN}{P + N}$$

Otras medidas

- Precisión. Es una medida de exactitud (¿qué porcentaje de las tuplas etiquetadas como positivas realmente lo son?)

$$precision = \frac{TP}{TP + FP}$$

- Recuerdo. Es una medida de completitud (¿qué porcentaje de las tuplas positivas están etiquetadas de esa manera?)

$$recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Otras medidas

- **Sensitivity.** Mide la probabilidad de tener actuales positivos.

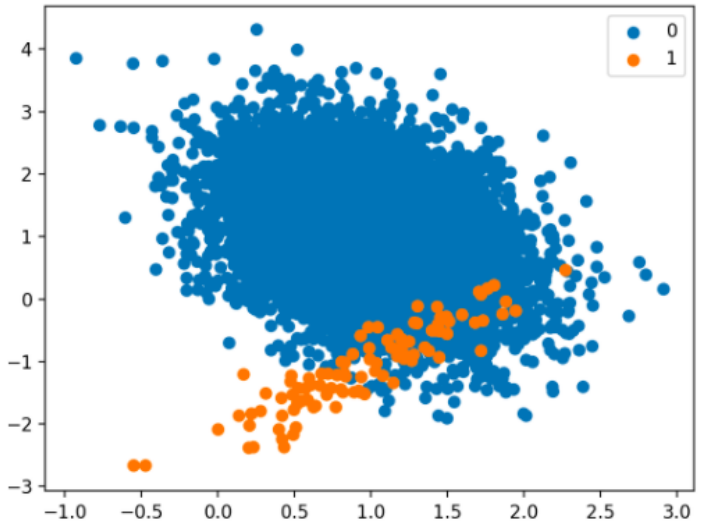
$$sensitivity = \frac{TP}{P}$$

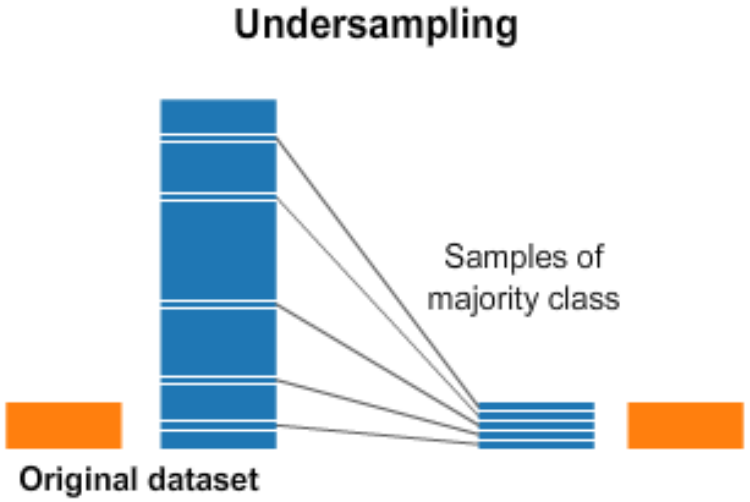
- **Specificity.** Mide la probabilidad de tener actuales negativos.

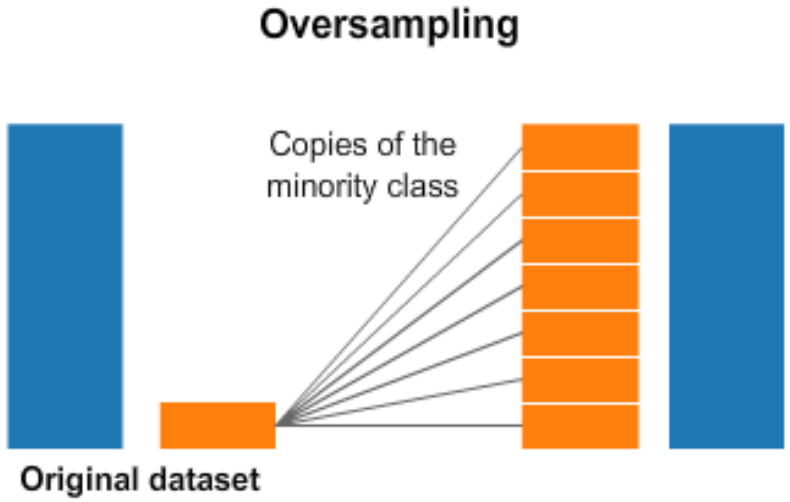
$$specificity = \frac{TN}{N}$$

El problema de clases mal balanceadas

- ¿Cuál es éste problema?







Resumen de medidas de clasificación

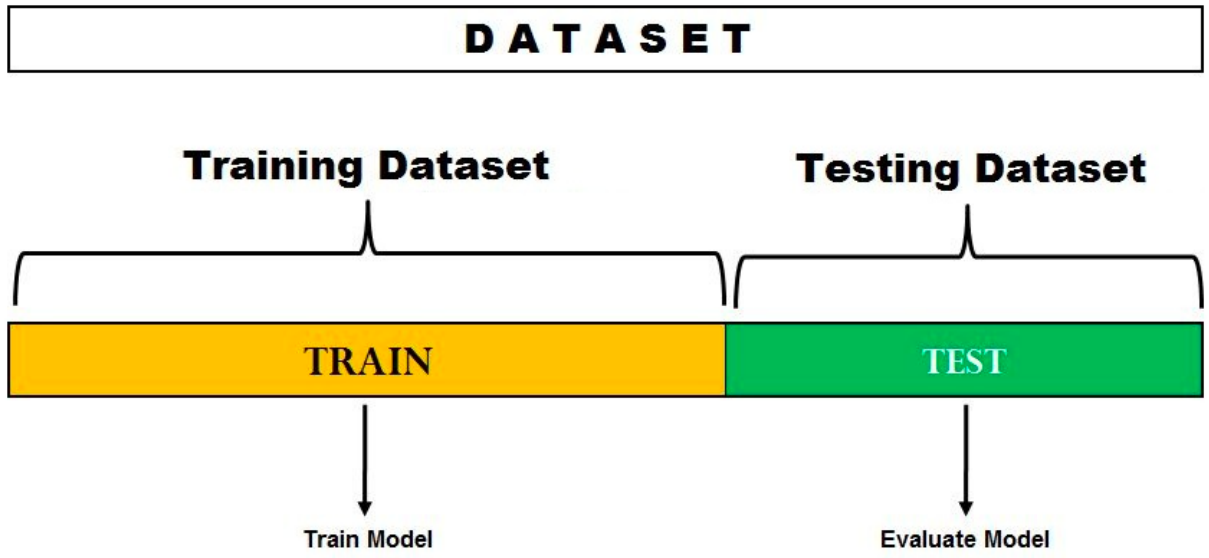
Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

Separación de datos

- Hold-Out
- Random subsampling
- K-Fold
- Stratified cross-validation
- Leave-one-out
- Bootstrap

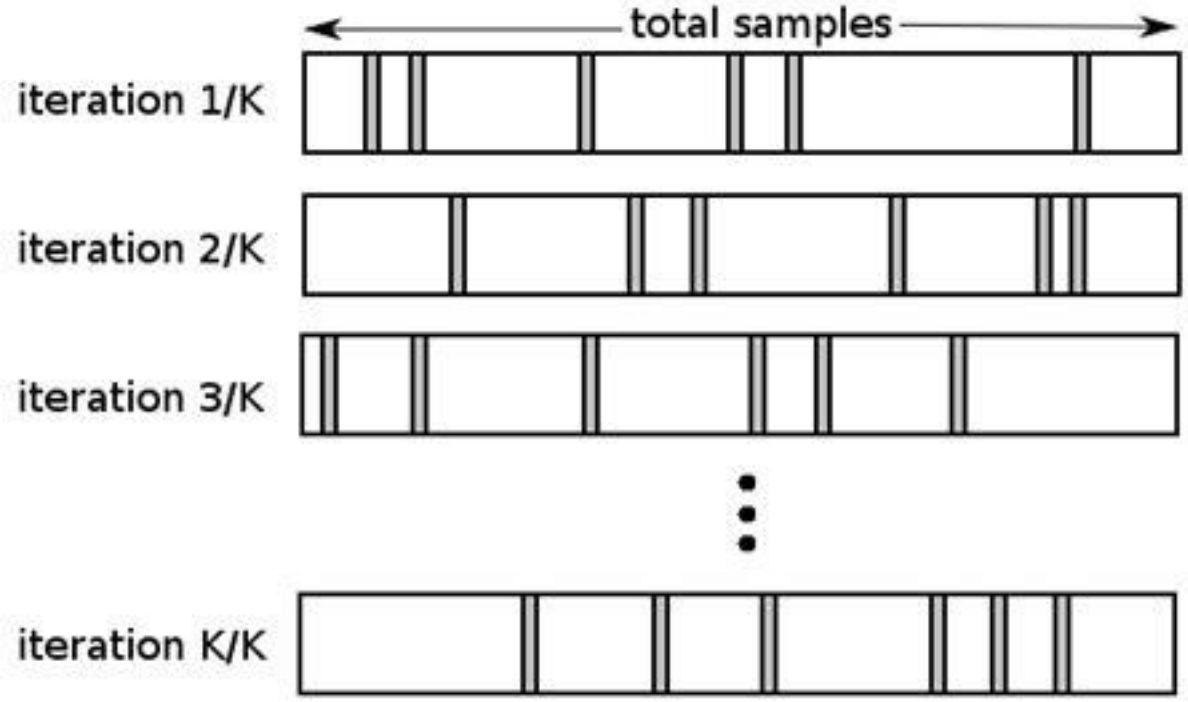
Separación de datos

- **Hold-Out.** Los datos aleatoriamente se particionan en dos conjuntos independientes, un conjunto de entrenamiento y un conjunto de prueba.



Separación de datos

- **Random subsampling.** Es una variación del método Hold-Out en donde el método Hold-Out se repite N veces. Y la exactitud general se estima promediando las exactitudes obtenidas en cada iteración.

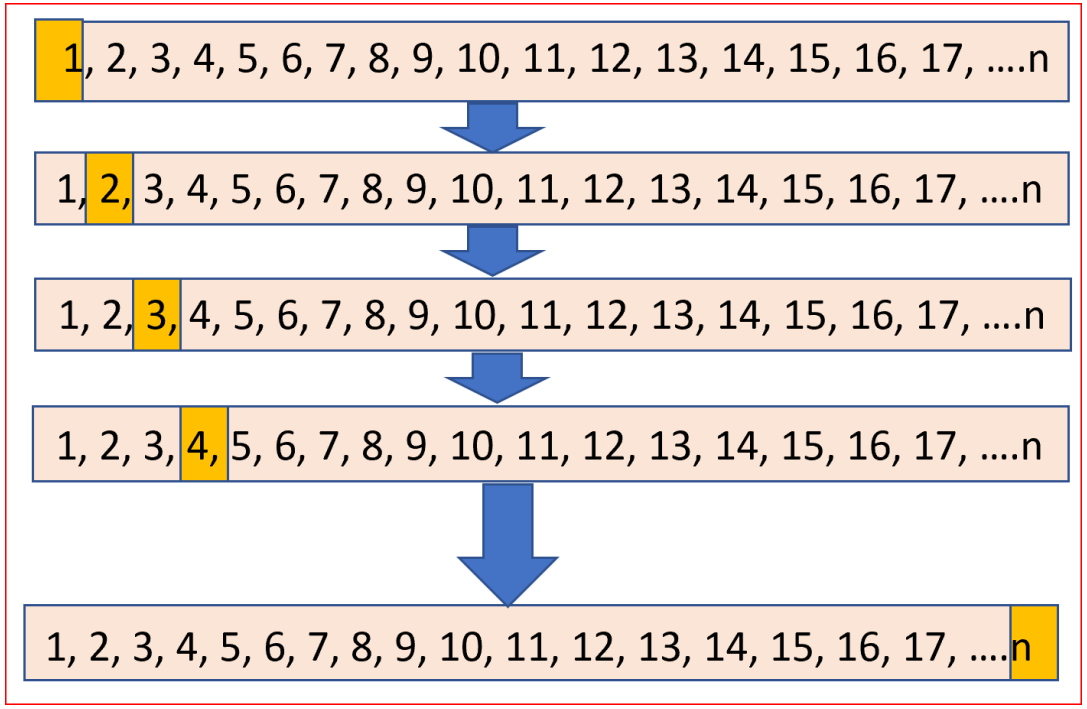


Separación de datos

- **K-Fold.** Los datos se separan en K particiones, donde se va la primer partición se usa para prueba y el resto para entrenamiento, y se va recorriendo dicha partición de prueba hasta ejecutar K experimentos. Al final se promedia el resultado de exactitud.

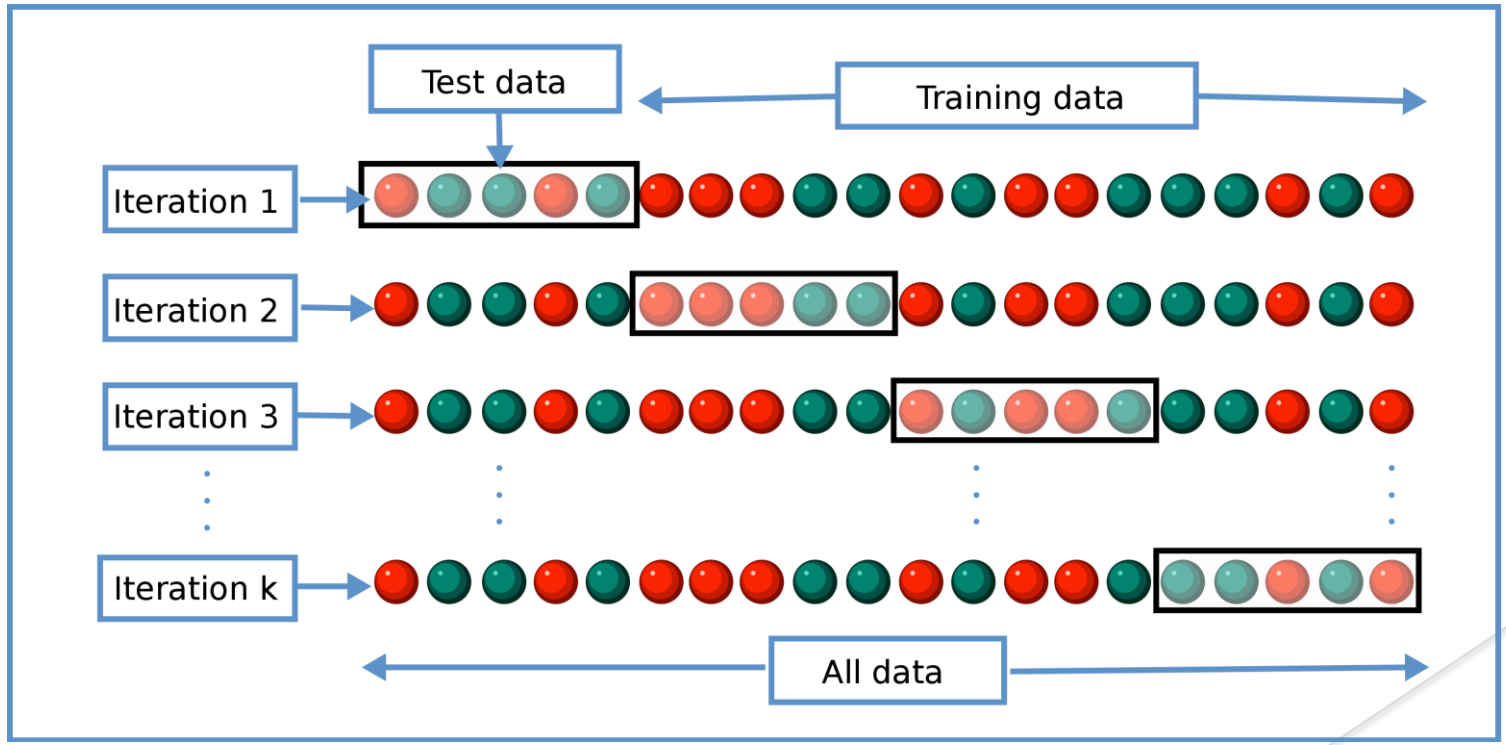
Separación de datos

- **Leave-one-out.** Igual que K-fold, pero K es igual al número de tuplas.



Separación de datos

- **Stratified cross-validation.** Igual que K-fold, pero las tuplas de clases dentro de cada K está balanceada para toda K.



Separación de datos

- **Bootstrap.** La selección aleatoria de tuplas asigna probabilidades a que en cada selección de muestras se repitan algunas tuplas previamente seleccionadas. *Nota. Aplica a cualquier otra técnica de separación de datos.*

Técnicas para mejorar la precisión de clasificación

- **Métodos ensemble**. K modelos se combinan para tomar una mejor decisión.
- **Bagging**. K modelos clasifican, donde el voto mayoritario gana.
- **Boosting**. Basado en el desempeño durante entrenamiento y validación de cada modelo, un peso se le asigna para ponderar su decisión final.
- **AdaBoost**. Similar al anterior, pero el peso es asignado de manera dinámica durante el entrenamiento y validación de cada modelo.
- **Random forest**. Comportamiento similar a Bagging, pero se limita exclusivamente a los clasificadores árboles de decisión.

Actividad

- Trabajar con 1 dataset previamente utilizada (no el del Iris)
- Aplicar validación a cada dataset
  - Utilizar 5 métricas diferentes
- Separando los datos mediante:
  - Hold-Out (60/40)
  - Hold-Out (100/100) Para comprobar overfitting
  - Random subsampling (N=30)
  - K-fold (K=10)
  - Leave-one-out
  - Stratified cross-validation (K=10)
- Utilizar 5 clasificadores diferentes para cada caso
- De los experimentos realizados, exponer:
  - Descripción de las técnicas utilizadas (ponerse de acuerdo en el foro de Bb para no repetir)
  - Explicar el procedimiento para la experimentación completa
  - Mostrar una tabla con toda la información para encontrar qué dio el mejor resultado
  - Resultados obtenidos