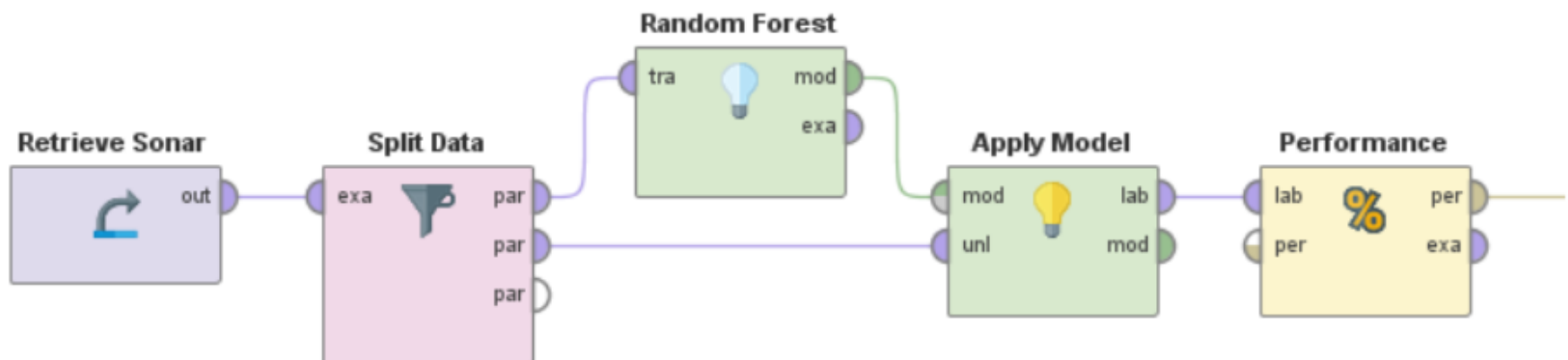


Validación utilizando conjuntos hold-out

- Ya vimos que considerar el error de entrenamiento es una mala idea...
- Lo primero a tener en cuenta es que obtener o generar más datos puede ser difícil y costoso!
- Es práctica común entonces utilizar una parte de los datos disponibles para entrenamiento, y otra para test (“hold-out set” – “data split”)



k-fold cross validation

- las muestras se particionan en k conjuntos de (aprox) el mismo tamaño
- el modelo se ajusta usando todas las muestras excepto el primer subset (*first fold*)
- las muestras excluidas se predicen con el modelo y se usan para estimar el rendimiento
- el subset se devuelve al dataset de entrenamiento y el proceso se repite con el segundo subset, y así sucesivamente
- los k resultados de estimaciones son integrados

Step 1: Divide the data set into k folds, here k is 10.



Step 2: Use one fold for testing a model built on all other data parts.



Step 3: Repeat the model building and testing for each of the data folds.



Step 4: Calculate the average of all of the k test errors and deliver this as result.

leave-one-out cross validation

LOOCV

- similar a k-fold cross-validation
- k = cantidad total de muestras
- k iteraciones
- en cada iteración, 1 muestra se deja afuera y se usa para medir el rendimiento del modelo entrenado con las $k-1$ restantes
- el rendimiento final es la media de los rendimientos obtenidos en cada una de las iteraciones

Evaluación de los modelos

- Métodos para determinar el desempeño de un modelo de clasificación:
 - Matriz de confusión
 - Curva ROC (receiver operator characteristic)
 - Lift Chart

Matriz de confusión

- Consideremos un caso de clasificación binario Y/N (yes/no)
 - Si el valor real es Y y el predicho Y, estamos ante un **“positivo verdadero” (TP)**
 - Si el valor real es N y el predicho Y, estamos ante un **“falso positivo” (FP)**
 - Si el valor real es N y el predicho N, estamos ante un **“negativo verdadero” (TN)**
 - Si el valor real es Y y el predicho N, estamos ante un **“falso negativo” (FN)**

Matriz de confusión

		Actual Class(Observation)	
		Y	N
Predicted Class (Expectation)	Y	TP (true positive) Correct result	FP (false positive) Unexpected result
	N	FN (false negative) Missing result	TN (true negative) Correct absence of result

- En un caso de clasificación perfecta, FN y FP serían 0
- El mismo razonamiento puede extenderse para múltiples clases

- ***habilidad*** del clasificador de seleccionar de seleccionar todos los casos que ***necesitan*** ser seleccionados
- Un clasificador ideal seleccionará todos los Y correctos y no perderá ninguno
- **$TP/(TP + FN)$**
- El caso ideal es $FN=0$
- En algunos problemas puede no ser suficiente:
 - Si la clase indica si una transacción es válida, y ocurren transacciones fraudulentas con **0.1%** de probabilidad, predecir todo como válido generaría un 99.9% de sensibilidad, lo cual no es útil

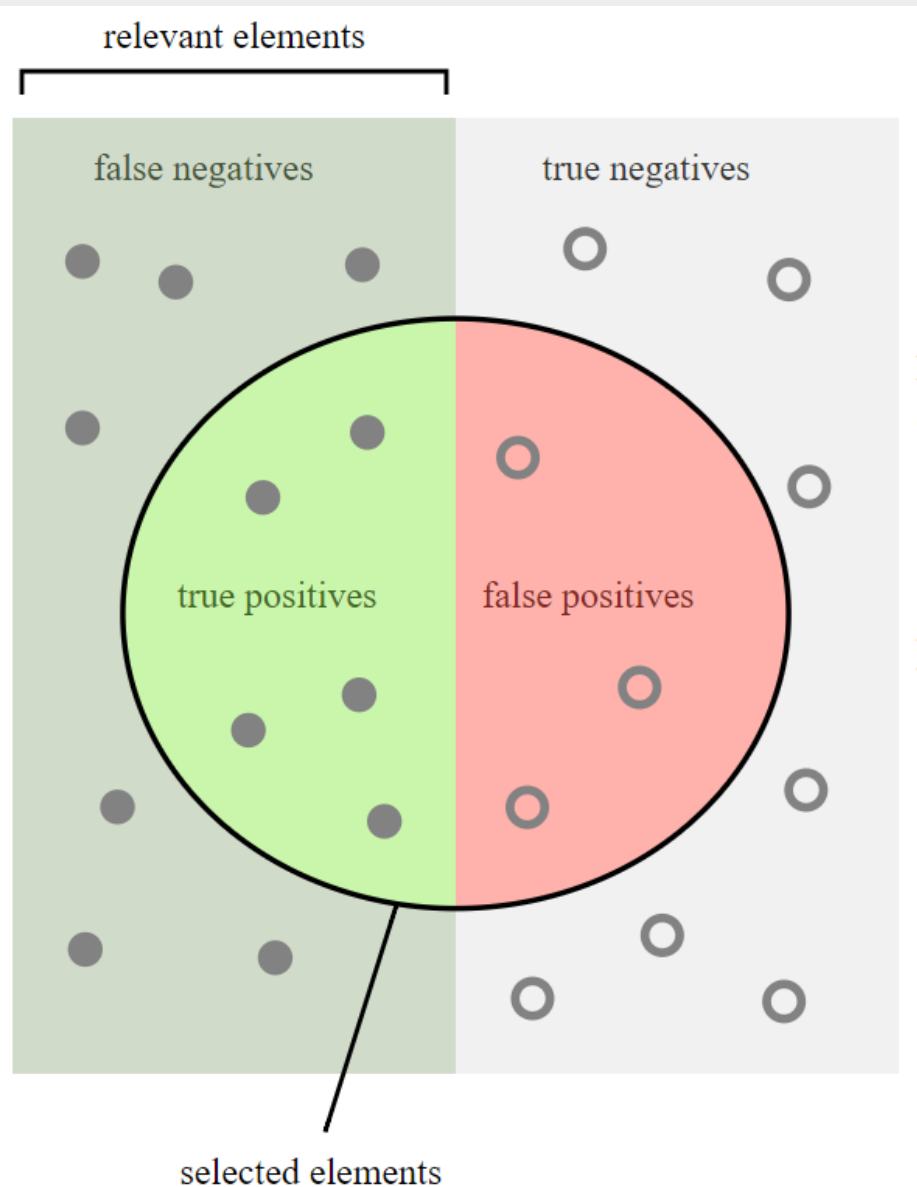
Otras métricas - Especificidad

- habilidad del clasificador de ***rechazar*** todos los casos que **deben** ser rechazados (es decir, **sin falsos positivos**)
- Un clasificador perfecto rechazará todos los **N** y no seleccionará ningún resultado no esperado (FP)
- $TN / (TN + FP)$

Otras métricas - Relevancia

- Fácil de entender en un escenario de búsqueda y recuperación en documentos
 - Corremos una búsqueda por un término específico y retorna **100** documentos
 - De estos, p.ej., sólo **70** son útiles porque eran *relevantes* para la búsqueda
 - La búsqueda perdió otros 40 que podrían haber sido útiles
- En este contexto se pueden definir algunos términos adicionales

Otras métricas: Precision y Recall



How many selected
items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant
items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Otras métricas - Precision

- Proporción, de todos los casos encontrados, que fueron *realmente relevantes*
- En el ejemplo anterior, el número era 70, entonces la precisión es 70/100
- **$TP/(TP+FP)$**

Otras métricas - **Recall**

- **Proporción** de casos **relevantes** que fueron encontrados **entre todos** los casos relevantes
- En el ejemplo anterior, sólo 70 casos relevantes de un total de 110 (= 70 encontrados + 40 perdidos) fueron encontrados, dando un **recall** de $70/110 = 63.33\%$
- **$TP/(TP+FN)$**
- Puede verse que es lo mismo que sensibilidad

Otras métricas - **Accuracy**

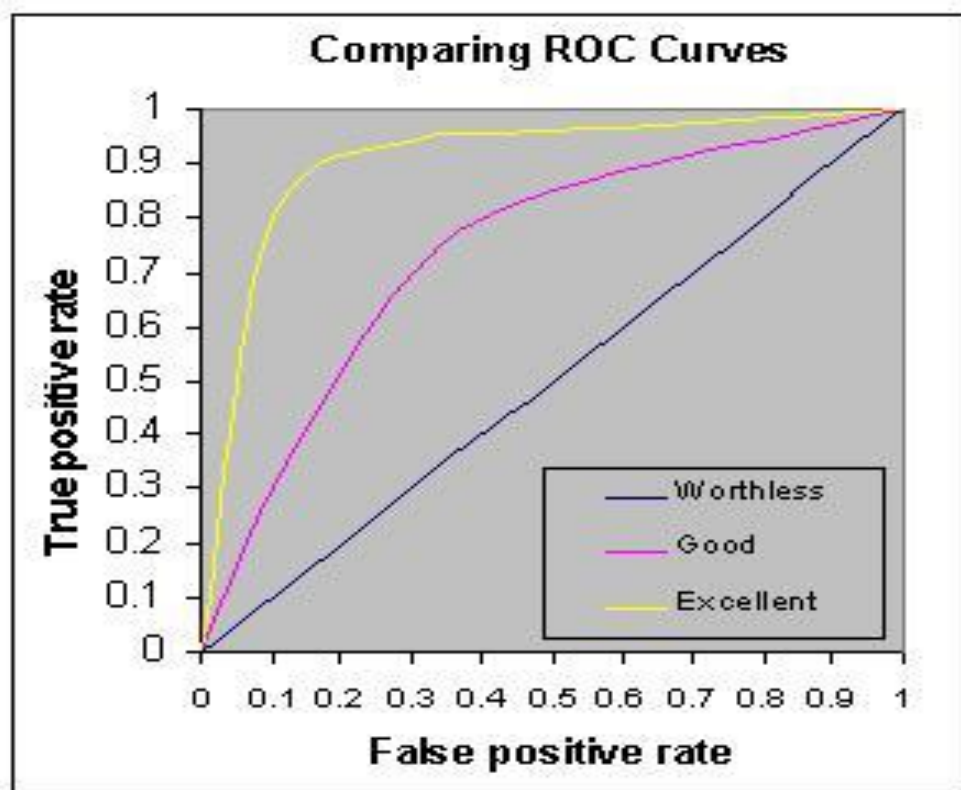
- Mide la capacidad del clasificador de predecir correctamente las clases
- Habilidad de **seleccionar** todos los casos que *deben ser seleccionados* y **rechazar** todos los que *deben ser rechazados*
 - 100% de accuracy implica $FN = FP = 0$
 - $(TP+TN)/(TP+FP+TN+FN)$
- **Error:** $(1 - \text{accuracy})$

Curva ROC

- Un clasificador puede tener un valor de *accuracy* alto pero valores bajos de *precision* y/o *recall*
- Ej: Sea un modelo de detección de fraude, donde indicamos con la clase positiva la ocurrencia de un fraude. Una medida de *recall* bajo (de todos los fraudes predecimos una proporción baja) es no deseable.
- ¿Podemos sacrificar un poco el *accuracy* para aumentar efectivamente el *recall*?
- La curva ROC (*receiver operator characteristic*) permite medir esto

Curva ROC

- En una curva ROC graficamos la fracción de **verdaderos positivos (sensitivity o recall)** que van apareciendo en los datos (eje vertical) vs. la fracción de **falsos positivos** (eje horizontal)

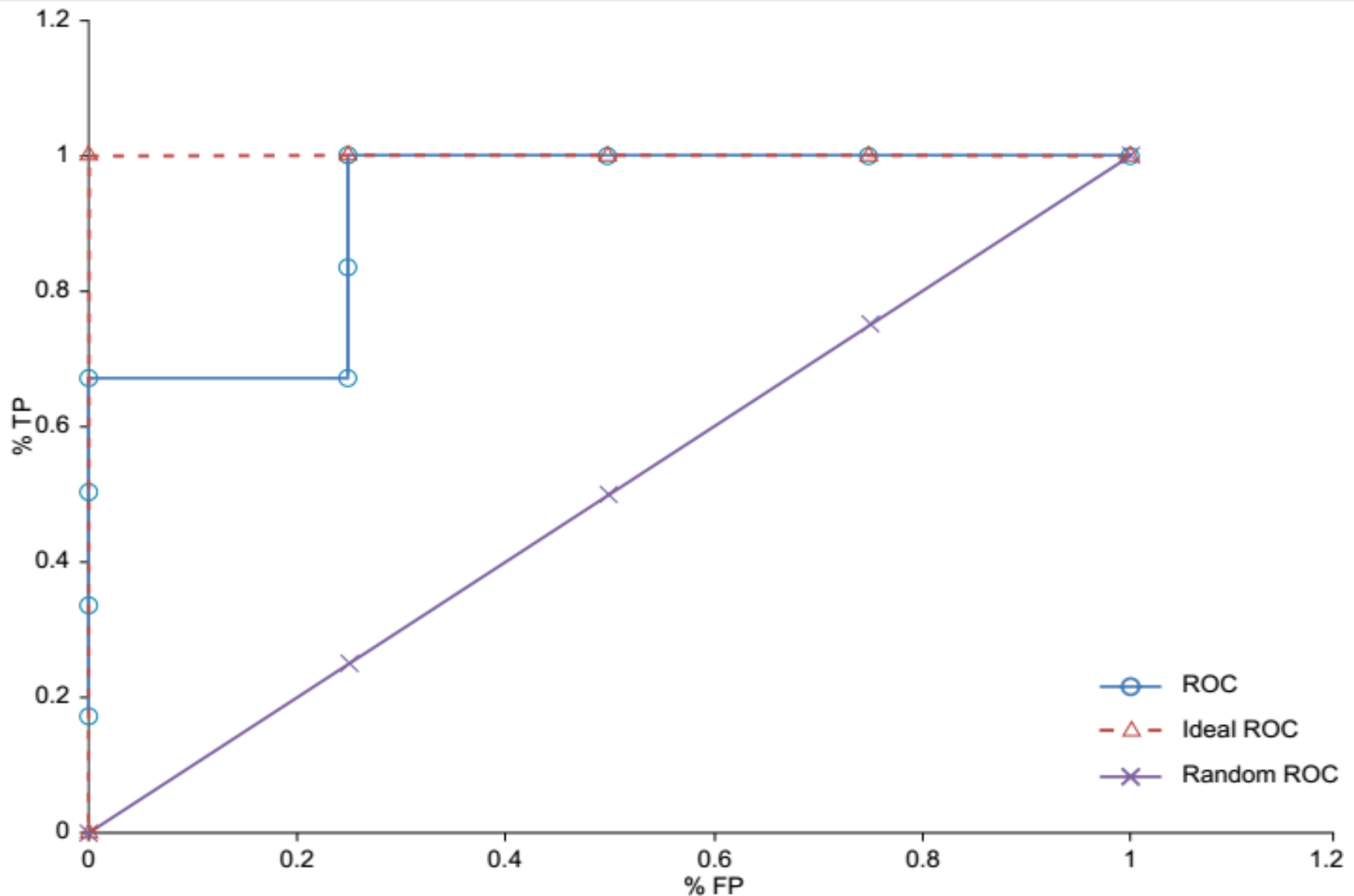


- La recta (0,0) (1,1) simboliza un modelo con un rendimiento similar a predicciones aleatorias (misma tasa de falsos positivos y verdaderos positivos) - AUROC=0.5
- Curvas por encima indican un mejor rendimiento del modelo

Table 8.3 Classifier Performance Data Needed for Building an ROC Curve

Actual Class	Predicted Class	Confidence of "response"	Type?	Number of TP	Number of FP	Fraction of FP	Fraction of TP
response	response	0.902	TP	1	0	0	0.167
response	response	0.896	TP	2	0	0	0.333
response	response	0.834	TP	3	0	0	0.500
response	response	0.741	TP	4	0	0	0.667
no response	response	0.686	FP	4	1	0.25	0.667
response	response	0.616	TP	5	1	0.25	0.833
response	response	0.609	TP	6	1	0.25	1
no response	response	0.576	FP	6	2	0.5	1
no response	response	0.542	FP	6	3	0.75	1
no response	response	0.530	FP	6	4	1	1
no response	no response	0.440	TN	6	4	1	1
no response	no response	0.428	TN	6	4	1	1
no response	no response	0.393	TN	6	4	1	1
no response	no response	0.313	TN	6	4	1	1
no response	no response	0.298	TN	6	4	1	1
no response	no response	0.260	TN	6	4	1	1
no response	no response	0.248	TN	6	4	1	1
no response	no response	0.247	TN	6	4	1	1
no response	no response	0.241	TN	6	4	1	1
no response	no response	0.116	TN	6	4	1	1

ROC y AUC



Preguntas...

