CRITERIOS DE SELECCIÓN DE MODELOS

Bases de Datos Masivas

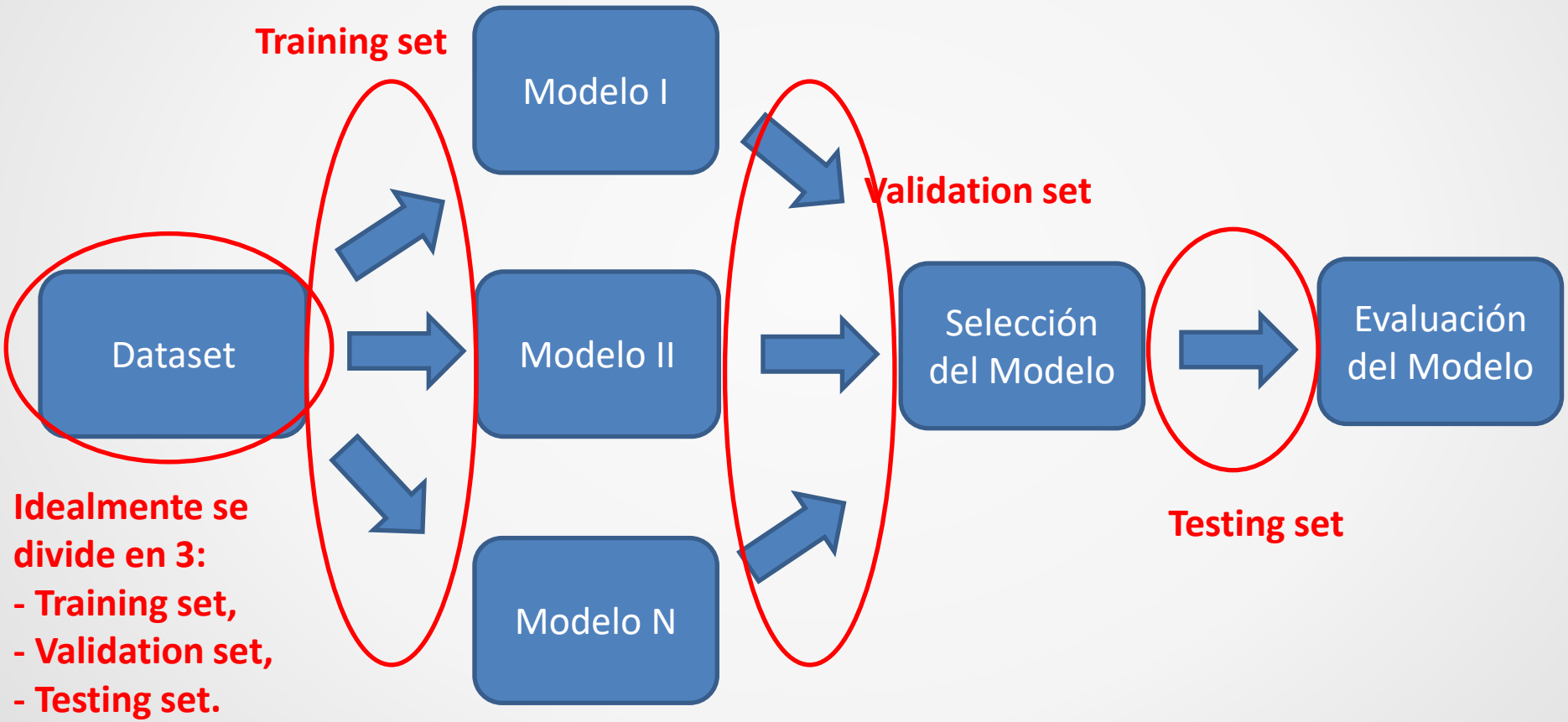
Selección de modelos

En estadística y *machine learning*, la "**selección de modelos**" es el problema de escoger entre diferentes modelos matemáticos que pretenden describir el mismo conjunto de datos.

Tenemos que seleccionar la “mejor” combinación de parámetros para nuestro algoritmo de aprendizaje.

El objetivo es optimizar una **medida de desempeño** del algoritmo en un conjunto de datos independiente.

Proceso de Selección de modelos



División del dataset

Si contamos con muchos datos, el mejor enfoque para ambos problemas es dividir aleatoriamente el conjunto de datos en tres partes:

Ajustar los modelos

Estimar el error de predicción de los modelos

Generalizar el error del modelo elegido

Idealmente, el conjunto de prueba debe mantenerse en una "bóveda", y ser sacado sólo al final del análisis de datos.

En la práctica, el conjunto de datos de validación & prueba suele ser un único set de datos y se elige el modelo que minimice el error. (Proporción: 2/3 y 1/3)

The Elements of Statistical Learning
Data Mining, Inference, and Prediction. 2 Edition. Stanford University

Selección de modelos

Para llevar adelante la selección del modelo, vamos a necesitar:

- Estrategia de evaluación: Es la forma en que dividimos los datos a efectos de llevar a cabo el entrenamiento, la validación y/o el testing del modelo.
- Métrica de evaluación: Las medidas de *performance* nos van a permitir evaluar de manera cuantitativa si uno de los modelos ajustados es mejor que otros.

Métricas de performance/evaluación

Algunas de las métricas de performance mas utilizadas son:

- Accuracy (clasificación),
- Curva ROC (clasificación),
- F-score (clasificación),
- Coeficiente de Kappa (clasificación).
- MSE (regresión) y RMSE (regresión),
- Criterio Akaike (regresión).

Accuracy

Accuracy de un clasificador M, **acc(M)**: Es el porcentaje de tuplas del conjunto de prueba que fueron correctamente clasificadas por el modelo M

- Error rate (tasa de mal clasificados) es: **M = 1 – acc(M)**

- Dadas m clases, $\boldsymbol{CM_{i,j}}$, una entrada en la **matriz de confusión**, indica el # de tuplas en la clase *i* que son etiquetadas por el clasificador como clase *j*

		Predicted class	
		C_1	C_2
Actual class	C_1	true positives	false negatives
	C_2	false positives	true negatives

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques. Elsevier. Cap. 6.12 en adelante.

Alternativas *Accuracy*

Existen alternativas o derivaciones de *Accuracy*

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precisión = \frac{TruePOS}{(TruePOS + FalsePOS)}$$

Accuracy en términos de Sensibilidad y Especificidad

$$Sensibilidad = \frac{TruePOS}{(TruePOS + FalsePOS)} \qquad Especificidad = \frac{TrueNEG}{(TrueNEG + FalseNEG)}$$

True POS, True NEG, False POS y False NEG pueden ser usados en un análisis de **Costo-Beneficio**

Curva ROC

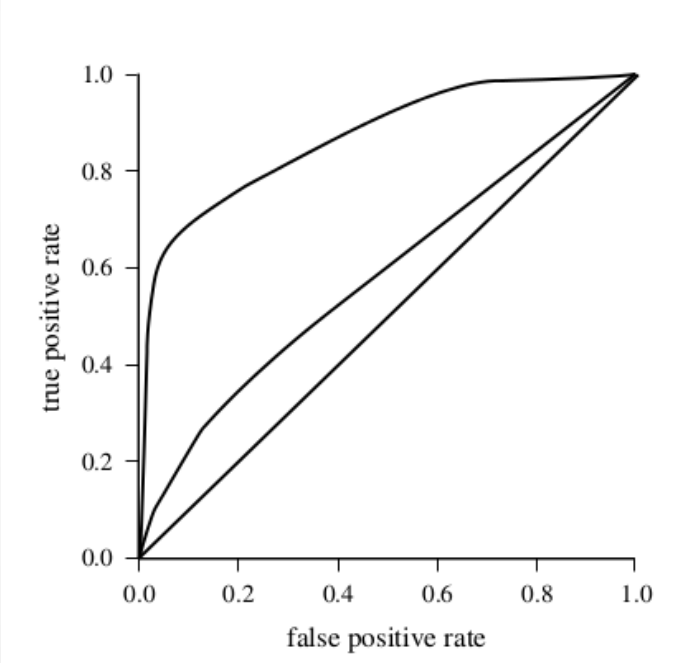
Curva ROC (*Receiver Operating Characteristics*): es una herramienta visual para comparar modelos que ajustan clases binarias.

Se originó a partir de la teoría de detección de señales

Muestra el ***trade-off*** entre la tasa de verdaderos positivos (**VP**) y la tasa de falsos positivos (**FP**).

El **área bajo la curva ROC** (**AUC**) es una medida de la precisión del modelo.

Cuanto más cerca de la línea diagonal, **menos preciso es el modelo.**



http://www.bioestadistica.uma.es/analisis/roc1/

F-Score

Es una medida de precisión muy utilizada en Information Retrieval & clasificación de documentos.

Se define a partir de la matriz de confusión como:

$$F_\beta = (1 + \beta^2) \frac{Precisión * Exhaustividad}{(\beta^2 * Precisión) + Exhaustividad}$$

Permite ponderar los conceptos de precisión y exhaustividad presentes en la métrica a través de β.

Con β = 1, tienen la misma ponderación.

Coeficiente Kappa

Resulta interesante dado que cuantifica el azar con respecto a la coincidencia observada (predicción & clase real), el cual simboliza como Pr(e).

El coeficiente es:

$$K = \frac{Accuracy \ - \ \Pr(e)}{1 \ - \Pr(e)}$$

$$Pr(SI\ al\ azar) = (PosPredicted/Total) * (PosClase/Total)$$

Donde Pr(e):

$$Pr(NO\ al\ azar) = (NegPredicted/Total) * (NegClase/Total)$$

$$Pr(e) = \Pr(Si\ al\ azar) * Pr(No\ al\ azar)$$

Por lo tanto, Pr(e) cuantifica la probabilidad de la coincidencia por azar.

		Predicted class	
		C_1	C_2
Actual class	C_1	true positives	false negatives
	C_2	false positives	true negatives

MSE y RMSE

Medir la precisión de la predicción, es medir qué tan lejos el valor pronosticado esta del valor real conocido.

Las Funciones de Perdida: Miden el error entre $\mathbf{y_i}$ y el pronosticado $\mathbf{y_i'}$

Absolute error: $| y_i – y_i' |$

Squared error: $(y_i – y_i')^2$

Test error (generalization error): El promedio de perdida sobre el conjunto de testing.

Mean absolute error: $\frac{\sum_{i=1}^{d} |y_i - y_i'|}{d}$

Mean Squared Error (MSE): $\frac{\sum_{i=1}^{d} (y_i - y_i')^2}{d}$

Relative absolute error: $\frac{\sum_{i=1}^{d} |y_i - y_i'|}{\sum_{i=1}^{d} |y_i - \bar{y}|}$

Relative squared error: $\frac{\sum_{i=1}^{d} (y_i - y_i')^2}{\sum_{i=1}^{d} (y_i - \bar{y})^2}$

Generalmente se utiliza $RMSE = \sqrt{\frac{\sum \left(X_{obs,i} - X_{model,i} \right)^2}{n}}$

Criterio de información de Akaike

Es una métrica para comparar modelos en regresión.

Es interesante porque penaliza la cantidad de variables predictoras.

El criterio de información de Akaike esta dado por:

$$AIC = \mathrm{n} \times \log\left(\frac{SS_R}{n}\right) + 2\mathrm{k}$$

donde n es el número de casos en el modelo, SS_R es la suma de cuadrados de los residuos del modelo y k es el número de variables predictoras.

$$SS\mathrm{R} = \sum_{i=1}^{n} \left(y - Prediccion(X)\right)^2$$

Estrategias de evaluación

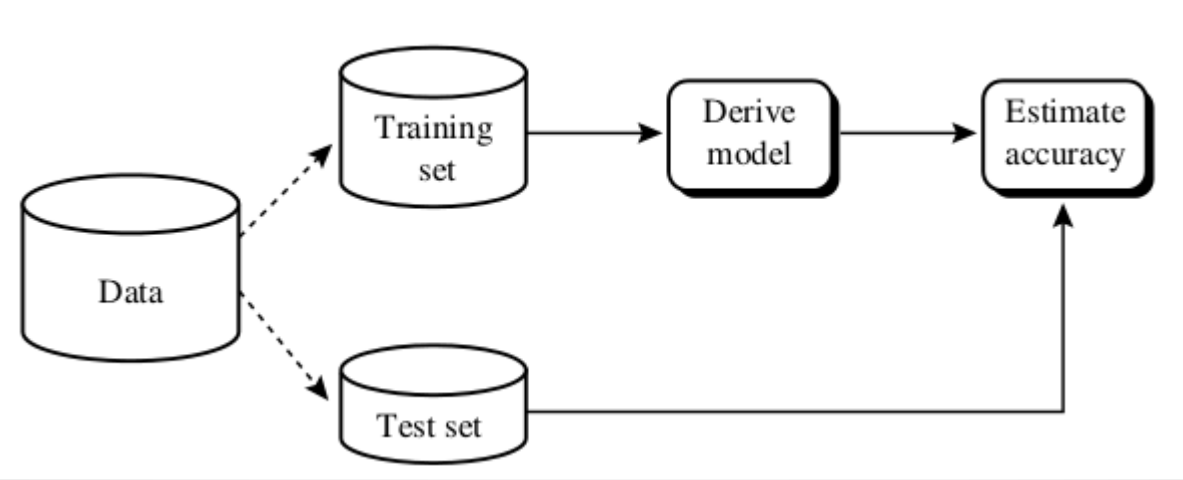
Algunas de las estrategias mas utilizadas son:

- Holdout,
- Subsampling,
- Cross-validation:
  - Leave-one-out (uno afuera),
  - K-fold (k carpetas).

Holdout method

Es el típico caso donde separamos al azar en *training* y *testing* sin solapamientos.

- Training set (e.g., 2/3) for model construction
- Test set (e.g., 1/3) for accuracy estimation



El problema de los holdout es que las **medidas de evaluación** van a tener una **varianza grande**.

Subsampling

Esta técnica es una variación de ***Holdout*** donde se repite la división aleatoria en dos grupos una k cantidad de veces.

El **Accuracy** del modelo es el promedio del valor obtenido para las métricas de evaluación en cada K.

Cross-validation

Consiste en particionar al azar el conjunto de entrenamiento en K subconjuntos mutuamente excluyentes y cada unos aproximadamente de igual tamaño.

Las versiones más conocidas son:

- K-Fold CV
- Leave-One-Out CV

K-Fold

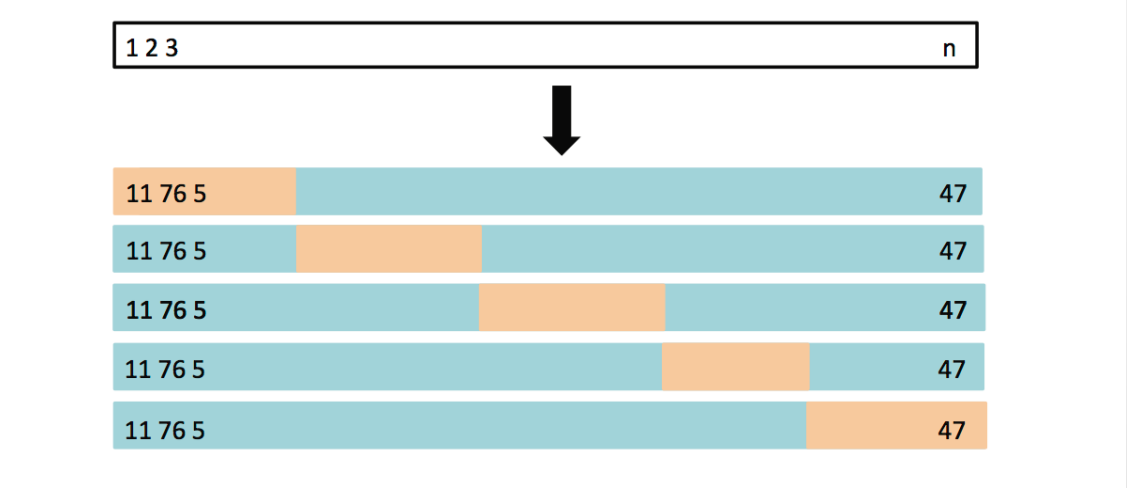
Con K-Fold, Dividimos el dataset en K diferentes partes de igual tamaño.

K = 5 ó K = 10 ← Son los K mágicos

Entonces, quitamos la primera parte, ajustamos el modelo en el resto de **K - 1 partes**, y evaluamos con la parte que dejamos afuera. Se repite *k veces.*

Rendimiento = Promediando los K.

Está demostrado que se obtienen estimaciones que no sufren de sesgo excesivamente alto ni varianza alta.



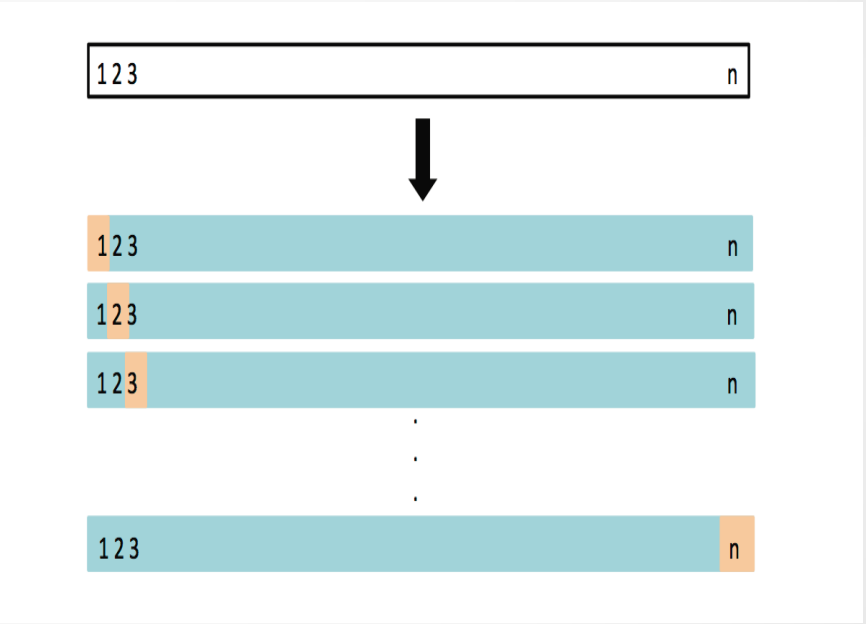
La diferencia con holdout es que todas las instancias participan las k veces.

Leave-One-Out

Es un caso especial de K-Fold donde k es igual a n.

Para cada modelo:

- Split the data set of size n into Training data set (blue) size: **n -1,**
- Validation data set (beige) size: **1,**
- Repeat this process n times.



Ventajas: Tiene menos sesgo y reduce la varianza. El modelo ve todos los casos durante el ajuste.

Desventaja: Es costoso computacionalmente.

Complejidad vs Error

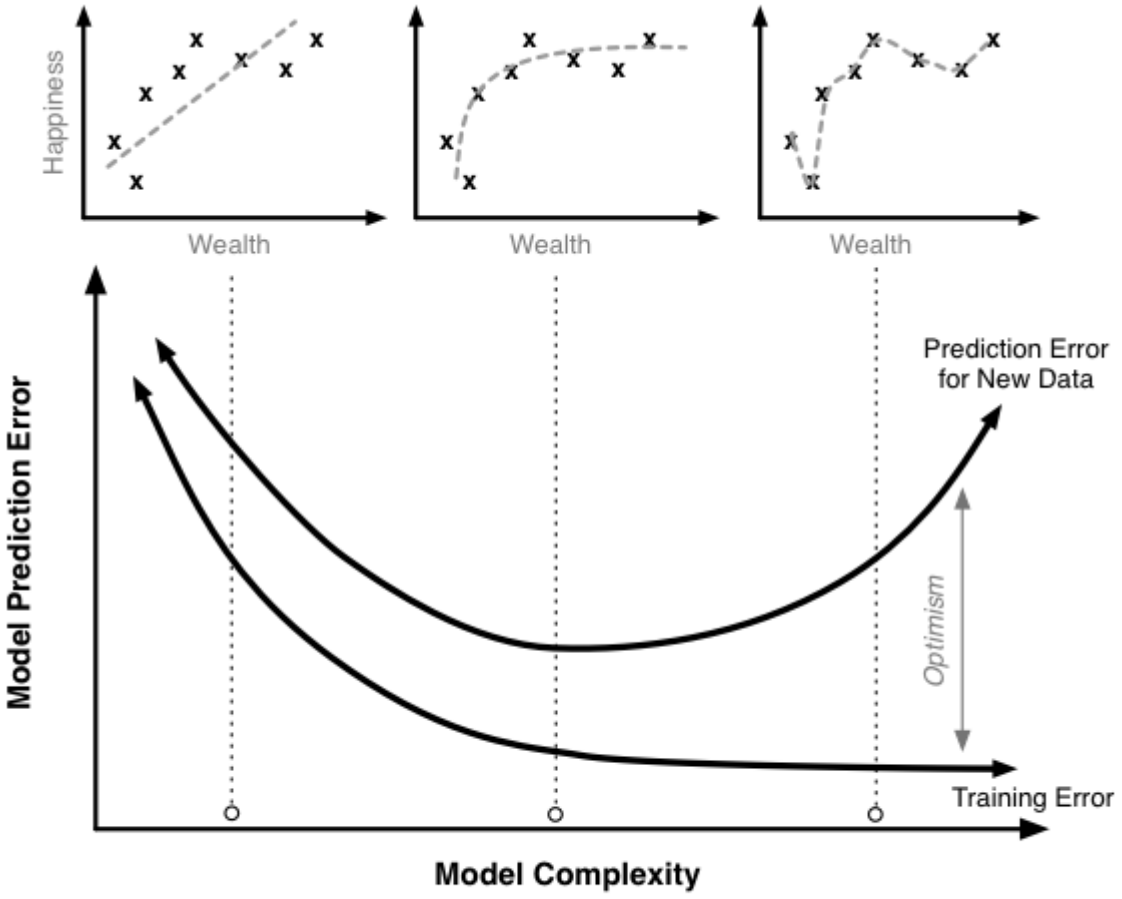
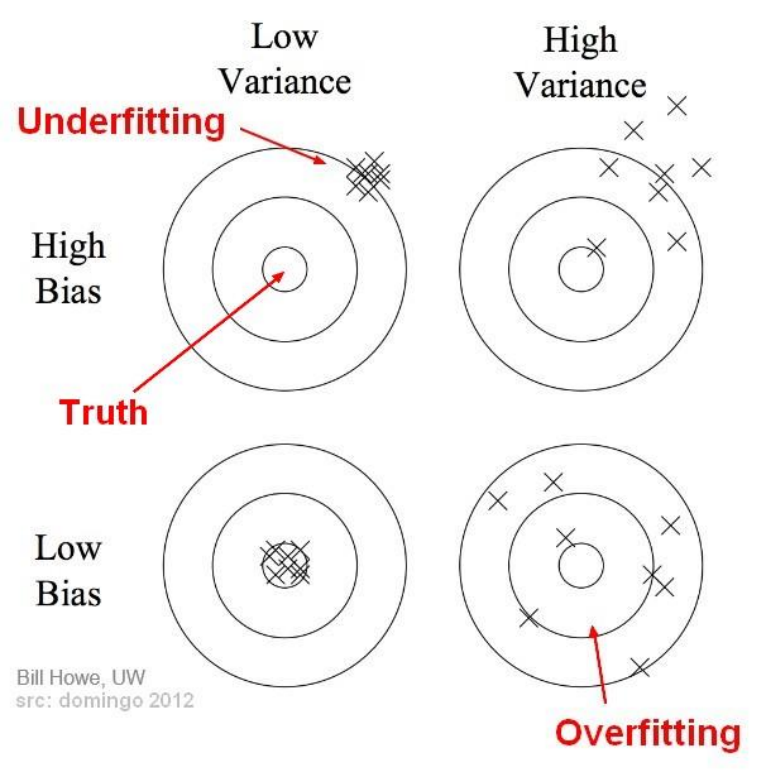


Fig. 1 Training, optimism and true prediction error.

Conceptos de Sesgo y Varianza



¿Qué funciones utilizar?

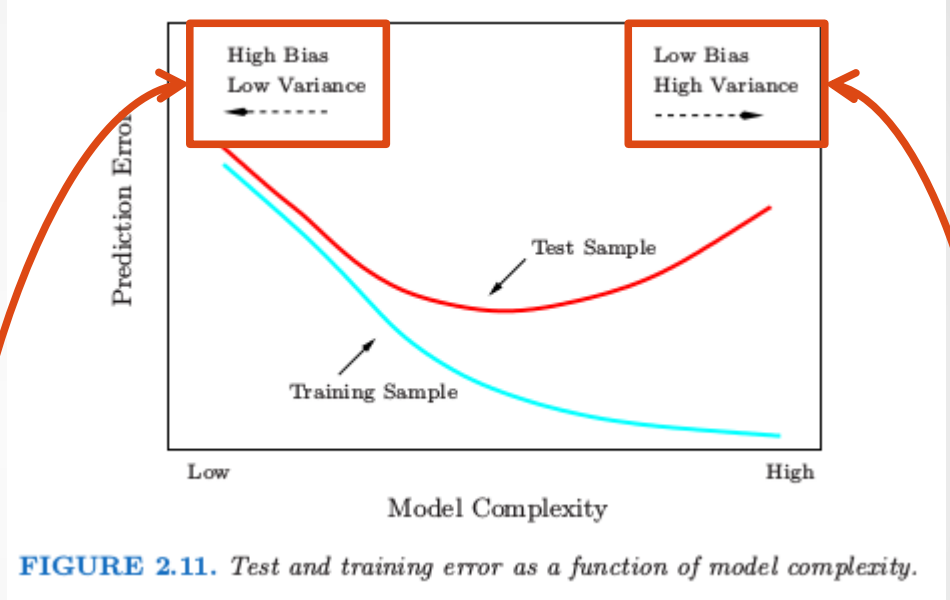


FIGURE 2.11. *Test and training error as a function of model complexity.*

Funciones rígidas:

Buena estimación de los parámetros óptimos

Poca flexibilidad

Funciones flexibles:
Buen ajuste
Mala estimación de los parámetros óptimos