ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

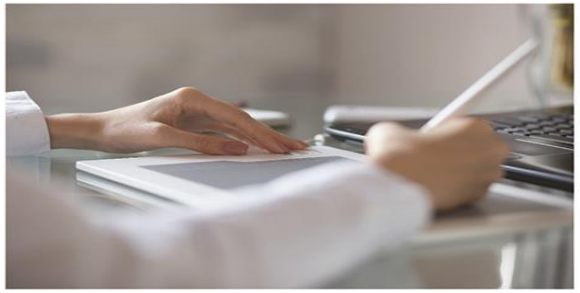
Diplomado en Big Data y Ciencias de Datos

Minería de Datos

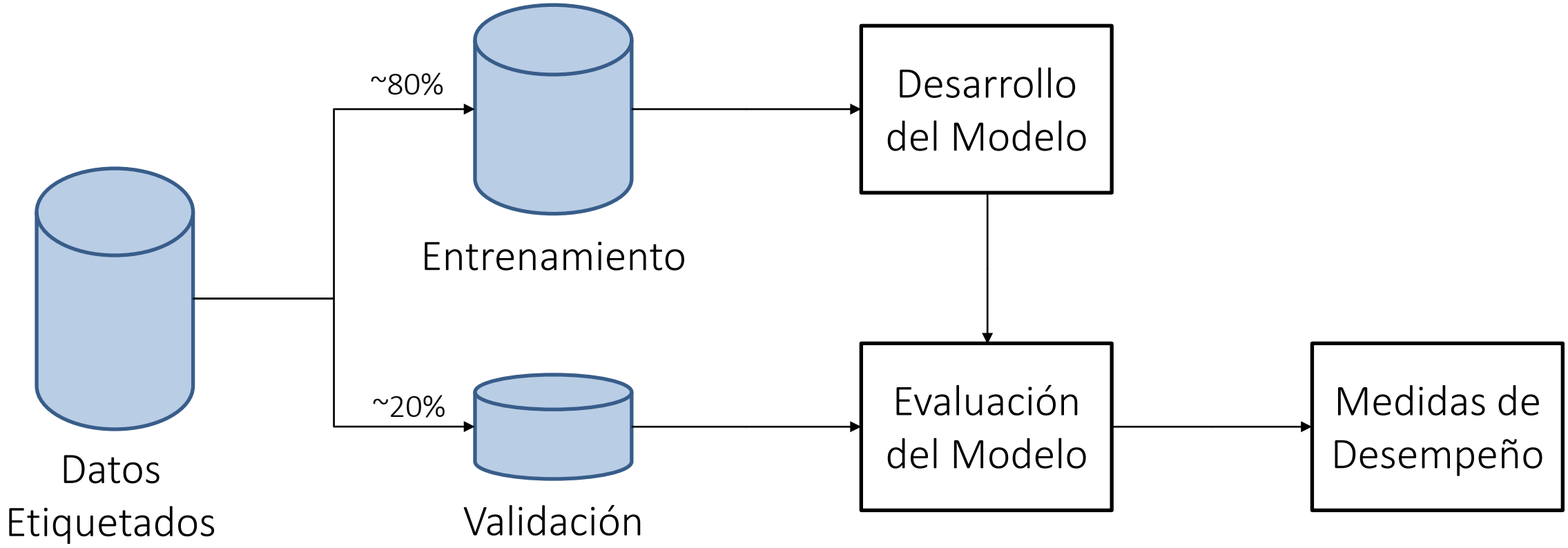
Selección de Modelos

Educación Profesional - Escuela de Ingeniería UC

Sebastián Raveau




Idea general



Tres formas de dividir los datos

Hold-out sample

Random sub-sampling

K-fold cross-validation

Hold-out sample

Es el método más sencillo, simplemente particionamos los datos en dos conjuntos:

uno de entrenamiento y otro de validación

La partición la decide el modelador, según la cantidad de datos disponibles

Ejemplos:	relación 2 a 1	67 % vs 33 %
	relación 3 a 1	75 % vs 25 %
	relación 4 a 1	80 % vs 20 %

Hold-out sample

La partición debe ser completamente al azar, de modo que las distribuciones de categorías y variables sean iguales en ambos conjuntos

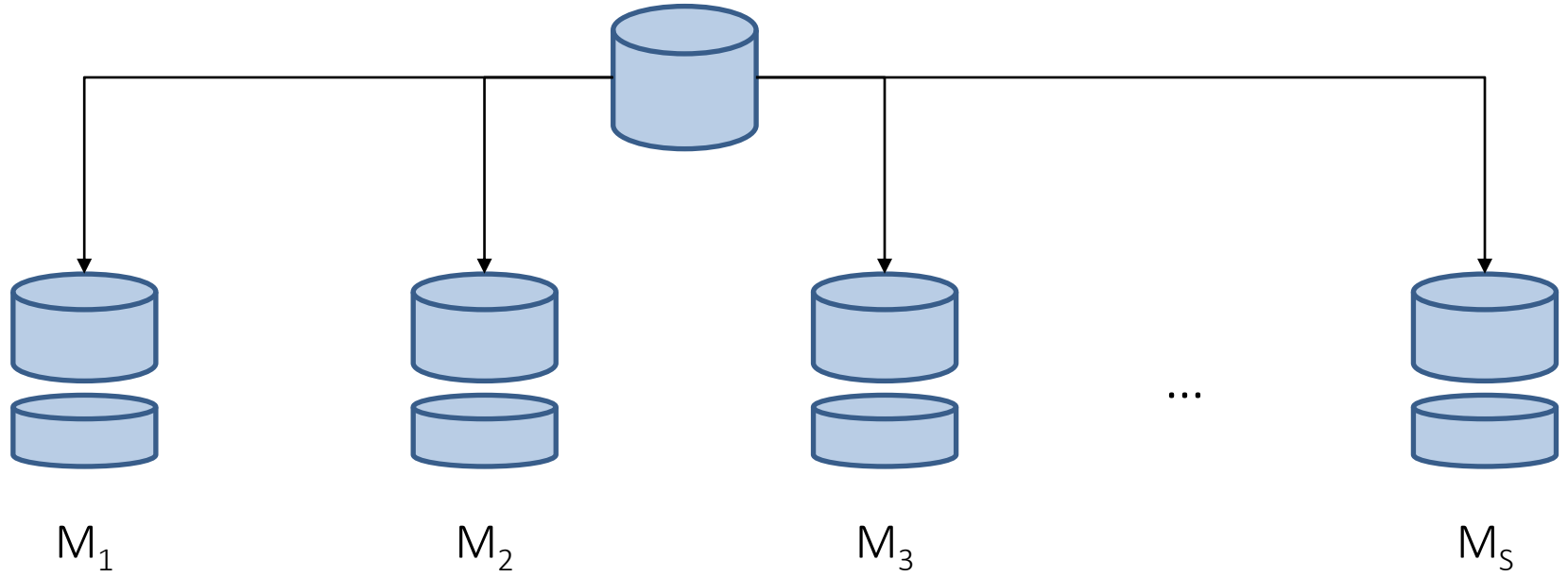
Si la partición está sesgada, el modelo entrenado estará sesgado

Si la partición está sesgada, la validación no será informativa

Random sub-sampling

Repetimos el método de hold out varias veces

Disminuye la probabilidad de tener una partición sesgada



Random sub-sampling

Obtenemos *S* modelos, lo que dificulta la interpretación

Cada modelo tiene cierto desempeño, los cuales se pueden promediar

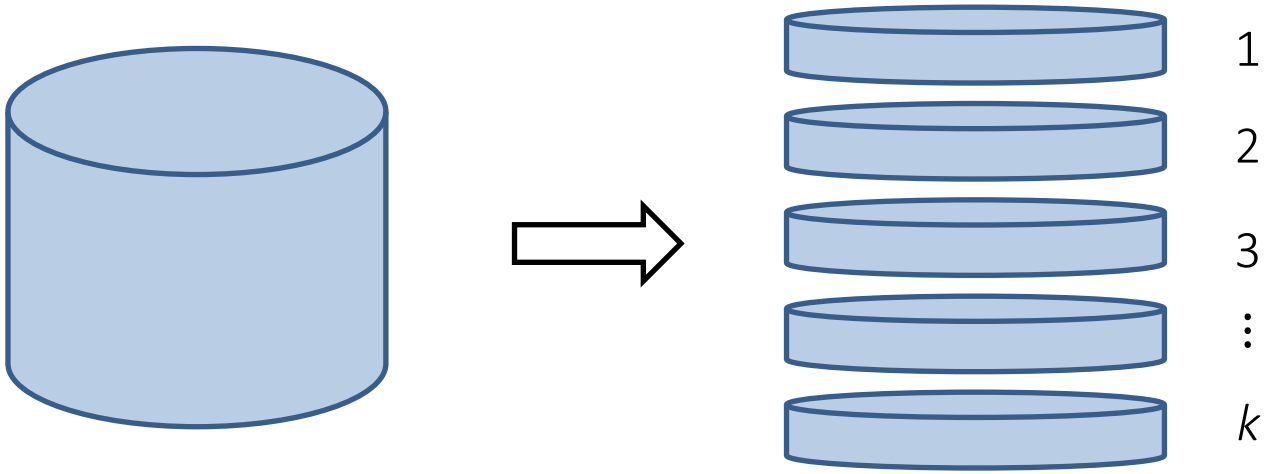
Cada modelo tiene cierta predicción, a partir de las cuales

- se promedia, para variables continuas
- se elige la mayoritaria, para variables discretas

K-fold cross-validation

Es, en cierta medida, una versión mejorada de random sub-sampling

La idea es asegurar que todos los datos han sido parte de la base de datos de entrenamiento y también de la base de datos de validación



K-fold cross-validation

Particionamos los datos en *k* sub-conjuntos

El proceso es iterativo:

en cada una de las *k* iteraciones elegimos un subconjunto como base de validación y entrenamos el modelo con los *k*-1 subconjuntos restantes

La predicción y evaluación de desempeño son análogos a random sub-sampling

K-fold cross-validation

La segmentación en *k* sub-conjuntos es aleatoria, por lo que podemos repetir el procedimiento varias veces

Un caso particular es cuando *k* es igual a la cantidad de observaciones, es decir, en cada modelo dejamos una observación afuera (LOOCV = leave one out cross-validation)