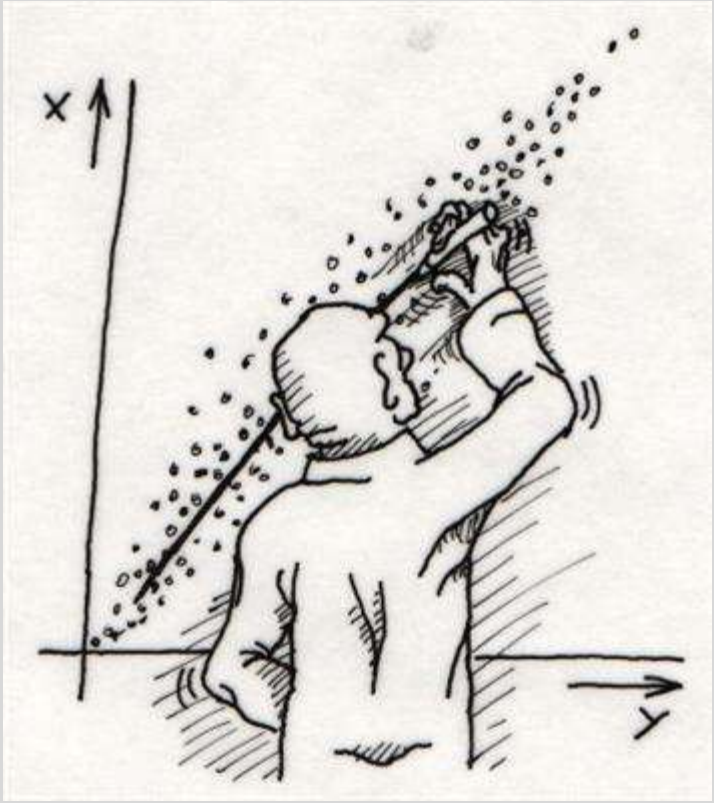


DigitalHouse > Coding School

DATA SCIENCE

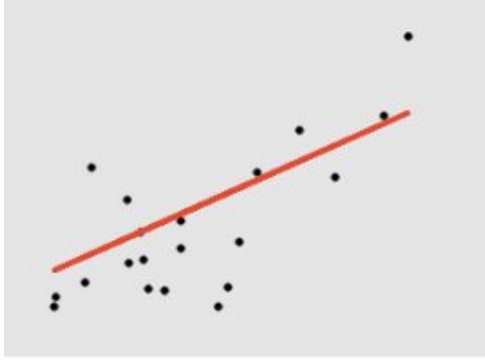
MÓDULO 3

Split Train Test Workflow

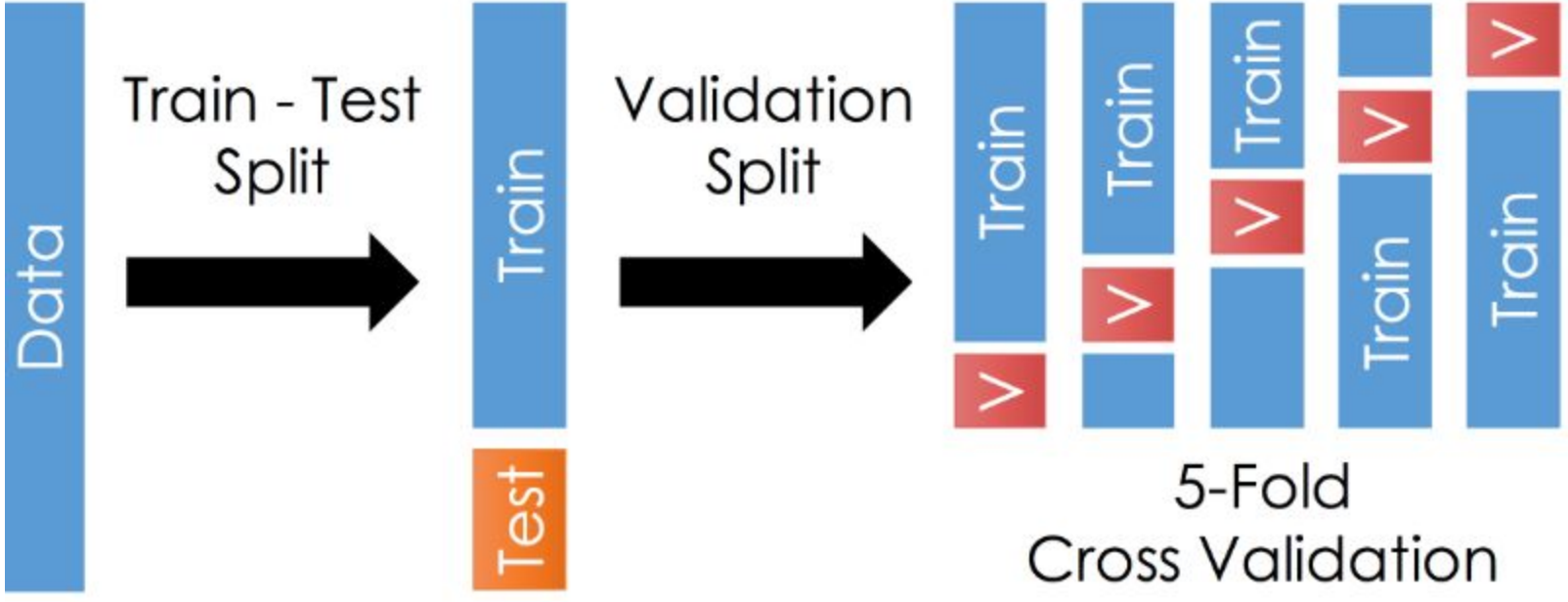
Split Train Test Workflow

OBJETIVOS DE LA CLASE

1. **Presentar las diferentes etapas de un proceso sencillo de entrenamiento**
2. **Comprender la lógica general del proceso de entrenamiento**
3. **Comprender las diferencias en el uso del Split Train Test y Cross Validation en este proceso**
4. **Evaluar tres modelos candidatos sobre un mismo dataset**



WorkFlow de Entrenamiento


Data
Train - Test Split
Train
Test
Validation Split
Train
V
Train
V
Train
V
Train
V
Train
5-Fold Cross Validation

- Tenemos tres modelos candidatos (esto es generalizable sin demasiados problemas)
  - Regresión lineal
  - Regresión lineal regularizada con Ridge
  - Regresión lineal regularizada con LASSO

- Queremos elegir el mejor modelo. El que mejor performa. ¿En dónde?...

- Entonces lo primero que hacemos es dividir el dataset en train/validación y test.

- Luego, tenemos que estimar los hiper parámetros: en nuestro caso los alpha de Ridge y LASSO
  - Usamos Validación Cruzada dentro de Training Set

- Finalmente, estimamos las versiones finales de los modelos candidatos sobre TODO el Training Set

- Nos interesa conocer cómo funciona el modelo “en general”, no en el único dataset que tenemos o conocemos.
- Buscamos modelos que generen buenas predicciones sobre datos “nuevos”.
- En algunos casos, tendremos acceso a datos “nuevos” (por ejemplo, clientes nuevos). Pero en otros, no tendremos acceso inmediato.

Práctica Guiada

WorkFlow Completo

Conclusión

- Entrenamos modelos diferentes.
- Hacemos una primer partición en train/validación y test.
- Entrenamos sobre train/validación y elegimos el modelo de mejor performance.
- Sobre el conjunto de test calculamos la performance del mejor modelo,entrenado con todo el conjunto de train/validación, y no lo volvemos a modificar!
- Si la performance sobre el conjunto de test no es aceptable tire el modelo y vuelva a empezar pero no optimize nuevamente en base a performance sobre test.