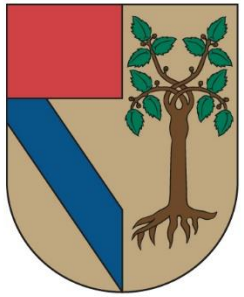


UNIVERSIDAD
PANAMERICANA

Machine Learning

(https://leonpalafox.github.io/mlclase/)

Leon F. Palafox PhD

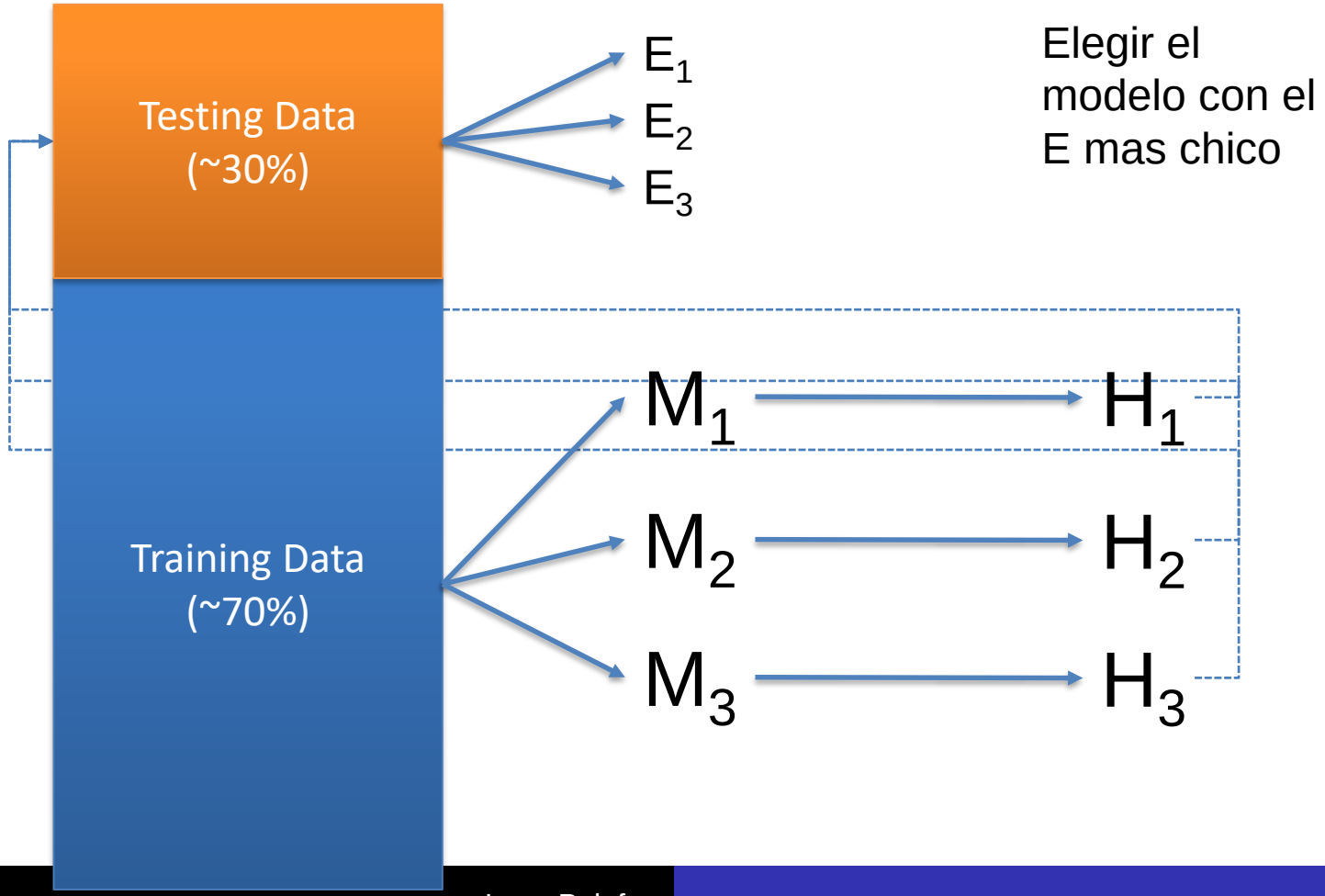
Métodos de Validación

- Cross validation (validación cruzada)
  - Probar diferentes modelos
  - Obtener estadísticas confiables
- Bias -- Variance Analysis
  - Regularización
  - Overfitting

Validación cruzada

- Buscamos errores chicos de entrenamiento?.
  - Por que?
  - Necesitamos probar en prueba (no en entrenamiento)
- La primera tecnica se llama Hold-Out Cross validation

Hold-out cross validation



Que es M?

- Todo lo que hast aahora hemos asignado arbitrariamente.
- Linear Regression
  - Orden del polinomio, parametro de regularización
- SVM
  - Kernel, variables del kernel

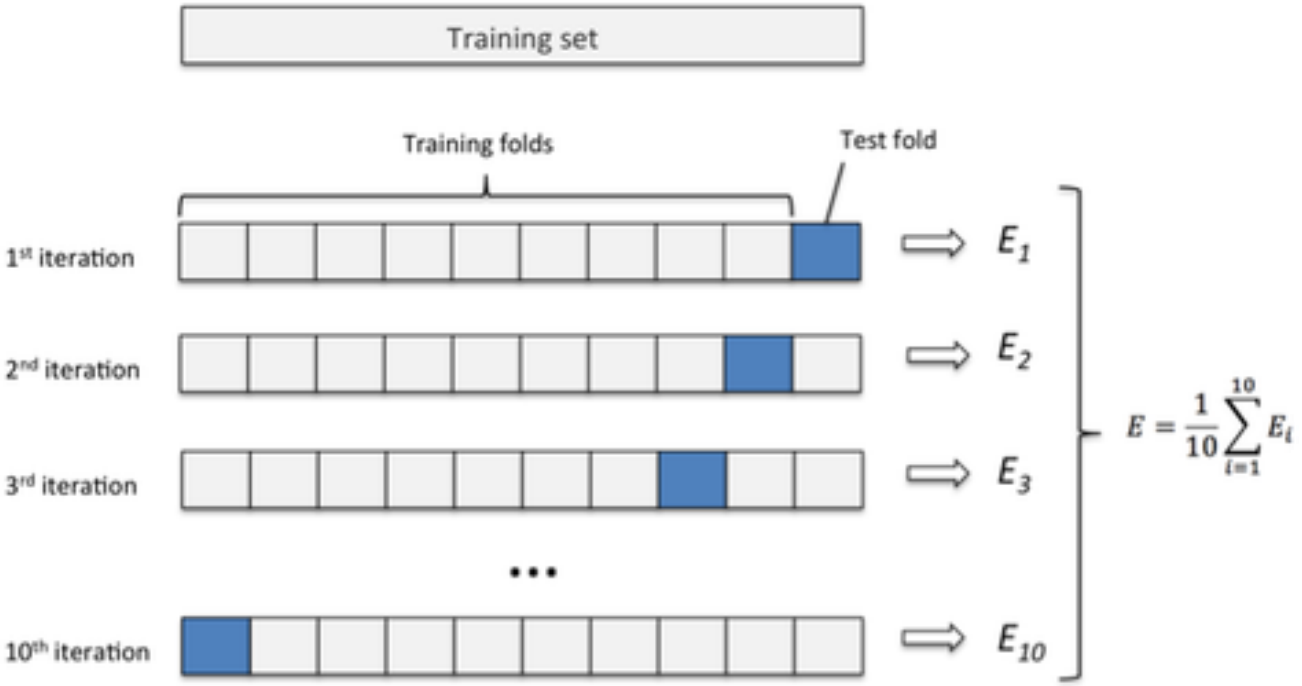
Problemas con Hold-out CV

- Estamos “desperdiciando” datos
- Problemas con pocos datos empiezan a ser complicados
- Tengan cuidado con artpiculos que hablan de CV con pocos datos.
  - Y aún más si ni siquiera hablan de CV.

Un CV incluso mejor

- K-fold CV
  - Divide los datos en K conjuntos(disjoint)
  - Para cada j = 1..k
    - Entrenar modelo (M_i) en cada subconjunto, except j
    - Obtener error (E_{ij}) para Modelo i en iteración j
  - Total error para M_i va a ser el promedio de errores (E_{ij})

K-Fold Cross validation

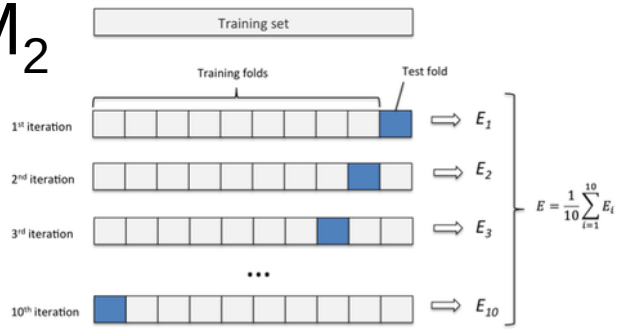


K-Fold Cross validation

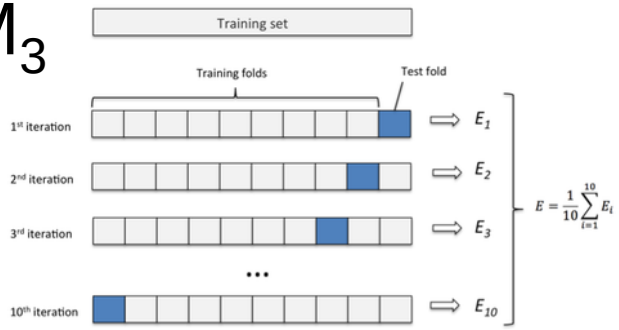
M_1

M_4

M_2



M_3



Revisión de Read team

- Que es un red team?
  - Independientes (revisores sin sesgo)
- Por que necesitamos un red team?.
  - Evitar overfitting en una revista.
  - Nuestro público no somos nostros, es una mayor audiencia.
- Idealmente debería haber red teams para todo.
  - Pláticas, presentaciones.

Movies

- Las taquilleras tartan de generalizar.
  - Audiencias de prueba
  - Actores de alto presupuesto (no necesariamente buenos).
  - “Wide appeal”
- Los ganadores de Oscar (por lo general) hacen overfit a los criticos.

Ventajas

- Podemos corer cada fold en cores distintos
  - Sklearn lo sabe hacer
- A diferencia de otros metodos, es muy facil de implementar
  - Las técnicas bayesianas son particularmente dificiles.

Desventajas

- Si no tenemos muchos datos, los folds van a estar muy correlacionados.
  - Esto hace que se generen overfits
- Toma mucho mas tiempo corer los algoritmos

- El mayor:

Cuantos folds se deben elegir?

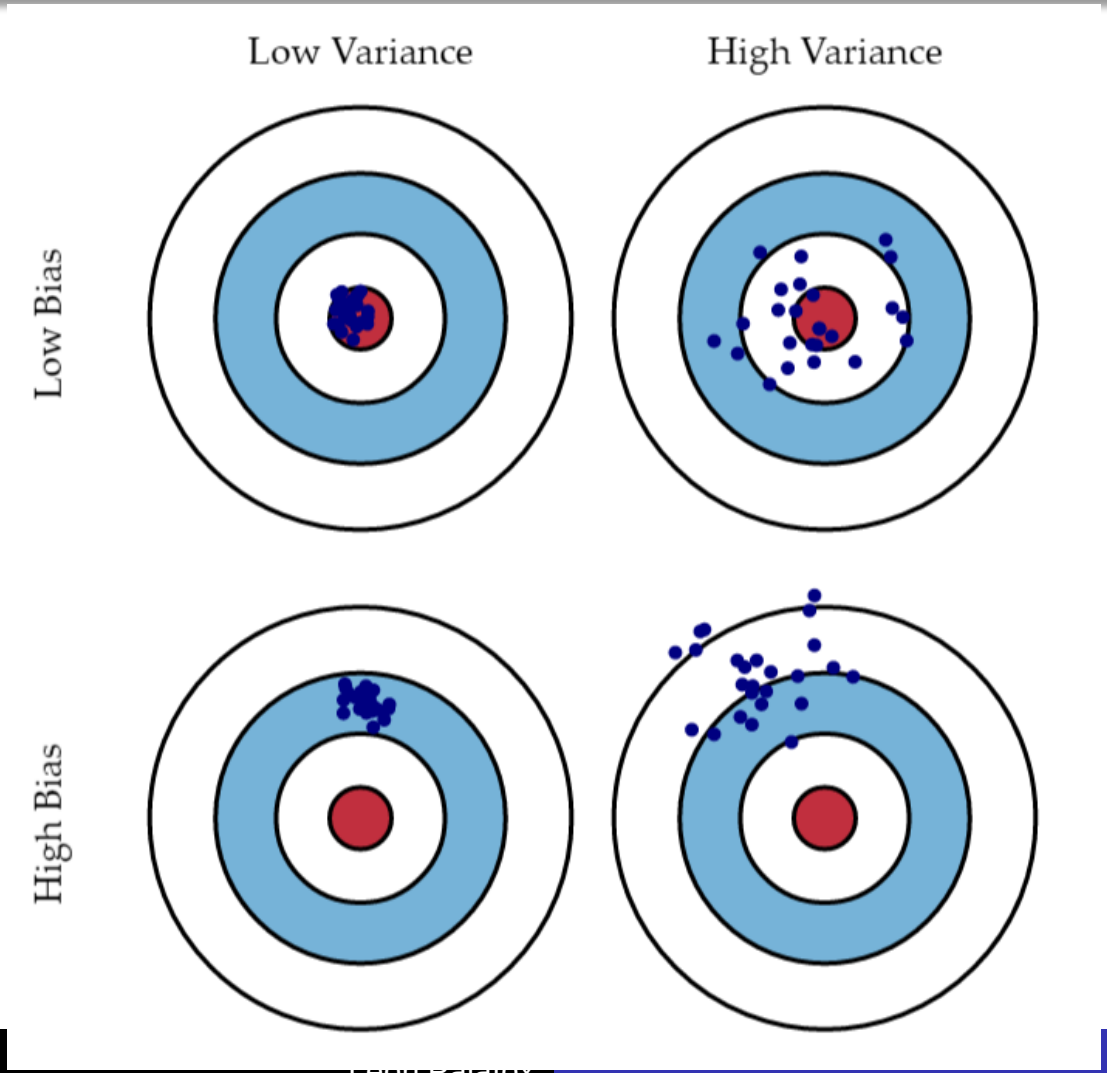
- K = 3 trabaja bien
- Cosas simples como SVMs, puedes usar 10.
- Leave-one out (K=N-1) es un mal chiste de los cientificos.
  - Tenemos tantos folds como datos.
  - Bootsraping glorificado.

Bias–Variance

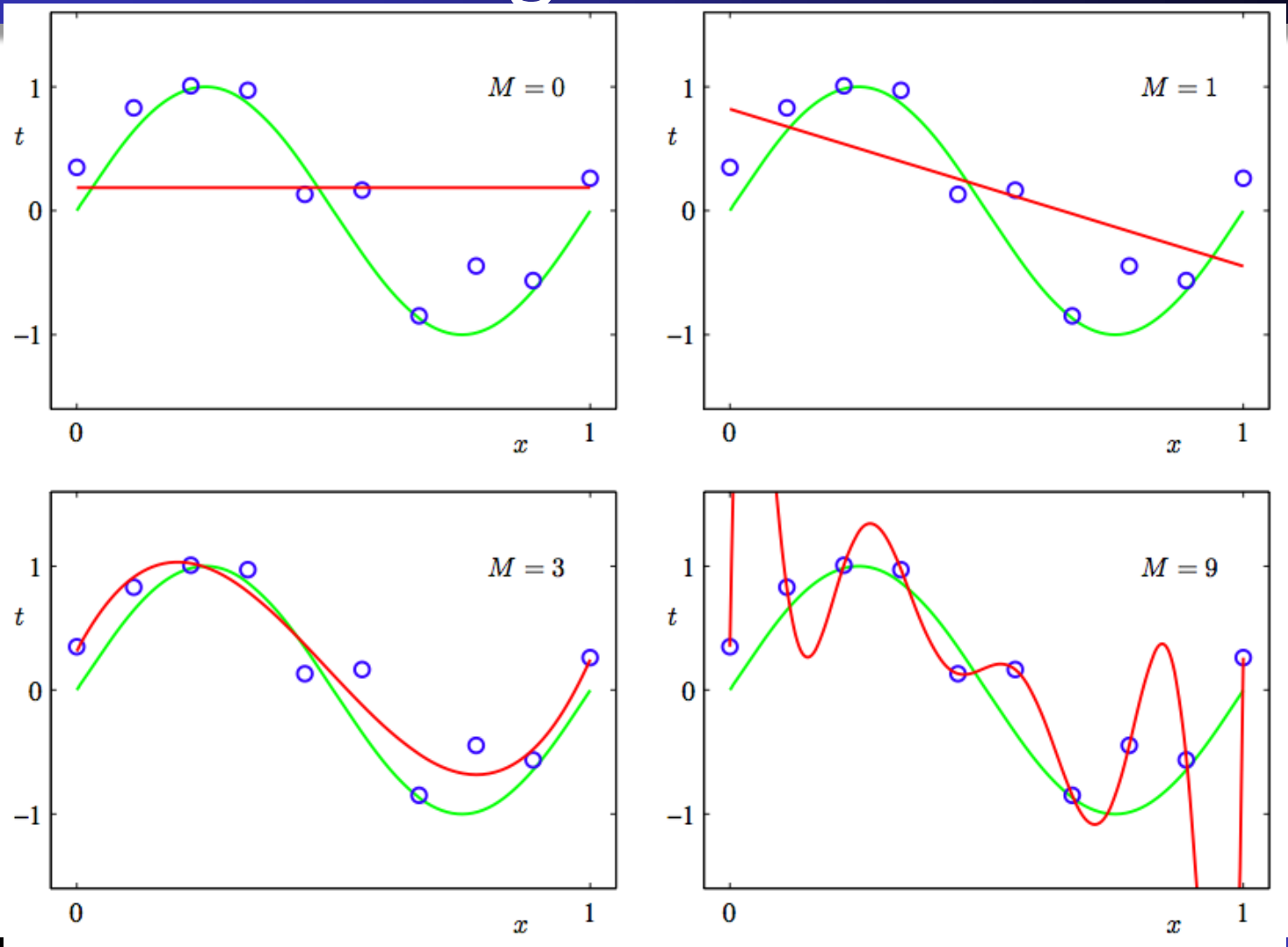
Concepts

- Bias: Que tan lejos estamos del set correcto de parámetros
- Variance: Que tan consistentes son las predicciones
- No se puede ganar siempre

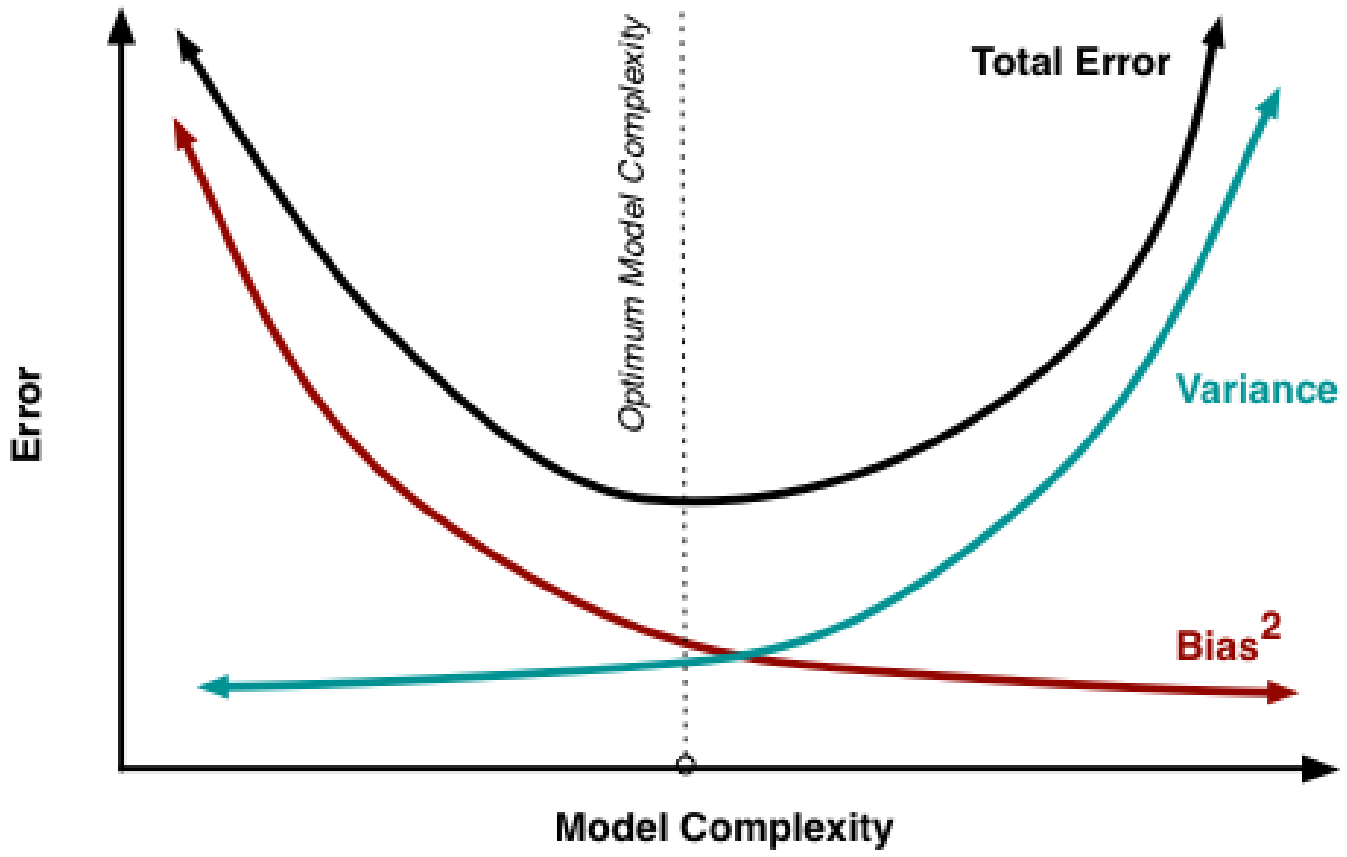
An example



Smallest training error



Bias-Variance



Notes

- Es un buen punto de paro cuando se usa CV
- Puedes usarlo sin CV y aun asi se obtienen Buenos resultados.
- Bias and Variance se definen de maneras distintas para distintos algoritmos, asi que es un poco mas dificil de implementar.

Examen!!!