Machine Learning

Hernán Aguirre

Universidad San Franciso de Quito

Universidad de Shinshu, Japón

Contenido del Curso

1. **Introducción**
2. **Aprendizaje estadístico**
3. **Regresión lineal**
4. **Clasificación**
5. **Métodos de remuestreo**
6. **Selección y regularización de modelos lineales**
7. **Más allá de la linealidad**
8. **Métodos basados en árboles**
9. **Máquinas de vectores de soporte**
10. Aprendizaje profundo
11. Análisis de supervivencia y datos censurados
12. Aprendizaje sin supervisión
13. Pruebas múltiples

5 Métodos de Remuestreo

1. Validación Cruzada
2. Bootstrap (Arranque?)
3. Laboratorio: Validación Cruzada y Bootstrap
4. Ejercicios

Que Implican?

- Extraer repetidamente muestras de un conjunto de entrenamiento y
- Reajustar un modelo de interés en cada muestra para obtener información adicional sobre el modelo ajustado
- Herramientas importantes en la aplicación práctica de muchos procedimientos de aprendizaje estadístico

Para que Sirven?

- Validación cruzada
  - Se usa para estimar el error de prueba asociado con un método de aprendizaje estadístico determinado
    - Evaluar su desempeño (evaluación de modelo) o
    - Seleccionar el nivel apropiado de flexibilidad (selección de modelo)
- Bootstrap
  - Comúnmente se usa para proporcionar una medida de precisión de una estimación de un parámetro o de un método de aprendizaje estadístico determinado

5.1

Validación Cruzada

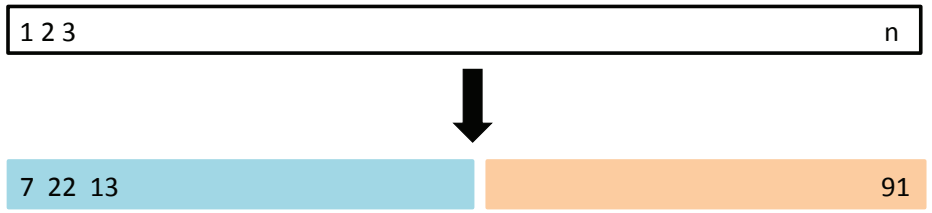
- El Enfoque del Conjunto de Validación
- Validación Cruzada dejar-uno-fuera
- Validación Cruzada k-veces
- Canje sesgo-varianza de la Validación Cruzada k-veces
- Validación Cruzada en Problemas de Clasificación

El Problema

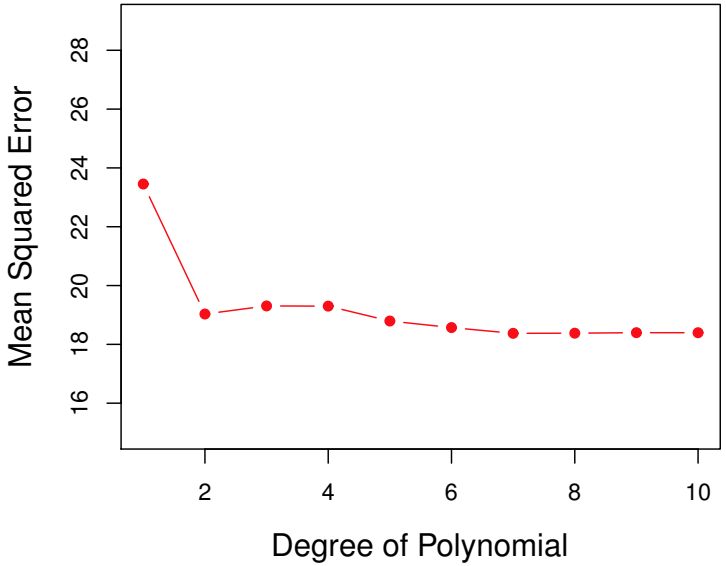
- La tasa de *error de entrenamiento* a menudo es bastante diferente de la tasa de *error de prueba*
- La tasa de *error de entrenamiento* puede subestimar drásticamente la tasa de *error de prueba*

5.1.1

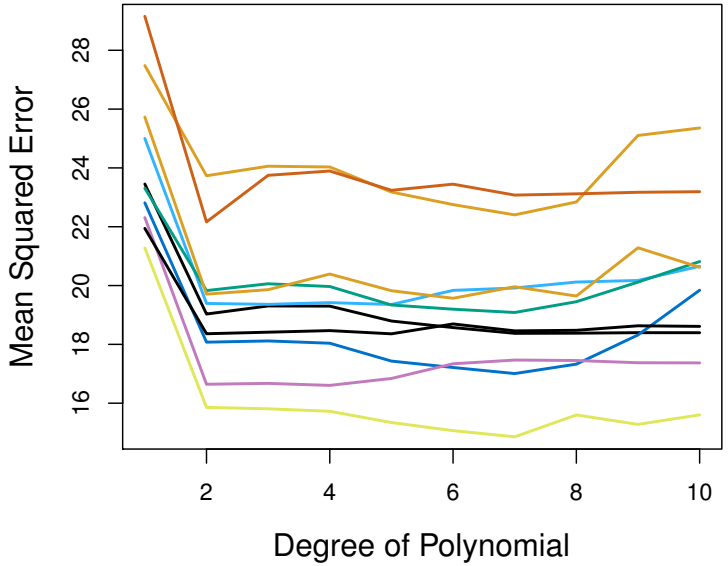
El Enfoque del Conjunto de Validación



dividir aleatoriamente la muestra



El método de validación se aplica una vez, i.e. una sola división de los datos en conjuntos de entrenamiento y validación



El método de validación se repitió diez veces, cada vez con una división aleatoria diferente de los datos en conjuntos de entrenamiento y validación

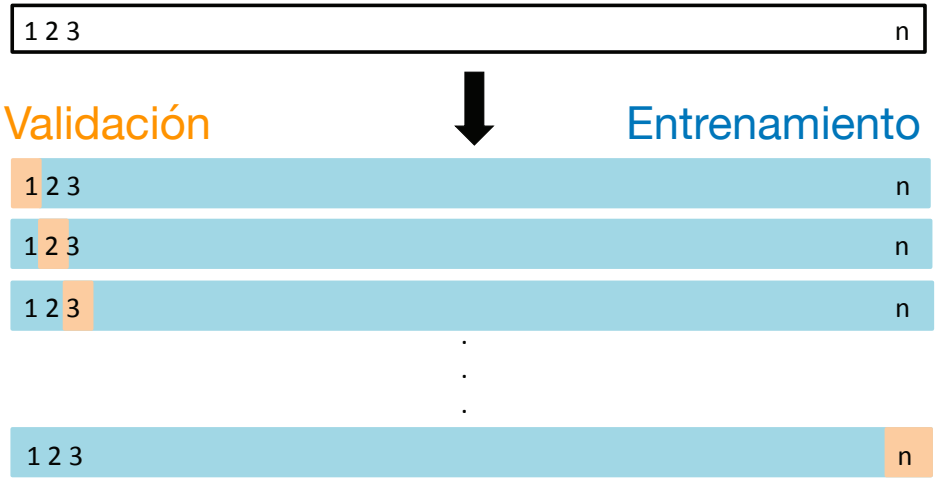
Dos Posibles Inconvenientes

- La estimación de la tasa de error de prueba puede ser muy variable
  - Depende de qué observaciones se incluyen en el conjunto de entrenamiento y qué observaciones en el de validación
- Sólo se utiliza un subconjunto de observaciones para ajustar el modelo
  - Los métodos estadísticos tienden a funcionar peor cuando se entrenan con menos observaciones

    ➡ la tasa de error del conjunto de validación puede tender a sobreestimar la tasa de error de prueba para el ajuste del modelo en todo el conjunto de datos

5.1.2

Validación Cruzada dejar-uno-fuera (LOOCV)



$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- LOOCV es un método muy general y se puede utilizar con cualquier tipo de modelado predictivo
- Por ejemplo, podríamos usarlo con regresión logística o análisis discriminante lineal, o cualquiera de los métodos discutidos en capítulos posteriores

Ventajas de LOOCV

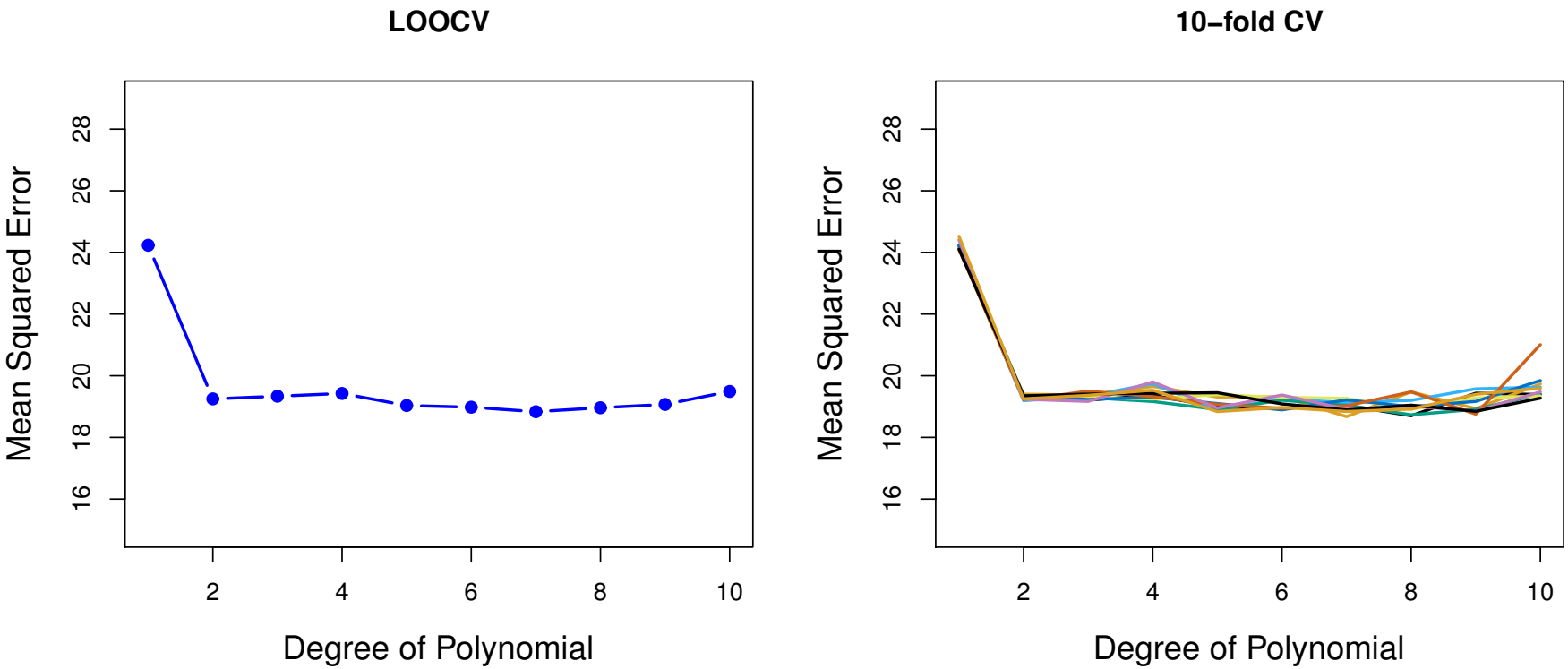
1. Tiene mucho menos sesgo
   - *LOOCV* ajusta repetidamente el método de aprendizaje con $n - 1$ observaciones, casi todo el conjunto de datos
   - *El enfoque del Conjunto de Validación* el ajuste se hace con una fracción de las observaciones, usualmente $n/2$
   - LOOCV sobreestima menos la tasa de error de prueba
2. No hay aleatoriedad
   - *El enfoque de Validación* produce resultados diferentes cuando se aplica repetidamente: aleatoriedad en las divisiones del conjunto de entrenamiento/validación

Desventaja de LOOCV

- LOOCV tiene el potencial de ser costoso de implementar, ya que el modelo debe ajustarse n veces

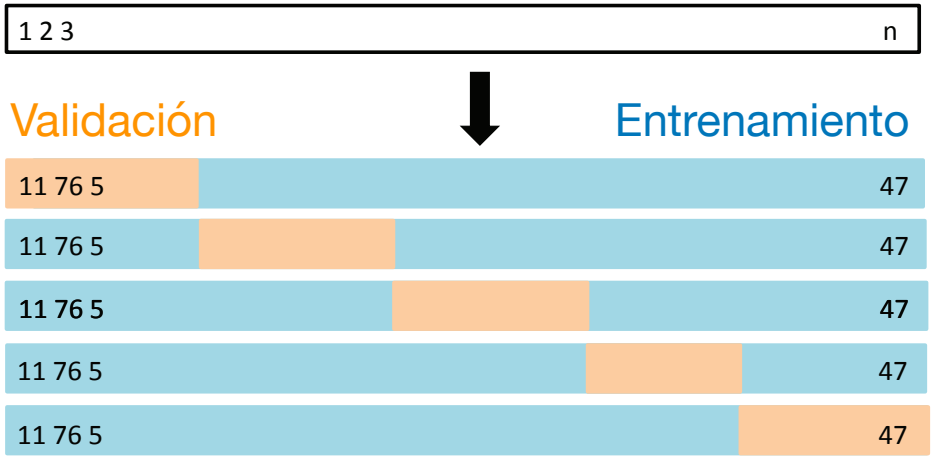
- Esto puede llevar mucho tiempo si n es grande y si cada modelo individual tarda en ajustarse.

Validación Cruzada en Datos de Autos



5.1.3

Validación Cruzada k-veces (k-veces CV)



$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

$$= \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2$$

- Alternativa a LOOCV
- Dividir aleatoriamente las n observaciones en k grupos, de aproximadamente el mismo tamaño n_i $(i = 1, \cdots, k)$
- El primer grupo se trata como un conjunto de validación y el método se ajusta a los $k - 1$ grupos restantes

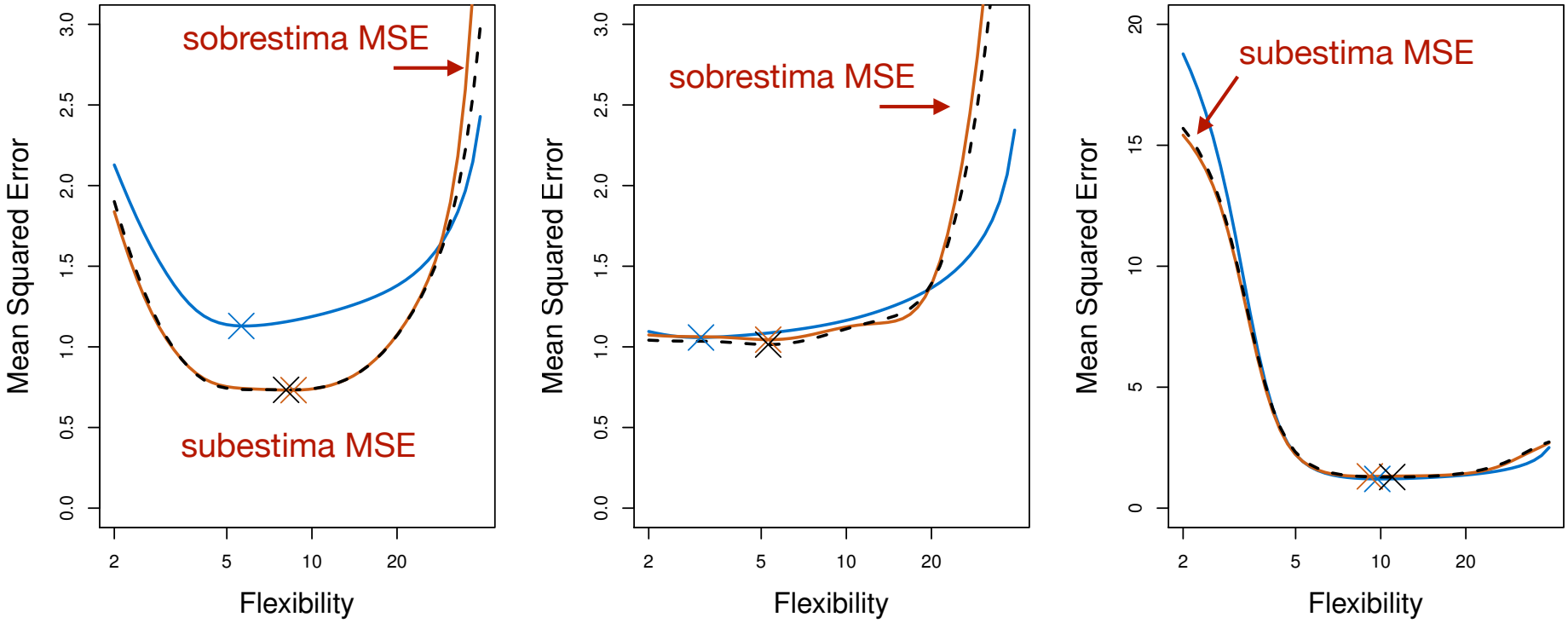
LOOCV y k-veces CV

- *LOOCV* es un caso especial de *k-veces CV*, $k = n$
- En la práctica,
  - Se realiza *k-veces CV* usando $k = 5$ o $k = 10$
- ¿Cuál es la ventaja de usar $k < n$?
  - La ventaja más obvia es computacional
  - Otras ventajas que implican el equilibrio entre sesgo y varianza.

MSE de Prueba Verdadero y Estimados por CV

Datos simulados

Splines suavizadas



- MSE de prueba verdadero y estimado por **LOOCV** y **10-veces CV.** Las cruces indican el mínimo de cada una de las curvas MSE
- Las curvas CV se acercan a identificar el nivel correcto de flexibilidad, correspondiente al MSE de prueba más pequeño

Canje Sesgo-Varianza
k-veces Validación Cruzada

- *Conjunto de Validación* → *sobreestimaciones* del *EP*
  - ∀ *CE* : $n/2$ observaciones
- *LOOCV* → *estimaciones* casi *insesgadas* del *EP*
  - ∀ *CE* : $n-1$ observaciones
- *k-veces CV*, $k = 5$ o 10 → *nivel intermedio de sesgo*
  - ∀ *CE* : $\frac{n}{2} \ll \frac{(k-1)}{k} n < n-1$ observaciones
- Desde la perspectiva de la reducción del sesgo, se debe preferir *LOOCV* a *k-veces CV*

EP : error de prueba
CE : conjunto de entrenamiento

Canje Sesgo-Varianza k-veces Validación Cruzada

- *LOOCV*
  - promedia resultados de *n* modelos ajustados en conjuntos casi idénticos de observaciones → resultados altamente correlacionados (+) entre sí
- *k-veces CV,* $k < n$
  - promedia resultados de *k* modelos ajustados en conjuntos de entrenamiento menos superpuestos → resultados menos correlacionados entre sí
- La media de muchas cantidades altamente correlacionadas tiene una varianza mayor que la media de muchas cantidades que no están tan correlacionadas
- La estimación del EP resultante de LOOCV tiende a tener una varianza mayor que la estimación del EP resultante de k-veces CV

Resumen

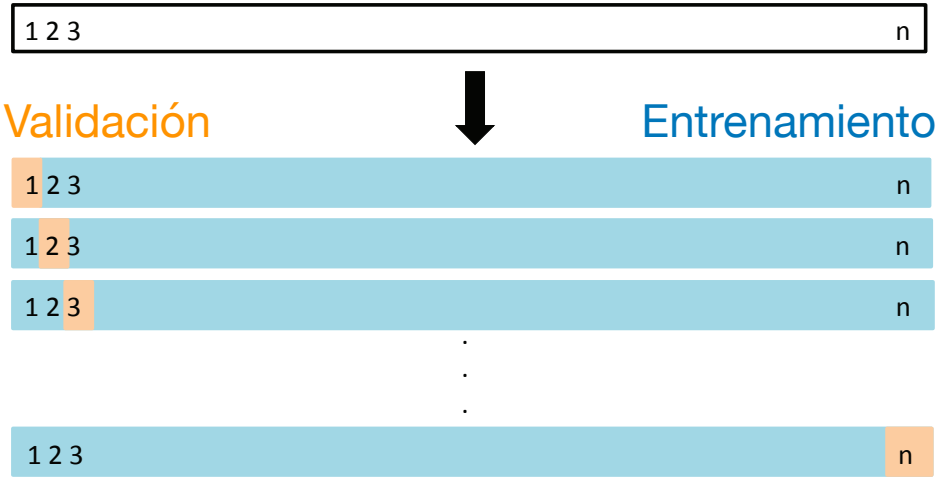
- *k-veces CV* a menudo proporciona estimaciones más precisas del *EP* que *LOOCV*
- Existe un canje entre sesgo y varianza asociada con la elección de k en *k-veces CV*
- Por lo general, se realiza *k-veces CV* usando $k = 5$ o 10
  - producen estimaciones del *EP* que no sufren ni un sesgo excesivamente alto ni una varianza muy alta

EP : error de prueba

5.1.5

Validación Cruzada en Problemas de Clasificación

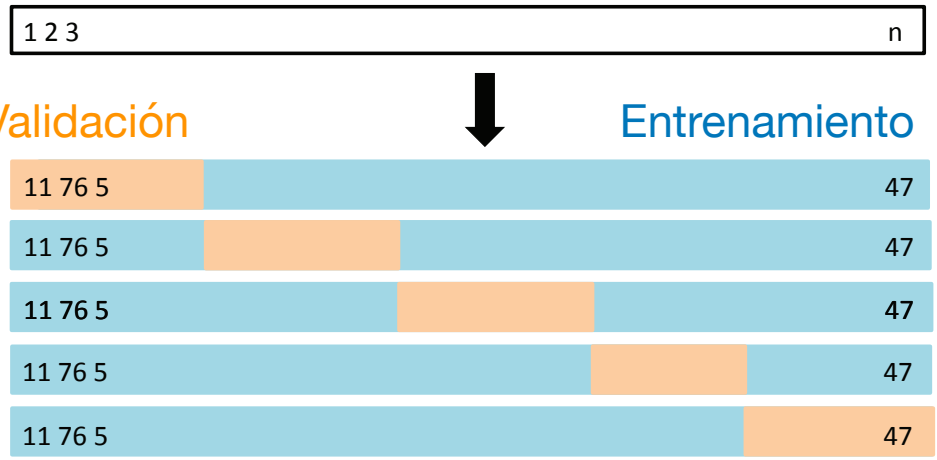
LOOCV



$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

K-veces CV



$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_i$$

$$= \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} I(y_{ij} \neq \hat{y}_{ij})$$

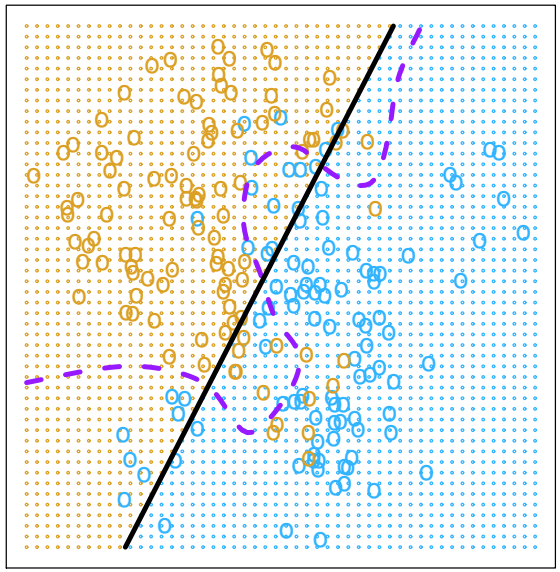
Bayes
error de prueba 0.133

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

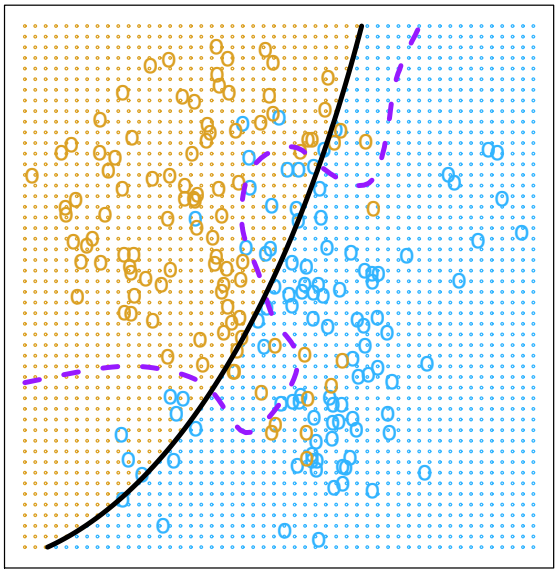
Error de prueba 0.201

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Degree=1



Degree=2



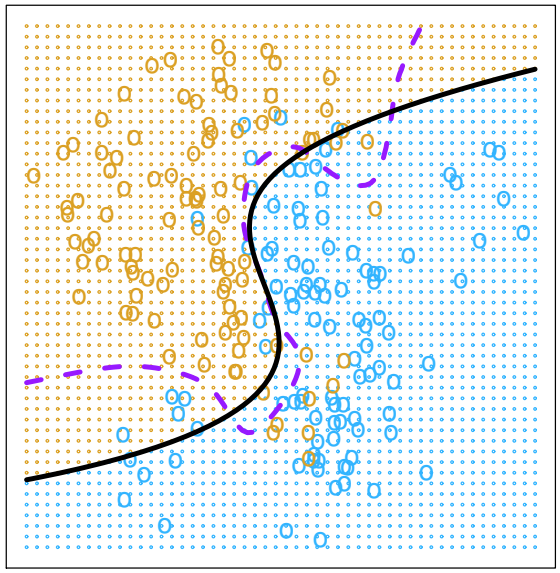
Error de prueba 0.197

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2$$

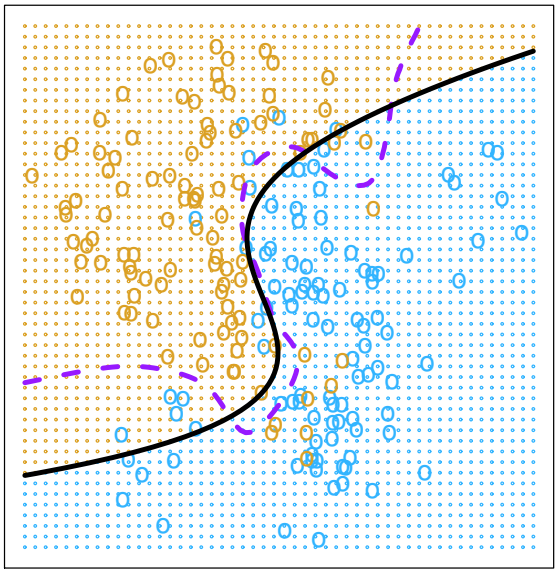
Error de prueba 0.160

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_3 x_1^3 + \beta_4 x_2 + \cdots + \beta_6 x_2^3$$

Degree=3



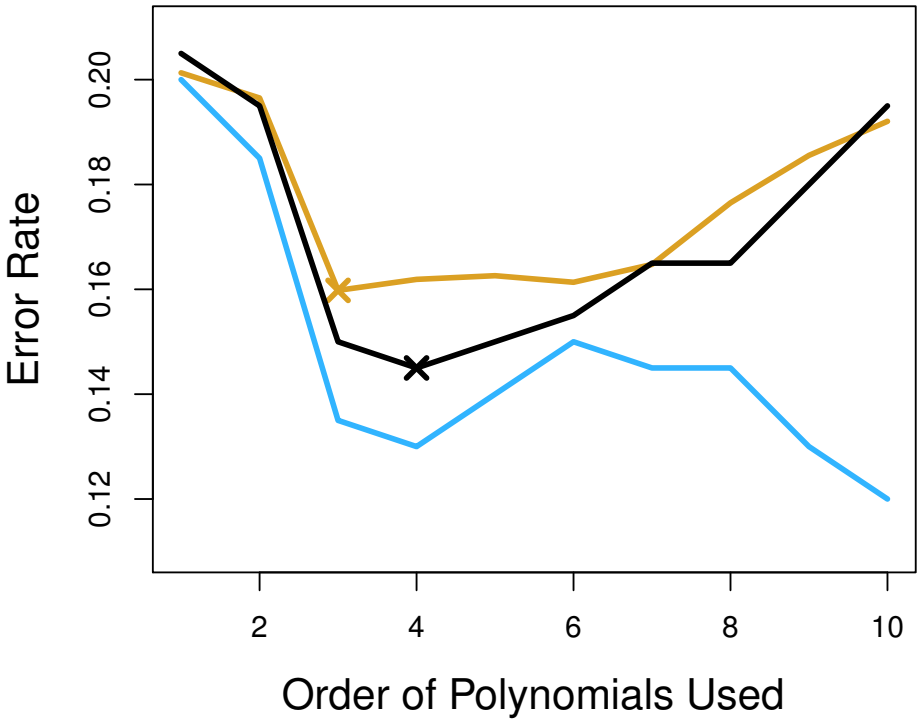
Degree=4

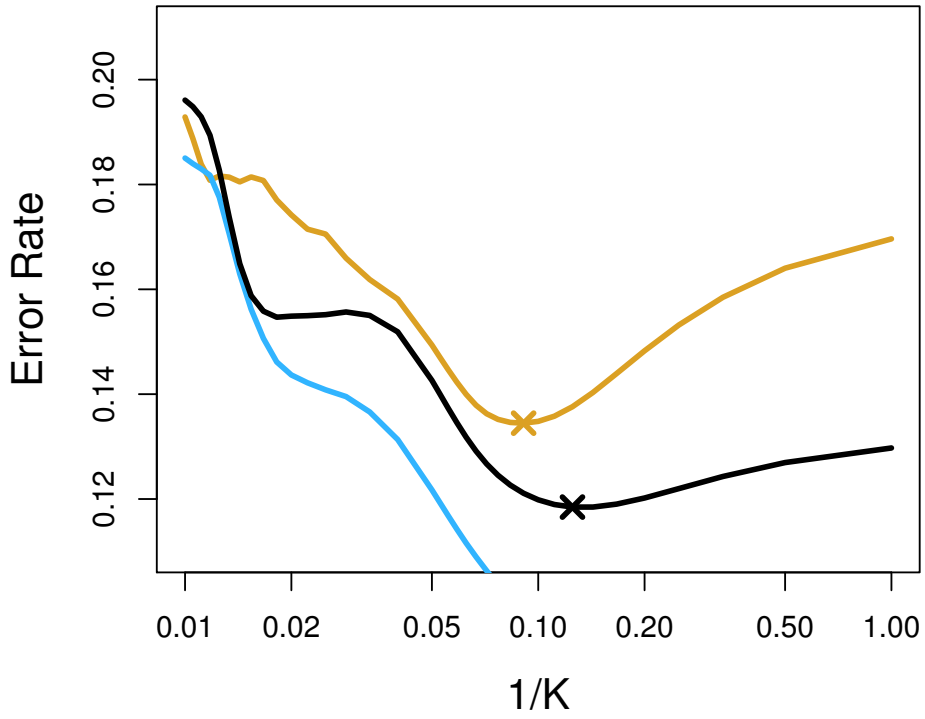


Error de prueba 0.162

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_4 x_1^4 + \beta_5 x_2 + \cdots + \beta_8 x_2^4$$

Validación Cruzada para Seleccionar Modelos





Regresión logística utilizando funciones polinómicas de los predictores

El clasificador KNN con diferentes valores del número de vecinos K

Error de prueba, error de entrenamiento y **error 10-veces CV**

Bootstrap

- Se usa para cuantificar la incertidumbre asociada con un determinado estimador o método de aprendizaje estadístico

  - Ejemplo, estimar los errores estándar de los coeficientes a partir de un ajuste de regresión lineal

Ejemplo

- Deseamos invertir una suma fija de dinero en dos activos financieros que producen rendimientos de X e Y, dos cantidades aleatorias

- Invertimos una fracción α del dinero en X y el resto $1 - \alpha$ en Y

- Existe una variabilidad asociada con los rendimientos de estos dos activos. Por lo tanto, deseamos elegir α para minimizar el riesgo total, o varianza, de la inversión.

  - Minimizar $Var(\ \alpha X + (1 - \alpha)Y\ )$.

Minimizar Riesgo

- Se puede demostrar que el valor que minimiza el riesgo está dado por

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

- $\sigma_X^2 = Var(X)$, $\sigma_Y^2 = Var(Y)$ y $\sigma_{XY} = Cov(X, Y)$

- Varianzas y covarianzas de los datos son desconocidas

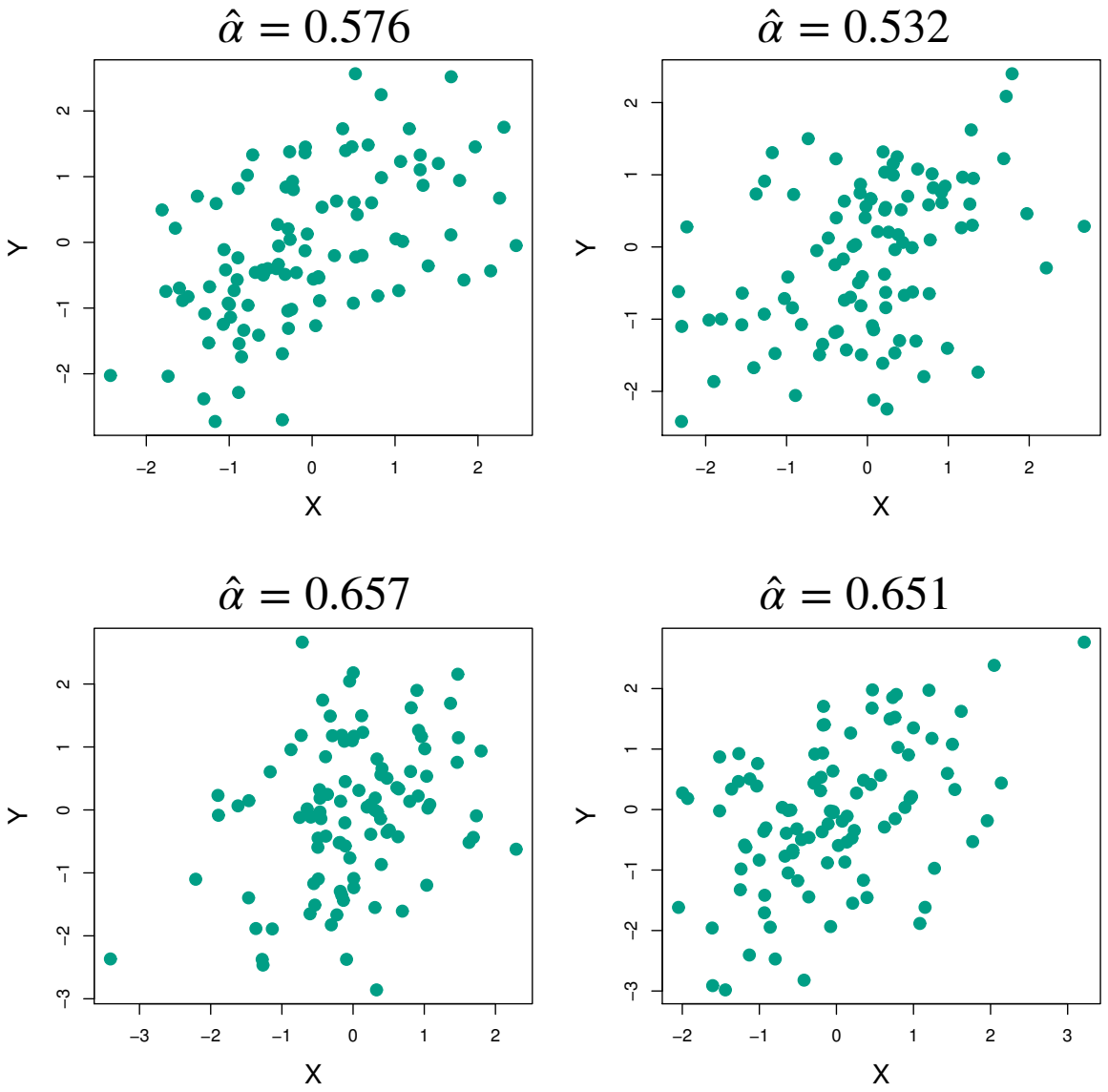
  ➡ Hay que estimarlas

Estimando α con Datos Simulados

A partir de los datos de X e Y se estiman $\sigma_X^2, \sigma_Y^2, \sigma_{XY}$, y de ellos se estima α

Para cuantificar la precisión de la estimación de α, repetimos 1000 veces el proceso de simular 100 observaciones pareadas de X e Y y estimar α

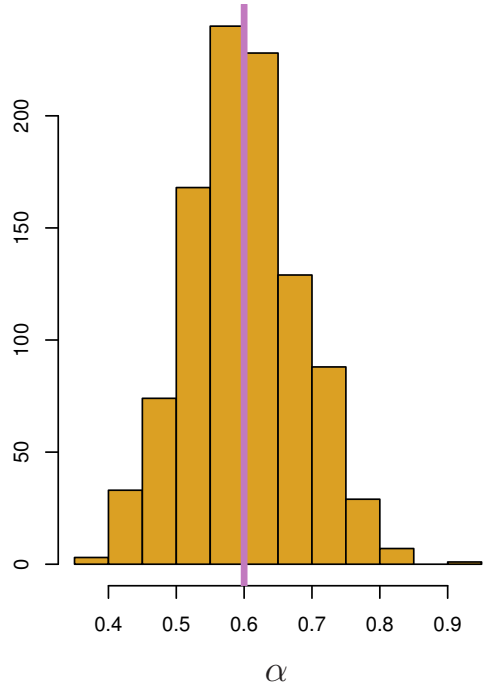
$$\sigma_X^2 = 1,\ \sigma_Y^2 = 1.25,$$
$$\sigma_{XY} = 0.5,\ \alpha = 0.6$$



Cada panel muestra 100 rendimientos simulados para las inversiones X e Y

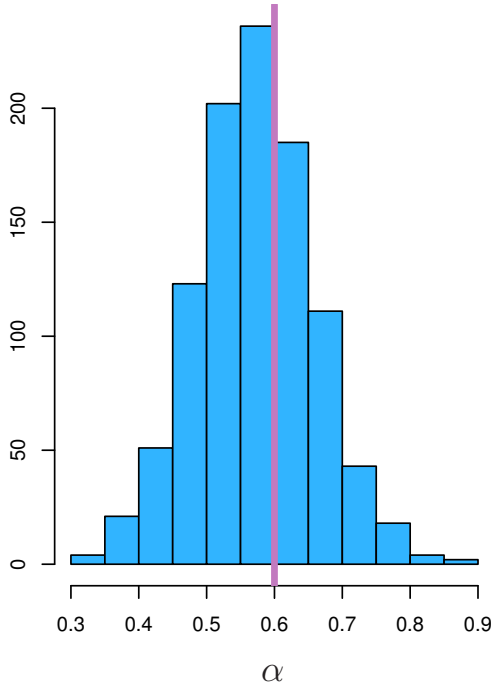
Valores Estimados

1000 muestras de la Población



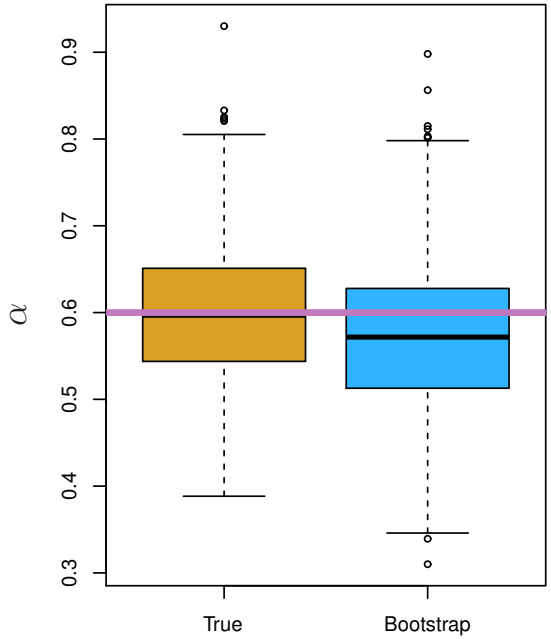
$SE(\hat{\alpha}) \approx 0.083$

1000 muestras Bootstrap

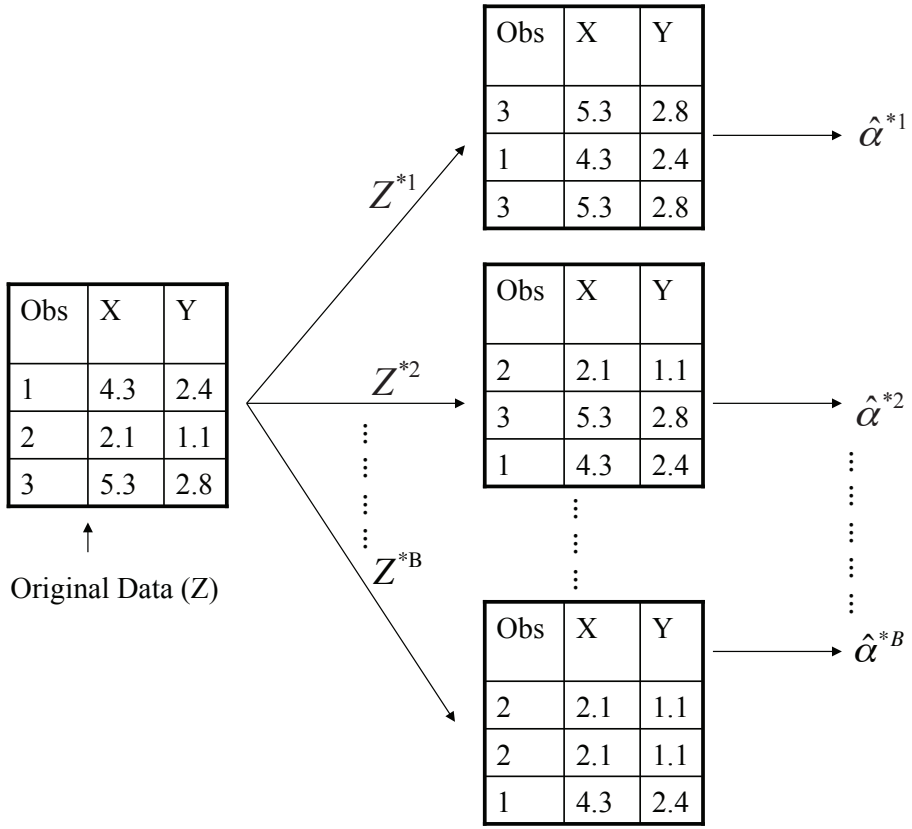


$SE(\hat{\alpha}) \approx 0.087$

Boxplots



Método Bootstrap



- Se obteniene conjuntos de datos distintos muestreando repetidamente del conjunto de datos original
  - En lugar de obtener repetidamente conjuntos de datos independientes de la población

5.3

Lab: Validación Cruzada y Bootstrap

- El enfoque del Conjunto de Validación
- Validación Cruzada
- Bootstrap (El arranque?)

5.4

Ejercicios