Tema 4 - Cross validación y bootstrap

Aprendizaje Estadístico

Lorena Romero Mateo - 77857300T

Email: lorena.romerom@um.es

Índice

1. Remuestreo - Validación de modelos **2**
1.1. Training error vs Test error 2
1.2. Más sobre estimaciones de error de predicción 2

2. Validación cruzada de K iteraciones **2**
2.1. Un caso especial, k=n 3

3. Validación cruzada en el conjunto de datos Auto **4**

4. Bootstrapping **4**
4.1. Un ejemplo simple 5
4.2. Bootstrapping: muestreo con repetición 7
4.3. ¿Puede el bootstrap estimar el error de predicción? 8

1. Remuestreo - Validación de modelos

En esta sección discutimos dos métodos de remuestreo: validación cruzada y bootstrap.

Estos métodos vuelven a ajustar un modelo de interés a muestras formadas a partir del conjunto de entrenamiento, con el fin de obtener información adicional sobre el modelo ajustado.

Por ejemplo, proporcionan estimaciones del error de predicción en el conjunto de prueba, y la desviación estándar y el sesgo de nuestras estimaciones de parámetros.

1.1. Training error vs Test error

Recordemos la distinción entre el error de prueba y el error de entrenamiento:

- El error de prueba es el error promedio que resulta de usar un método de aprendizaje estadístico para predecir la respuesta en una nueva observación, una que no se utilizó en el entrenamiento del método.
- En contraste, el error de entrenamiento se puede calcular fácilmente aplicando el método de aprendizaje estadístico a las observaciones utilizadas en su entrenamiento.
- Pero la tasa de error de entrenamiento a menudo es bastante diferente de la tasa de error de prueba, y en particular, la primera puede subestimar dramáticamente a la segunda.

1.2. Más sobre estimaciones de error de predicción

- Mejor solución: un gran conjunto de prueba designado. A menudo no disponible.
- Algunos métodos hacen un ajuste matemático a la tasa de error de entrenamiento para estimar la tasa de error de prueba. Estos incluyen la estadística Cp, AIC y BIC. Se discuten en otra parte de este curso.
- Aquí, en cambio, consideramos una clase de métodos que estiman el error de prueba dejando fuera un subconjunto de las observaciones de entrenamiento del proceso de ajuste, y luego aplicando el método de aprendizaje estadístico a esas observaciones dejadas fuera.

2. Validación cruzada de K iteraciones

- Enfoque ampliamente utilizado para estimar el error de prueba.
- Las estimaciones se pueden usar para seleccionar el mejor modelo y para dar una idea del error de prueba del modelo final elegido.
- La idea es dividir aleatoriamente los datos en K partes de igual tamaño. Dejamos fuera la parte k, ajustamos el modelo a las otras K-1 partes (combinadas) y luego obtenemos predicciones para la parte k dejada fuera.

- Esto se hace a su vez para cada parte k = 1, 2, ..., K, y luego se combinan los resultados.

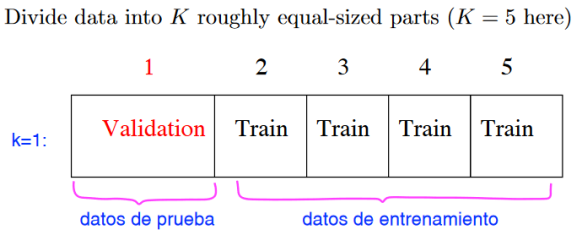


Figura 1: K-fold Cross Validación

Sea K partes C1, C2, ..., CK, donde Ck denota los índices de las observaciones en la parte k. Hay nk observaciones en la parte k: si N es un múltiplo de K, entonces nk = n/K.

- Calcular

$$CV(K) = \sum_{k=1}^{K} \frac{n_k}{n} \text{MSE}_k$$

donde

$$\text{MSE}_k = \sum_{i \in C_k} \frac{(y_i - \hat{y}_i)^2}{n_k}$$

y $\hat{y}_i$ es el ajuste para la observación i, obtenido de los datos con la parte k eliminada.

- Establecer K = n produce validación cruzada de n iteraciones o dejar-uno-fuera (LOOCV).

2.1. Un caso especial, k=n

Con la regresión lineal o polinómica de mínimos cuadrados, un atajo sorprendente hace que el costo de LOOCV sea el mismo que el de un solo ajuste de modelo. La siguiente fórmula se cumple: (**solo cuando k = n y es un modelo lineal.**)

$$CV(n) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2,$$

donde $\hat{y}_i$ es el valor ajustado i-ésimo del ajuste original de mínimos cuadrados, y h_i es la influencia (diagonal de la matriz "hat"; ver libro para más detalles). Esto es como el MSE ordinario, excepto que el residuo i-ésimo se divide por $1 - h_i$.

- LOOCV a veces es útil, pero típicamente no sacude los datos lo suficiente. Las estimaciones de cada pliegue están altamente correlacionadas y, por lo tanto, su promedio puede tener alta varianza.

- Una mejor opción es $K = 5$ o 10.

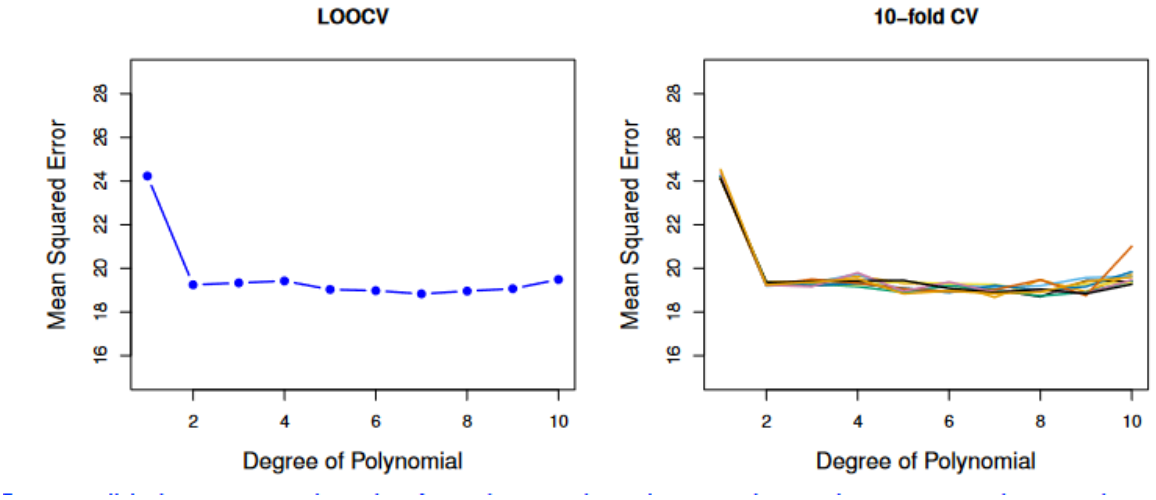


Figura 2: Auto data revisited

3. Validación cruzada en el conjunto de datos Auto

La validación cruzada se utilizó en el conjunto de datos Auto para estimar el error de prueba que resulta de predecir mpg usando funciones polinómicas de la potencia.

- Izquierda: La curva de error de LOOCV.
- Derecha: Se ejecutó la validación cruzada de 10 iteraciones nueve veces por separado, cada una con una división aleatoria diferente de los datos en diez partes.

La figura muestra las nueve curvas de error de CV ligeramente diferentes.

4. Bootstrapping

El uso del término bootstrap se deriva de la frase "levantarse por sus propios medios", que se cree ampliamente que está basada en una de las aventuras del siglo XVIII "Las sorprendentes aventuras del Barón Munchausen" de Rudolph Erich Raspe:

"El Barón había caído al fondo de un lago profundo. Justo cuando parecía que todo estaba perdido, pensó en levantarse por sus propios medios. "

- No es lo mismo que el término "bootstrap" utilizado en informática, que significa "iniciar" una computadora desde un conjunto de instrucciones básicas, aunque la derivación es similar.
- Los métodos de bootstrapping fueron introducidos a finales de los años 1970 por el estadístico Bradley Efron.

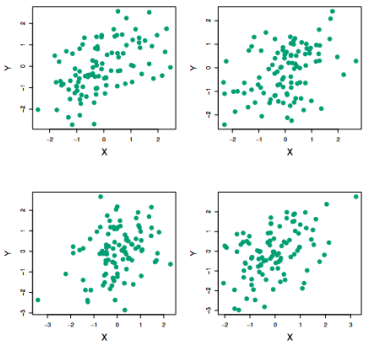
4.1. Un ejemplo simple

- Supongamos que deseamos invertir una suma fija de dinero en dos activos financieros que generan rendimientos de X e Y, respectivamente, donde X e Y son cantidades aleatorias.
- Invertiremos una fracción α de nuestro dinero en X, y el restante $1 - \alpha$ en Y.
- Deseamos elegir α para minimizar el riesgo total, o la varianza, de nuestra inversión. En otras palabras, queremos minimizar $\text{Var}(\alpha X + (1 - \alpha)Y)$.
- Se puede demostrar que el valor que minimiza el riesgo está dado por

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

  donde $\sigma_X^2 = \text{Var}(X)$, $\sigma_Y^2 = \text{Var}(Y)$, y $\sigma_{XY} = \text{Cov}(X, Y)$.
- Pero los valores de σ_X^2, σ_Y^2 y σ_{XY} son desconocidos.
- Podemos calcular estimaciones para estas cantidades, $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$ y $\hat{\sigma}_{XY}$, usando un conjunto de datos que contiene mediciones para X e Y.
- Luego podemos estimar el valor de α que minimiza la varianza de nuestra inversión usando

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$



Each panel displays 100 *simulated returns for investments* X *and* Y. *From left to right and top to bottom, the resulting estimates for* α *are* 0.576, 0.532, 0.657, *and* 0.651.

Figura 3: Example of Bootstrapping

Para estimar la desviación estándar de $\hat{\alpha}$, repetimos el proceso de simular 100 observaciones pareadas de X e Y, y estimamos α 1,000 veces.

- De este modo, obtuvimos 1,000 estimaciones para α, que podemos llamar $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{1000}$.

- El panel izquierdo de la Figura en la diapositiva 29 muestra un histograma de las estimaciones resultantes.
- Para estas simulaciones, los parámetros se establecieron en $\sigma_X^2 = 1$, $\sigma_Y^2 = 1{,}25$, y $\sigma_{XY} = 0{,}5$, por lo que sabemos que el valor verdadero de α es 0.6 (indicado por la línea roja).

El promedio de las 1,000 estimaciones para α es

$$\bar{\alpha} = \frac{1}{1000}\sum_{r=1}^{1000}\hat{\alpha}_r = 0{,}5996,$$

muy cercano a $\alpha = 0{,}6$, y la desviación estándar de las estimaciones es

$$\sqrt{\frac{1}{1000}\sum_{r=1}^{1000}(\hat{\alpha}_r - \bar{\alpha})^2} = 0{,}083.$$

- Esto nos da una muy buena idea de la precisión de $\hat{\alpha}$: $SE(\hat{\alpha}) \approx 0{,}083$.
- Así que, hablando en términos generales, para una muestra aleatoria de la población, esperaríamos que $\hat{\alpha}$ difiera de α en aproximadamente 0.08, en promedio.

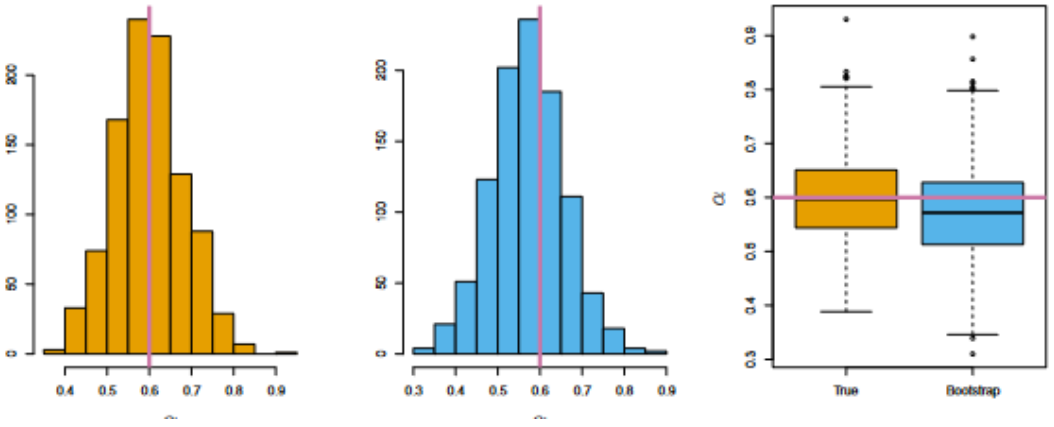


Figura 4: Izquierda: Un histograma de las estimaciones de α obtenidas generando 1,000 conjuntos de datos simulados de la población verdadera. Centro: Un histograma de las estimaciones de α obtenidas de 1,000 muestras bootstrap de un único conjunto de datos. Derecha: Las estimaciones de α mostradas en los paneles izquierdo y central se presentan como diagramas de caja. En cada panel, la línea rosa indica el valor verdadero de α.

4.2. Bootstrapping: muestreo con repetición

El procedimiento descrito anteriormente no se puede aplicar, porque para datos reales no podemos generar nuevas muestras de la población original.

- Sin embargo, el enfoque bootstrap nos permite usar una computadora para imitar el proceso de obtención de nuevos conjuntos de datos, de modo que podamos estimar la variabilidad de nuestra estimación sin generar muestras adicionales.
- En lugar de obtener repetidamente conjuntos de datos independientes de la población, obtenemos conjuntos de datos distintos muestreando repetidamente observaciones del conjunto de datos original con reemplazo.
- Cada uno de estos "conjuntos de datos" bootstrap se crea mediante muestreo con reemplazo, y tiene el mismo tamaño que nuestro conjunto de datos original. Como resultado, algunas observaciones pueden aparecer más de una vez en un conjunto de datos bootstrap dado y algunas no aparecer en absoluto.

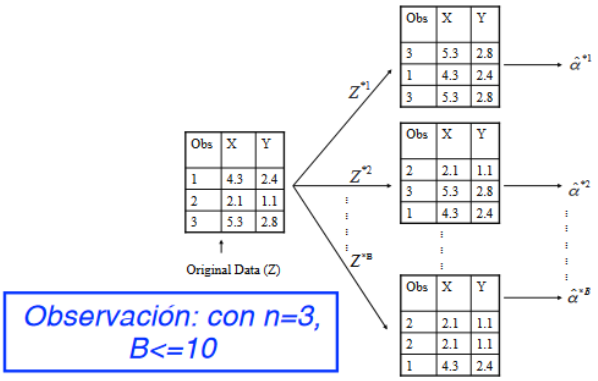


Figura 5: Una ilustración gráfica del enfoque bootstrap en una pequeña muestra que contiene n = 3 observaciones. Cada conjunto de datos bootstrap contiene n observaciones, muestreadas con reemplazo del conjunto de datos original. Cada conjunto de datos bootstrap se utiliza para obtener una estimación de α

Denotando el primer conjunto de datos bootstrap por Z^{1*}, usamos Z^{1*} para producir una nueva estimación bootstrap para α, que llamamos $\hat{\alpha}_1^*$.

- Este procedimiento se repite B veces para algún valor grande de B (digamos 100 o 1000), con el fin de producir B conjuntos de datos bootstrap diferentes, $Z^{1*}, Z^{2*}, \ldots, Z^{B*}$, y B estimaciones correspondientes de α, $\hat{\alpha}^{1*}, \hat{\alpha}^{2*}, \ldots, \hat{\alpha}^{B*}$.
- Estimamos el error estándar de estas estimaciones bootstrap usando la fórmula

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}^{r*} - \bar{\hat{\alpha}}^*)^2}.$$

- Esto sirve como una estimación del error estándar de $\hat{\alpha}$ estimado a partir del conjunto de datos original. Ver los paneles central y derecho de la Figura en la diapositiva 29. Los resultados del bootstrap están en azul. Para este ejemplo $SE_B(\hat{\alpha}) = 0{,}087$.

4.3. ¿Puede el bootstrap estimar el error de predicción?

- En la validación cruzada, cada uno de los K pliegues de validación es distinto de los otros K-1 pliegues utilizados para el entrenamiento: no hay superposición. Esto es crucial para su éxito. ¿Por qué?

- Para estimar el error de predicción usando el bootstrap, podríamos pensar en usar cada conjunto de datos bootstrap como nuestra muestra de entrenamiento, y la muestra original como nuestra muestra de validación.

- Pero cada muestra bootstrap tiene una superposición significativa con los datos originales. Aproximadamente dos tercios de los puntos de datos originales aparecen en cada muestra bootstrap. ¿Puedes probar esto?

- Esto hará que el bootstrap subestime seriamente el verdadero error de predicción. ¿Por qué?