Herramientas de aprendizaje supervisado

Clase 3 - Validación cruzada y Selección de variables

Manuel Benjamín

October 28, 2023

Universidad de Buenos Aires

Selección de modelos

¿Cómo comparamos el ajuste de los distintos modelos?

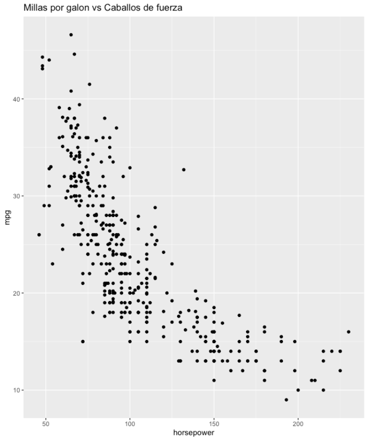
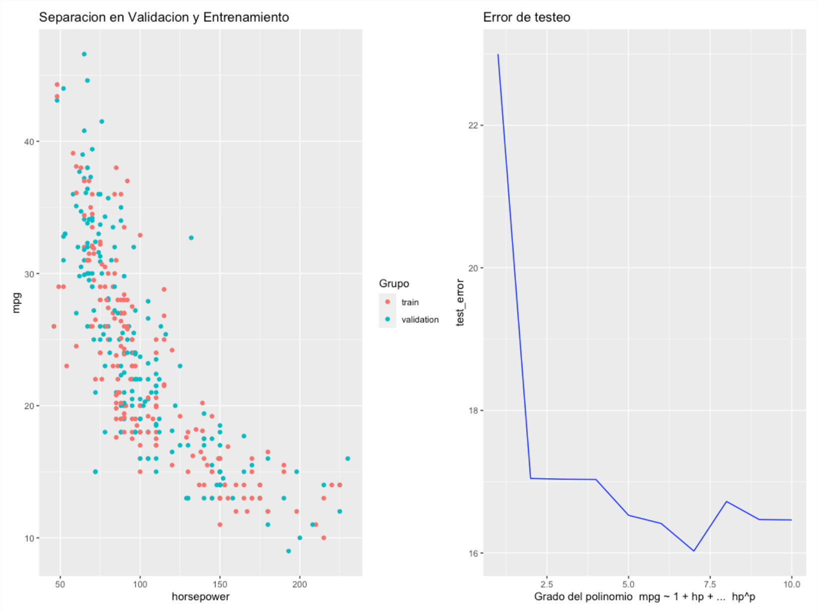


Figure 1: Dataset *Auto*: Millas por galón versus Caballo de fuerza de 392 modelos de autos. ¿ Que grado debemos usar para un ajuste polinomial?

Si el objetivo es predecir...

Metodología 1: Entrenamiento - Validación

- Partimos el data set en un grupo de entrenamiento y otro de validación.
- Estimamos cada uno de los modelos con el set de entrenamiento.
- Comprobamos la capacidad predictiva de cada modelo evaluando el MSE sobre el conjunto de validación.


Separacion en Validacion y Entrenamiento
mpg
horsepower
Grupo
train
validation
Error de testeo
test_error
Grado del polinomio mpg ~ 1 + hp + ... hp^p

¿Que sucede si volvemos a sortear la partición en validación y entrenamiento?

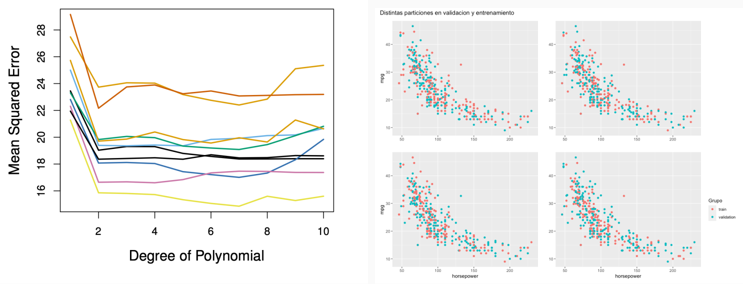


Figure 2: Izquierda: Curvas de MSE para distintas separaciones en validación y entrenamiento. El proceso fue repetido 10 veces y cada curva corresponde al MSE de esa partición aleatoria. Derecha: Distintas asignaciones en validacion y entrenamiento.

Problemas de elegir un modelo separando en validación y entrenamiento

- La estimación del error de testeo puede tener mucha variabilidad, dependiendo de que observaciones quedan en el entrenamiento y en el de validación
- Al ajustar con la mitad de los datos uno espera que estemos SOBRESTIMANDO el error de predicción.

Validación cruzada K cruces - Receta

1. Separar las observaciones en K grupos disjuntos $C_1, C_2, \ldots, C_K$ de cantidad n_k lo mas parecidas posible.
2. Para cada $1 \leq k \leq K$,

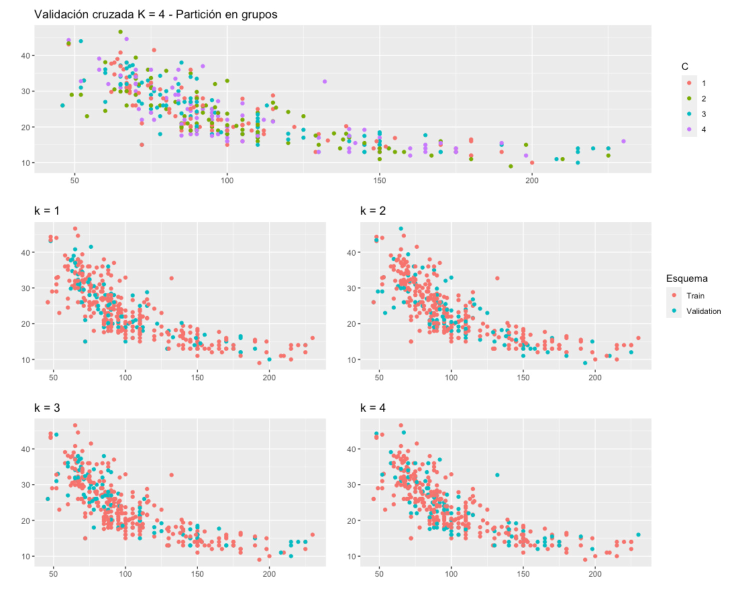
$$\mathrm{T}_k = \cup_{j \neq k} C_j$$
$$\mathrm{V}_k = C_k$$

3. Definimos MSE_k como el error de validación del cruce k al ajustar el modelo con T_k y validar con V_k

$$MSE_k = \frac{1}{n_k} \sum_{i \in C_k} (y_i - \hat{y}_i)^2.$$

4. Calculamos el error de validación cruzada

$$\mathrm{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$$


Validación cruzada K = 4 - Partición en grupos
C
1
2
3
4
k = 1
k = 2
k = 3
k = 4
Esquema
Train
Validation
10
20
30
40
50
100
150
200

En general tenemos que ajustar K veces el modelo.

- Con $K = n$ tenemos *Leave- one-out-cross-validation* (LOOCV).
- En caso de regresión lineal o polinomial, vale que

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

  Donde $\hat{y}_i$ es el el valor predicho con las n variables y h_i es el *leverage* o elemento de la diagonal de la matriz de proyección (matriz sombrero) $H = X(X^t X)^{-1} X^t$.

  Esto requiere un único ajuste en vez de n!

- Como el conjunto de entrenamiento de cada cruce contiene una proporción de $\frac{K-1}{K}$ del total de los datos, en general el error de validación cruzada sera una sobre estimación del error de predicción.
- Cuando $K = n$ (LOOCV) el sesgo se minimiza. Pero tiene mucha varianza.
- En general se usan $K = 5$ o $K = 10$ como un buen balance entres sesgo y variabilidad.

Validación cruzada para clasificación

- Dividimos los datos en K grupos disjuntos $C_1, \dots, C_K$ de cantidad de datos n_k similares
- Calculamos

$$CV_K = \sum_{j=1}^{K} \frac{n_k}{n} Err_k$$

  Donde

$$Err_k = \sum_{j \in C_k} I(y_i \neq \hat{y}_j)$$

  con $\hat{y}_j$ obtenido ajustando con el $T_k = \cup_{j \neq k} C_j$.
- Podemos usar otra métrica de ajuste en vez del accuracy.

Usos de validación cruzada

- Selección de modelos Entre M modelos distintos elegir el mas apropiado en términos de predicción.
  Por ejemplo: Un modelo lineal generalizado para una bernoulli con M distintas funciones de link.
- Tuneo de hiper-parámetros.
  Se tiene una familia de motodologías de ajuste que dependen de un parámetro y se quiere elegir el mejor parámetro.
  Ejemplos: Vecinos cercanos, Ridge, Lasso, Elastic Net...

Ejercicio

Considere un modelo de regresión logística con $p = 1000$ variables explicativas.

Se sabe que el modelo es esparso, es decir solo algunos de los coeficientes son distintos de cero. Además solo se disponen $n = 50$ observaciones.

Se consideran entonces la siguiente familia de metodologías de ajuste: Para cada m entre 1 y 20, se elijen las m variables con mayor correlación con Y y se ajusta la regresión logística con las m variables seleccionadas.

1. Para cada $1 \leq m \leq 20$ calcule el error de validación cruzada con $K = 5$ cruces.
2. Grafique.
3. ¿Qué valor de m minimiza el CV?
4. Repita varias veces lo anterior con nuevos cruces. Se minimiza en el mismo m?
5. Haga los mismo con LOOCV.

Para la simulación genere una muestra de $n = 50$ con un modelo donde $P(x)$ depende

$$\beta_i = \begin{cases} 1 & \text{si } 1 \leq i \leq 7 \\ 0 & \text{caso contrario} \end{cases}$$

Selección de variables en el modelo lineal

- A pesar de la simplicidad el modelo lineal tiene ventajas en términos de interpretabilidad e incluso suele mostrar buen desempeño en predicción.
- ¿Cómo podemos mejorar el modelo lineal al reemplazar el ajuste de cuadrados mínimos por otras alternativas de ajuste?

- Mejorar su capacidad predictiva en casos donde $p > n$
- Interpretabilidad del modelo: Fijando estimaciones de coeficientes a cero, de manera que sea mas fácil de interpretar. Esta tarea se conoce como selección de variables
    - Selección de subconjuntos
    - Regularización Lasso (L1).

Selección del mejor subconjunto de variables

1. Para cada subconjunto de covariables posible elegir el de menor error de validación cruzada u otra medida de performance predictiva (AIC, BIC).

Extensión a otros modelos

Aunque contamos las ideas para regresión lineal, esto se puede extender a otros modelos estadísticos (GLM)

- Por motivos computacionales, hallar el mejor subconjunto no puede ser aplicado para p grandes
- Existen problemas estadísticos cuando p es grande. Al haber tantos modelos posibles, existe una posibilidad de encontrar un modelo que se vea bien pero que no tenga buenos modelos predictivos.
- Al haber tantos modelos posibles, esto puede llevar a **sobreajustes** y a una alta variabilidad en la estimación de los coeficientes.
- Por esta razones utilizamos metodologias de a pasos (stepwise) que restringen la cantidad de modelos evaluados y disminuyen la variabilidad de las estimaciones.

Forward stepwise

1. Sea M_0 el modelo sin ninguna variable explicativa.
2. para cada $k = 0, \dots, p - 1$
    2.1 Considerar todos los $p - k$ que incluyen una variable extra al modelo M_k.
    2.2 Elegir el mejor de los $p - k$ modelos con CV, AIC o BIC.
3. De la sucesión de modelos $M_0, \dots, M_p$ elegir el mejor en términos de CV, AIC, o BIC.

- ¿Cuántos modelos ajustamos con esta metodología?
- No podemos garantizar que encontramos el modelo óptimo entre los 2^p modelos.

Backward stepwise

1. Sea M_p completo, es decir con las p variables explicativas.
2. para cada $k = p - 1, \ldots, 1$
   2.1 Considerar todos los k que incluyen todas las variables salvo una de M_k
   2.2 Elegir el mejor de los k modelos con CV, AIC o BIC.
3. De la sucesión de modelos $M_0, \ldots, M_p$ elegir el mejor en términos de CV, AIC, o BIC.

- Como forward stepwise, recorremos solo $1 + \frac{p(p+1)}{2}$ modelos.
- Como en forward no podemos asegurar que vamos a encontrar el modelo óptimo.
- A diferencia de forward, requiere $n > p$.