Índice general

Capítulo 1

Métodos de resampleo

Para poder evaluar la performance de cualquier método de aprendizaje estadístico dado un set de datos, necesitamos una medida que nos diga cuan bien se ajustan nuestras predicciones a los datos. En general, esta medida es el error cuadrático medio,

$$ECM = \mathbf{E}[(\theta - \hat{\theta})^2]$$

Idealmente, nos gustaría contar con infinitas muestras de manera que podamos usar algunas para crear nuestros modelos y otras para testearlos, pero en la realidad solo contamos con una. Si tenemos una muestra que usamos para crear el modelo, y el modelo seleccionado es el que mejor se ajusta a los datos, es bastante lógico pensar que si usamos esa misma muestra para evaluar el modelo el error será pequeño, o al menos estaremos subestimado el verdadero error. Trataremos entonces, a partir de nuestra única muestra, proponer diferentes escenarios que simulen el caso de tener más muestras para poder así validar nuestro modelo con una muestra "fresca".

1.1. Entrenamiento y testeo

Consiste en partir la muestra original en 2 muestras: una que se utilizará para entrenar el modelo, y la otra para testear su rendimiento (se suele usar la tasa 70/30 pero depende de la cantidad de datos que se tenga). A partir de esta separación se utilizará la muestra de entrenamiento (la porción más grande) para construir todos los modelos posibles, y luego para seleccionar el mejor modelo se calculará por ejemplo el error cuadrático medio sobre la muestra de testeo. De esta manera el modelo no se verá sobreajustado a los datos que usamos para estimar el error y obtendremos una mejor estimación del mismo.

1.2. K-Fold cross validation

Recordemos que el error cuadrático medio es una esperanza, y que en la práctica lo que buscamos es una estimación a partir de los datos observados. Si tuvieramos muchas muestras de entrenamiento y muchas muestras de testeo, calcularíamos muchas estimaciones del *ecm* y al promediarlas tendríamos una estimación aún mejor (ya que el promedio es una variable con igual media a lo que se busca estimar y menor varianza a medida que aumenta el tamaño de la muestra). Esta parece ser

una muy buena idea pero, ¿como consigo más muestras? La idea que propone *k-fold* es la siguiente: Para un valor de k fijo, partimos la muestra original en k partes, obteniendo las muestras $m_1, \ldots, m_k$. En el primer paso usaremos como muestra de testeo a m_1 y para entrenamiento la union de todas las demás muestras, desde m_2 hasta m_k. De esa manera conseguimos nuestra primera estimación del error ECM_1, de la misma forma que lo hicimos con el método de entrenamiento y testeo. En el siguiente paso, m_2 será nuestra muestra de testeo y todo el resto la de entrenamiento, nuevamente calculamos el error ECM_2. Repetimos este procedimiento k veces, para finalmente conseguir una estimación del error de la siguiente manera

$$\widehat{ECM}_{kf} = \frac{1}{k}\sum_{i=1}^{k} ECM_i$$

Ahora nos quedamos pensando: si esta partición nos lleva a una mejor estimación , ¿porqué no llevarla al extremo?

1.3. Leave-one-out cross validation

La idea es la misma que con *k-fold* pero ahora con $k = n$. Para cada observación desde 1 hasta n, dejamos fuera esa observación para construir el modelo con las $n - 1$ restantes, y mediremos el error de estimación solo para esa observación que dejamos afuera. Luego, calcularemos el error de validación cruzada como

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} ECM_i$$

y dependiendo del caso en el que estamos estudiando, el $\widehat{ECM}_i$ tendrá su fórmula correspondiente de cálculo.

Este enfoque devuelve un valor con menor sesgo ya que promediamos muchos más valores y ajustamos muchos más modelos con una gran cantidad de observaciones pero tiene mayor varianza y es mucho más caro computacionalmente porque debo ajustar muchos más modelos. Estudiemos un poco más en detalle lo que significa, estudiando el equilibrio entre sesgo y varianza.

1.4. Balance entre sesgo y varianza

Este será un tema recurrente para cada estimación que hagamos. El error cuadrático medio, tal como lo vimos en el capítulo anterior lo podemos descomponer en dos términos, uno que corresponde a la varianza del estimador, y otro que corresponde al sesgo. A medida que el modelo aumenta su complejidad, el sesgo disminuye pero la varianza aumenta. Todo el tiempo estamos buscando encontrar un equilibrio entre esos dos componentes. En el caso de la regresión, lo podríamos ver de esta forma:

Si $y = f(x) + \epsilon$, donde ϵ es un error aleatorio del cual no tenemos ningún control, entonces

$$ECM = \mathbf{E}[(y - \hat{f}(x))^2] = \mathbf{var}(\hat{f}(x)) + B^2(\hat{f}(x)) + \sigma^2$$

donde σ^2 es la varianza del error aleatorio. Para el caso de estimación no paramétrica habíamos visto que tanto la varianza como el sesgo del estimador depemndían de h, de manera que cuando h aumentaba aumentaba el sesgo pero disminuía la varianza. Esto lo vemos en un **ejemplo simulado en R** para poder entender visualmente a que se refiere.

Si calculamos el ECM para cada valor de h, podemos observar una curva decreciente para el sesgo, y una curva creciente para la varianza, encontrando el equilibrio en el punto en el que el error estimado es el mínimo.

Para resampleo, ya sea *k-fold* o *leave-one-out* cross validation, la complejidad del modelo estará dada por el valor de k, a un menor valor de k el modelo será más simple, teniendo un mayor sesgo y menor varianza, está en nosotros encontrar el equilibrio.

1.5. Bootstrap

Es una herramienta muy usada y extremadamente poderosa que se suele usar para estudiar ciertos aspectos de los estimadores difíciles de estudiar si recurrimos a la teoría. Por ejemplo, si estudiamos el promedio de la muestra, $\bar{X}$, es una variable aleatoria de la cual podemos deducir fácilmente su esperanza y varianza, pero, ¿es tán fácil si estudiamos la mediana? ¿conocemos su esperanza y varianza? No, es súper difícil.

En un mundo ideal, podríamos tener muchas muestras, para cada una de ellas podríamos calcular la estimación del parámetro desconocido y con todas esas estimaciónes podríamos aproximar la distribución del estimador.

Ahora tenemos una sola muestra, ¿que hacemos? el método Bootstrap propone extraer de nuestra muestra de tamaño n nuevas muestras de tamaño n realizando extracciones con reposición. Se puede probar empíricamente que si la muestra original es representativa de la población en estudio, las muestras bootstrapeadas también lo serán.

Por lo tanto, si tenemos $X_1, \ldots, X_n \overset{i.i.d}{\sim} F_\theta(x)$ a partir de la cual queremos estudiar un estimador $\hat{\theta}$, tendremos ahora N_{boot} muestras y construiremos $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_{Nboot}$ estimaciones. A partir de ellos podremos, por ejemplo, estimar la varianza de $\hat{\theta}$, que es en general muy difícil. Podemos hacer histogramas, boxplot, etc, y tener una idea del comportamiento de la distribución del estimador. También podemos comparar las varianzas de diferentes estimadores para un mismo parámetro.

Analicemos dos tipos de Bootstrap: Paramétrico y no paramétrico.

1.5.1. Bootstrap paramétrico

En este caso, podremos suponer como conocida la distribución de la población, ya sea por datos históricos o mediante algún test de hipótesis evaluado sobre nuestros datos, es decir, estaremos asumiendo un modelo paramétrico, por ejemplo que $X_i \sim F_\theta(x)$. Para proceder con nuestra estimación bootstrap realizaremos los siguientes pasos:

1. Estimamos al parámetro $\hat{\theta}$

2. Generamos muchos (digamos Nboot) conjuntos de datos de tamaño n, utilizando $F_{\hat{\theta}}(x)$..Para cada uno de esos conjuntos de datos encontraremos una estimación puntal $\hat{\theta}^*$ de nuestro parámetro desconocido θ.

3. La distribución empírica de los valores resultantes $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_{Nboot}$ es una aproximación a la distribución de $\hat{\theta}$

Ejemplo lámparas en R

1.5.2. Bootstrap no paramétrico

Para este caso no haremos ningún supuesto con respecto a la distribución de X, por lo tanto solo contamos con una muestra para poder entender como se comporta nuestro estimador. Los pasos para conseguir una muestra bootstrap de estimaciones puntuales del estimador son

1. Generar Nboot muestras de tamaño n, realizando extracciones *con reposición* de la muestra original.

2. Para cada una de las muestras generadas encontraremos una estimación puntal $\hat{\theta}^*$ de nuestro parámetro desconocido θ.

Por último, es una herramienta súper útil para encontrar intervalos de confianza, veamos 2 tipos.

1.5.3. Estimadores asintóticamente normales

Decimos que $\hat{\theta}$ es asintóticamente normal *si*

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\mathbf{var}(\hat{\theta}_n)}} \overset{(a)}{\sim} \mathcal{N}(0,1)$$

Si por ejemplo $\hat{\theta}_n = \bar{X}$ se prueba usando el teorema central del límite.

De aqui podemos deducir el intervalo de confianza *bootstrap normal* de nivel asintótico α usando el método del pivote. Si suponemos que vale

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\mathbf{var}(\hat{\theta}_n)}} \overset{(a)}{\sim} \mathcal{N}(0,1)$$

y tenemos un estimador para la varianza $\hat{se}$ tal que $\frac{\sqrt{\mathbf{var}(\hat{\theta}_n)}}{\hat{se}} \to 1$, usando slutsky podemos construir un intervalo de la forma

$$IC = (\hat{\theta}_n - z_{1-\alpha/2} * \hat{se}; \hat{\theta}_n + z_{1-\alpha/2} * \hat{se})$$

Este será un intervalo de confianza de nivel asintótico $1-\alpha$ para θ, donde $\hat{se}$ es el valor que calcularemos a partir de bootstrap.

1.5.4. Intervalo de confianza bootstrap percentil

En este caso no se quiere asumir ningún tipo de distribución subyacente. Se consideran realizaciones $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_{n_{boot}}$ del estimador $\hat{\theta}$ obtenidos haciendo Bootstrap como en el método anterior. Luego, un posible intervalo de nivel aproximado $1-\alpha$ es

$$IC = (\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)})$$

donde $\hat{\theta}^*_{(\gamma)}$ es el $\gamma-$percentil de la muestra $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_{n_{boot}}$.

Ejemplo en R