Reporte 2 (Anexos)

Taller Multidisciplinario

Joshué Helí Ricalde Guerrero

2 de julio de 2020

1. Regresión Penalizada

Recordando que en el modelo lineal

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\underset{p\times 1}{\beta} + \underset{n\times 1}{e},$$

con $\mathbb{E}[e] = 0$ y $\mathrm{Var}(e) = \sigma^2 I$, el *estimador de mínimos cuadrados* cumple que

$$\widehat{Y} = X\widehat{\beta} = \underset{\mathrm{col}(X)}{\mathrm{proy}}\, Y.$$

Explícitamente,

$$\widehat{\beta} = \underbrace{\left(X^T X\right)^-}_{\substack{\text{Inversa}\\ \text{generalizada}}} X^T Y.$$

De ser invertible, la inversa generalizada es única y $\left(X^T X\right)^{-1}$.

Algúnas de las dificultades que se tienen en éste método son las siguientes:

1. Si $\left(X^T X\right)$ no es invertible, $\left(X^T X\right)^-$ no es única y $\widehat{\beta}$ no está definida unívocamente.
2. Si $\left(X^T X\right)$ es invertible pero mál condicionada, el error cuadrático medio es grande:

$$\mathrm{tr}\left(\mathrm{Var}\left(\widehat{\beta}\right)\right) = \mathrm{tr}\left(\sigma^2\left(X^T X\right)^{-1}\right) = \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_j},$$

con λ_j's autovalores de $\left(X^T X\right)$; luego,

$$\begin{aligned}\mathrm{ecm}\left(\widehat{\beta}\right) :&= \mathbb{E}\left[\left\|\widehat{\beta}-\beta\right\|^2\right]\\ &= \left\|\mathbb{E}\left[\widehat{\beta}\right]-\beta\right\|^2 + \mathrm{tr}\left(\mathrm{Var}\left(\widehat{\beta}\right)\right)\\ &= 0 + \mathrm{tr}\left(\mathrm{Var}\left(\widehat{\beta}\right)\right),\end{aligned}$$

donde la última igualdad se dá gracias a que el estimador es insesgado ($\mathbb{E}\left[\widehat{\beta}\right] = \beta$).

Para solucionar estas problemáticas, se sigue la idea de "penalizar" aquellos estimadores que se alejan del origen.

Definición 1.1. Para $\lambda > 0$, se define el estimador *Ridge* o *regularizado* como

$$\widetilde{\beta}(\lambda) := \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_2^2 \right\},$$

con $\|\cdot\|_2$ la norma euclidiana.

Definición 1.2. Se define el estimador *Lasso* $\widehat{\beta}(\lambda)$ del vector de coeficientes $\beta (\in \mathbb{R}^P)$ como cualquier solución del problema de optimización

$$\widehat{\beta}(\lambda) := \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

para $\lambda \geq 0$ fijo.

Definición 1.3. Se define el estimador de *Red elástica* al vector de coeficientes $\beta (\in \mathbb{R}^P)$ que soluciona el problema de optimización

$$\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2 + \lambda \left[\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2 \right] \right\},$$

para $\lambda \geq 0$ y $\alpha \in [0,1]$ fijos.

2. Validación Cruzada

Como se ha visto hasta ahora, para regresión se ha tomado $\lambda > 0$ definido por el investigador. Sin embargo, queda la cuestión de ¿cómo escoger este λ? En general, este problema no sólo se reduce a modelos de regresión penalizada, sino a otros contextos que hacen uso de parámetros de "aprendizaje" o *hiperparametros*, y comúnmente es conocido *Determinación de (hiper)parámetros* ó *Selección de modelos*.

Aunque hay varios metodologías para abordar el tema, por lo pronto únicamente se revisa *Validación cruzada* (Cross-Validation o CV), aúnque su uso puede llevarse en paralelo a desigualdades de concentración. También se señala que se revisa CV en general, y no GCV (Generalized Cross-Validation), que aunque tiene "aplicabilidad en Ridge, puede ser limitante."

La idea principal de CV es básicamente tomar la muestra de datos y segmentarla, de manera que se utilice una parte de la muestra para encontrar el estimador $\beta(\lambda)$, y el resto de la muestra para evaluar que tan bueno resulto el estimador. Como puede verse, una de las ventajas de CV es que no se requiere de cálculo teórico adicional; en cambio, la desventaja es que requiere de más cálculo computacional.

Aquí aparece la pregunta de "¿porqué es necesario tomar otros modos de selección?" La respuesta: porque $\|Y - X\widehat{\beta}(\lambda)\|$ no es bueno para estimar λ, pues simplemente se elegiría $\lambda = 0$.

2.1. Half-Cross-Validation

Consiste en dividir la muestra original M en 2, M_1, M_2, $M_1 \cap M_2 =$, $M_1 \cup M_2 = M$, una para estimación (a veces conocida también como *entrenamiento*), y otra para validación.

1. Con M_1 obtener $\widehat{\beta}^{M_1}(\lambda)$.
2. Con M_2 obtener

$$\widehat{C}(\lambda) := \sum_{i \in M_2} \left(y_i - x_i^T \widehat{\beta}^{M_1}(\lambda) \right).$$

3. Seleccionar $\widehat{\lambda}$ tal que

$$\widehat{\lambda} = \arg\min_{\lambda \geq 0} \widehat{C}(\lambda). \tag{2.1}$$

2.2. Ordinary Cross-Validation

También conocida como *Leave-One-Out Cross-Validation*, consiste básicamente en dividir la muestra original M, de n datos, en n diferentes sub-muestras en vez de sólo dos.

1. Para todo $i = 1, \dots, n$, obtener $M_{(i)} = M \setminus \{(x_i, y_i)\}$.
2. Con $M_{(i)}$ estimar $\widehat{\beta}^{(i)}(\lambda)$.
3. Obtener el i-ésimo predictor

$$\widehat{y}_i = \sum_{j=1}^{p} x_{ij} \widehat{\beta}_j^{(i)}(\lambda),$$

obtener la función objetivo (a veces llamada *oráculo*)

$$\begin{aligned} \widehat{C}(\lambda) :&= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{y}_i \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \widehat{\beta}_j^{(i)}(\lambda) \right). \end{aligned}$$

4. Resolver (2.1).

Una de las ventajas de este tipo de CV es que se tiene poco sesgo; sin embargo, por el mismo motivo se tiene mucha varianza. Además, el cálculo numérico es pesado, pues se requiere estimar n parámetros distintos.

2.3. K-fold CV

Se sigue la misma idea que se ha estado llevando hasta ahora, particionando la muestra original M en K partes. ¿Cómo determinar estos pedazos? Al azar, ya que puede haber factores relacionados (por ejemplo, el tiempo, en el sentido que "las muestras tomadas al comienzo no son las mismas que las tomadas al final").

1. Se especifíca K, y con ello se particiona la muestra M en K partes distintas M_k, $k = 1, \dots, K$.
2. Con M_k estimar $\widehat{\beta}^{(k)}(\lambda)$.
3. Para el i-ésimo dato (que está en la k-ésima partición), se calcula el predictor

$$\widehat{y}_i = \sum_{j=1}^{p} x_{ij} \widehat{\beta}_j^{(k)}(\lambda).$$

4. Sumando sobre todos los índices de todas las particiones, se obtiene la función oráculo

$$\begin{aligned} \widehat{C}(\lambda) :&= \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{\#M_k} \sum_{i \in M_k} \left(y_i - \widehat{y}_i \right). \right] \\ &= \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{\#M_k} \sum_{i \in M_k} \left(y_i - \sum_{j=1}^{p} x_{ij} \widehat{\beta}_j^{(k)}(\lambda) \right) \right], \end{aligned}$$

donde el factor $\frac{1}{K}$ es el que contribuye a la disminución de la varianza.

5. Resolver (2.1).

Comentario: Otro método popular (para selección de modelos) es Bootstrap. Sin embargo, éste no se aborda porque *Lasso* (y en general penalización l_p) de por sí tiene mucha complejidad computacional.

3. Modelo de Mezclas Dirichlet-Multinomial (DMM)

3.1. Muestreo Multinomial

Sea $X = \left(X_{ij}\right)_{N\times S}$ una matriz de abundancias, donde X_{ij} es la abundancia observada de la taxa j en el individuo i, y se denota por X_i a la fila i-ésima de X, $1 \leq i \leq N$. Se asume que cada uno de éstos renglones es la realización de una variable aleatoria multinomial de parámetro $p_i = (p_{i1}, \ldots, p_{iS})$. Entonces, la verosimilitud de la muestra está dada por

$$L(X|p_1, \ldots, p_N) = \prod_{j=1}^{S} L_i(X_i|p_i),$$

donde

$$L_i(X_i|p_i) = J_i! \prod_{j=1}^{S} \frac{p_{ij}^{X_{ij}}}{X_{ij}}, \qquad J_i = \sum_{j=1}^{S} X_{ij}, \qquad i = 1, \ldots, N.$$

Puesto que se trata de un enfoque Bayesiano, se supondrá que los parámetros mísmos p_i son variables aleatorias.

3.2. Mezclas Dirichlet como distribuciones previas

Recordando que una variable aleatoria Dirichlet tiene como realizaciones vectores con entradas no negativas, cuya suma es igual a 1, se puede considerar a ésta como una distribución de probabilidad sobre distribuciones de probabilidad:

$$Dir(p_i|\overline{\alpha} = \theta\overline{m}) = \Gamma(\theta) \prod_{j=1}^{S} \frac{p_{ij}^{\theta m_j - 1}}{\Gamma(\theta m_j)} \delta\left(\sum_{j=1}^{s} p_{ij} - 1\right),$$

donde Γ denota la función Gamma, $\overline{\alpha} = (\alpha_1, \ldots, \alpha_S)$, $\alpha_j > 0$, $\theta = \sum_{j=1}^{S} \alpha_j$, y $m = (m_1, \ldots, m_S)$ tal que $\sum_{j=1}^{S}$.

Ahora, supóngase que se trabajan con K meta comunidades; entonces, se tienen $\overline{\alpha}_k$ parámetros de distribuciones Dirichlet, cada una con un peso π_k, $k = 1, \ldots, K$. Para cada observación i se tiene un vector $z_i = (z_{ik})$ con entradas binarias, que indica la metacomunidad a la que pertenece (es decir, es un vector de ceros con un sólo 1 en alguna entrada). Las distribuciones previas de éste vector no son más que la mezcla de los pesos,

$$P(z_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}};$$

en consecuencia, la distribución competa de la mezcla es

$$P(p_i|Q) = \sum_{k=1}^{K} Dir(p_i|\overline{\alpha}_k)\pi_k,$$

donde $Q = (K, \overline{\alpha}_1, \ldots, \overline{\alpha}_K, \pi_1, \ldots, \pi_K)$ es el vector de hiperparámetros.

3.3. Distribución posterior de parámetros multinomiales

Añadiendo la información de la muestra (la verosimilitud multinomial) a la previa, se obtiene directamente que la posterior de los parámetros p_i es

$$P(p_i|X_i) = \frac{\sum_{k=1}^{K} L_i(X_i|p_i) Dir(p_i|\overline{\alpha}_k)\pi_k}{\sum_{k=1}^{K} Dir(p_i|\overline{\alpha}_k)\pi_k}.$$

Pero, de las propiedades de ambas distribuciones, tras hacer las manipulaciones necesarias, se tiene que

$$P(p_i|X_i) = \sum_{k=1}^{K} Dir(p_i|\overline{\alpha}_k + X_i) P(z_{ik} = 1|X_i Q).$$

3.4. Comparación de Modelos por aproximación de Laplace

Sea H_K el modelo DMM con K componentes (metacomunidades). Puesto que se está trabajando en un contexto Bayesiano, denótese por $p(\mathcal{H}_K|X)$ la probabilidad de observar el modelo dada la matriz de datos (de abundancia) X. Del teorema de Bayes,

$$p(\mathbb{H}_K) \propto p(\mathcal{H}_K)p(X_K|\mathcal{H}_K),$$

donde $p(\mathcal{H}_K)$ es la probabilidad previa por un modelo de K clusters y $p(X|\mathcal{H}_K)$ es la evidencia a favor de tal modelo. Dicho de otra manera, la primera es la preferencia del investigador por el modelo dado, mientras que la segunda es la preferencia *de los mismos datos* por el mismo modelo. Ésta a su vez está dada por

$$p(X|\mathcal{H}_K) = \int p(X|Q,\mathcal{H}_k)p(Q|\mathcal{H}_K)\mathrm{d}Q.$$

Entonces, el criterio que se busca para seleccionar el número de componentes adecuado es *escoger el K máximo a partir del cual los incrementos en la preferencia $p(X|\mathcal{H}_K)$ sean despreciables*; i.e., a partir del cual incluir más clusters aporte muy poco. Lamentablemente, la integral no puede ser resuelta analíticamente, pero sí puede ser estimada usando la aproximación de Laplace

$$\log p(X|\mathcal{H}_K) \approx \log p(X|\hat{Q},\mathcal{H}_K) + \log p(\hat{Q}|\mathcal{H}_K) + \frac{M}{2}\log 2\pi - \frac{1}{2}|H|,$$

donde M es el número de parámetros en Q, $\hat{Q}$ son los parámetros que maximizan la distribución posterior, y H es la matriz Hessiana del negativo de la posterior en $\hat{Q}$:

$$H = -\nabla^2 \log p(\hat{Q}|X).$$

Tomando el negativo de la log-aproximación, el criterio resulta comparable con el de Akaike (AIC) y Bayes (BIC).