Actividad:

Elaborar un ensayo sobre el problema “big p – little n”

Desarrollo:

En general, cuando trabajamos en problemas de predicción, en la cual contamos con una variable que necesitamos predecir y depende de otras (***p*** variables), conocidas como predictores, contamos con una cantidad de muestras ***n*** que nos permitirán hacer uso de algoritmos de Machine Learning, cumpliendo la condición de que ***n*** es más grande que ***p***.

Sin embrago, en algunas ocasiones y dependiendo del tipo de problema, este no es el caso, es decir ***p > n***, y a veces ***p*** es mucho mayor que ***n*** (o ***p >> n***). Esto representa un “problema” adicional a resolver y se le conoce como “***big p – little n***” o “***the short, fat data problem***”.

Una de las maneras de abordar este inconveniente parte de suponer que solo una pequeña proporción de las variables ***p*** son las que efectivamente pueden explicar la variable independiente, dicha suposición puede ser una de las siguientes[1]:

• Entre las variables explicativas, del total de ***p***, sólo un pequeño número ***q*** son relevantes, conocido como ***Classic sparsity.***

• En el caso de que todas las variables explicativas sean importantes, se puede dividir el espacio para que en cualquier región local, solo un pequeño número de características sea relevante (***Locally-Low Dimension***).

• Aunque todas las variables explicativas de ***p*** sean importantes, se puede encontrar un pequeño número de combinaciones lineales de esas variables que explican la mayor parte de la variación en la respuesta. (***Nearly Dark Representation***).

La primera aproximación en la comunidad de minería de datos para solucionar el problema ***p >> n*** fue con el método ***LASSO***[2]***,*** también conocido como ***L1-norm regularization***. Este es un método de regresión penalizado que realiza simultáneamente la reducción y la selección de variables. La salida producida por el método ***LASSO*** consiste en una ruta de solución lineal por piezas, iniciando con el modelo nulo y terminando con el ajuste mínimo de cuadrados completos, ya que el valor de un parámetro de ajuste disminuye. Por lo tanto, el rendimiento del modelo seleccionado depende en gran medida de la elección de este parámetro.[3]

[1] http://www2.stat.duke.edu/~banks/218-lectures.dir/dmlect9.pdf

[2] least angle shrinkage and selection operator

[3] Kirkland, Lisa & Kanfer, Frans & Millard, Sollie. (2015). LASSO Tuning Parameter Selection. Annual Proceedings of the South African Statistical Association Conference: Proceedings of the 57th Annual Conference of the South African Statistical Association for 2015 (SASA 2015). 49-56.

La fórmula del método ***LASSO*** es[4]:

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

Cost function for Lasso regression

Otro enfoque es el de ***feature selection*** o selección de características, el cual consiste en seleccionar un subconjunto de predictores para usarlos como entrada a los modelos predictivos. Las técnicas comunes incluyen métodos de filtro que seleccionan características en función de su relación estadística con la variable objetivo (p.e., correlación) y métodos *wrapper* que seleccionan características en función de su contribución a un modelo al predecir la variable objetivo (por ejemplo, RFE[5]).

Un área en la cual se requiere recurrir a la aplicación de métodos de solución de ***p >> n*** son las ciencias ómicas, como se muestra en el artículo de Zhang et al.[6] y otra es la de la bioinformática en general, p.e. en el artículo de Wu & Ma[7]. En ambos casos recurren a dichos métodos de solución con resultados favorables, en este último se manejan valores de ***p*** > 40.000 y valores de ***n*** entre 10 y 1000.

[4] https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

[5] Recursive feature elimination

[6] Zhang, M., Zhang, D. & Wells, M.T. Variable selection for large p small n regression models with incomplete data: Mapping QTL with epistases. BMC Bioinformatics 9, 251 (2008). https://doi.org/10.1186/1471-2105-9-251

[7] Wu, C., & Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics. Briefings in bioinformatics, 16(5), 873–883. https://doi.org/10.1093/bib/bbu046