El problema de la dimensionalidad

José A. Guerrero
CEO Datrik Intelligence, S.A.

Las tres características que mejor definen al *Big Data* son el volumen, la heterogeneidad y la velocidad de generación de los datos. Existe la creencia de que, en lo que respecta al volumen, siempre 'más datos es mejor', y suele ser cierto, aunque debemos matizar dicha aseveración: Un mayor número de observaciones redunda en un mejor modelo, pero un mayor número de variables no necesariamente lo hace.

Para entender el problema debemos reflexionar sobre el proceso de ajuste de un modelo predictivo. La palabra que mejor define este proceso es 'equilibrio'. Equilibrio entre la complejidad y la capacidad de generalización. Un modelo excesivamente complejo será capaz de captar toda la información pero inevitablemente hará lo mismo con el ruido existente. Como resultado, las predicciones que se realicen sobre un nuevo conjunto de datos serán mediocres. A esto se le denomina *overfitting* o **sobreajuste**.

Para evitar el sobreajuste fijaremos una serie de restricciones a la complejidad de los modelos, de forma directa, limitando la profundidad de los árboles de clasificación o regresión, o indirecta, añadiendo un factor de penalización que se sume al término de error de ajuste, como se hace en Regresión de Ridge o en máquinas de vectores de soporte. Entendemos por **regularización** a la inclusión de cualquier tipo de restricción a la complejidad de un modelo, y la selección del nivel de estas penalizaciones se determina mediante validación cruzada.

El **problema de la dimensionalidad** lo definiremos como los potenciales efectos negativos derivados del aumento del número de variables respecto a las observaciones. Una gran dimensionalidad se traduce frecuentemente en sobreajuste, ya que aumentan los grados de libertad del sistema. También, en el caso de que haya colinealidad entre los predictores, puede impactar en la convergencia de los algoritmos y la estabilidad de las soluciones. La regularización suele ser necesaria para acotar el problema, aunque a veces no es suficiente, necesitándose actuaciones directas sobre la dimensión de los predictores.

El problema de la dimensionalidad lo definiremos como los potenciales efectos negativos derivados del aumento del número de variables respecto a las observaciones

Realizaremos a continuación una revisión de los métodos de abordaje del problema. Estos se dividen en métodos de selección y de extracción. Los primeros seleccionan un subconjunto de variables con cierto criterio, mientras los segundos generan nuevas variables mediante una transformación de las originales.

MÉTODOS DE FILTRADO

Se encuentran entre los métodos de selección y se caracterizan por utilizar un criterio para seleccionar las variables que es independiente del algoritmo con el que se ajuste el modelo. Los ejemplos más básicos son los que utilizan criterios univariables, como la correlación o asociación de cada predictor con la variable respuesta, criterios de información mutua o test de contraste de hipótesis, uni o multivariables, para calcular el nivel de significación de la relación entre las variables.

Estos métodos tienen la ventaja de que son rápidos de ejecutar. Entre sus inconvenientes están que, en el caso de los univariables, rechacemos una variable cuyo efecto principal no sea significativo pero que pudiera tener una interacción con otra variable. Otra situación frecuente es que se incluyan en la selección varios predictores que estén muy correlacionados entre sí. Para evitar esto último se han desarrollado dos métodos que tienen en cuenta no solo la relación de las variables con la respuesta sino con el resto de variables.

Una de estas técnicas es **mRMR (mínima redundancia, máxima relevancia).** El objetivo es medir no solo la información mutua entre la variable y la respuesta (relevancia) sino también la información mutua entre los predictores (redundancia).

Otro método que comparte concepto aunque utiliza diferente técnica para su resolución es **QPFS**

(*Quadratic Programming Feature Selection*). La idea es transformar la selección de variables en un problema de optimización cuadrática, donde los términos de segundo orden corresponden a la interrelación de los predictores (matriz de covarianzas) y los términos lineales a las correlaciones con la variable respuesta.

MÉTODOS ENVOLVENTES

Son otros métodos de selección que, a diferencia de los métodos de filtrado, intentan seleccionar un subconjunto de variables que obtenga el mejor ajuste posible con un algoritmo determinado. Para intentar que estos subconjuntos sean lo más generalizable posible se utilizan algoritmos base que sean robustos y cuyos parámetros sean fáciles de sintonizar, en concreto *random forest* o modelos lineales.

Los clásicos **procedimientos de *stepwise*** se encuadrarían en esta categoría. Partiendo de todas las variables se determina en cada paso el candidato que menos contribución aporta al modelo, y se elimina. Alternativamente se puede construir de forma creciente, comenzando sin ninguna variable e incluyendo en cada iteración la que más aporte. El criterio para decidir qué variable entra o sale puede ser la importancia relativa por permutación en el caso de *random forest* o un F-test (Snedecor), basado en la descomposición de la suma de cuadrados (varianza explicada) en el caso de regresión múltiple.

Boruta es un método que utiliza *random forest* como algoritmo subyacente. La idea es generar en cada iteración una serie de variables sombra a partir de los predictores, copiando cada uno de ellos y permutando entre sí los elementos de cada nueva columna. Se ajusta un modelo por *random forest* y se calculan las importancias relativas de cada variable. Si una variable sistemáticamente queda por debajo de las sintéticas (ruido), será indicativo de que su aportación al modelo será dudosa y por tanto se elimina. El proceso continua hasta que todas las variables son aceptadas, rechazadas o se alcanza un número de iteraciones límite.

Otra alternativa que podemos encuadrar en esta categoría es la ***Regresión Lasso***. Cuando hacemos una regresión múltiple con un gran número de variables los coeficientes tienen tendencia a aumentar de tamaño, lo cual es signo de sobreajuste. La regresión con regularización de Ridge utiliza un término de penalización sobre la norma L2 de los coeficientes de manera que se controla dicho sobreajuste. La regresión de Lasso es similar a la de Ridge, pero utiliza la norma L1 para la penalización. Uno de los efectos que tiene este cambio es que conforme aumentamos el factor de penalización, algunos de los coeficientes de las variables se optimizan en cero, y a partir de ese punto, si seguimos aumentando el factor, no vuelven a tomar valores no nulos. Esta interesante propiedad hace que la regresión de Lasso se utilice más como método de selección de variables más que como un modelo propiamente dicho.

MÉTODOS DE EXTRACCIÓN

El enfoque de los métodos de extracción es radicalmente distinto: transformar el conjunto de variables iniciales en un conjunto de menor dimensión que sea capaz de retener la mayor parte de información.

Los métodos de extracción tienen una gran ventaja y un pequeño inconveniente. La ventaja es que, como veremos, no utilizan información de la variable respuesta, y por tanto se pueden utilizar datos no etiquetados (análisis semisupervisado). Haciendo esto conseguimos una mejor representación de la información, ya que los modelos de extracción aprenden relaciones entre predictores que luego serán de utilidad para la fase del modelado predictivo.

A cambio, el inconveniente es que tras la transformación de los datos estos dejan de tener cualquier interpretación sencilla, ya que de hecho ni siquiera van a tener expresiones funcionales de las variables originales, salvo las combinaciones lineales que obtendremos en PCA.

- **PCA (Análisis de componentes principales)** es sin duda el método clásico más conocido. De forma intuitiva, dado un conjunto de puntos en un espacio multidimensional, PCA realiza un cambio del sistema de coordenadas de manera que las primeras dimensiones en dicho sistema recojan la mayor variabilidad posible de los datos. PCA asume que los datos siguen distribuciones normales y que tienen una representación lineal en cierta base. En la práctica la relajación de dichas hipótesis suele funcionar relativamente bien.
- **tSNE (*t-distributed stochastic neighbor embedding*)** es otro más reciente método para reducción de dimensionalidad. A diferencia de PCA, es una técnica no lineal, cuyo objetivo es que puntos similares o próximos en el espacio multidimensional queden repre-

sentados como próximos en el espacio reducido. tSNE modela dos distribuciones de probabilidad en ambos espacios y minimiza la divergencia de Kullback–Leibler de estas distribuciones. El resultado es una potente herramienta que suele obtener mejor separabilidad que PCA en proyecciones gráficas en dos o tres dimensiones y que mejora también su rendimiento cuando se utiliza como extractor de variables para modelos predictivos.

- **K-means** es de sobra conocido como un método de *clustering* por su potencia y escalabilidad, pero su aplicación como técnica de reducción de dimensionalidad no lo es tanto. Los métodos de *clustering* actúan agrupando casos en base a una medida de similitud o una métrica. Una vez que los clústeres están calculados podemos representar cada observación por su distancia, o una función de esta, a los centroides de los mismos. En la práctica obtenemos una reducción no lineal de la dimensionalidad. El uso de la distancia es suficiente para la aplicación de técnicas basadas en árboles, dado que estas son invariantes respecto de transformaciones monótonas de los predictores. Sin embargo, para la aplicación de otros métodos de ajuste, será necesaria la transformación de las distancias. Podemos, por ejemplo, utilizar el inverso de la distancia al cuadrado y posteriormente normalizar respecto a la suma de los valores para cada observación. El resultado final es una representación con tanta dimensiones como clústeres y que puede ser interpretada como una ponderación en base a los centroides de estos.

Las redes neuronales son actualmente, junto a los métodos basados en árboles, uno de los grupos de técnicas más utilizadas en aprendizaje automático. Los ***autoencoders***, una de sus posibles configuraciones, se pueden utilizar para la reducción de dimensionalidad. Básicamente un *autoencoder* consiste en una red neuronal cuya entrada es igual a la salida, e internamente tiene varias capas ocultas, la central de las cuales tiene un número de neuronas sensiblemente inferior a la de la entrada. Esta estructura fuerza la reconstrucción de los datos de entrada después de pasar por la capa central, de manera que las neuronas centrales deben aprender a capturar el máximo de información posible. En una variante de esta estructura, los ***denoising autoencoders***, se añade ruido aleatorio a los datos de entrada manteniéndose el objetivo de recuperar la señal previa a esta perturbación. Con esta técnica se consigue que la codificación de la información sea más robusta. La capa central del *autoencoder* puede ser utilizada como una nueva representación (comprimida) de los datos.

Un modelo bastante extendido para la clasificación de textos es **LDA (*Latent Dirichlet Allocation*).** La hipótesis de LDA es que cada texto pertenece con una distribución de probabilidad a un conjunto de categorías, y que la probabilidad de aparición de las palabras en un texto depende de dichas categorías. La distribución a priori de las categorías sigue una distribución de Dirichlet y de ahí deriva el nombre. En cierta forma LDA es una reducción de dimensionalidad, ya que la representación de un texto como *BoW (bag of words)* se traduce a una ponderación del conjunto de categorías que suele ser varios órdenes de magnitud menor. Aunque el método surge y se aplica en *text mining*, nada impide usarlo con datos que puedan ser representados matricialmente de una manera similar (matrices binarias o matrices de frecuencia, por ejemplo).

Incluso ***random forest*** se puede utilizar también como técnica para la reducción de dimensionalidad. La idea es utilizar el nodo final en el que recae cada observación como una variable categórica más: una nueva variable por cada árbol con tantos niveles como nodos finales. Limitando el número de nodos a un número razonable y fijando el número de árboles como la dimensión final deseada obtendremos una nueva representación del conjunto de datos.

REFLEXIÓN

Hemos visto hasta aquí que no solo podemos abordar el problema con técnicas específicas, sino que, con ciertas dosis de imaginación, es posible usar métodos que originariamente se pensaron para otros problemas.

Ideas como regularizar los coeficientes de una regresión multivariable de una manera no homogénea, sino teniendo en cuenta aspectos como la dependencia entre sí de las variables (regularización por bloques), *subsampling* de variables y ensamblado de los modelos resultantes, *clustering* de variables como alternativa al *subsampling* aleatorio..., pueden leerse en foros especializados; algunas con más fundamento, otras más alocadas, pero en definitiva plantean retos a la comunidad científica que sin duda harán que en el futuro próximo dispongamos de más herramientas para abordar el problema de la dimensionalidad y, por extensión, del *Big Data*.