Tecnológico de Monterrey

Aprendizaje automático

Selección de variables y regularización

Dr. Omar Mendoza Montoya

14/08/2024

Reducción de dimensionalidad es el proceso de representar los datos originales en un espacio de menor dimensionalidad sin que haya una pérdida de información significativa.

Razones para reducir la dimensionalidad

- Simplificación del modelo.
- Reducción del tiempo de entrenamiento.
- Eliminar información ruidosa.
- Remover dependencias entre variables.
- Evitar la maldición de la dimensionalidad.
- Reducir el sobreajuste y mejorar la capacidad del modelo de generalización.

La maldición de la dimensionalidad

- La maldición de la dimensionalidad hace referencia a diversos problemas que surgen cuando se trabaja con modelos matemáticos en espacios de alta dimensionalidad (cientos o miles de variables).

- En espacio de alta dimensionalidad, se requieren una enorme cantidad de observaciones para entrenar apropiadamente un modelo.

Como regla más o menos aceptada, se deben tener al menos entre 3 y 5 observaciones por cada dimensión en el modelo.

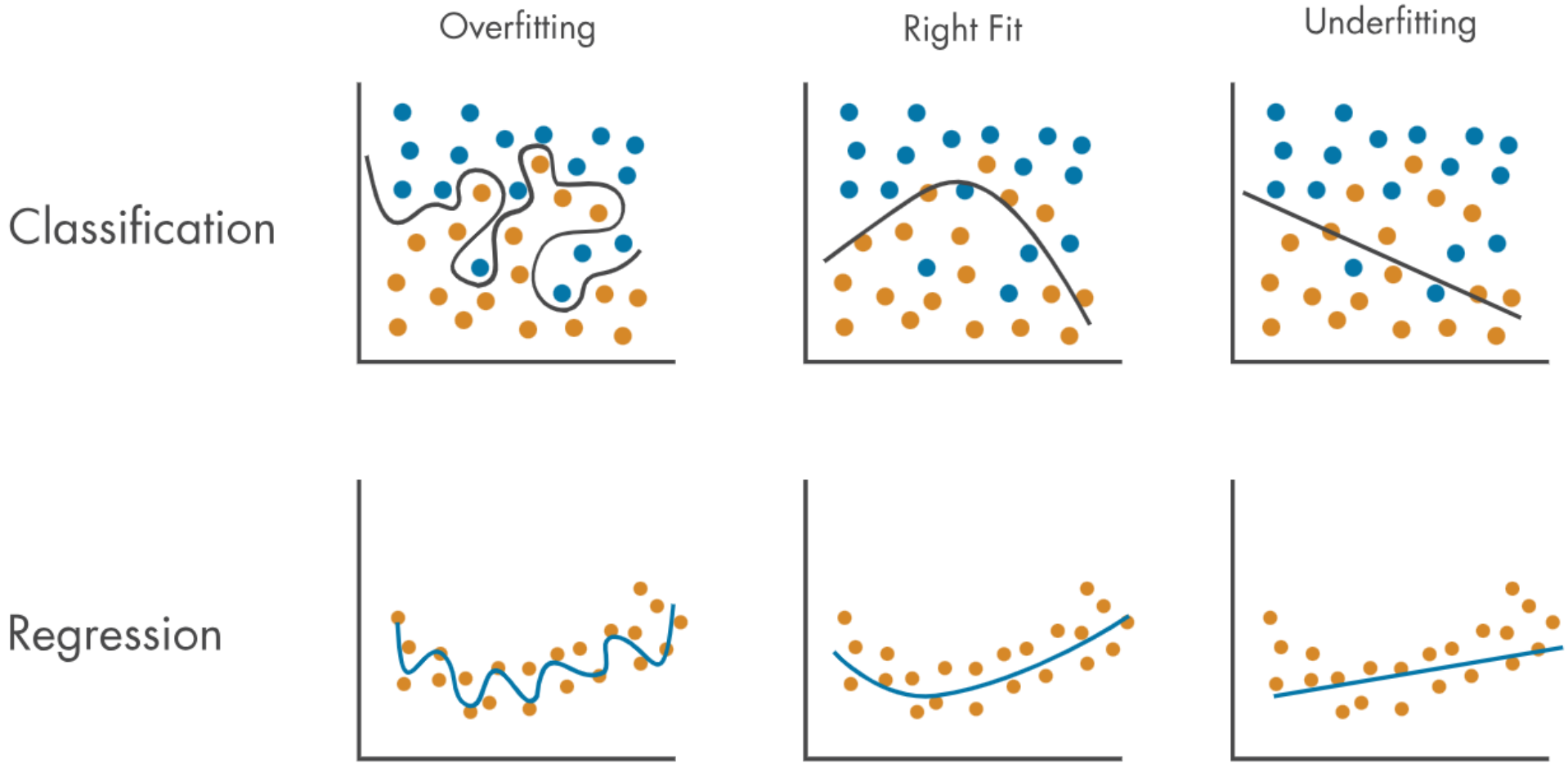
Sobreajuste

- Sobreajuste se refiere a la situación en que el modelo es demasiado bueno en hacer predicciones correctas con el conjunto de entrenamiento, pero tiene un rendimiento muy pobre con datos nuevos.

- Esto puede ocurrir cuando hay demasiadas variables en un problema, o el modelo tiene una cantidad considerable de parámetros para ajustar.

- También puede ocurrir por una mala selección del conjunto de entrenamiento, el cual no representa a la población de las observaciones posibles.

Sobreajuste



https://www.mathworks.com/discovery/overfitting.html

Métodos de reducción de dimensionalidad

- Transformación de características. Los datos pasan por una función transformadora que mapea las observaciones del espacio $\mathcal{R}^p$ a otro espacio $\mathcal{R}^q$ $(q < p)$.

  Análisis de componentes principales (PCA), Kernel PCA, LDA, análisis de correlación canónica (CCA), mínimos cuadrados parciales (PLS), etc.

- Selección de características. Se determina un subconjunto de características del total que se tienen disponible para formar parte del modelo de aprendizaje, descartando el resto.

  Filters, wrappers, filter-wrappers, embebidos.

Filters

- En los métodos tipo filters, el conjunto de variables se ordena de acuerdo a algún criterio que evalúa individualmente cada una y se seleccionan las mejores k variables.

- El criterio indica algo que nos interesa sobre el problema de aprendizaje que se desea resolver (mayor correlación con la variable a predecir, menor entropía, mejor separabilidad).

- Usualmente el criterio es independiente al modelo de clasificación o regresión a utilizar.

Medidas para ordenar variables

- Índice de separabilidad de Fisher
- Correlación entre cada variable con la variable de respuesta
- Índice Davies-Bouldin
- Entropía
- Ganancia en la información

Para regresión, el índice de correlación entre cada variable y la variable de respuesta es una buena medida para hacer la selección con la técnica de Filters.

Wrappers

- En los métodos tipo wrappers, se prueban diferentes combinaciones de variables para determinar el mejor subconjunto (problema NP).

- Si el conjunto de variables es demasiado grande, se utilizan métodos de optimización basados en heurísticas (recocido simulado, el algoritmo genético).

- Otras estrategias están basadas en búsqueda voraz (stepwise feature selection o sequential feature selection).

La evaluación del rendimiento del modelo con cada conjunto de predictores se debe realizar particionando el conjunto de entrenamiento en dos subconjuntos, uno para entrenar cada modelo, y otro para evaluar el error de los modelos.

Filter-Wrappers

- En los métodos tipo filter-wrappers, se combinan las dos estrategias anteriores:
  1. Las características se ordenan usando alguna métrica tal como se hace con filters.
  2. Se comienza con el clasificador con el mejor predictor de acuerdo al ordenamiento.
  3. Se agrega la siguiente mejor variable al modelo, y se prueba si el modelo tuvo una mejora. Si es el caso, se repite este paso, de lo contrario, se termina con el conjunto que se tiene sin la última variable agregada.

- Es posible realizar la búsqueda comenzando con todas las variables, y luego eliminar una por una hasta que no se observen mejoras.

Métodos embebidos

- En los métodos embebidos, el conjunto de características se encuentra en el algoritmo de aprendizaje. Es decir, la tarea de encontrar los parámetros del modelo y el subconjunto de características se lleva a cabo al mismo tiempo.

- Ejemplos de este tipo son métodos con funciones de costo con un término de regularización.

- Ejemplos de estos métodos son la regresión Ridge y Lasso.

Regularización

- En el caso de regresión, se modifica la función de costo de mínimos cuadrados:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

por

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda R(\beta)$$

$\lambda =$ parámetro de regularización.

$R(\beta) =$ Función que penaliza la estructura del modelo.

Regresión Ridge y Lasso

- Regresión Ridge:

$$MSE^{Ridge}(D,\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{p} \beta_i^2$$

Regresión Lasso:

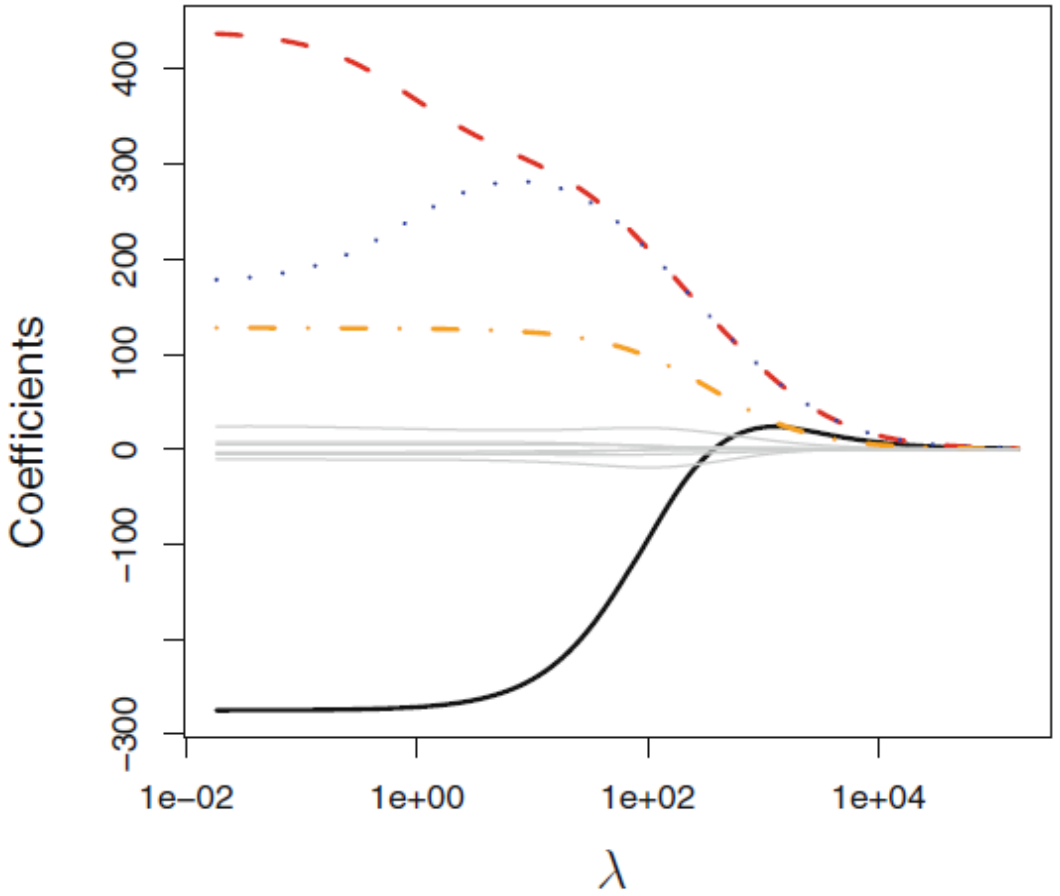
$$MSE^{LASSO}(D,\beta)\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i;w)\right)^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

En regresión Ridge y Lasso, cuando la constante de regularización λ es alta, los coeficientes del modelo tienden a un valor de cero.

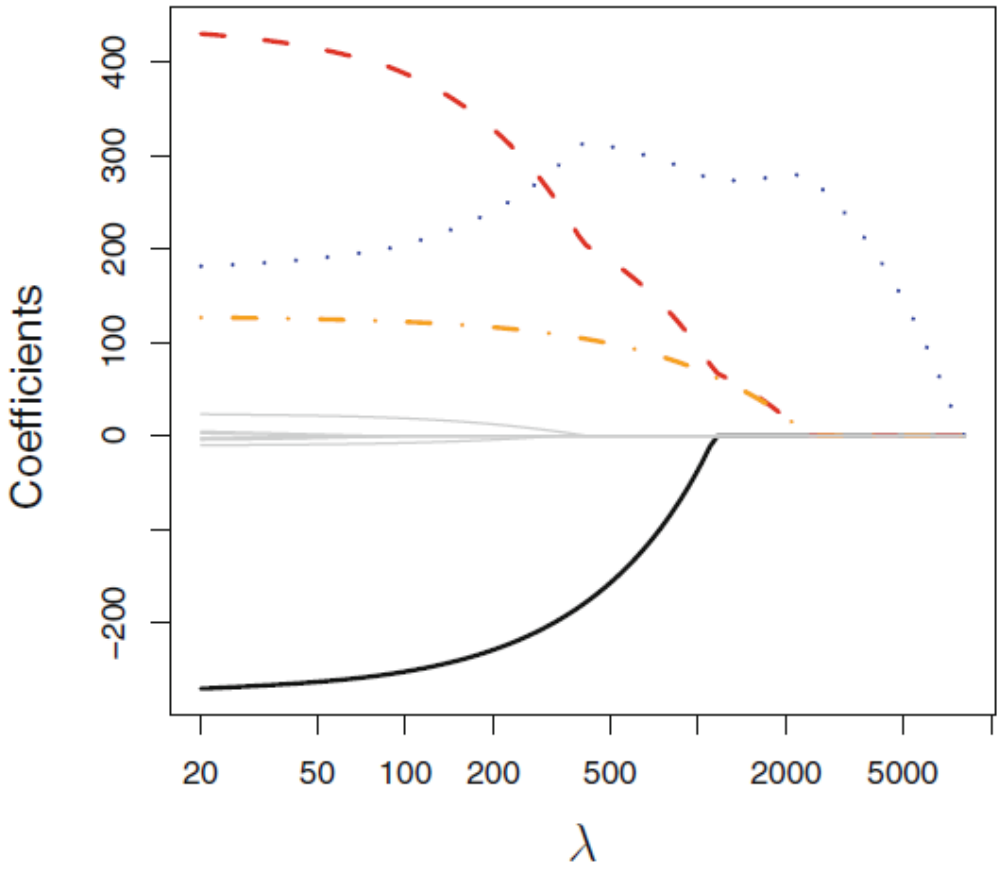
Solo aquellas variables importantes tienden a sobrevivir cuando se incrementa la constante de regularización.

Graficos Ridge y Lasso

Ridge regression



Lasso



Hastie, Tibshirani, Friedman, 2009

Optimización para regresión Ridge

- Para regresión Ridge, es posible utilizar descenso de gradiente para encontrar los parámetros del modelo.

$$\nabla MSE^{Ridge}(D, \beta) = -\frac{2}{n}\sum_{i=1}^{n} r_i x_i^T + 2\lambda\beta$$

- Regla de aprendizaje:

$$\beta \leftarrow (1 - 2\alpha\lambda)\beta + \alpha\left(\frac{2}{n}\sum_{i=1}^{n} r_i x_i^T\right)$$

¿Cómo debe evaluarse el rendimiento de un modelo con validación cruzada cuando hay selección de características?

Consejos para la reducción de dimensionalidad

- La selección de características debe realizarse con el conjunto de entrenamiento en todo momento.
- El conjunto de prueba nunca debe estar involucrado en la reducción de dimensionalidad.
- En caso de validación cruzada, se debe realizar la selección de características en cada partición con los subconjuntos de entrenamiento.

Sin elección de características

- Por cada partición aleatoria
  - Entrenar con el subconjunto de entrenamiento.
  - Evaluar con el subconjunto de prueba.
- Reportar el rendimiento promedio de todas las particiones.

Con selección de características

- Por cada partición aleatoria
  - Seleccionar características con el subconjunto de entrenamiento.
  - Entrenar con el subconjunto de entrenamiento.
  - Evaluar con el subconjunto de prueba.
- Reportar el rendimiento promedio de todas las particiones.

Cuando se tenga evaluado el modelo, podemos utilizar el conjunto de datos completo para seleccionar características y entrenar el modelo. Este modelo es el que se utilizaría en producción.

Bibliografía

- James , G., Witten, D., Hastie, T. & Tibshirani, R. (2023). *An introduction to statistical learning: with applications in Python* (2da ed.). Springer.
  - Capítulo 6

- Hastie , T., Tibshirani , R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2da ed.). Springer.
  - Capítulo 3
  - Capítulo 7