Subject: Probability and Statistics

Class XIII: Validación de modelos y Regresión regularizada

UdeA

Bioengineering

Francisco José Campuzano Cardona

Bioengineerer, MSc in Engineering

Model Validation

UdeA

Antes se mencionó que la precisión de un modelo se podía medir buscando minimizar el error, usando métricas como el MAE, RMSE o maximizando el R2. Ampliemos un poco el tema de validación de modelos en general, no solo de regresión.

1. **Cómo no validar un modelo:** Usar los datos de entrenamiento para la validación es una forma incorrecta de validar un modelo. Evidentemente un modelo será más preciso frente a los datos que "vio" comparado con los que "no vio", en su entrenamiento.

2. **Como sí validarlo:** La validación siempre debe usar dato independientes de los datos de entrenamiento, se debe dejar siempre un set de datos aparte para la validación. O se puede realizar validación cruzada *k-folds.*

UdeA

¿Cómo se selecciona el mejor modelo?

La respuesta a esta pregunta no es trivial, y como resultado aparece otro pregunta. Si mi estimador tiene un mal desempeño, ¿qué debo hacer? Las siguientes podrían ser respuestas:

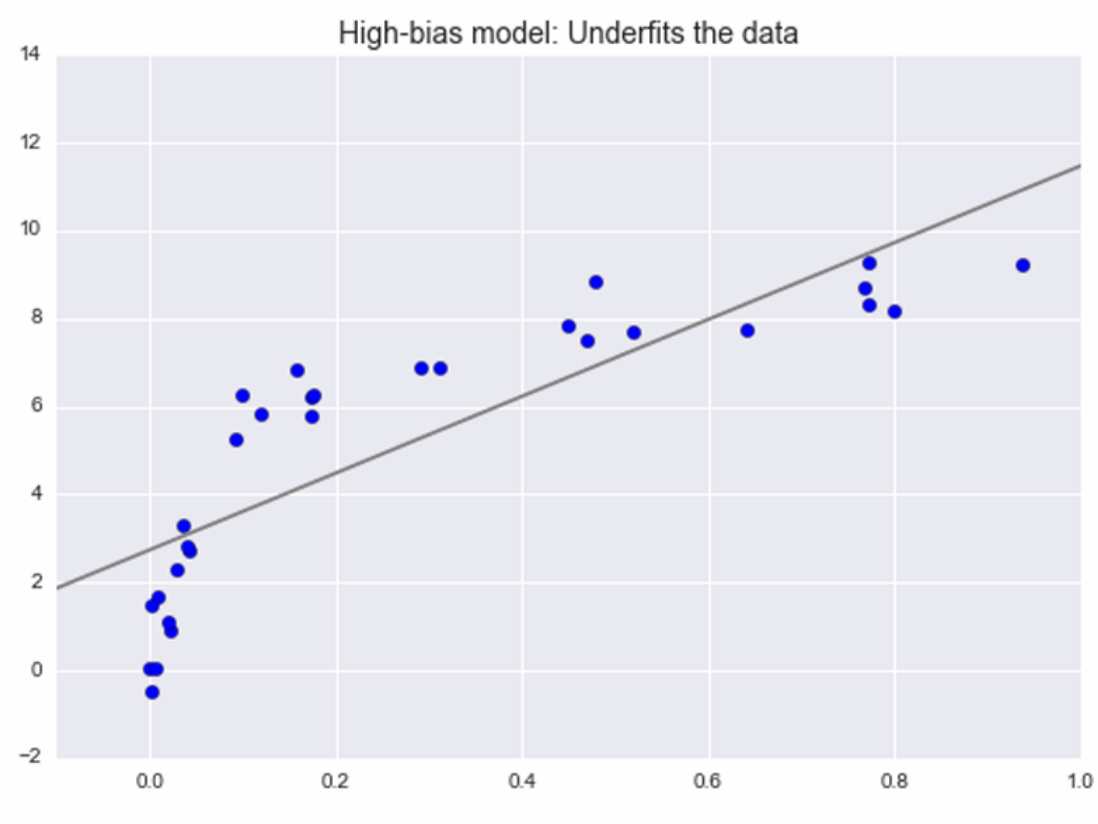
- Usar un modelo más complicado
- Usar un modelo menos complicado
- Usar más datos para el entrenamiento
- Usar más variables descriptoras o características.

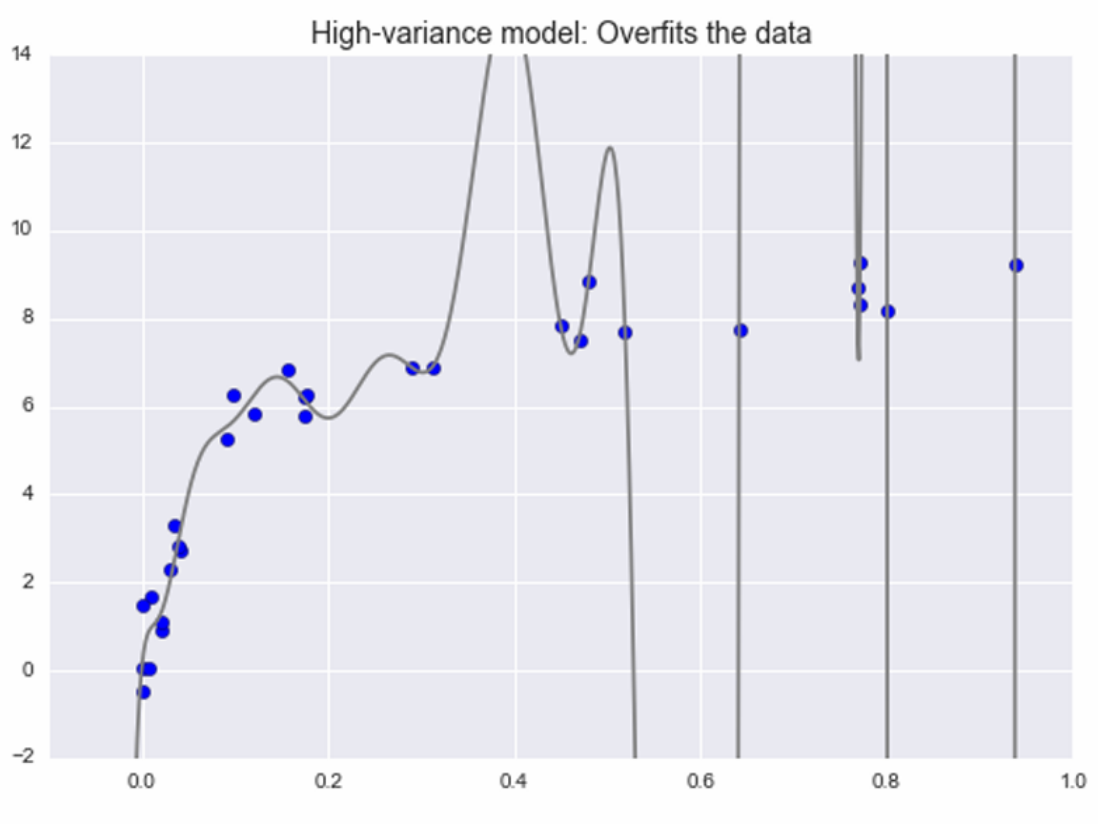
Pero cuál es la respuesta, depende.

Model validation

UdeA

Veamos: Ninguno de los dos modelos se ajusta bien.





Model Validation

UdeA

Ahora incluyendo los datos de prueba...

Entonces:

Cuando un modelo está infraajustado, las métricas de precisión son igual de malas si se calculan con los datos de entrenamiento o con los de prueba. También se llama modelo de alto sesgo

Cuando el modelo está sobreajustado, las métricas de precisión calculadas con los datos de entrenamiento son buenas, pero muy malas cuando se calculan con los datos de prueba. También se llama modelo de alta varianza.

Curva de Validación

UdeA

Entonces:

Si variamos el modelo desde uno sesgado hasta uno de alta varianza, y analizamos el comportamiento de las métricas de precisión, obtendremos, que la curva de la métrica de precisión de los datos de prueba deberían tener un máximo, qué sería el mejor modelo.

Veamos un ejemplo con sklearn

Curvas de Aprendizaje

UdeA

Las curvas de validación pueden cambiar conforme lo hace el número de datos de entrenamiento. Entonces para un modelo con una complejidad fija, se puede mostrar cómo es el comportamiento del Score del modelo en relación al numero de datos de entrenamiento, lo que se conoce como curva de aprendizaje.
Veamos en Sklearn.

Regularización

UdeA

El sobreajuste como vimos se da por modelos muy complejos, o de alta varianza, y esta complejidad se da por pesos grandes de los *features,* es decir, coeficientes muy grandes. Cuando uno o más coeficientes son demasiado altos, la salida del modelo se vuelve sensible a alteraciones menores en los datos de entrada

Entonces obtener modelos más simples, se puede lograr haciendo que los coeficientes sean menores, lo que se conoce como regularización. Para esto, se debe recordar que el modelo se construye buscando minimizar el error residual, y para lograr esto, los modelos pueden inflar ciertos coeficientes, para ajustarse a los datos, lo que genera un sobreajuste.

UdeA

Regularización Ridge.

En este caso lo que se busca es adicionar un valor de penalización en el cálculo de error, y esta penalización depende del cuadrado de la suma de los coeficientes. Lo que obliga a que el modelo deba disminuir el valor de los coeficientes, para lograr minimizar el error.

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{P} (\beta_j)^2$$

UdeA

Regularización Lasso

La regularización Lasso se hace de la misma forma pero el valor de penalización se hace con el valor absoluto de los coeficientes.

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{P} |\beta_j|^2$$

UdeA

Entonces, ¿qué se logra con cada tipo de regularización?

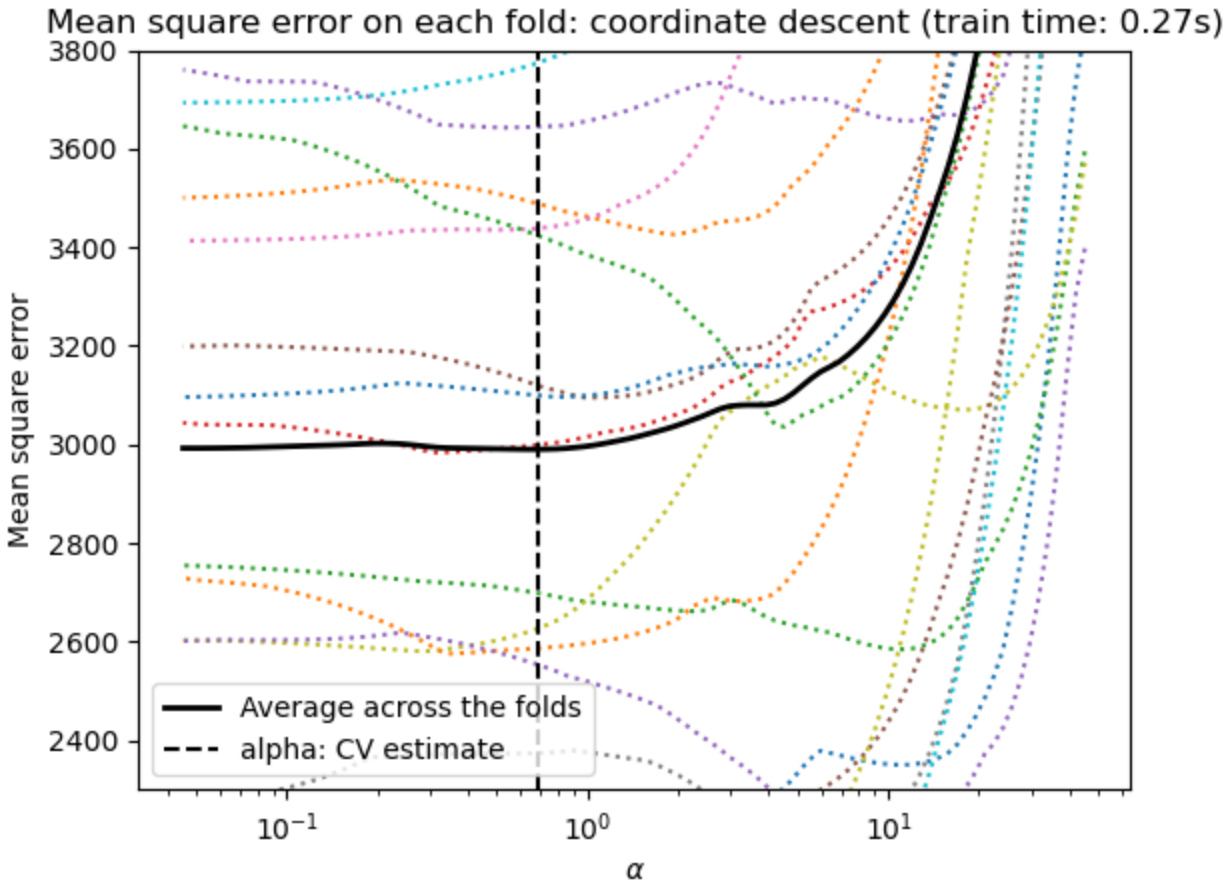
La regularización Ridge, mantiene todos los coeficientes, pero con un menor peso, lo que genera que el modelo sea más simple, aunque tenga un poco más de error. Así el modelo baja el sobreajuste.

La regularización Lasso logra disminuir hasta cero algunos coeficientes, lo que hace que salgan del modelo algunas características. Esto resulta en un modelo más simple, no solo por tener coeficientes más bajos, sino menos regresores.

UdeA

Y ¿cómo se selecciona alfa?

Esta valor puede seleccionarse con varios métodos, por ejemplo, con una validación cruzada por *k-folds*. En este caso, se determina el error para cada *fold* y cada alfa, y se promedian las curvas obtenidas, y el alfa que causa el menor valor de error promedio, es el alfa elegido.



Subject: Probability and Statistics

UdeA

Bioengineering

¡Thanks!

Francisco José Campuzano Cardona

Bioengineering. MSc in Engineering