Tempo restante 0:58:27

Pregunta **1**

Aínda non respondido

Puntúa como 5,00

Razona si la siguiente afirmación es verdadera o falsa:

Queremos predecir la edad de una persona a partir de la información obtenida de un escáner cerebral utilizando regresión. En la práctica sólo disponemos de 10 individuos para cada uno de los cuales registramos su edad y su actividad cerebral medida en 20000 regiones del cerebro. En este caso sería preferible utilizar un modelo de regresión lineal múltiple en lugar de un modelo de regresión Lasso.

FALSO.

En este caso estaríamos ante un problema muy claro de altas dimensiones, al contar con 20.000 variables predictoras (p) pero sólo 19 individuos (n). En este tipo de problemas, el modelo obtenido mediante un método de regresión lineal demasiado simple y es muy probable que ocurra *overfitting*, lo que nos va a llevar a un rendimiento muy pobre del modelo por con un conjunto de test independiente.

También se nos puede presentar el problema de la multicolinearidad entre variables predictoras, debido al elevado número de estas que aumenta la probabilidad de que ocurra, más aún teniendo en cuenta que la actividad en diversas regiones del cerebro se puede activar por la misma causa.

Por esto es mejor usar un modelo de regresión Lasso en lugar de un modelo de regresión múltiple. Esto se debe a que mediante Lasso un gran número de estas 20.000 variables predictoras, aquellas que sean poco significativas, tenderán a 0. De esta manera se reducirá en gran medida la dimensión del problema con el que estamos tratando, pasando a un ajuste de un modelo más simple.

Pregunta **2**

Aínda non respondido

Puntúa como 5,00

Disponemos de una muestra de $n = 200$ observaciones de una variable respuesta cuantitativa Y y una variable predictora X. Supongamos que en realidad la relación entre X e Y es polinómica de orden 3 (es decir, un polinomio de orden 3 ajusta perfectamente a la nube de puntos). Explica como se comporta en términos de sesgo y varianza un ajuste lineal.

El sesgo se refiere al error introducido por aproximar mediante un modelo más simple (menos flexible) que el modelo real, por lo tanto, a mayor flexibilidad menor será el valor del sesgo.

Como un ajuste lineal es más simple que un ajuste polinómico de orden 3, el valor del sesgo será mayor que el que sería si usásemos un modelo más complejo que se ajustase más al modelo real.

En términos de varianza:

La varianza se refiere a la cantidad en la que cambiaría el modelo descubierto si usásemos un conjunto de entrenamiento diferente, por lo tanto, a mayor flexibilidad del modelo que estemos usando para estimar la variable de respuesta **Y** a partir de la variable predictora **X**, mayor será el valor de la varianza.

Como un ajuste lineal tiene una flexibilidad menor que un ajuste polinómico de orden 3, la varianza para nuestro ajuste lineal será menor que si estuviésemos intentando descubrir un ajuste polinómico de orden 3, de mayor flexibilidad.

En conclusión, el sesgo pesará mucho más que la varianza para nuestro ajuste lineal.

Pregunta **3**

Aínda non respondido

Puntúa como 10,00

El conjunto de datos Auto, incluido en la librería *ISLR* contiene información correspondiente a consumo de combustible, potencia y otros datos técnicos sobre automóviles. Utiliza la función *lm* para ajustar un modelo de regresión lineal múltiple que explique el consumo *mpg* en función de las variables *cylinders, displacement, horsepower* y *weigh* utilizando las primeras 300 observaciones (reserva las 92 observaciones restantes como muestra test). Contesta razonadamente a las siguientes preguntas:

1. ¿Qué predictores son estadísticamente significativos?
2. Calcula el error cuadrático medio (MSE) para la muestra de entrenamiento y la muestra test obtenidos con el ajuste lineal.
3. ¿Cuáles son los coeficientes estimados al ajustar un modelo de regresión a la muestra de entrenamiento con regularización Lasso usando $\lambda = 3$? Calcula el error cuadrático medio (MSE) para la muestra test.
4. Al ajustar un modelo de regresión con regularización Lasso, ¿en qué orden se hacen cero los coeficientes del modelo ajustado al incrementar el valor de la penalización λ?

1. Respuesta:

Utilizando el siguiente código de R:

```
# Cargamos el dataset
auto <- Auto
# Dividimos en conjunto de entranamiento y conjunto de test
conjunto_train <- auto[1:300,]
conjunto_test <- auto[301:392,]
# Ajustamos el modelo de regresión lineal múltiple
z <- lm(mpg ~ cylinders + displacement + horsepower + weight, data = conjunto_train)
# Obtenemos un resumen
summary(z)
```

Que nos da la siguiente salida:

```
Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight,
   data = conjunto_train)
Residuals:
   Min      1Q  Median      3Q     Max
-9.0350 -1.9082 -0.0353  1.6746 14.6485
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.934241   1.193316  33.465   <2e-16 ***
cylinders    -0.166091   0.326780  -0.508   0.6116
displacement -0.002809   0.006940  -0.405   0.6859
horsepower   -0.023264   0.009816  -2.370   0.0184 *
weight       -0.004760   0.000546  -8.719   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pregunta **4**

Aínda non respondido

Puntúa como 5,00

Supongamos que ajustamos un modelo de regresión lineal mediante el procedimiento de estimación Ridge, para un valor determinado de λ. Explica cómo afecta a nivel de varianza en la estimación de los coeficientes, el valor de λ.

Un modelo de regresión lineal mediante el procedimiento de estimación de Ridge va a producir un conjunto diferente de los coeficientes estimados para cada posible valor de λ. El hiper-parámetro λ multiplica el sumatorio de los coeficientes estimados, al cuadrado, en la función que buscamos minimizar, por lo que cuanto mayor sea el valor de λ mayor será el número de coeficientes estimados que tiendan a 0 (*shrinkage* o contracción) y por esto el modelo descubierto será menos flexible.

Por lo tanto, cuanto **mayor sea el nivel de λ**, **menor será la flexibilidad** del modelo descubierto, lo que nos lleva a una **varianza menor**. Por otra parte cuanto **menor sea el valor de λ más flexible** será el modelo descubierto (ya que un número menor de los coeficientes estimados tenderá a 0), lo que nos lleva a un **valor mayor para la varianza** en la estimación de los coeficientes.

◄ Análisis de Componentes Principales con R

Ir a...

Evaluación y selección de modelos; métodos basados en vecinos más próximos ►