Solución a los ejercicios conceptuales del libro

An Introduction to Statistical Learning with Applications in R

Juan Rosendo González Feria

17 de marzo de 2022

Resumen

El presente archivo pretende ser un compendio de las soluciones a los ejercicios conceptuales del libro *An Introduction to Statistical Learning with Applications in R*. El libro consiste en una serie de ejercicios al final de cada capítulo. Estos ejercicios están separados en dos tipos: conceptual y aplicado. Este archivo contendrá solamente los ejercicios que se encuentran en la categoría de conceptuales, dejando los ejercicios aplicados para otro documento aparte en el que el código y el texto congenien mejor.

1. Capítulo 1: *Introduction*

El capítulo es meramente introductorio a los objetivos del libro, por lo que no hay una sección de ejercicios aquí.

2. Capítulo 2: *Statistical Learning*

Los ejercicios conceptuales del capítulo 2 son los siguientes:

1. For each of parts a) through d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

   a) The sample size n is extremely large, and the number of predictors p is small.

   b) The number of predictors p is extremely large, and the number of observations n is small.

   c) The relationship between the predictors and response is highly non-linear.

   d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

Para responder el ejercicio es necesario entender que quiere decir que un modelo sea flexible. Un modelo se dice que es flexible si para estimar $\hat{f}$ dispone de muchas formas posibles. Así los modelos de regresión lineal no son flexibles pues solo generan "rectas", mientras que en un modelo SVM $\hat{f}$ dispone de una amplia gama de formas.

Así en el caso donde se cuentan con pocos predictores y muchas observaciones es de esperar que un modelo inflexible se desempeñe mejor, pues al ser demasiadas observaciones un modelo flexible podría sobreajustar sus datos con más facilidad.

En el caso inverso es de esperar que un modelo flexible tenga un desempeño mejor, ya que al haber muchos predictores es altamente probable que la función f no tenga una forma "clara".

Por otro lado si la relación entre predictores es altamente no lineal, un modelo flexible se puede desempeñar mejor que uno inflexible al no asumir una forma parámetrica de la función f

Finalmente, si la varianza de los errores es extremadamente alta un modelo no flexible podría desempeñarse mejor que uno flexible pues estos últimos podría ajustarse de manera errónea dada la alta varianza de los errores.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Nuevamente para responder esta pregunta debemos ser capaces de distinguir las diferencias entre problema de regresión y uno de clasificación. Así también debemos saber las diferencias entre predicción e inferencia. Comencemos con la diferencia entre regresión y clasificación:

Un problema es de	
Regresión	**Clasificación**
Si la variable respuesta y es numérica	Si la variable respuesta y es categórica

Cuadro 1: Principales características entre un problema de regresión y uno de clasificación.

Ahora precisemos las diferencias entre predicción e inferencia:

El fin de un modelo es la	
Predicción	**Inferencia**
Cuando solamente queremos hacer predicciones a partir del modelo. Tipicamente los modelos son usados aquí como cajas negras.	Cuando además de buscar hacer buenas predicciones se busca conocer a fondo la relación funcional entre la variable respuesta y sus predictores, la importancia de los predictores y su relación con la respuesta.

Cuadro 2: Predicción vs Inferencia.

Realizadas las precisiones ahora ya podemos responder adecuadamente:

- *Para el primer problema se tiene que el* ***salario CEO es la variable respuesta*** *y el resto son los predictores. Al ser una variable numérica se trata de un problema de* ***regresión****. Además, al estar interesado de los factores que afectan el salario entonces el modelo se crea con el fin de hacer* ***inferencia****.*
- *En el segundo problema se tiene que el veredicto* ***éxito/fracaso es la variable respuesta****, claramente esta variable es de tipo categórico, por lo que estamos frente a un problema de* ***clasificación****. Finalmente, dado que solo nos importa si el nuevo producto será un éxito o no, el modelo se implementa con el fin de hacer* ***predicción****.*

- *El último problema tiene al* ***porcentaje del tipo de cambio USD/Euro como variable respuesta****. Ahora, al se este porcentaje una variable numérica estamos frente a un problema de* ***regresión****. Finalmente, el problema especifica que estamos unicamente interesados en hacer predicciones, por lo que el fin del modelo es hacer* ***predicción****.*

3. We now revisit the bias-variance decomposition.

   a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

   b) Explain why each of the five curves has the shape displayed in part a).