APRENDIZAJE SUPERVISADO, I: CONCEPTOS GENERALES Y LÍNEA DE REGRESIÓN

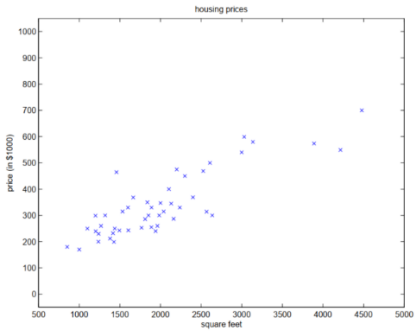
Contenido

1.- Introducción y Nomenclatura

Imaginemos que tenemos un conjunto de datos que consiste en:

- Precio de algunas casas en una zona
- Pies cuadrados (metros cuadrados) de cada uno de ellos

Living area ($feet^2$)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



Si determinamos que existe una asociación entre pies cuadrados y precio, podremos predecir el precio de otras casas en esa zona (Ng, 2012).

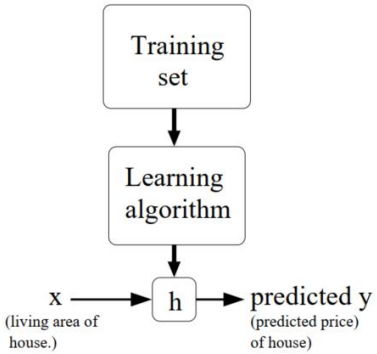
La variable de entrada son los pies cuadrados **(X)** (también conocida como **predictor, variable independiente, característica o simplemente variable**) . Cada una de las muestras de la entrada se denomina como **$x^{(i)}$**

La variable de salida **(Y)** (**también conocida como respuesta, variable dependiente o variable** objetivo) es el precio. Cada una de las muestras del output se denota como **$y^{(i)}$.**

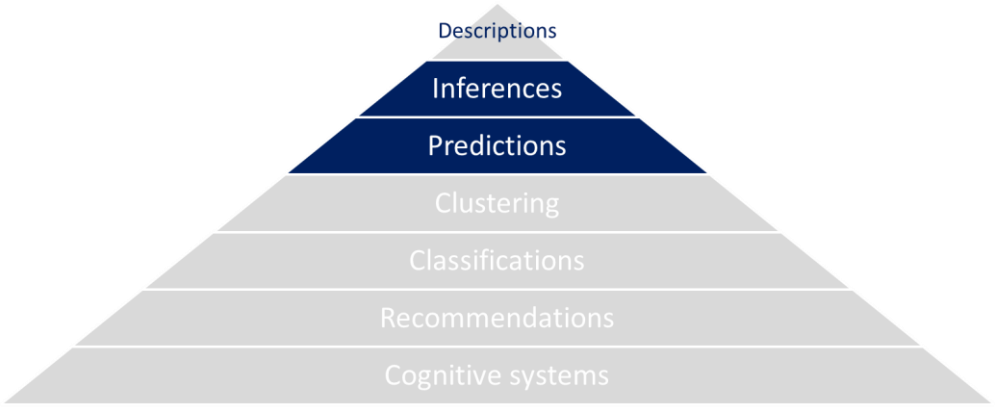
Un par **$(x^{(i)}, y^{(i)})$** se denomina **ejemplo de entrenamiento (training example),** y el conjunto de datos que usaremos para aprender: una lista de m ejemplos de entrenamiento **$\{(x^{(i)}, y^{(i)}); i = \{1, . . ., m\}$** se denomina **conjunto de entrenamiento (training set).**

Téngase en cuenta que el **superíndice** **"(i)"** en la notación es simplemente un índice en el conjunto de entrenamiento y no tiene nada que ver con la exponenciación. También usaremos X para denotar el espacio de valores de entrada, e Y el espacio de valores de salida. En este ejemplo, **X = Y = R**

Para describir el problema de aprendizaje supervisado un poco más formalmente, nuestro objetivo es, dado un conjunto de entrenamiento, aprender una función h: X → Y para que **h(x)** sea un predictor "bueno" para el valor correspondiente de y. Por razones históricas, esta función h se llama hipótesis. Visto gráficamente, el proceso es, por lo tanto, así (Ng, 2012):



Cuando la variable objetivo que estamos tratando de predecir es continua, como en nuestro ejemplo de vivienda, llamamos al problema de aprendizaje un problema **de regresión**. Cuando y puede asumir solo un pequeño número de valores discretos (como si, dada la superficie habitable, quisiéramos predecir si una vivienda es una casa o un apartamento, por ejemplo), lo llamamos un **problema de clasificación.**



El aprendizaje supervisado es un conjunto de algoritmos de aprendizaje automático que se pueden entrenar conociendo pares de entrada y salida.

2.- Generalización del problema de la regresión

De manera más general, supongamos que observamos una respuesta cuantitativa Y y p predictores diferentes, X_1, X_2, . . ., Xp. Suponemos que existe alguna relación entre Y y X = (X_1, X_2, . . . , Xp), que se puede escribir de forma muy general como (James, Witten, et al., 2013):

$$Y = f(X) + \epsilon.$$

Aquí f es una función fija pero desconocida de **X_1, . . ., X_p**, y **ε** es un **término de error aleatorio**, que es independiente de **X** y tiene promedio cero. En esta formulación, **f** representa la información sistemática que **X** proporciona sobre **Y.**

En esencia, el aprendizaje estadístico paramétrico se refiere a un conjunto de enfoques para estimar **f.**

Hay dos razones principales por las que tal vez deseemos estimar f: **predicción** e **inferencia.**

3.- Predicción: error reducible e irreducible (James, Witten, et al., 2013, p. 34)

Dado que el término de error es cero en promedio, podemos predecir Y usando una aproximación

$$\hat{Y} = \hat{f}(X)$$

donde **f^** representa nuestra estimación para f, e **Y^** representa la predicción resultante para **Y**. En este contexto, **f^** a menudo se trata como una **caja negra**, en el sentido de que uno no se preocupa típicamente por la forma exacta de f^, siempre que produzca predicciones precisas para Y.

La precisión de **Y^** como predicción para Y depende de dos cantidades, que llamaremos el **error reducible** y el **error irreducible**

Este **error es reducible** porque potencialmente podemos mejorar la precisión de **f^** mediante el uso de la técnica de aprendizaje estadístico más adecuada para estimar f.

Sin embargo, incluso si fuera posible formar una estimación perfecta para f, de modo que nuestra respuesta estimada tomara la forma **Y^ = f(X)**, ¡nuestra predicción todavía tendría algún error!

Esto se debe a que, Y es también una función de **ε,** que, por definición, no se puede predecir usando **X**. Por lo tanto, la variabilidad asociada con **ε** también afecta la precisión de nuestras predicciones. Esto se conoce como el **error irreducible**, porque no importa lo bien que estimemos f, no podemos reducir el error introducido por **ε.**

La cantidad **ε** puede contener:

- Variables medidas que son útiles para predecir **Y:** ya que no las medimos, f no puede usarlas para su predicción.
- También puede contener variaciones inconmensurables, no medibles.

Consideremos una estimación dada **f^** y un conjunto de predictores **X**, que produce la predicción **Y^ = f^(X).** Supongamos por un momento que tanto f^ como X son fijos. Entonces, es fácil demostrar que

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}, \end{aligned}$$

donde **E (Y – Y^)²** representa el promedio, o valor esperado, de la diferencia esperada al cuadrado entre el valor previsto y el valor real de Y, y Var(ε) representa la varianza asociada con el término de error **ε**

4.- Inferencia

A menudo estamos interesados en comprender la forma en que, **Y** se ve afectado por el cambio de $\mathbf{X_1, \ldots, X_p}$. En esta situación deseamos estimar **f**, pero nuestro objetivo no es necesariamente hacer predicciones para **Y**. En cambio, queremos entender la relación entre **X e Y,** o más específicamente, entender cómo **Y** cambia en función de $X_1, \ldots, X_p$. Ahora **f^ no puede ser tratado como una caja negra**, porque necesitamos saber su forma exacta.

La inferencia se trata de comprender la relación causa-efecto. Queremos saber el por qué.

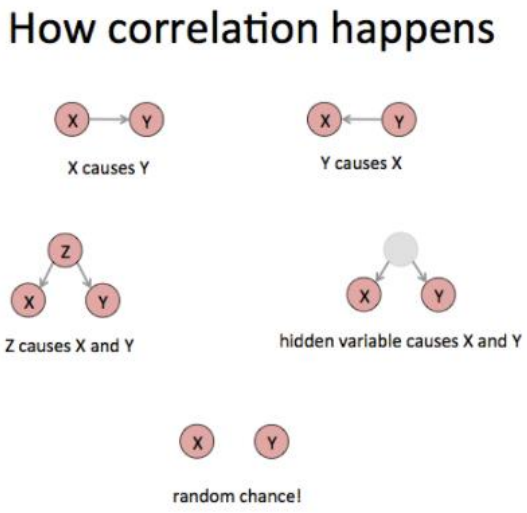
En una regresión lineal, por ejemplo, los parámetros que ponderan las entradas representan la relación estimada.

En cualquier caso, es importante distinguir la diferencia entre causalidad y correlación.

Si bien la causalidad y la correlación pueden existir al mismo tiempo, la correlación no implica causalidad. **La causalidad** se aplica explícitamente a los casos en que la acción **A** causa el resultado **B**. Por otro lado, la correlación es simplemente una relación. La acción A se relaciona con la acción B, pero un evento no necesariamente hace que ocurra el otro evento (Madhaven, 2020).

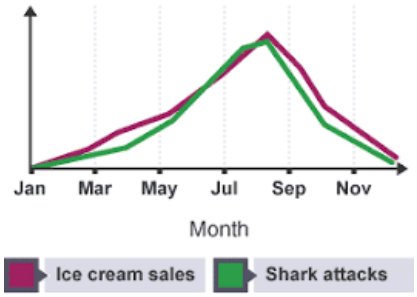
La correlación y la causalidad a menudo se confunden porque a la mente humana le gusta encontrar patrones incluso cuando no existen. A menudo fabricamos estos patrones cuando dos variables parecen estar tan estrechamente asociadas que una depende de la otra. Eso implicaría una relación de causa y efecto donde el evento dependiente es el resultado de un evento independiente. Hay muchas otras posibilidades para una asociación, incluyendo (Madhaven, 2020):

- Lo contrario es cierto: B causa A.
- Los dos están correlacionados, pero hay más: A y B están correlacionados, pero son causados por C.
- Hay otra variable involucrada: A causa B, siempre y cuando D suceda.
- Hay una reacción en cadena: A causa E, lo que lleva a E a causar B (pero sólo se vió que A causa B).



(*Por qué la correlación no es causalidad*, s.f.)

Por ejemplo (*Correlation Does Not Necessarily Mean Causation*, 2021), se realizó un estudio que encontró una fuerte correlación entre las ventas de helados y el número de ataques de tiburones para varias playas que fueron muestreadas.



(*Por qué la correlación no es causalidad*, s.f.)

Conclusión: El aumento de las ventas de helados causa más ataques de tiburones (a los tiburones les gusta comer personas que están llenas de helado).

Mejor explicación: La variable de confusión es la temperatura. Las temperaturas más cálidas hacen que las ventas de helados aumenten. Las temperaturas más cálidas también traen más personas a las playas, lo que aumenta las posibilidades de ataques de tiburones. Esto se conoce como respuesta común, donde dos variables (ventas de helados y ataques de tiburones) responden a los cambios en alguna tercera variable (temperatura).

Las pruebas de hipótesis o la experimentación A / B / nos revelan la diferencia entre la correlación a la causalidad, si bien están fuera del ámbito de esta nota introductoria

5.- Enfoques paramétricos y no paramétricos

Podemos distinguir dos enfoques diferentes: Paramétrico y No Paramétrico

- **Enfoques paramétricos**: reduce el problema a
    - Asumimos una forma de función dada. Por ejemplo, podemos suponer en nuestro ejemplo anterior que la relación entre el precio y los pies al cuadrado es lineal, siendo h_Θ nuestra salida estimada, **X** la entrada y Θ_0 **y** Θ_1 los parámetros a entrenar

      $$h_\Theta=\Theta_0+\Theta_1 X$$

    - Luego entrenamos el modelo con nuestros pares de muestras **$(x^{(i)}, y^{(i)})$[1] y calculamos el Θ_i.** En este sentido vamos a ajustar la función asumida a los ejemplos de entrenamiento.
    - El problema surge la función asuminda. Si está lejos de lo real, la estimación será muy pobre.
    - La relación y por lo tanto la función estimada podría ser muy compleja, multivariante (varias entradas) y no lineal (no solo una recta sino cualquier otra función posible). Esto conducirá al concepto de sobreajuste y a la compensación sesgo-varianza que cubriremos más adelante.
- **Enfoques no paramétricos**:
    - No hacemos suposiciones sobre la función. Solo tratamos de encontrar una estimación que se acerque lo más posible a los puntos de datos sin ser demasiado áspero o ondulante (James, Witten, et al., 2013).
    - El problema es la falta de interpretabilidad del modelo. No conocemos las relaciones de causalidad
    - De todos modos, los mismos conceptos de sobreajuste y compensación sesgo-varianza son aplicables aquí.
    - Por ejemplo, las redes neuronales pueden considerarse no paramétricas.

[1] Recuerda que Y es la real salida y **$y^{(i)}$ en nuestro caso el precio real que corresponde al real $x^{(i)}$el pies** [2]

5.- Paramétrico: modelo para ajustar la función a los datos de entrenamiento (Ng, 2012, p. 4,5)

Vamos a describir la metodología para ajustar un modelo lineal con una sola entrada o variable a unos datos de entrenamiento, pero la metodología se puede generalizar. Volvamos al ejemplo del precio-tamaño

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

Aquí, las **X** son vectores unidimensionales en R. Por ejemplo, $\mathbf{x^{(i)}}_1$ es la superficie de la i-ésima casa en el conjunto de entrenamiento. En general, al diseñar un problema de aprendizaje, dependerá del científico de datos qué características elegir. Si se está recopilando datos de vivienda, también puede decidir incluir otras características como si cada casa tiene una chimenea, el número de baños, etc. Diremos más sobre la selección de características más adelante, pero por ahora tomemos la función como dada.

Para realizar un aprendizaje supervisado, debemos decidir cómo vamos a representar **funciones/hipótesis** h_θ en un ordenador. Como elección inicial, digamos que decidimos aproximarnos a **Y (la relación real)** como una función lineal de x (la variable de entrada):

$$h_\theta(x) = \theta_0 + \theta_1 x_1$$

Aquí, los θ_i son los parámetros (también llamados pesos) que parametrizan el espacio de las funciones lineales que se mapean de X a Y.

Ahora, dado un conjunto de entrenamiento, ¿cómo elegimos, o aprendemos, los parámetros θ? Un método razonable parece ser hacer **h(x)** esté cerca de **Y**, al menos para las muestras de entrenamiento. Para formalizar esto, definiremos una función que mide, para cada valor de los **θ,** cómo de cerca están los $\mathbf{h(x^{(i)})}$ **de** los $\mathbf{y^{(i)}}$ correspondientes. Definimos pues la **función de coste**

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Si se ha estudiado anteriormente este tema, se puede reconocer esta función de coste como la función de mínimos cuadrados que da lugar al **modelo de regresión de mínimos cuadrados ordinarios.**

Queremos elegir **θ** para minimizar **J(θ).** Para hacerlo, usaremos un algoritmo de búsqueda que comience con alguna "suposición inicial" para **θ**, y que cambie repetidamente θ para hacer **J(θ)** más pequeño, hasta que con suerte converjamos a un valor de θ que minimice J(θ). Específicamente, consideremos el algoritmo de **descenso de gradiente (gradiente descent)**, que comienza con alguna **θ** inicial y realiza repetidamente la actualización.

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta).$$

(Esta actualización se realiza simultáneamente para todos los valores de **j = 0, . . . , n.)**. en la función del gradiente descendente, **α se llama tasa de aprendizaje (learning paramenter**) . Este es un algoritmo muy natural que repetidamente da un paso en la dirección de la disminución más pronunciada de **J**.

Para implementar este algoritmo, debemos averiguar cuál es el término de la derivada parcial que indicamos en el lado derecho.

$$\begin{aligned}\frac{\partial}{\partial\theta_j}J(\theta) &= \frac{\partial}{\partial\theta_j}\frac{1}{2}\left(h_\theta(x)-y\right)^2 \\ &= 2\cdot\frac{1}{2}\left(h_\theta(x)-y\right)\cdot\frac{\partial}{\partial\theta_j}\left(h_\theta(x)-y\right) \\ &= \left(h_\theta(x)-y\right)\cdot\frac{\partial}{\partial\theta_j}\left(\sum_{i=0}^{n}\theta_i x_i - y\right) \\ &= \left(h_\theta(x)-y\right)x_j\end{aligned}$$

Para un solo ejemplo de entrenamiento, esto proporciona la regla de actualización:

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}.$$

La regla se llama regla de actualización de **LMS (LMS significa "mínimos cuadrados medios". "least square mean")** y también se conoce como la regla de aprendizaje de Widrow-Hoff.

Esta regla tiene varias propiedades que parecen naturales e intuitivas. Por ejemplo, la magnitud de la actualización es proporcional al término de error **($y^{(i)} - h\theta(x^{(i)})$;** así, por ejemplo, si nos encontramos con un ejemplo de entrenamiento en el que nuestra predicción casi coincide con el valor real de $y^{(i)}$, entonces encontramos que hay poca necesidad de cambiar los parámetros; en contraste, se realizará un cambio mayor en los parámetros si nuestra predicción $h_\theta(x^{(i)})$ tiene un gran error (es decir, si está muy lejos de $y^{(i)}$).

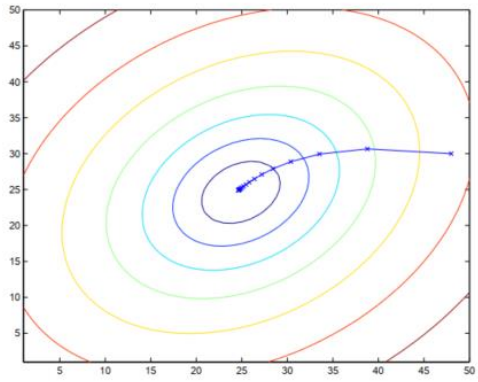
Habíamos derivado la regla LMS para cuando solo había un solo ejemplo de entrenamiento. Para más de un ejemplo lo reemplazamos por el siguiente algoritmo:

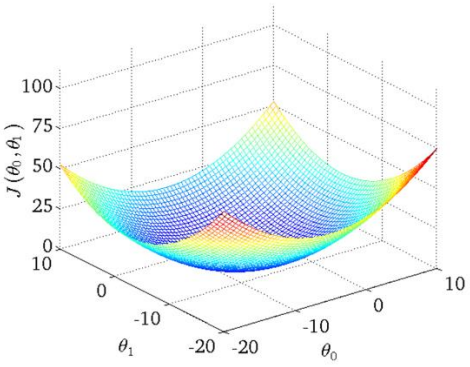
$$\begin{aligned}&\text{repeat until convergence } \{ \\ &\quad \theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right) \\ &\quad \theta_1 := \theta_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)\cdot x^{(i)} \\ &\}\end{aligned}$$

Esto es simplemente el descenso de gradiente en la **función de coste original J**. Este método analiza cada ejemplo en todo el conjunto de entrenamiento en cada paso y se denomina **descenso de gradiente por lotes[2] ("batch gradient descent")**. Tengase en cuenta que, si bien el descenso del gradiente puede ser susceptible a mínimos locales en general, el problema de optimización que hemos planteado aquí para la regresión lineal tiene solo un óptimo global y

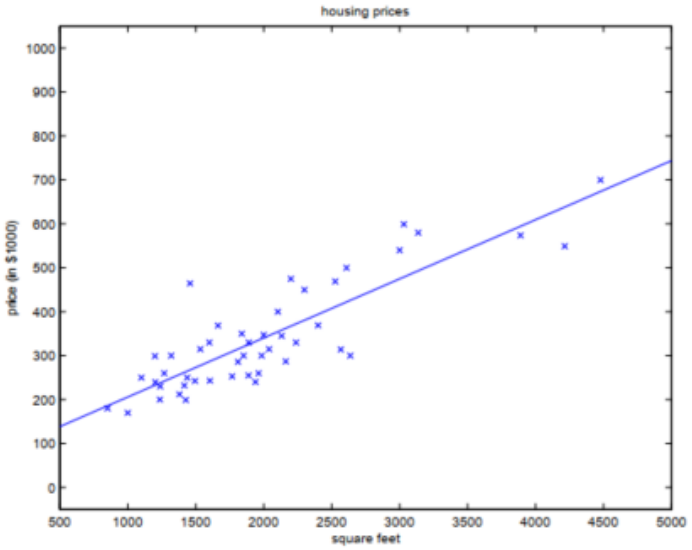
[2] "Batch": Cada paso de descenso de gradiente utiliza todos los ejemplos de entrenamiento.

ningún otro local; por lo tanto, el **descenso del gradiente siempre converge** (**suponiendo que la tasa de aprendizaje α no sea demasiado grande**) al mínimo global. De hecho, **J** es una función cuadrática convexa. En el gráfico se muestra el descenso de gradiente a medida que se ejecuta para minimizar una función cuadrática.





Cuando ejecutamos el descenso del gradiente por lotes para ajustar **θ** en nuestro conjunto de datos anterior, para aprender a predecir el precio de la vivienda en función del área habitable, obtenemos $\theta_0 = 71.27$, $\theta_1 = 0.1345$. Si trazamos **hθ(x)** en función de x (área), a lo largo con los datos de entrenamiento, obtenemos la siguiente figura:



6.- Flexibilidad e interpretabilidad

En el apartado anterior hemos ajustado una línea recta al conjunto de entrenamiento, pero por supuesto podemos añadir más características y podemos cambiar la forma de la curva para ajustarnos mejor a los datos y por tanto el error reducible será menor. Asimismo, podemos elegir una metodología no paramétrica que se ajuste muy bien a los datos sin asumir una función específica y finalmente consiguiendo una "superficie" compleja.

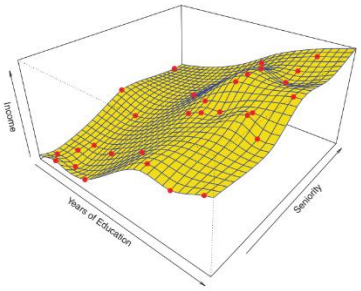


FIGURE 2.6. *A rough thin-plate spline fit to the* Income *data from Figure 2.3. This fit makes zero errors on the training data.*

Por ejemplo, abajo en la figura más a la izquierda se muestra el resultado de ajustar una $\mathbf{y = \theta_0 + \theta_1 x}$ a un conjunto de datos (**gráfico A)**. Vemos que los datos no se encuentran realmente en línea recta, por lo que el ajuste no es muy bueno (Ng, 2012, p. 14).

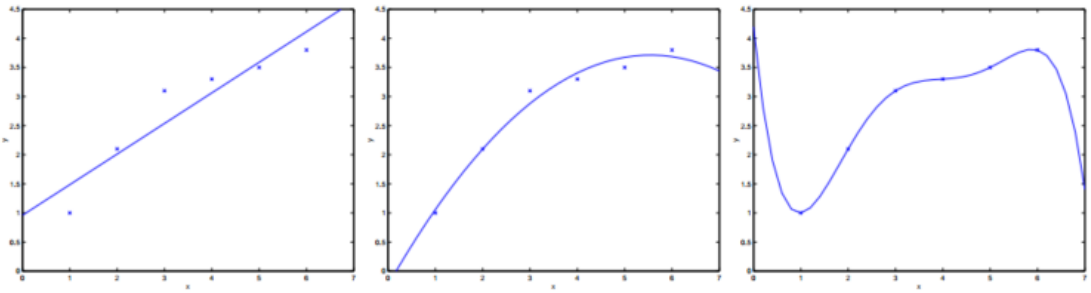


Gráfico A **Gráfico B** **Gráfico C**

Si hubiéramos añadido una característica extra $\mathbf{x_2}$, y ajustado (gráfico **B**) $\mathbf{y = \theta_0 + \theta_1 x + \theta_2 x^2}$, entonces obtenemos un ajuste ligeramente mejor a los datos. (Ver figura del medio B) Ingenuamente, podría parecer que cuantas más características agreguemos, mejor. Sin embargo, también existe el peligro de agregar demasiadas características: la figura más a la derecha es el resultado de encajar un polinomio de orden 5 (gráfico **C)** $\mathbf{y = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots\ldots + \Theta_5 x^5}$.

Vemos que a pesar de que la curva ajustada pasa a través de los datos perfectamente, no esperaríamos que esto sea un muy buen predictor de, digamos, los precios de la vivienda (y) para diferentes superficies (x). Sin definir formalmente lo que significan estos términos, diremos que la figura de la izquierda muestra un caso de **subajuste**, en el que los datos muestran claramente la estructura no capturada por el modelo, y la figura de la derecha es un ejemplo de **sobreajuste** (Ng, 2012, p. 14). Volveremos a estos conceptos más adelante.

De los muchos métodos existentes, algunos son menos flexibles, o más restrictivos, en el sentido de que pueden producir solo una gama relativamente pequeña de formas para estimar **f**. Por ejemplo, la regresión lineal es un enfoque relativamente inflexible, porque solo puede generar funciones lineales. Otros métodos, como las splines de placa delgada (no paramétricas), son considerablemente más flexibles porque pueden generar una gama mucho más amplia de formas posibles para estimar **f.** (James, Witten, et al., 2013, p. 25)

Se podría razonablemente hacer la siguiente pregunta: ¿por qué elegiríamos usar un método más restrictivo en lugar de un enfoque muy flexible? Hay varias razones por las que podríamos preferir un modelo más restrictivo pero la esencial es la interpretación. Si nos interesa principalmente **la inferencia**, entonces los modelos restrictivos son mucho más interpretables.

(James, Witten, et al., 2013, p. 25).. En el ejemplo dado es difícil interpretar el significado de la superficie a la potencia de 5. En el siguiente gráfico podemos ver la relación entre flexibilidad e interpretabilidad de algunos métodos.

(James, Witten, et al., 2013, p. 25)

En algunos entornos, sin embargo, solo nos interesa la predicción, y la interpretabilidad del modelo predictivo simplemente no es de interés. En este contexto, podríamos esperar que sea mejor utilizar el modelo más flexible disponible. Sorprendentemente, ¡este no es siempre el caso! A menudo obtendremos predicciones más precisas utilizando un método menos flexible. Este fenómeno, que puede parecer contradictorio a primera vista, tiene que ver con el potencial de sobreajuste en métodos altamente flexibles.

7.- Evaluación de la precisión del modelo: MSE para diferentes ajustes (James, Hastie, et al., 2013, pp. 29-33)

Para evaluar el rendimiento de un método de aprendizaje estadístico en un conjunto de datos dado, necesitamos alguna forma de medir cómo de buenas son sus predicciones por la coincidencia con los datos observados. Es decir, necesitamos cuantificar hasta qué punto el valor de respuesta predicho para una observación dada está cerca del valor de respuesta verdadero para esa observación. En el contexto de la regresión, la medida más utilizada es el **error cuadrático medio (MSE) que hemos visto**, dado por:

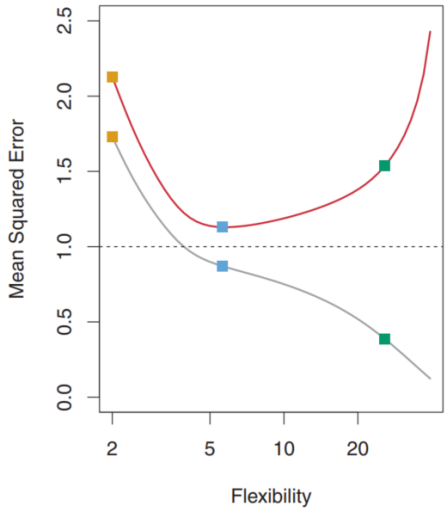
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

El MSE se calcula utilizando los datos de entrenamiento que se utilizaron para ajustarse al modelo, por lo que debe denominarse con mayor precisión el **MSE de entrenamiento**. Más bien, estamos interesados en la precisión de las predicciones que obtenemos cuando aplicamos nuestro método a datos de prueba nunca antes vistos. **Queremos elegir el método que da la prueba más baja MSE, en lugar de la MSE de entrenamiento más baja.**

En algunas configuraciones, podemos tener un conjunto de datos de prueba disponible, es decir, podemos tener acceso a un conjunto de observaciones que no se utilizaron para entrenar el método de aprendizaje estadístico.

Veamos cuál es la relación entre la MSE y la flexibilidad. La **curva gris** muestra el **MSE de entrenamiento promedio** en función de la flexibilidad, o más formalmente los grados de libertad, para una serie de splines de suavizado (un tipo de funciones no paramétricas). **Los grados de libertad** es una cantidad que resume la flexibilidad de una curva. Una curva más restringida y, por lo tanto, más suave tiene menos grados de libertad que una curva ondulada; tenga en cuenta que en la figura la regresión lineal está en el extremo más restrictivo, con dos grados de libertad. **El entrenamiento MSE disminuye monótonamente a medida que aumenta la flexibilidad.**

La curva roja (conjunto de pruebas) en comparación con **el gris (conjunto de entrenamiento)** muestra la evolución, pero ahora en el conjunto de pruebas. Observamos que el **MSE de entrenamiento disminuye** monótonamente a medida que aumenta la flexibilidad del modelo, y que hay una **forma de U en el MSE de prueba.** La línea discontinua es el error irreducible, el mínimo posible para el conjunto de prueba.



8.- The bias-Variance Trade-off (James, Hastie, et al., 2013, pp. 33-36)

La forma de U observada en las curvas MSE de prueba resulta ser el resultado de dos propiedades competitivas de los métodos de aprendizaje estadístico.

Es posible demostrar que la prueba esperada MSE, para un valor dado x_0, siempre se puede descomponer en la suma de tres cantidades fundamentales: **la varianza** de $\hat{f}(x_0)$, el **sesgo** de $\hat{f}(x_0)$ y **la varianza del ε de error.**

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$ [3]

[3] Está fuera del ámbito de esta nota técnica para explicarla en profundidad, así que no debe preocuparse si no comprende planamente la formulación. La intuición explicada es el elemento relevante

La ecuación nos dice que, para minimizar el error de prueba esperado, necesitamos seleccionar un método de aprendizaje estadístico que simultáneamente logre **una baja varianza y un bajo sesgo.**

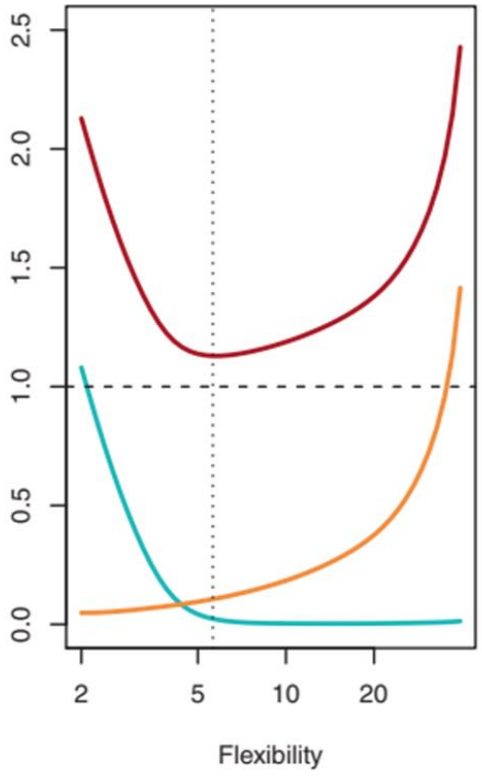
La varianza (variance) se refiere a la cantidad por la cual **f^** cambiaría si la estimamos **utilizando un conjunto de datos de entrenamiento diferente**. Dado que los datos de entrenamiento se utilizan para ajustarse al método de aprendizaje estadístico, diferentes conjuntos de datos de entrenamiento darán como resultado un **f^** diferente**. Pero lo ideal es que la estimación para f no varíe demasiado entre series de entrenamiento.** Si un método tiene una alta varianza, entonces pequeños cambios en los datos de entrenamiento pueden resultar en grandes cambios en **f^. En general, los métodos estadísticos más flexibles tienen una mayor varianza.**

El sesgo (Bias) se refiere al **error** que se introduce al **aproximar un problema de la vida real**, que puede ser extremadamente complicado, **por un modelo mucho más simple. En general, los métodos más flexibles resultan tener menos sesgo.**

En general, a medida que usamos métodos más flexibles, la varianza aumentará y el sesgo disminuirá. La tasa relativa de cambio de estas dos cantidades determina si la PRUEBA MSE aumenta o disminuye.

A medida que aumentamos la flexibilidad de una clase de métodos, el sesgo tiende a disminuir inicialmente más rápido de lo que aumenta la varianza. Sin embargo, en algún momento el aumento de la flexibilidad tiene poco impacto en el sesgo, pero comienza a aumentar significativamente la varianza. Cuando esto sucede, la prueba MSE aumenta.

En el siguiente gráfico, **la línea azul el sesgo** y la **línea naranja representan la varianza.** La línea discontinua horizontal representa Var(ε), el error irreducible. **El rojo representa el MSE**, la precisión del método.



La relación entre el sesgo, la varianza y el conjunto de prueba MSE dada en la ecuación anterior y mostrada en la figura se conoce como la **compensación sesgo-varianza (bias-variance trade-off)**

Bibliografía

Correlación no significa necesariamente causalidad. (2021). https://www.statsmedic.com/correlation-does-not-mean-causation

James, G., Hastie, T., Tibshirani, R., Witten, D., & Friedman, J. (2013). Una introducción al aprendizaje estadístico - con aplicaciones en R. En *Elementos* (Vol. 1). https://doi.org/10.1007/978-1-4614-7138-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *Una introducción al aprendizaje estadístico*.

Madhaven, A. (2020). Correlación vs causalidad: Comprenda la diferencia para su producto. En *amplitud*. https://amplitude.com/blog/causation-correlation

Ng, A. (2012). 1. Aprendizaje supervisado. *Aprendizaje automático*, 1–30.

Por qué la correlación no es causalidad. (s.f.). Consultado el 18 de septiembre de 2021 en https://www.ibpsychmatters.com/why-correlation-is-not-causation