

Herramientas de aprendizaje supervisado

Clase 1 - Introducción

Manuel Benjamín

October 7, 2023

Universidad de Buenos Aires

1. Generalidades del curso
2. ¿Qué es el aprendizaje supervisado?
3. ¿Quién es y como estimamos f ?

Generalidades

Docentes

Manuel Benjamín

Carlos Pita

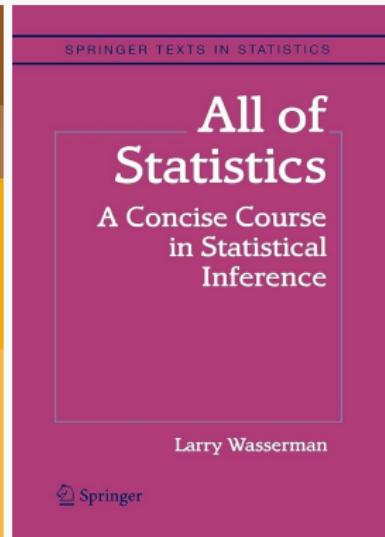
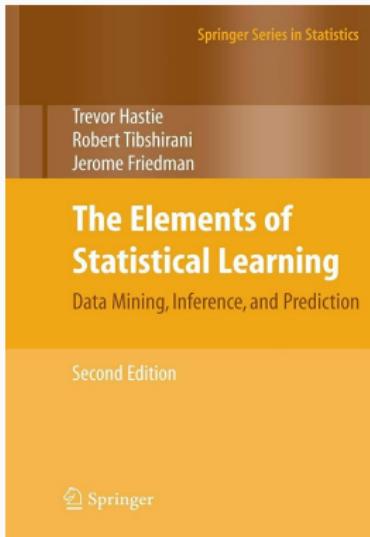
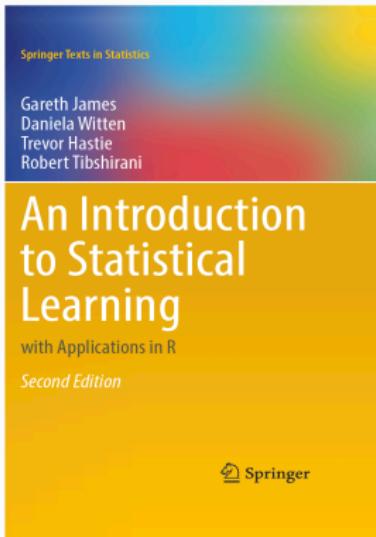
Régimen de aprobación

Examen parcial - Sábado 2 de Diciembre.

Exposición de trabajo final - Sábado 9 de Diciembre.

Recuperatorio - Sábado 16 de Diciembre.

Bibliografía



Estructura de las clases

Exposición ($\approx \%40$)

Simulaciones y aplicaciones ($\approx \%25$)

Trabajo en clase + consultas ($\approx \%35$)

Todas las clases subiremos ejercicios al campus que servirán como práctica de cada tema que veamos.

¿Qué es el aprendizaje
supervisado?

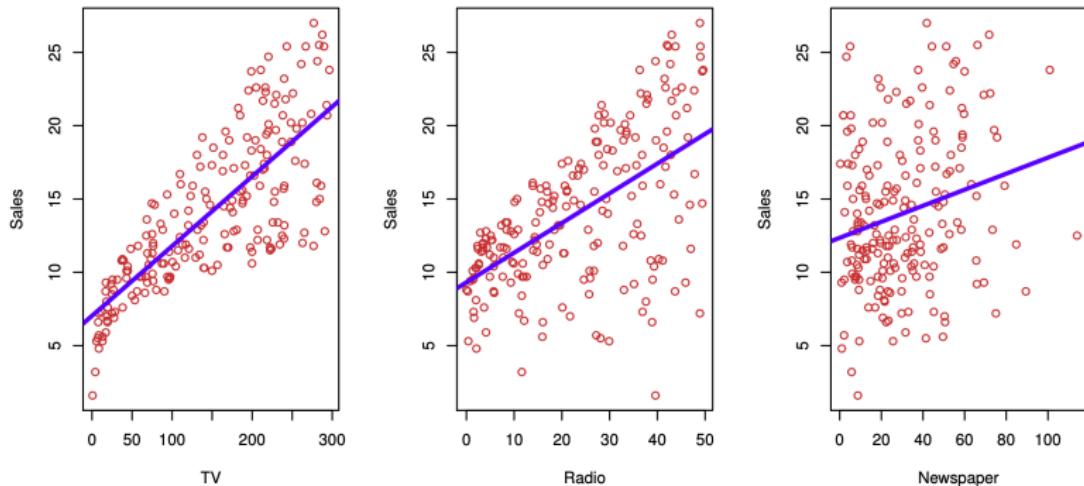


Figure 1: *Advertising* data set. *Sales* en miles de unidades como función de presupuestos en miles de dólares en *TV*, *Radio* y *Newspaper* para 200 mercados. En azul el ajusto del modelo lineal simple por cuadrados mínimos.

Observamos una variable cuantitativa Y y p variables predictoras X_1, \dots, X_p .

Asumimos una relación entre $X = (X_1, \dots, X_p)$ e Y

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

La función f es fija pero *desconocida*.

El término ε corresponde a un error aleatorio.

El aprendizaje estadístico refiere a un conjunto de metodologías para estimar la función f y herramientas para evaluar las estimaciones obtenidas

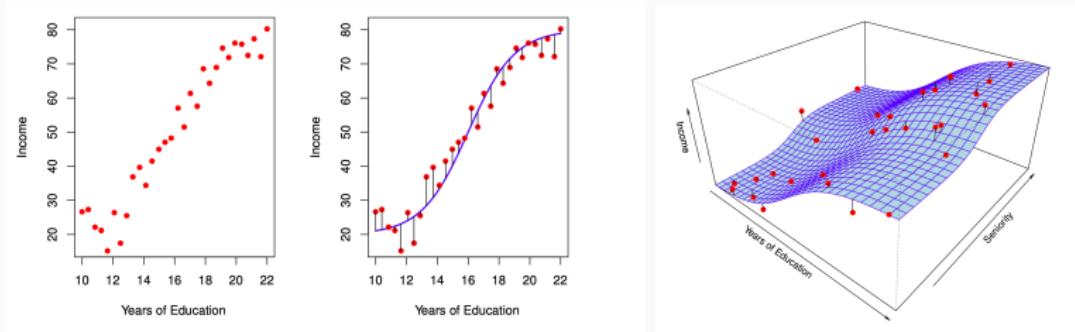


Figure 2: *Income* data. Conjunto de datos simulados con Ingresos en miles de usd, años de educación y Seniority en su trabajo. La curva azul y la superficie azul son las respectivas f . Al ser datos simulados las conocemos exactamente. Las rectas verticales corresponden a los errores ε de cada observación.

Ejercicio

Identificar X , Y y f asumida en el ejemplo de *Advertising* y sus ajustes por cuadrados mínimos

¿Por qué estimar f ?

- Predicción.

Dado un nuevo valor de X queremos poder predecir el valor de Y

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

- Inferencia.

¿Que variables predictoras están asociadas con Y ?

¿Qué relación existe entre la respuesta y cada variable predictora?

La metodología a usar para estimar f va a depender de que se busca (predecir, hacer inferencia o una mezcla de ambas cosas)

Error al predecir

La exactitud al predecir Y con $\hat{Y} = \hat{f}(X)$ depende de dos cantidades

$$Y - \hat{Y} = \underbrace{f(X) - \hat{f}(X)}_{\text{Error reducible}} + \underbrace{\varepsilon}_{\text{Error irreducible}}$$

Error esperado al predecir

Si X y \hat{f} estan fijos

$$E(Y - \hat{Y})^2 = \underbrace{(f(X) - \hat{f}(X))^2}_{\text{reducible}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible}}$$

- El error irreducible es una cota inferior a la exactitud de la predicción de Y . Esta cota suele ser desconocida en la práctica.
- Nos vamos a enfocar en técnicas para estimar f con el objetivo de minimizar el error cuadrático medio reducible.

¿Cómo estimamos f ?

Disponemos de n datos de entrenamiento

$$\{(x_1, y_1), \dots, (x_n, y_n)\},$$

donde $x_i = (x_{i1}, \dots, x_{in})$.

Queremos encontrar una f que se ajuste bien a los datos, para esto existen dos tipos de métodos:

- Paramétricos
- No paramétricos.

Métodos paramétricos:

1. Hacemos una suposición de la forma de f a través de una familia identificada con finitos parámetros. e.g

Modelo Lineal:

$$f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

2. Utilizamos los datos de entrenamiento para 'ajustar', 'entrenar' o 'estimar' el modelo.

En el modelo lineal podemos ajustar por cuadrados mínimos.
(Aunque no es la única posible).

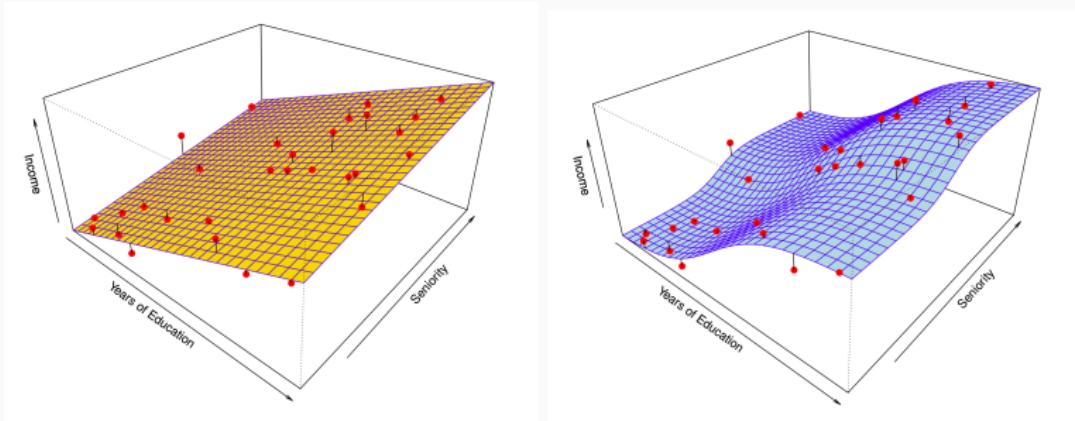


Figure 3: Modelo lineal ajustado a los datos de *Income*. En el primer grafico, en amarillo \hat{f} , en el segundo y en azul f

Ventaja

El modelo paramétrico reduce el problema de estimar f a estimar tres parámetros

$$\text{income} \approx \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$$

Desventaja

El modelo que uno elige no suele contener a la verdadera forma de f .

¿Que sucede si la verdadera forma de f esta muy lejos de ser capturada por el modelo paramétrico?

Modelos no paramétricos

Buscan estimaciones de f que se mantengan lo mas cerca posible de los datos de entrenamiento tratando de ser 'suave'.

Ventajas

Evitando asumir formas particulares de f tiene el potencial de ajustar muchas mas formas posibles.

Desventajas

Se necesitan muchas observaciones (muchas mas que en un modelo paramétrico) para producir buenas estimaciones de f .

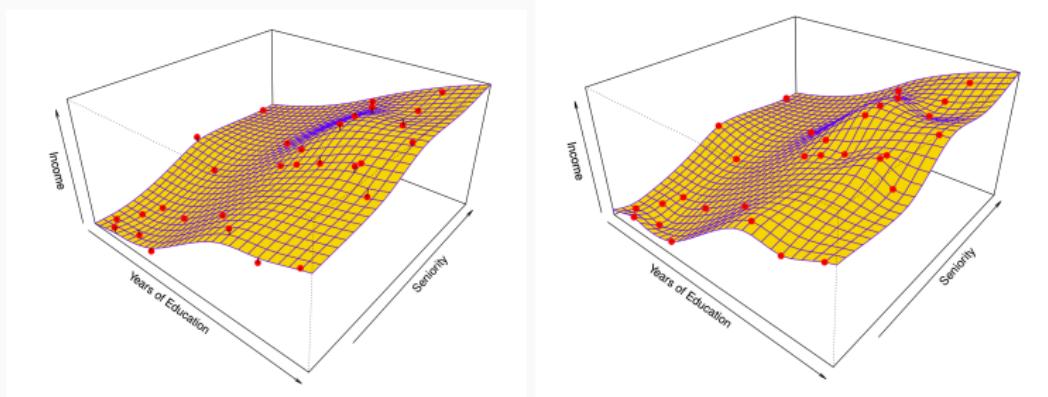


Figure 4: Ajustes no parametricos de Income data. A la izquierda un ajuste suave. A la derecha un ajuste poco suave.

Balance entre poder de predicción e interpretabilidad de un modelo

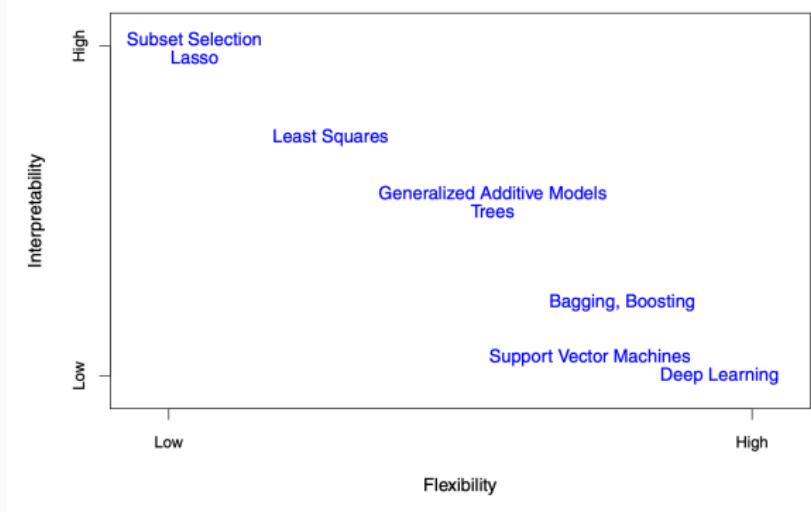


Figure 5: Tradeoff entre flexibilidad e interpretabilidad para distintas metodologías de aprendizaje estadístico. En general si la flexibilidad aumenta, la interpretabilidad baja.

Problemas de clasificación

Las variables cualitativas (categóricas) toman valores en K clases distintas. Estas clases no tienen una relación de orden implícito.

Examples

Color de ojos (negro, marrón, azul,...)

Tipo de sangre (A, AB, B, O)

Si una persona entró en cesación de pagos (si o no).

Predecir una variable categórica se conoce como problema de clasificación.

Existen metodologías que solo tienen sentido para ambos problemas (árboles, vecinos cercanos), otras que solo sirve para clasificación (regresión logística) y otras que solo para regresión (regresión lineal).

En general todas las metodologías que veremos aceptan variables explicativas cualitativas y cuantitativas.

Problemas no supervisados Se disponen de n observaciones x_i pero no tienen una respuesta asociada y_i

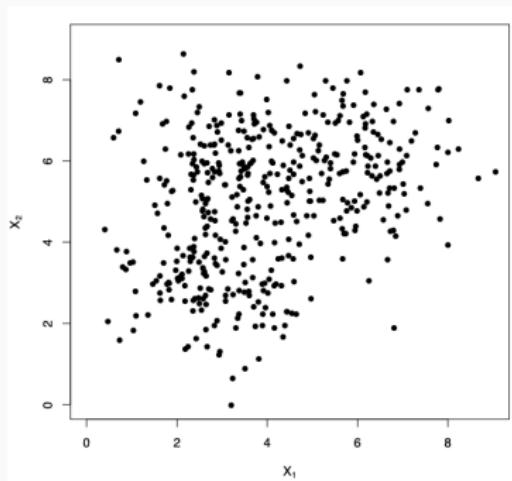
Reducción de la dimensión

Clustering.

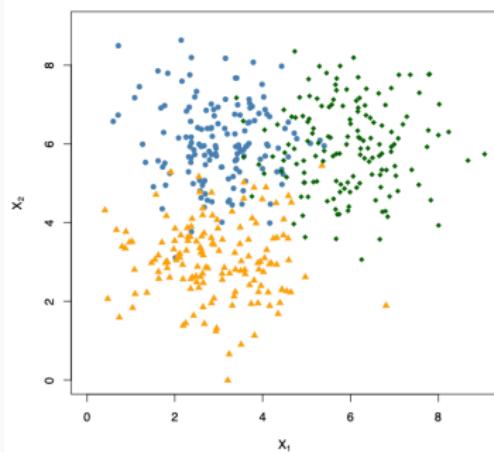
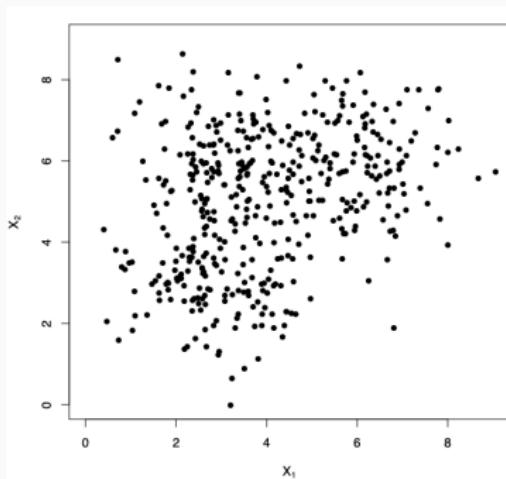
Detección de anomalías.

Ingeniería de features.

Clustering ¿Cuántos grupos hay?



Clustering ¿Cuántos grupos hay?



¿Quién es y como estimamos f ?

Teoría de la decisión - Modelos cuantitativos - ¿Quién es f?

Existen infinitas funciones que cumplen

$$Y = f(X) + \varepsilon$$

Si buscamos $f(X)$ para predecir Y , necesitamos identificarla de alguna manera. Para esto elegimos una función de perdida $L(Y, f(X))$ para penalizar los errores de predicción. La opción más común (pero no la única) es elegir la *pérdida cuadrática*

$$L(Y, f(X)) = (Y - f(X))^2$$

Esto es una variable aleatoria!

Teoría de la decisión - Modelos cuantitativos - ¿Quién es f?

El criterio de decisión es la función que minimiza

$$EPE(f) = E(Y - f(X))^2.$$

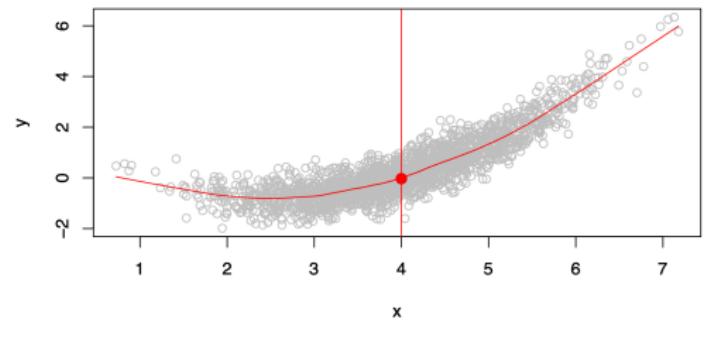
La esperanza condicional es quien minimiza!

$$f(x) = E(Y|X = x)$$

Ejercicio opcional

Demostrar el resultado anterior.

Pueden encontrar los argumentos en ESL pag 18.



Tenemos definida la función de regresión

$$f(4) = E(Y|X = 4)$$

En caso de X multivariada

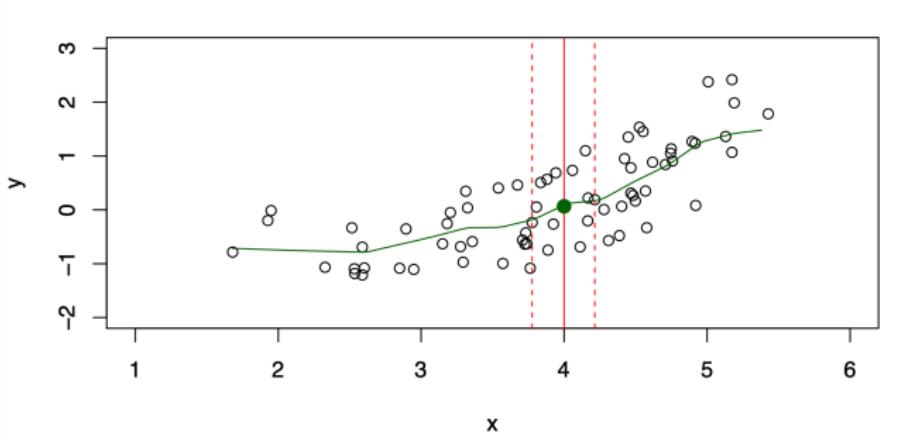
$$f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

¿ Cómo estimamos f ?

- Si tuviésemos muchas observaciones con $X = x$ podríamos promediar sus valores de Y .
- En general tenemos pocos o ninguno!
- Podemos relajar la definición considerando las observaciones con X 'cerca' a x

$$\hat{f}(x) = \text{Prom}\{Y | X \in N(x)\},$$

donde $N(x)$ es un vecindario de x .



- **K vecinos mas cercanos**
Promediamos las K observaciones mas cercanas al punto que queremos predecir.
- **Kernels**
Vecindad definida por un radio (y promedio ponderado).

¿Cómo evaluamos el desempeño de un modelo ajustado?

Queremos medir que tan bien las predicciones se ajustan a datos nuevos (con los que el modelo no fue ajustado)

Los datos nuevos o 'no vistos' los llamamos datos de testeo.

Error cuadrático medio

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2,$$

Siempre distinguimos si lo calculamos sobre los datos de entrenamiento o de testeo.

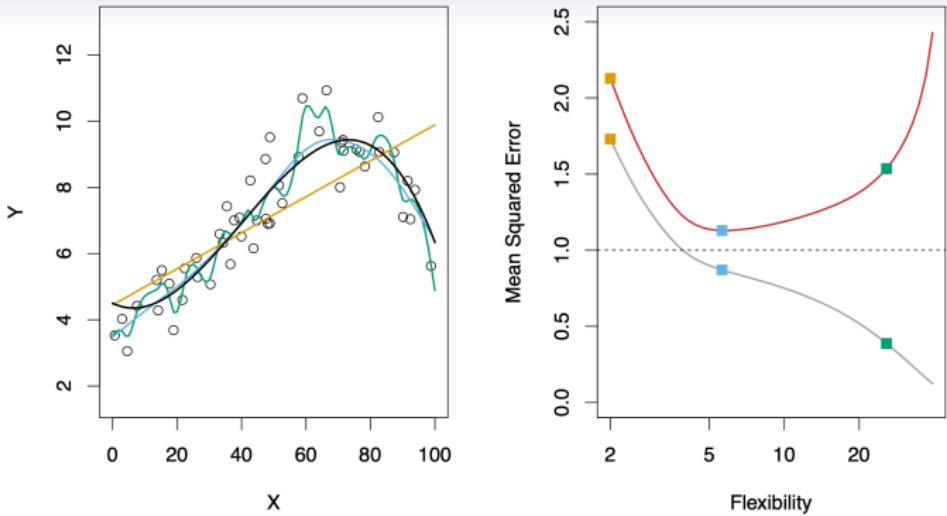
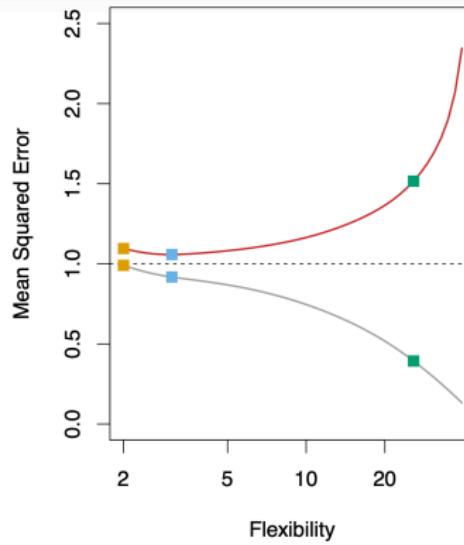
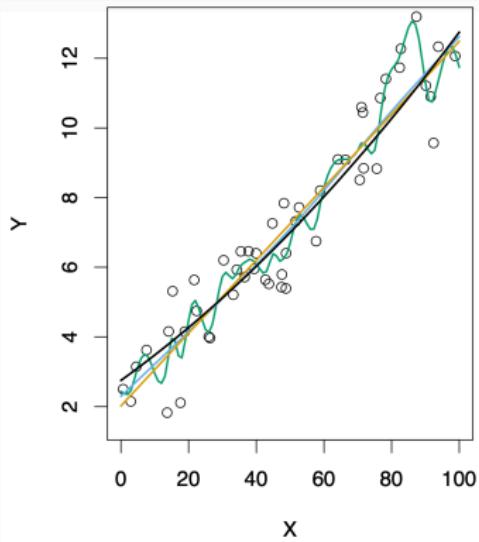
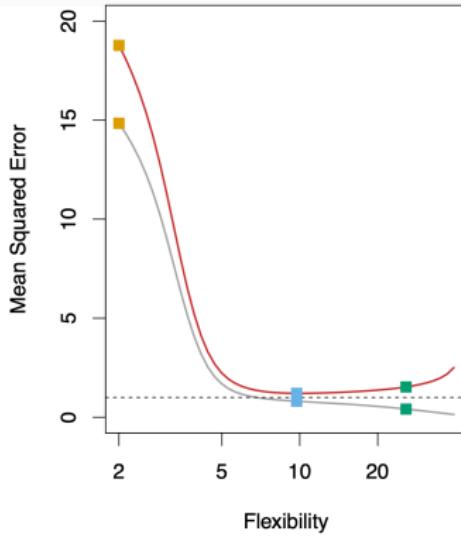
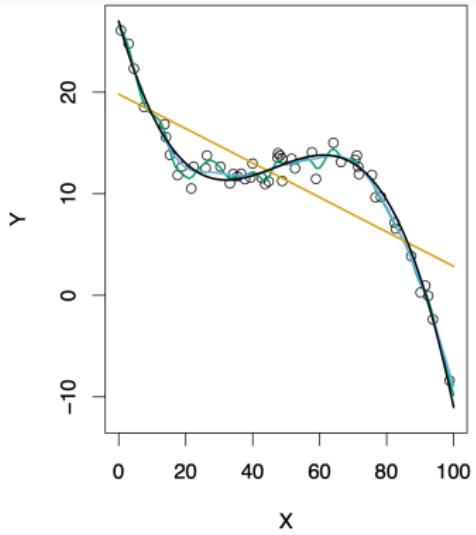


Figure 6: La curva negra es $f(x)$. La curva roja corresponde al MSE de testeо y la gris al MSE de entrenamiento. Las curvas naranja, verde y celeste son ajustes de diferente flexibilidad





Ejercicio

Recrear los gráficos anteriores utilizando vecinos mas cercanos con distintos valores de K para $f(x) = 5x(x - 1)(x + 1)$ y una muestra uniforme de $n = 50$ en el intervalo $(-1.5, 1.5)$.

Descomposición en sesgo y varianza

Supongamos que tenemos una metodología para ajustar f , un conjunto de entrenamiento y (X_0, Y_0) una observación de la población independiente de la muestra de entrenamiento. Si el modelo es $Y = f(X) + \varepsilon$ (con $f(x) = E(Y|X=x)$),

$$E(Y_0 - \hat{f}(X_0))^2 = \text{Var}(\hat{f}(X_0)) + [\text{SESGO}(\hat{f}(X_0))]^2 + \text{Var}(\varepsilon).$$

La esperanza es sobre la variabilidad de (X_0, Y_0) y del conjunto de entrenamiento.

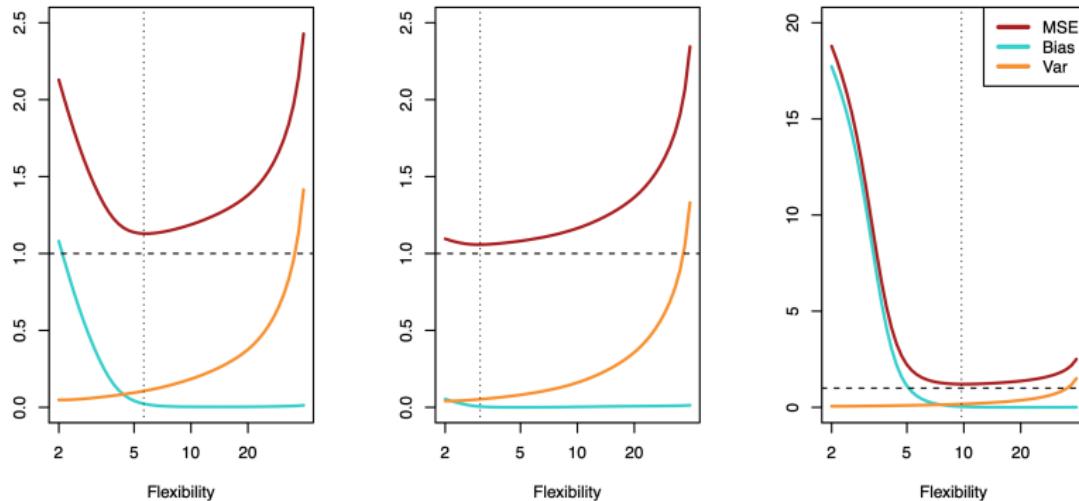


Figure 7: Curvas de sesgo y varianza para los tres graficos anteriores

En general, mientras mas flexible sea \hat{f} mayor sera su varianza pero menor su sesgo.

Ejercicios

- Obtener las curvas de sesgo y varianza del problema anterior.
- Resuelva los ejercicios de la sección 2.4 del ISLR.
- Considere la función de regresión

$$f(x) = x + 0.1 * x^{10}.$$

Cree una muestra $(x_1, y_1), \dots, (x_n, y_n)$ con $n = 20$, $X_i \sim U(-1, 1)$ e $y_i = f(x_i) + \varepsilon_i$ donde $sd(\varepsilon) = 1$.

Considere ambos ajustes lineales

$$y \sim \beta_0 + \beta_1 x \quad y \sim \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_{10} x^{10}.$$

Responda

1. Alguno de los dos modelos es insesgado?
2. ¿Cuál de los dos modelos espera que funcione mejor en términos de predicción?
3. Evalúe ambos modelos en un set de entrenamiento lo suficientemente grande.