Machine Learning

Hernán Aguirre

Universidad San Franciso de Quito

Universidad de Shinshu, Japón

Contenido del Curso

1. **Introducción**
2. **Aprendizaje estadístico**
3. **Regresión lineal**
4. **Clasificación**
5. **Métodos de remuestreo**
6. **Selección y regularización de modelos lineales**
7. **Más allá de la linealidad**
8. **Métodos basados en árboles**
9. **Máquinas de vectores de soporte**
10. Aprendizaje profundo
11. Análisis de supervivencia y datos censurados
12. Aprendizaje sin supervisión
13. Pruebas múltiples

Aprendizaje Estadístico

1. ¿Qué es el Aprendizaje Estadístico?
2. Evaluación de la Precisión del Modelo
3. Laboratorio: Introducción a Python
4. Ejercicios

2.1

¿Qué es el Aprendizaje Estadístico?

1. Terminología, Definición y Ejemplos
2. ¿Por qué estimar f ?
3. ¿Cómo estimamos f ?
4. Interpretabilidad y flexibilidad del modelo
5. Aprendizaje supervisado versus no supervisado
6. Problemas de regresión versus clasificación

2.1.1

Terminología, Definición y Ejemplos

- Variables de entradas $X = (X_1, X_2, \cdots, X_p)$
  - *predictores*, *variables independientes*, *características* o simplemente *variables*.
- Variables de salida Y
  - *respuesta* o *variable dependiente*

Definición

- Supongamos que observamos una *respuesta* **cuantitativa** Y y p *predictores* diferentes $X = (X_1, X_2, \cdots, X_p)$
- Suponemos que existe alguna **relación** entre Y y X

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

  - f es alguna **función** fija pero **desconocida**
  - ϵ es un término de **error** aleatorio, independiente de X con media cero
- El *aprendizaje estadístico* se refiere a un conjunto de enfoques para **estimar** f **a partir de datos observados**

Publicidad y Ventas

- Supongamos que queremos investigar la asociación entre publicidad y ventas de un producto en particular.

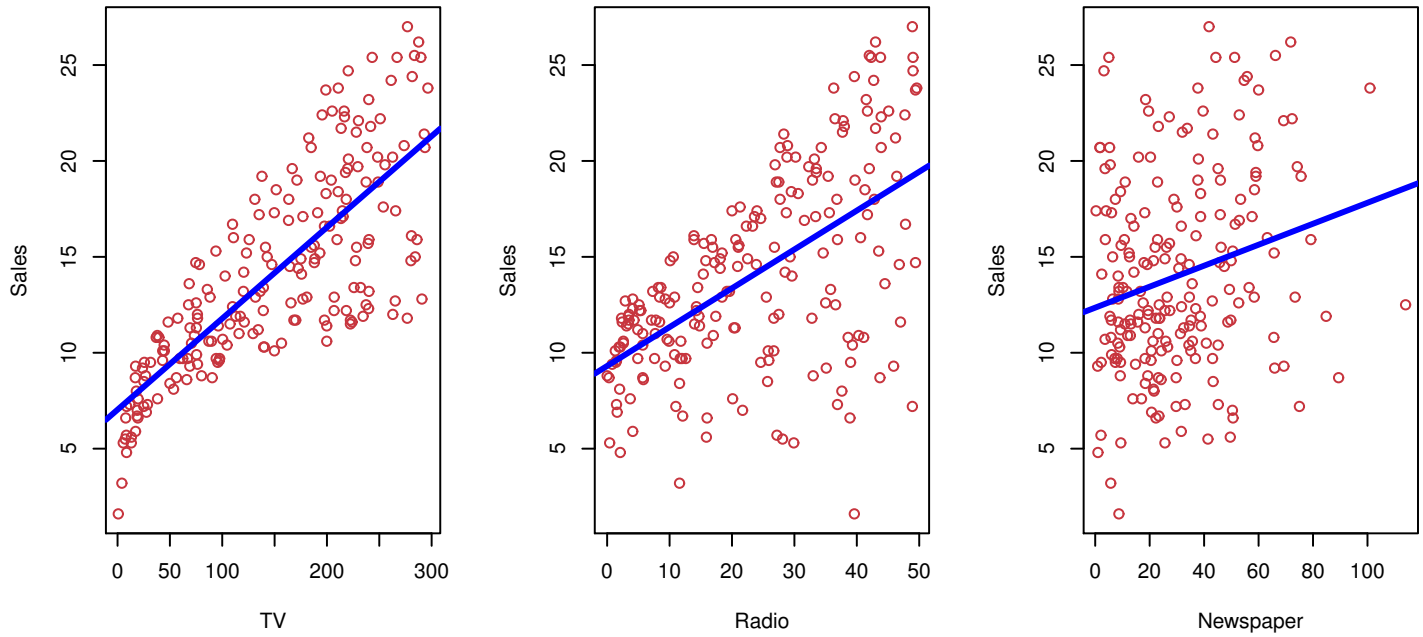
- El conjunto de datos de publicidad consta

  - ventas del producto en 200 mercados diferentes

  - presupuestos de publicidad del producto en cada uno de esos mercados para tres medios diferentes: TV, radio y periódicos.

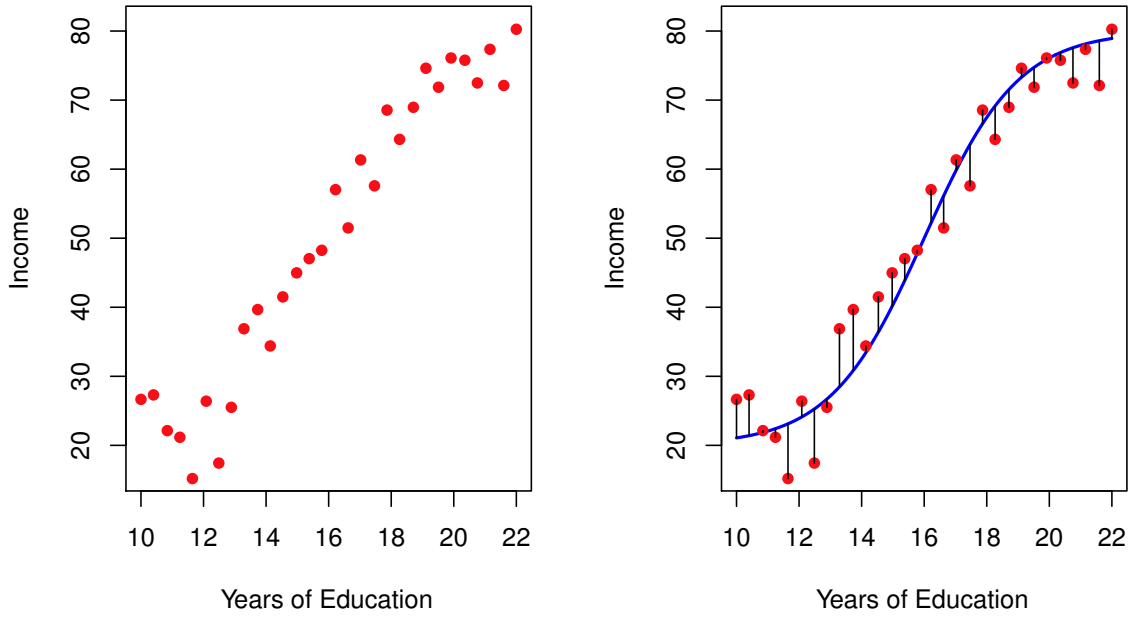
ventas	TV	radio	periódicos

Publicidad y Ventas (2)



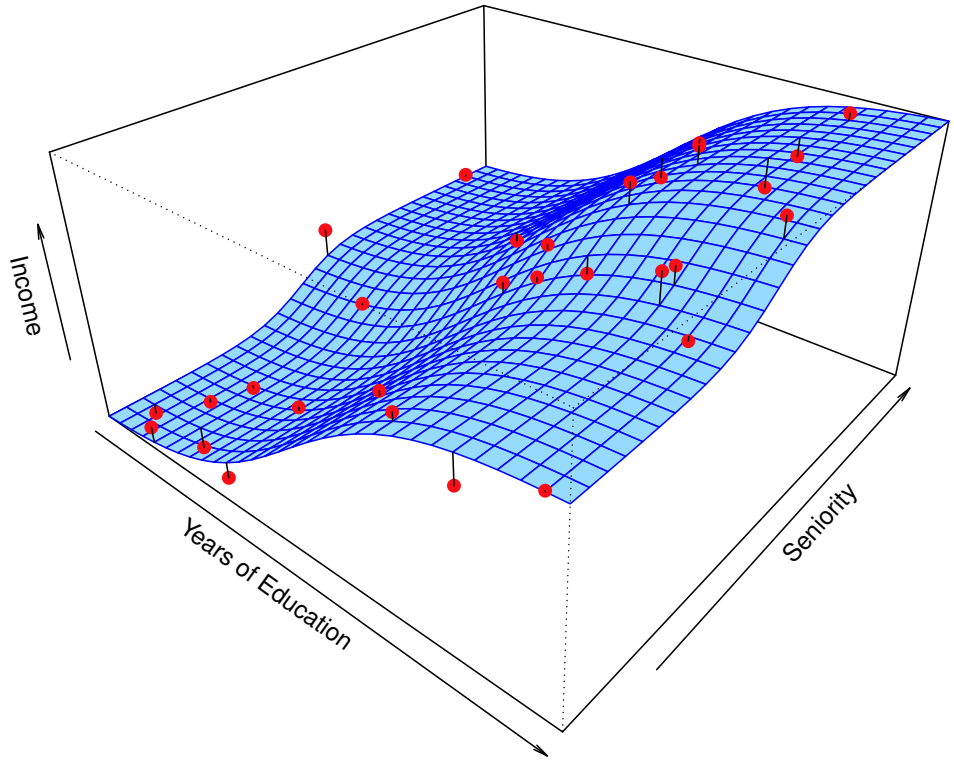
- Ventas en función de los presupuestos de TV, radio y periódicos para 200 mercados diferentes (en miles de dólares)
- Ajuste de *mínimos cuadrados* de las ventas a cada una de las variables
- Línea azul representa un modelo simple que se puede utilizar para **predecir** las ventas con base en el presupuesto en TV, radio y periódicos.

Ingresos



- Valores observados de ingresos (en miles de dólares) y años de educación para 30 personas.
- La curva representa la **verdadera relación subyacente** entre los ingresos y los años de educación.
- Las **líneas negras** representan el **error** asociado con cada observación

Datos de Ingresos (2)



- Ingresos en función de los años de educación y antigüedad
- Superficie azul representa la **verdadera relación subyacente**, conocida en vista que se simulan los datos
- Puntos rojos indican los valores observados para 30 individuos

2.1.2

¿Por qué estimar f ?

- Hay dos razones principales por las que deseamos estimar f :
  - predicción
  - inferencia

Predicción

- Hay situaciones donde un conjunto de entradas X y salidas Y están disponibles, pero no sabemos cómo obtener Y a partir de X
- En esta configuración, podemos predecir Y usando

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

  - $\hat{f}$ es nuestra estimación de f
  - $\hat{Y}$ es la predicción resultante para Y
- no nos preocupamos por la forma exacta de $\hat{f}$ (**caja negra**), siempre que produzca predicciones precisas para Y

Predicción (2)

- La precisión de $\hat{Y}$ como predicción de Y depende de dos cantidades, *error reducible* y *error irreducible*

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

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

$$E(Y - \hat{Y})^2 = E\,[\; f(X) + \epsilon - \hat{f}(X)\;]^2$$

$$= E\,[\; f(X) - \hat{f}(X)\;]^2 + Var(\epsilon)$$

- *Poner un ejemplo mara mostrar de donde proviene el error irreducible y de donde el reducible*

Inferencia

- Interesa comprender la asociación entre Y y $X_1, \cdots, X_p$
- Deseamos estimar f, pero nuestro objetivo no es necesariamente hacer predicciones para Y
- $\hat{f}$ no puede tratarse como una caja negra, porque necesitamos conocer su forma exacta

Inferencia (2)

- ¿Qué predictores están asociados con la respuesta?
  - Identificar los pocos predictores importantes
- ¿Cuál es la relación entre la respuesta y cada predictor?
  - ¿Positiva, negativa, la relación entre la respuesta y un predictor dado depende de los valores de otros predictores?
- ¿Se puede resumir adecuadamente la relación entre Y y cada predictor utilizando una ecuación lineal, o la relación es más complicada?

Inferencia (3)

- Dependiendo de si el objetivo es la predicción, la inferencia o una combinación de ambas, pueden ser apropiados diferentes métodos para estimar f.
- Modelos lineales
  - Inferencias relativamente simples e interpretables
  - predicciones no tan precisas como otros enfoques
- Modelos altamente no lineales
  - predicciones bastante precisas
  - menos interpretable, la inferencia es más desafiante

Ej. Predicción

- Campaña de marketing directo
- Identificar personas que respondan positivamente a un correo (*resultado*) basándose en observaciones de variables demográficas medidas en cada individuo (*predictores*)
- Se quiere predecir con precisión la respuesta utilizando los predictores
- No interesa obtener una comprensión profunda de las relaciones entre cada predictor y la respuesta;

Ej. Inferencia

- Considere los datos de publicidad
- A alguien le puede interesar responder preguntas como:
  - ¿Qué medios están asociados a las ventas?
  - ¿Qué medios generan mayor impulso en las ventas?
  - ¿Qué magnitud de aumento en las ventas se asocia con un aumento determinado en la publicidad televisiva?

Ej. Inferencia (2)

- Modelar la marca de un producto que un cliente podría comprar en función de variables como el precio, la ubicación de la tienda, los niveles de descuento, el precio de la competencia, etc.

- Lo que realmente podría interesar es la asociación entre cada variable y la probabilidad de compra.

- Por ejemplo, ¿en qué medida el precio del producto está asociado con las ventas?

Ej. Predicción e Inferencia

- En un entorno inmobiliario, se puede tratar de relacionar los valores de las viviendas con factores como la tasa de criminalidad, la zonificación, la distancia a un río, la calidad del aire, las escuelas, el nivel de ingresos de la comunidad, el tamaño de las casas, etc.

- Inferencia: interés en la asociación entre cada variable y el precio de la vivienda; por ejemplo, ¿cuánto más valdrá una casa si tiene vista al río?

- Predicción: interés en predecir el valor de una casa dadas sus características: ¿esta casa está infravalorada o sobrevalorada?

2.1.3

¿Cómo estimamos f ?

- Nuestro objetivo es aplicar un método de aprendizaje estadístico a los datos de entrenamiento para estimar la función desconocida f
  - Encontrar una función f tal que Y≈f(X) para cualquier observación (X,Y)
- Enfoques lineales y no lineales para estimar f
- Se pueden caracterizar como
  - paramétricos o
  - no paramétricos

Métodos Paramétricos

- Primero, hacemos una suposición sobre la forma funcional de f
  - Ejemplo, suponemos que $f$es lineal en X

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

  - El problema de estimar f se simplifica enormemente
  - En lugar de tener que estimar una función p-dimensional $f(X)$ completamente arbitraria, solo es necesario estimar $p + 1$ coeficientes $\beta_0, \beta_1, \cdots, \beta_p$

Métodos Paramétricos (2)

- Segundo
  - una vez seleccionado un modelo, necesitamos un procedimiento que utilice los datos de entrenamiento para ajustar o entrenar el modelo.

Métodos Paramétricos (3)

- Ventaja
  - reduce el problema de estimar f a uno de estimar un conjunto de parámetros.
- Desventaja potencial
  - el modelo que elegimos generalmente no coincidirá con la forma verdadera desconocida de f
  - Si el modelo elegido está demasiado alejado de la f verdadera, entonces nuestra estimación será deficiente

Métodos Paramétricos (4)

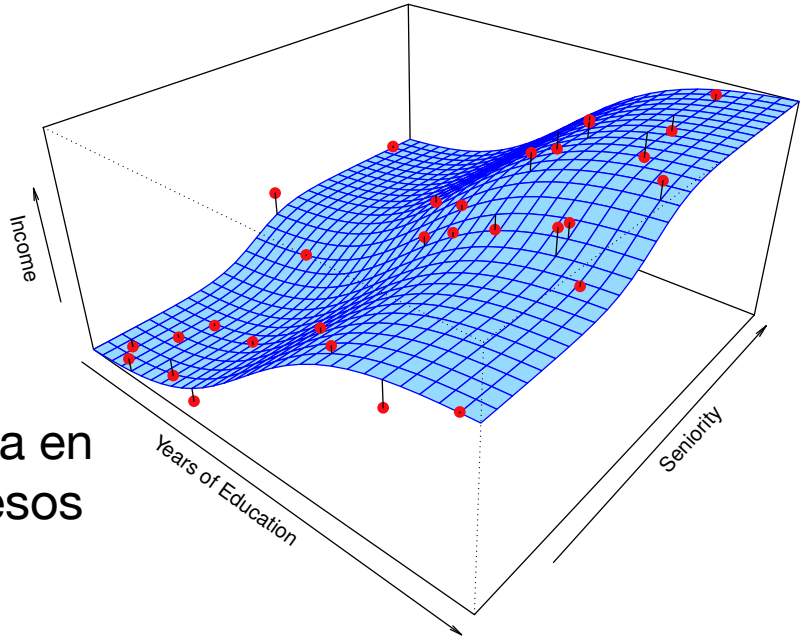
- Podemos intentar resolver este problema eligiendo modelos flexibles que puedan adaptarse a muchas formas funcionales diferentes posibles para f.

- Pero, en general, ajustar un modelo más flexible requiere estimar un mayor número de parámetros.

- Estos modelos más complejos pueden conducir a un fenómeno conocido como *sobre ajuste de datos*, que esencialmente significa que siguen los errores o el ruido demasiado de cerca.

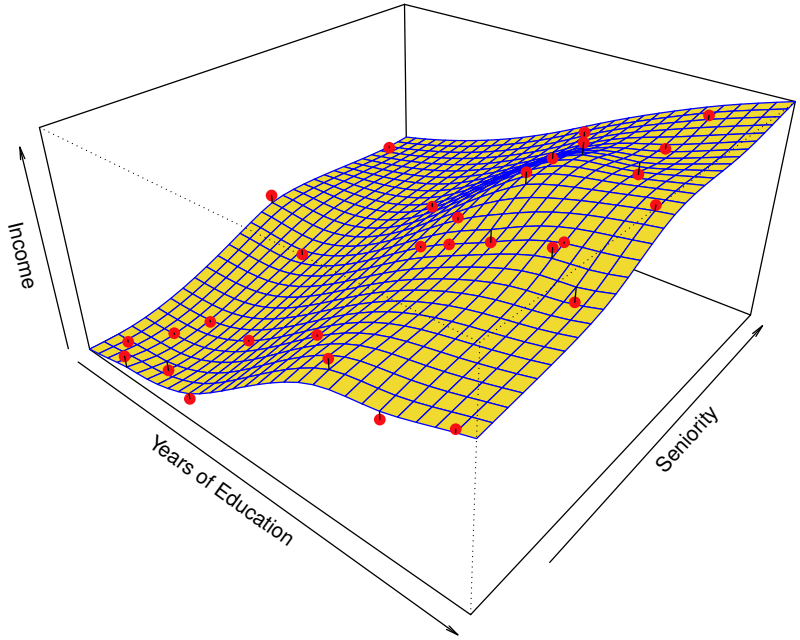
Métodos no Paramétricos

- No hacen suposiciones explícitas sobre la forma funcional de f.
- Buscan una estimación de f que se acerque lo más posible a los puntos de datos
- Tienen el potencial de ajustarse con precisión a una gama más amplia de formas posibles para f
- Dado que no reducen el problema de estimar f a un pequeño número de parámetros, se requiere un número muy grande de observaciones (muchas más de las que normalmente se necesitan para un enfoque paramétrico) para obtener una estimación precisa de f

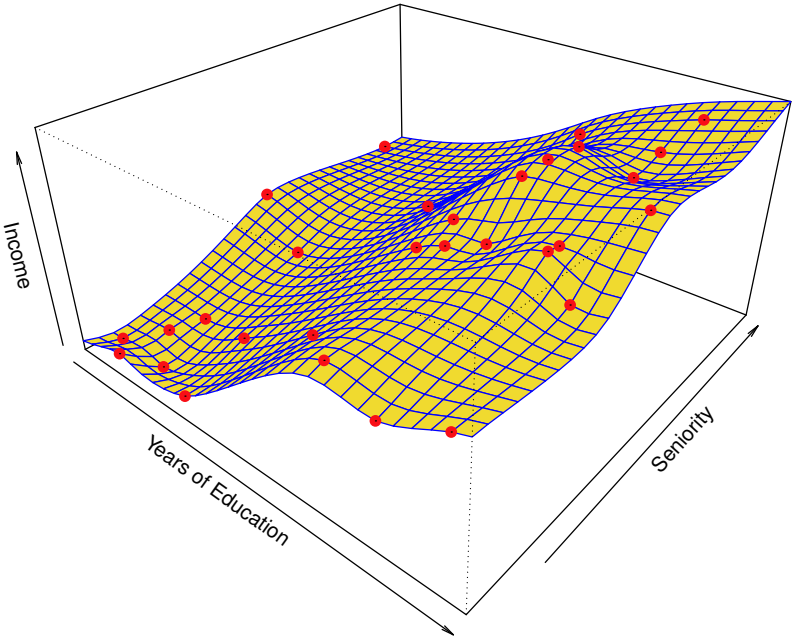


Relación verdadera en los datos de Ingresos

ajuste perfecto! → sobreajuste



Spline de placa delgada y suave ajustada a los datos de Ingresos



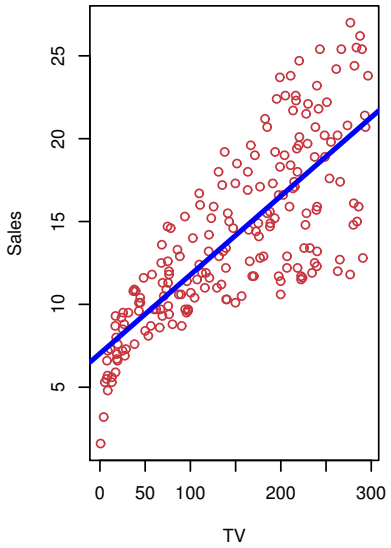
Spline de placa delgada **más áspera** ajustada a los datos de Ingresos

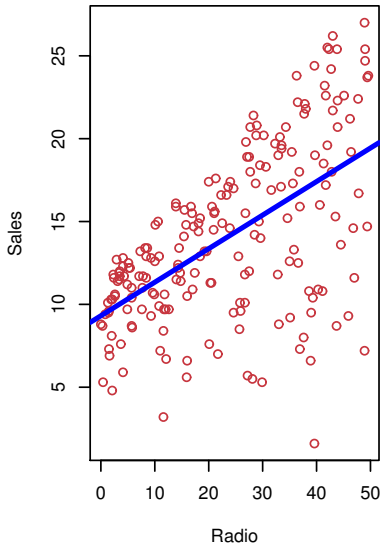
2.1.4

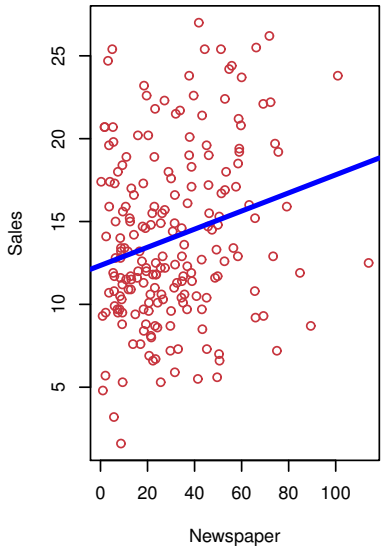
Interpretabilidad y Flexibilidad

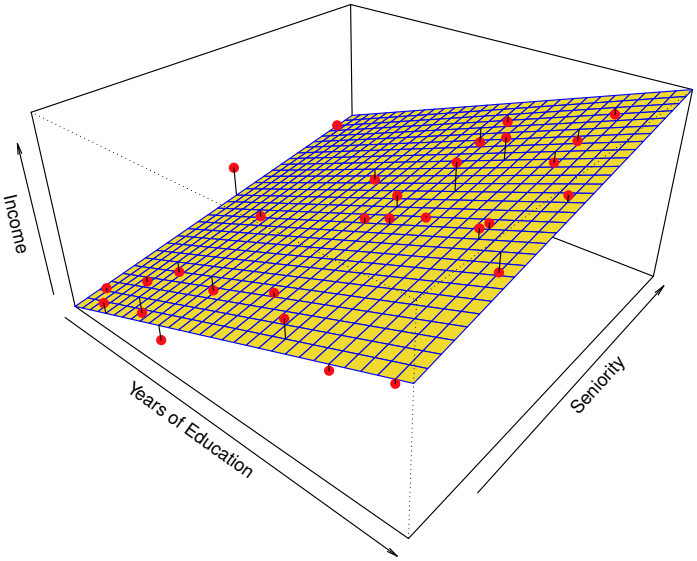
- Algunos métodos son menos flexibles que otros,
  - Pueden producir un rango relativamente pequeño de formas funcionales para estimar f
- La *regresión lineal* es un enfoque relativamente inflexible
  - Solo puede generar funciones lineales o planos
- Los *splines* son considerablemente más flexibles
  - Pueden generar una gama mucho más amplia de formas posibles para estimar f

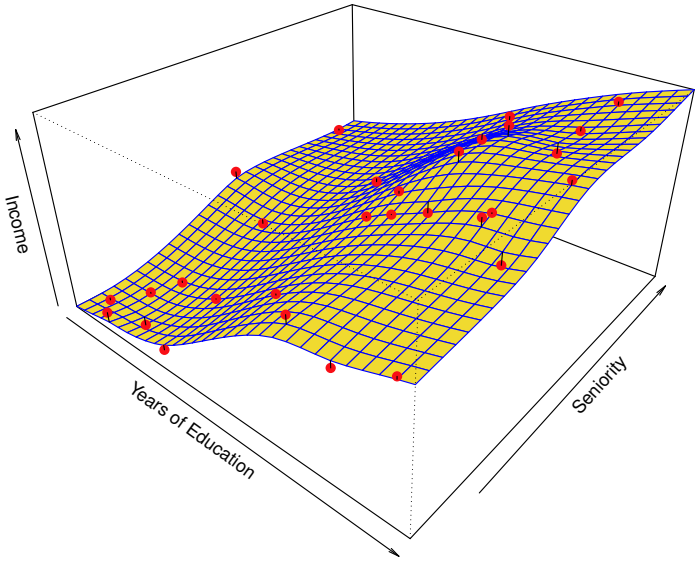
Interpretabilidad y Flexibilidad (2)

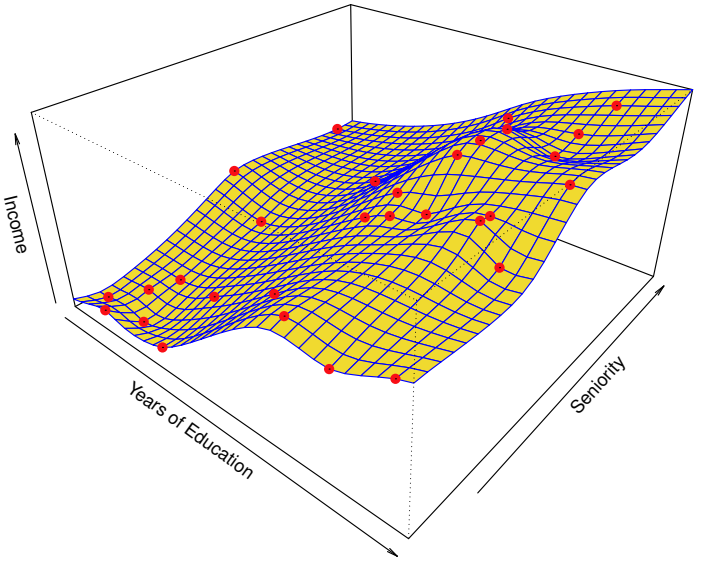












Interpretabilidad y Flexibilidad (3)

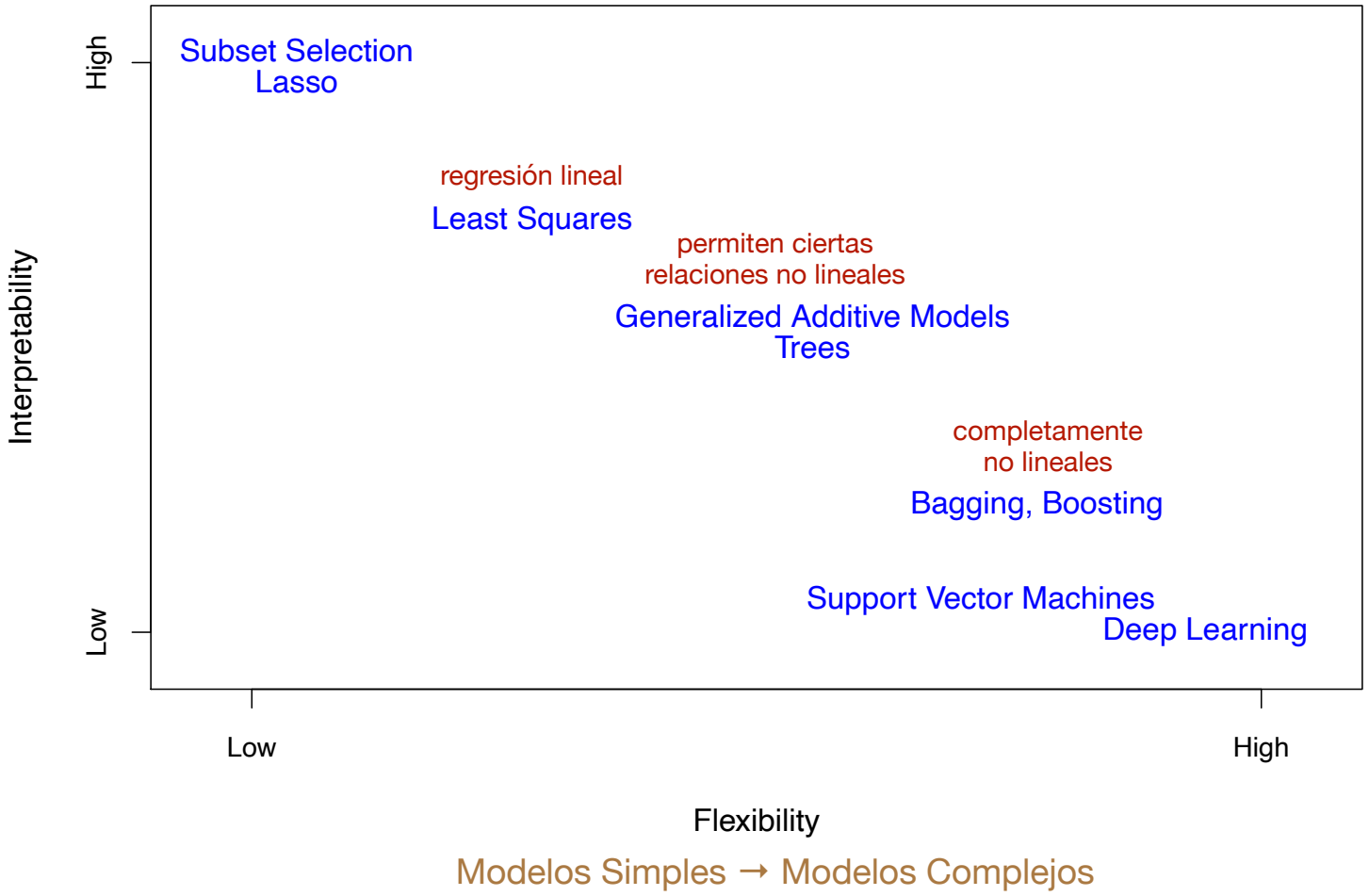
- ¿Por qué elegiríamos utilizar un método más restrictivo en lugar de un enfoque muy flexible?

- Si nos interesa principalmente la inferencia:

  - modelos restrictivos (y simples) son mucho más interpretables

  - Enfoques muy flexibles pueden conducir a estimaciones tan complicadas de f que es difícil entender cómo un predictor individual está asociado con la respuesta

Interpretabilidad y Flexibilidad (4)



Modelos Simples → Modelos Complejos

En general, a medida que aumenta la flexibilidad de un método, disminuye su interpretabilidad.

Interpretabilidad y Flexibilidad (5)

- Nos interesa la predicción (ej. el precio de una acción) :
  - requisito para el algoritmo es que prediga con precisión
- Mayor flexibilidad podría mejorar la precisión
- ¿Es mejor utilizar el modelo más flexible disponible?
  - ¡No, no es siempre el caso!
  - A menudo obtendremos predicciones más precisas con un método menos flexible
  - Este fenómeno tiene que ver con la posibilidad de *sobre ajuste* en métodos altamente flexibles

2.1.5 Aprendizaje Supervisado y No Supervisado

- La mayoría de los problemas de aprendizaje se clasifican en una de dos categorías:
  - supervisados o
  - no supervisados
- Aprendizaje por refuerzo

Aprendizaje Supervisado

- Para cada observación de las mediciones de los predictores hay una medición de respuesta asociada

$$\forall\ x_i,\ i = 1, \cdots, n\ \exists\ y_i$$

- La respuesta y_i *supervisa* el aprendizaje, el ajuste, de un modelo que relaciona la respuesta con los predictores
  - Predicción: obtener con precisión la respuesta para observaciones futuras
  - Inferencia: comprender la relación entre la respuesta y los predictores
- *Regresión lineal* y *regresión logística* (clásicos), *GAM*, *boosting* y *máquinas de vectores de soporte* (más modernos), *redes neuronales*, operan en el dominio del aprendizaje supervisado

Aprendizaje
No Supervidado

- La situación es algo más desafiante: para cada observación tenemos un vector de mediciones de predictores pero ninguna respuesta asociada

$$\forall\ x_i,\ i = 1, \cdots, n\ !\exists\ y_i$$

- *No supervisado*: porque carecemos de una variable de respuesta y_i que pueda supervisar el aprendizaje y el análisis
- No es posible ajustar un modelo, ej. de *regresión lineal*, ya que no hay ninguna variable de respuesta que predecir

Aprendizaje No Supervidado (2)

- ¿Qué tipo de análisis estadístico es posible?
- Podemos buscar comprender las relaciones entre las variables o entre las observaciones.
  - Ej.: Análisis de conglomerados o agrupamiento (*clustering*)
  - Determinar, sobre la base de $x_1, \cdots, x_n$ si las observaciones se dividen en grupos distintos

Estudio de Segmentación de Mercado

- Podríamos observar múltiples características (variables) de los clientes potenciales: código postal, ingresos familiares y hábitos de compra

    - Sospechamos que los clientes se dividen en diferentes grupos: los que gastan mucho y los que gastan poco

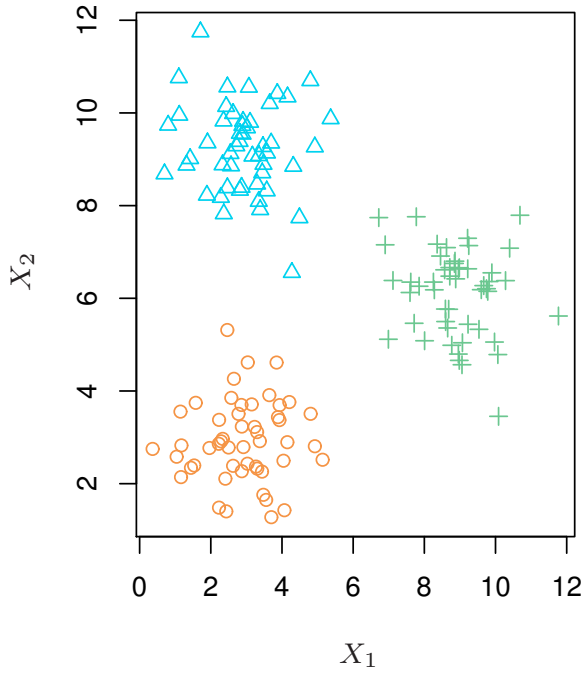
- Información sobre patrones de gasto de cada cliente no está disponible (no sabemos cuanto gasta cada cliente potencial) → no es posible un análisis supervisado

Estudio de Segmentación de Mercado (2)

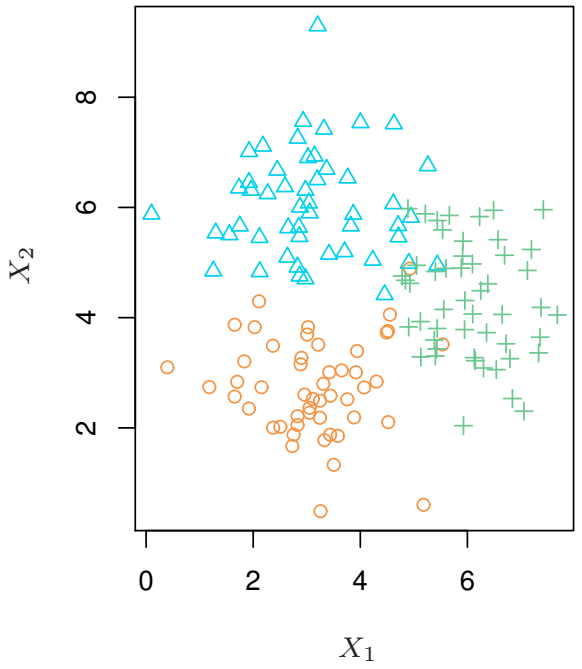
- Podemos intentar agrupar a los clientes en función de las variables medidas, para identificar distintos grupos de clientes potenciales

- Identificar dichos grupos puede ser interesante porque podría ser que los grupos difieran con respecto a alguna propiedad de interés, como los hábitos de gasto

Datos de Agrupamiento (Ej.)



- Tres grupos bien separados
- Un método de agrupación debería identificar exitosamente a los grupos



- Cierta superposición entre los grupos
- Tarea de agrupación es más desafiante

Regresión vs. Clasificación

- Variables cuantitativas: valores numéricos
  - La edad, la altura o los ingresos de una persona, el valor de una casa, el precio de una acción, …
- Variables cualitativas (categóricas): valores en una de K clases o categorías diferentes.
  - El estado civil (casada o no), la marca del producto (A, B o C), si una persona incumple una deuda (sí o no) o un diagnóstico de cáncer (mielogenosis aguda, leucemia, leucemia linfoblástica o sin leucemia), …

Regresión vs. Clasificación (2)

- Problemas de regresión
  - respuesta cuantitativa
- Problemas de clasificación
  - respuesta cualitativa
- La distinción entre métodos de aprendizaje no siempre es tan clara

Regresión vs. Clasificación (3)

- La *regresión lineal* de mínimos cuadrados se utiliza con una respuesta cuantitativa
- La *regresión logística* se utiliza típicamente con una respuesta cualitativa (de dos clases o binaria)
    - La *regresión logística* es un método de clasificación
    - Pero como estima probabilidades de clase, también se puede considerar como un método de regresión
- Métodos como *K-vecinos más cercanos* y *boosting*, se pueden utilizar en el caso de respuestas cuantitativas o cualitativas

Regresión vs. Clasificación (4)

- Tendemos a seleccionar métodos de aprendizaje en función de si la respuesta es cuantitativa o cualitativa
  - *Regresión lineal* cuando sea cuantitativa y la *regresión logística* cuando sea cualitativa (ej.)
- Sin embargo, generalmente se considera menos importante si los **predictores** son cualitativos o cuantitativos
- La mayoría de los métodos de aprendizaje *se pueden aplicar independientemente del tipo de variable predictiva*, siempre que los predictores cualitativos se codifiquen adecuadamente antes de realizar el análisis

2.2

Evaluación de la Precisión del Modelo

1. Medir la calidad del ajuste
2. El canje entre sesgo y varianza
3. El entorno de clasificación

Varios Métodos

- Amplia gama de métodos de aprendizaje estadístico
- ¿Por qué son necesarios tantos enfoques diferentes, en lugar de un único método óptimo?
  - Ningún método domina a los demás sobre todos los conjuntos de datos posibles (*no free lunch*)
- Una tarea importante es *seleccionar qué método produce los mejores resultados para un conjunto de datos* dado
  - Una de las partes más desafiantes del aprendizaje estadístico en la práctica

2.2.1

Medir la Calidad del Ajuste

- Para evaluar el rendimiento de un método de aprendizaje necesitamos alguna forma de medir qué tan bien sus predicciones coinciden realmente con los datos observados.

- Necesitamos cuantificar en qué medida el valor de respuesta previsto para una observación determinada se acerca al valor de respuesta real para esa observación

Error Cuadrático Medio

- En el ámbito de la *regresión*, la medida más utilizada es el *error cuadrático medio* (MSE)

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

$\hat{f}(x_i)$ es la predicción que da $\hat{f}$ para la i-ésima observación

- El MSE será pequeño si las respuestas previstas son muy cercanas a las respuestas verdaderas. De otra forma será grande

Error Cuadrático Medio (2)

- Interesa la precisión de las predicciones cuando aplicamos el método a *datos de prueba* nunca antes vistos
  - No importa qué tan bien funcione con los *datos de entrenamiento*
- Preocupa qué tan bien predecirá el precio de las acciones de mañana o del próximo mes. No el precio de ayer
- Queremos predecir con precisión el riesgo de diabetes para futuros pacientes en función de sus mediciones clínicas. No interesa predecir el riesgo de diabetes para los pacientes que ya sabemos sufren de diabetes

Entrenamiento y Prueba

Datos Entrenamiento

$$x_i = (x_{i1}, \cdots, x_{ip})$$
$$y_i = (y_{i1}, \cdots, y_{ir})$$

$$\begin{matrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{matrix} \qquad \begin{matrix} \hat{f}(x_1) \\ \hat{f}(x_2) \\ \vdots \\ \hat{f}(x_n) \end{matrix}$$

MSE entrenamiento

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

Datos Prueba

no vistos durante el entrenamiento

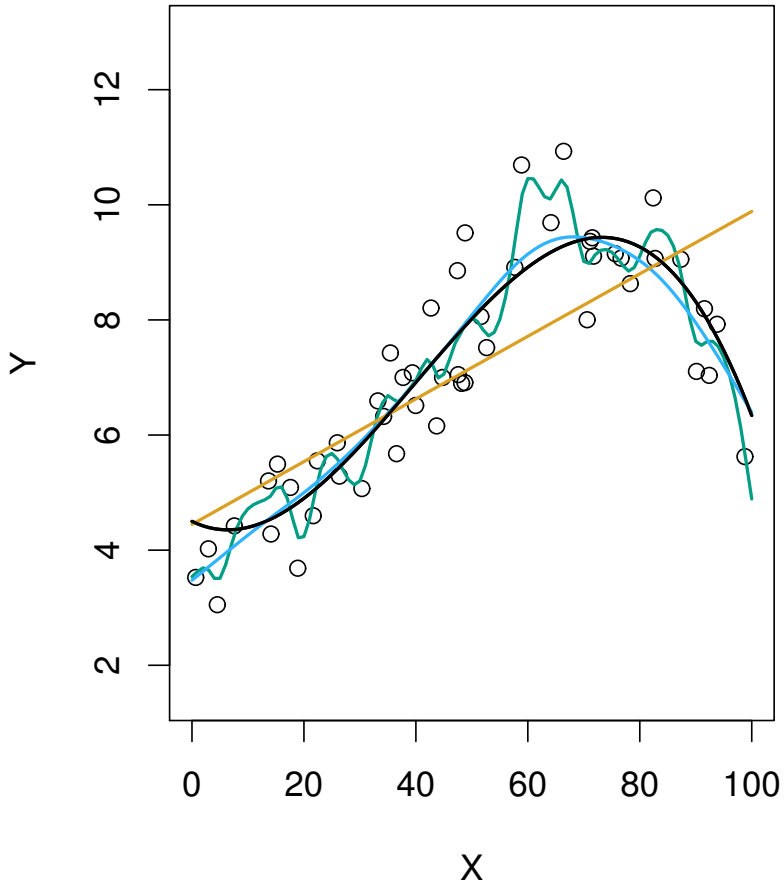
$$\begin{matrix} x_1^p & y_1^p \\ \vdots & \vdots \\ x_m^p & y_m^p \end{matrix} \qquad \begin{matrix} \hat{f}(x_1^p) \\ \vdots \\ \hat{f}(x_m^p) \end{matrix}$$

MSE prueba lo más bajo posible

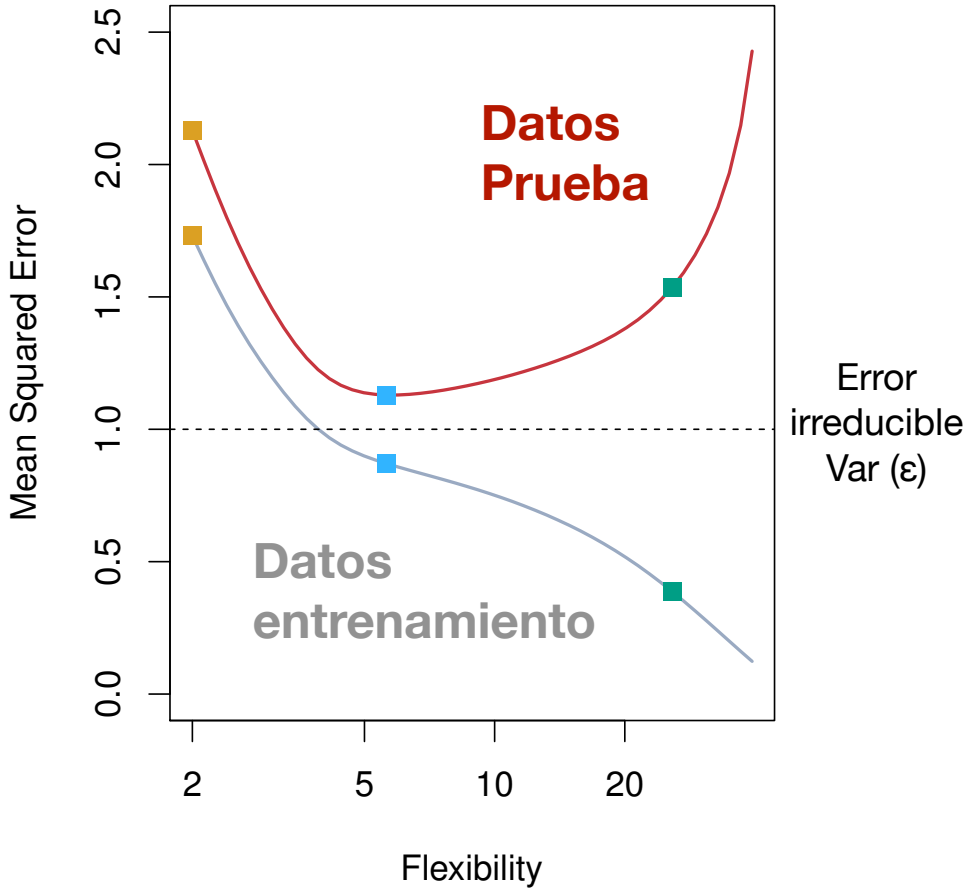
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i^p - \hat{f}(x_i^p))^2$$

No hay garantía de que el método con el MSE de entrenamiento más bajo también tenga el MSE de prueba más bajo

Entrenamiento y Prueba (2)



Datos simulados de f en negro. Estimaciones de f : la línea de regresión lineal y dos ajustes de spline suavizados (curvas azul y verde)



MSE de entrenamiento, MSE de prueba y MSE de prueba mínimo posible Var(ε) (línea discontinua)

Entrenamiento y Prueba (3)

- A mayor flexibilidad del método de aprendizaje estadístico, menor es el MSE de entrenamiento
- Una forma de U en el MSE de prueba.
- Esta es una propiedad fundamental del aprendizaje estadístico que se mantiene independientemente del conjunto de datos particular que se tenga a mano y del método estadístico que se utilice.
- A medida que aumenta la flexibilidad del modelo, el MSE de entrenamiento disminuirá, pero es posible que el MSE de prueba no.

Entrenamiento y Prueba (4)

- Cuando un método dado produce un MSE de entrenamiento pequeño, pero un MSE de prueba grande, se dice que estamos sobreajustando los datos

    - El procedimiento de aprendizaje está trabajando demasiado para encontrar patrones en los datos de entrenamiento y puede estar detectando algunos patrones que son causados simplemente por el azar en lugar de por propiedades verdaderas de la función desconocida f

- Cuando sobre ajustamos los datos de entrenamiento, el MSE de la prueba será muy grande porque los supuestos patrones que el método encontró en los datos de entrenamiento simplemente no existen en los datos de prueba

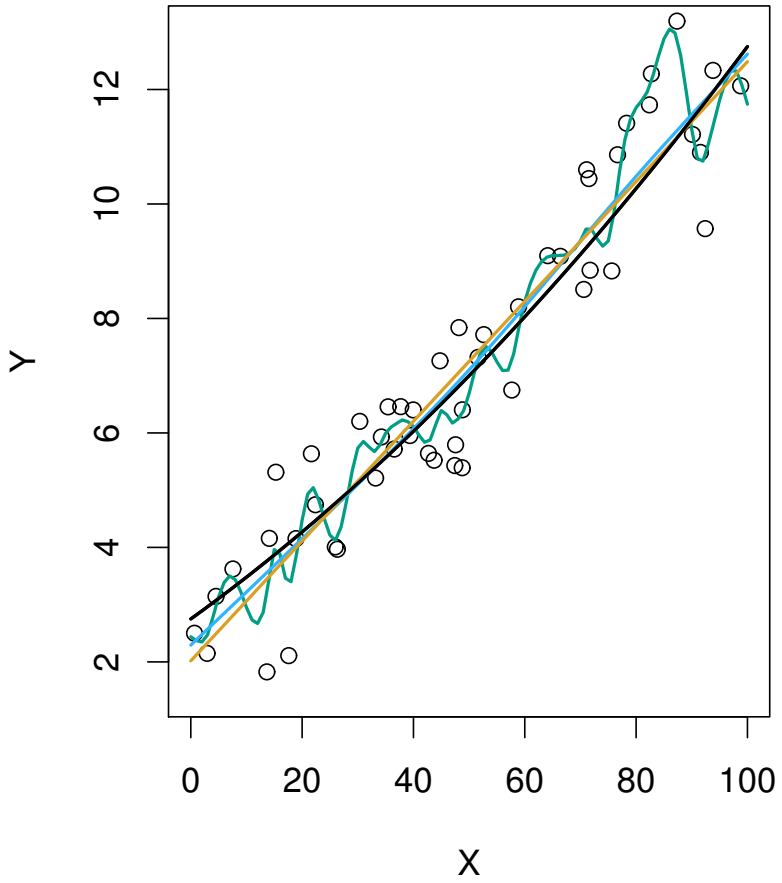
Entrenamiento y Prueba (5)

- La mayoría de los métodos de aprendizaje estadístico, buscan minimizar el MSE de entrenamiento.

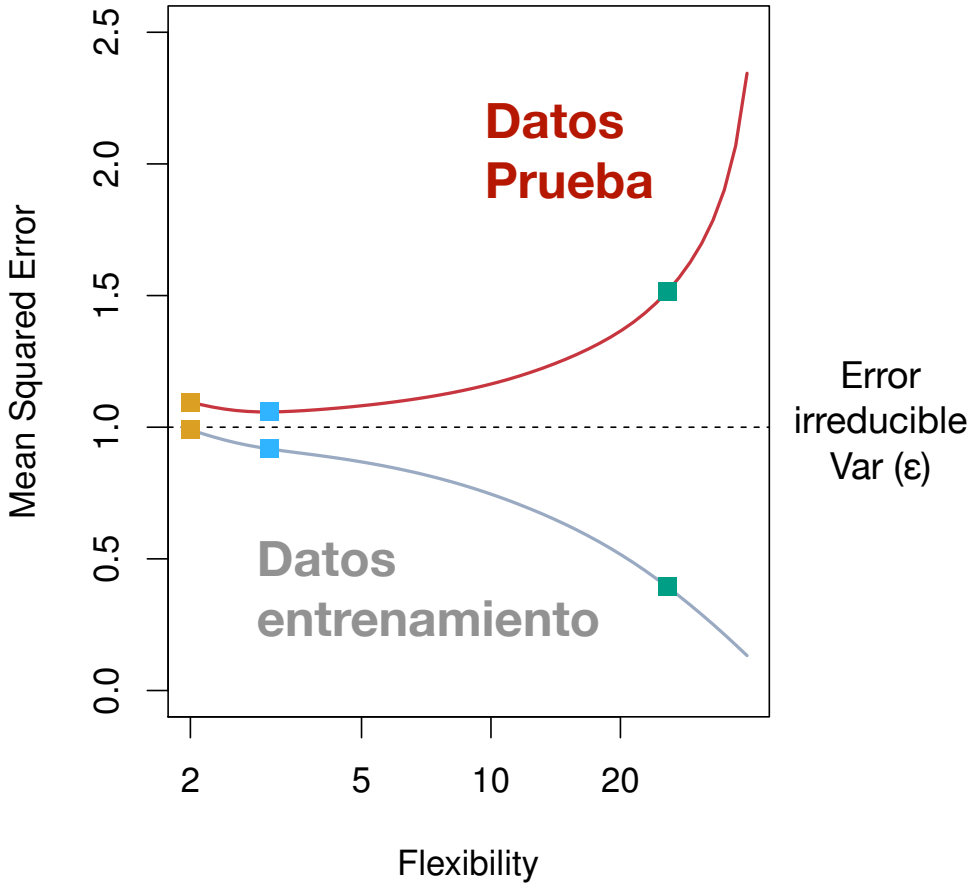
  - Independientemente de si se hay *sobre ajuste* esperamos: *MSE de entrenamiento < MSE de prueba*

- ***Sobre ajuste*****:** un modelo menos flexible produce un MSE de prueba más pequeño.

Entrenamiento y Prueba (6)

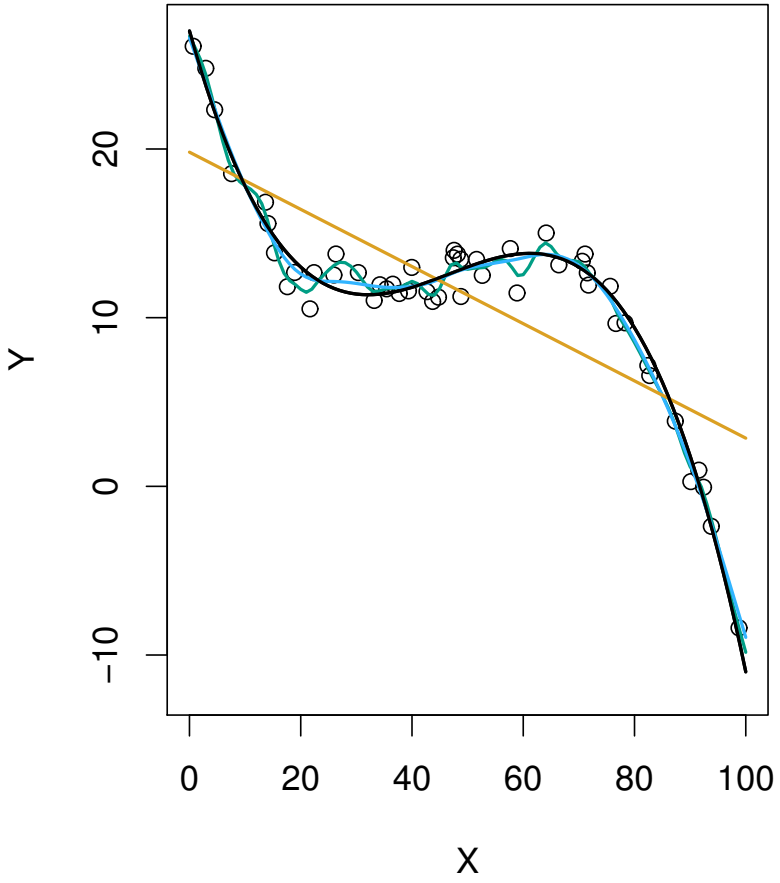


Datos simulados de f en negro. Estimaciones de f : la línea de regresión lineal y dos ajustes de spline suavizados (curvas azul y verde)

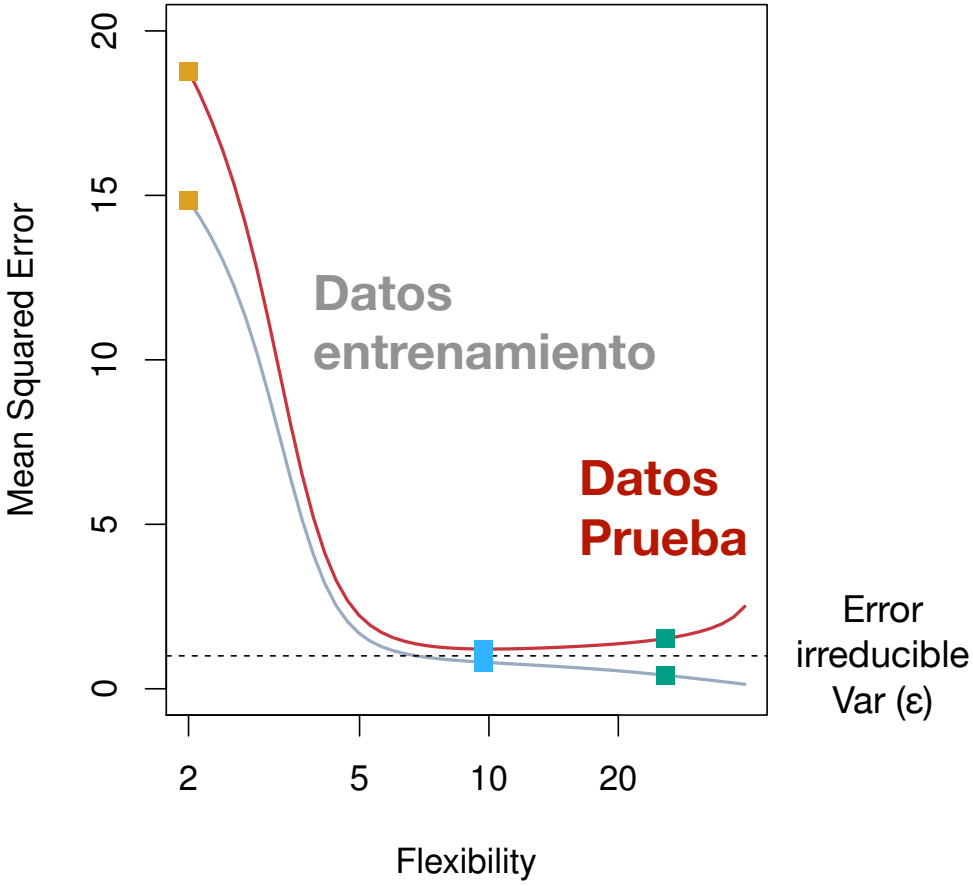


MSE de entrenamiento, MSE de prueba y MSE de prueba mínimo posible Var(ε) (línea discontinua)

Entrenamiento y Prueba (7)



Datos simulados de f en negro. Estimaciones de f : la línea de regresión lineal y dos ajustes de spline suavizados (curvas azul y verde)



MSE de entrenamiento, MSE de prueba y MSE de prueba mínimo posible Var(ε) (línea discontinua)

Entrenamiento y Prueba (8)

- En la práctica, se puede calcular el MSE de entrenamiento con relativa facilidad
    - Estimar el MSE de prueba es considerablemente más difícil porque normalmente no hay datos de prueba disponibles
- Como ilustran los tres ejemplos anteriores, el nivel de flexibilidad correspondiente al modelo con el MSE de prueba mínimo puede variar considerablemente entre conjuntos de datos
- Hay varios enfoques que pueden usarse en la práctica para estimar este punto mínimo. Por ejemplo, la validación cruzada que permite estimar el MSE de prueba utilizando los datos de entrenamiento

2.2.2

El Canje Sesgo -Varianza

- La forma de U observada en las curvas *MSE de prueba* es el resultado de dos propiedades en competencia de los *métodos de aprendizaje* estadístico
  - Sesgo
  - Varianza

MSE de Prueba Esperado en x^p

$$E(y^p - \hat{f}(x^p))^2$$

- Es el *MSE de prueba* promedio
  - si estimáramos repetidamente f usando una gran cantidad de conjuntos de entrenamiento diferentes y probáramos cada $\hat{f}$ obtenida en x^p

Descomposición de MSE

- *MSE de prueba* en x^p se puede descomponer en la suma de tres cantidades fundamentales
  - varianza de $\hat{f}(x^p)$,
  - sesgo de $\hat{f}(x^p)$ (*bias*)
  - varianza de los términos de error ϵ

$$E(y^p - \hat{f}(x^p))^2 = Var(\hat{f}(x^p)) + [Bias((x^p))]^2 + Var(\epsilon)$$

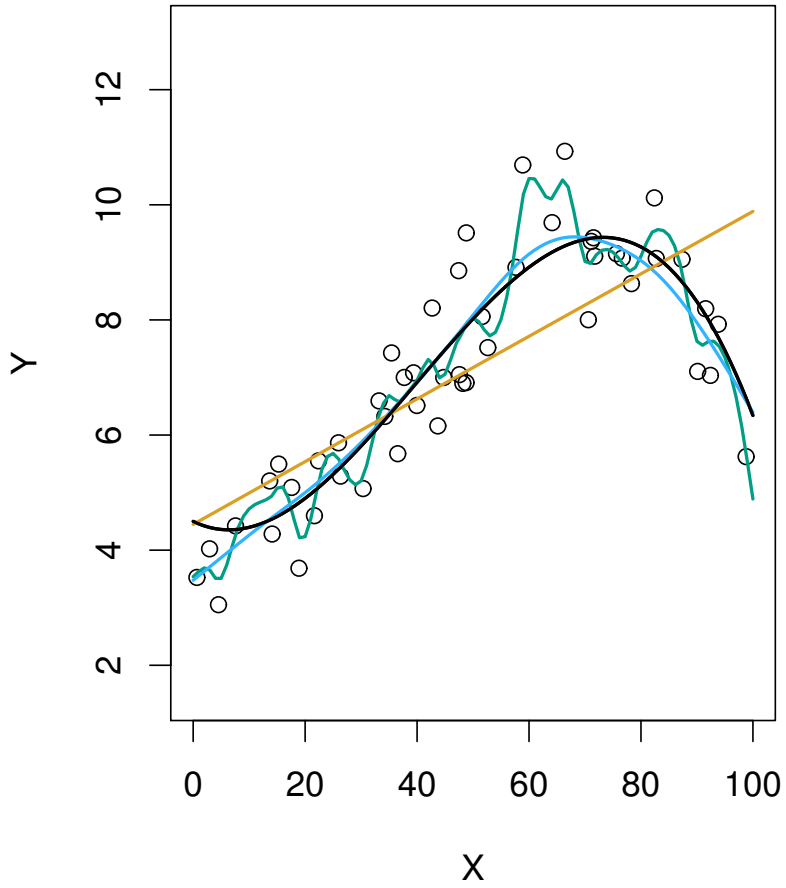
$$MSE \geq Var(\epsilon)$$

- Para minimizar el *MSE de prueba esperado* → seleccionar un método de aprendizaje que logre simultáneamente una ***varianza baja*** y un ***sesgo bajo***

Varianza

- ***Varianza*:** cantidad en la que $\hat{f}$ cambiaría si la estimáramos utilizando un conjunto de datos de entrenamiento diferente.
- Idealmente: la estimación de f no debería variar demasiado entre conjuntos de entrenamiento.
- Si un método tiene una alta varianza, entonces pequeños cambios en los datos de entrenamiento pueden resultar en grandes cambios en $\hat{f}$.
- En general, los métodos estadísticos más flexibles tienen una mayor varianza.

Varianza (2)



Varianza alta

La curva verde flexible sigue muy de cerca las observaciones y tiene una varianza alta. La estimación $\hat{f}$ cambia considerablemente si cualquiera de estos puntos de datos cambian

Varianza baja

La línea naranja es relativamente inflexible y tiene una varianza baja. Mover cualquier observación probablemente provocará solo un pequeño cambio en la posición de la línea

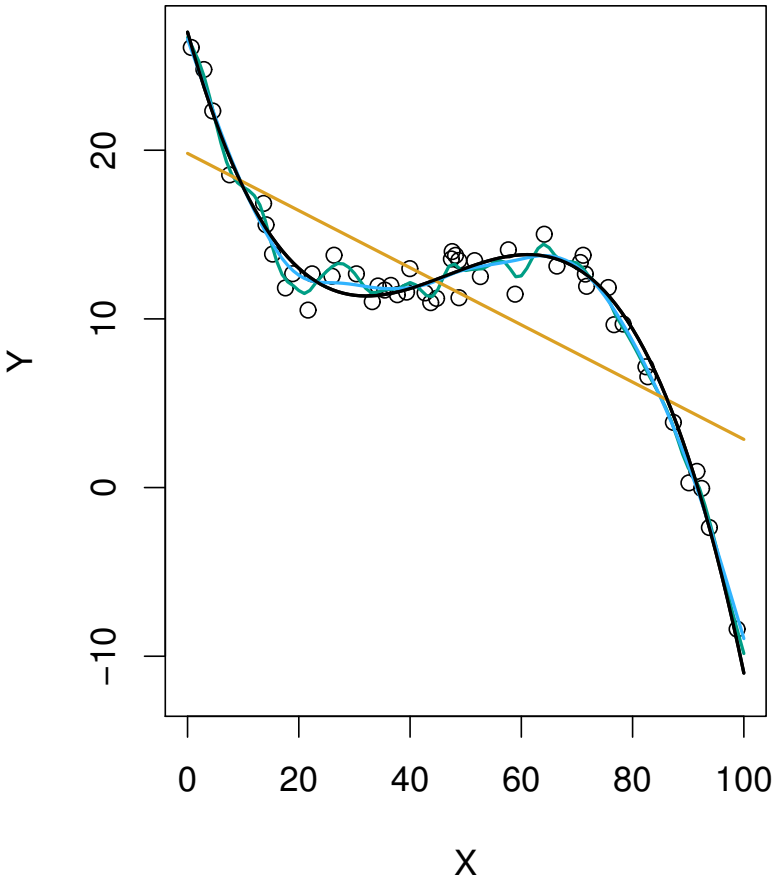
Sesgo

- El sesgo se refiere al error que se introduce al aproximar un problema de la vida real, que puede ser complejo, mediante un modelo mucho más simple

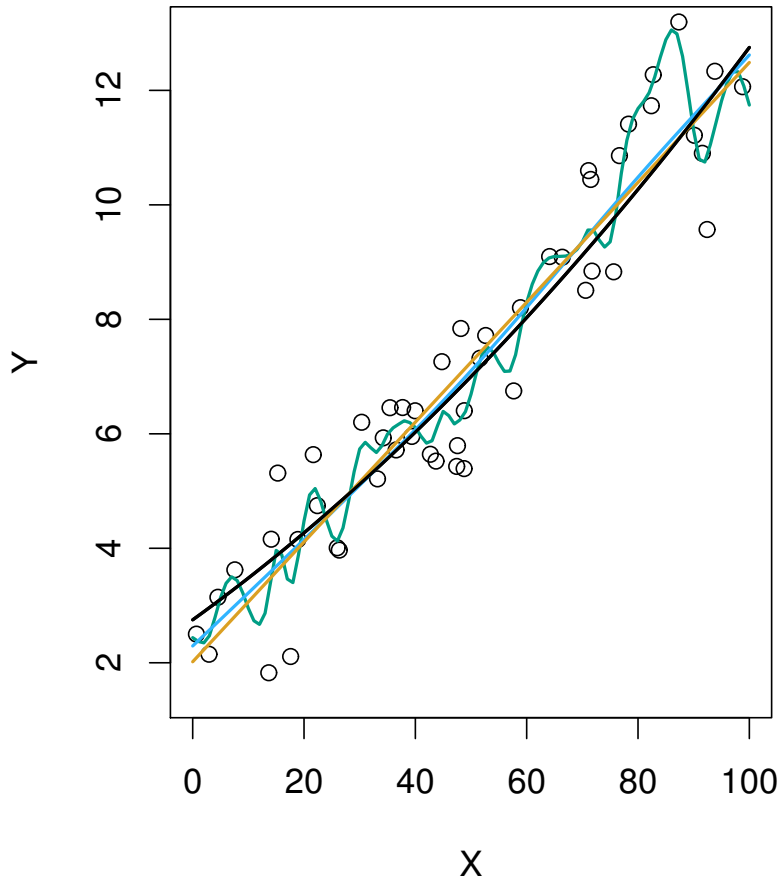
- Por ejemplo, es poco probable que cualquier problema de la vida real realmente tenga una relación lineal simple, por lo que realizar una regresión lineal (RL) sin duda dará como resultado algún sesgo en la estimación de f

- Generalmente, los métodos más flexibles dan como resultado menos sesgo

Sesgo (2)

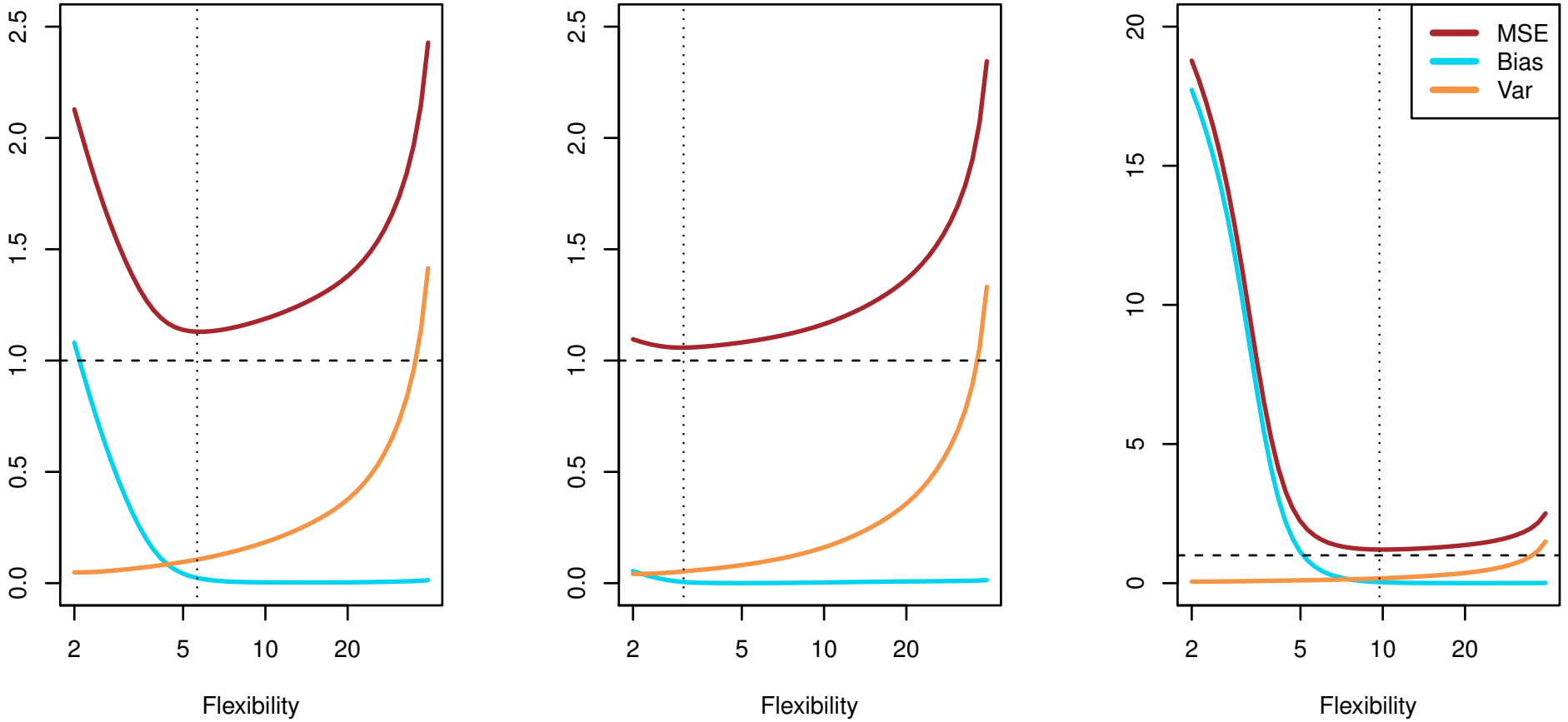


- La f verdadera es no lineal
- No importa cuántas observaciones de entrenamiento recibamos, la RL no producirá una estimación precisa
- RL da como resultado un alto sesgo

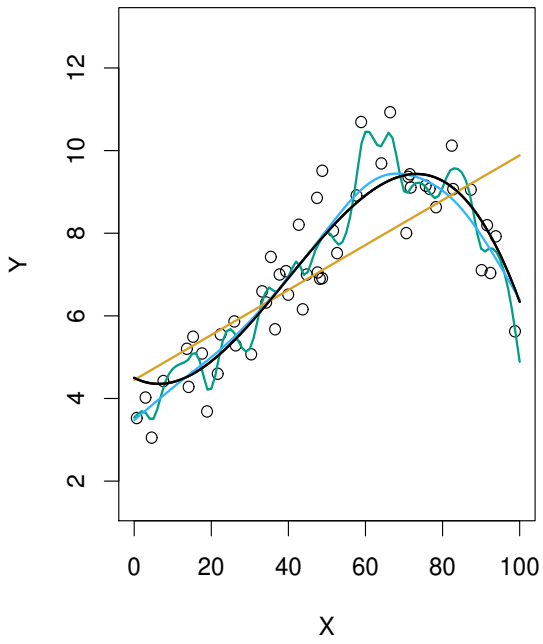


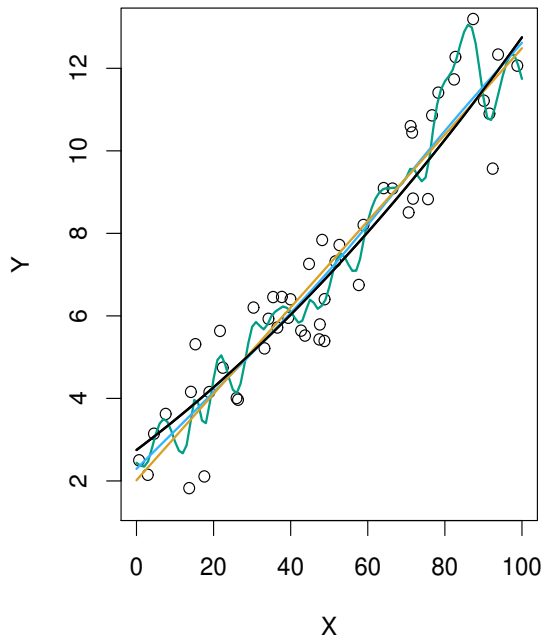
- La f verdadera está muy cerca de ser lineal
- Con suficientes datos, debería ser posible que la RL produzca una estimación precisa

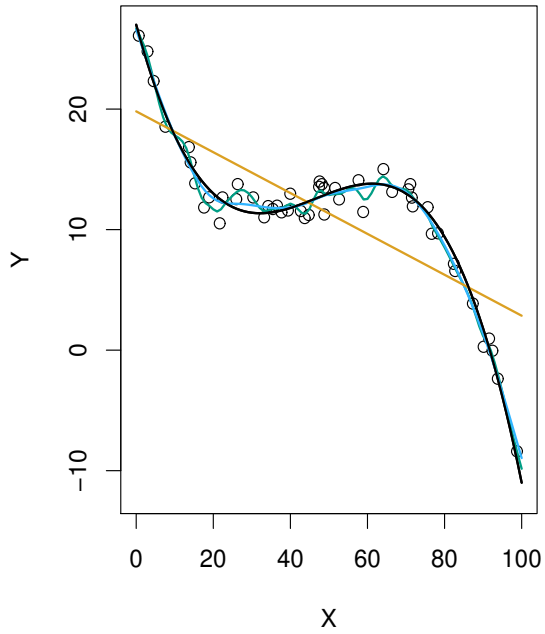
Canje Sesgo -Varianza

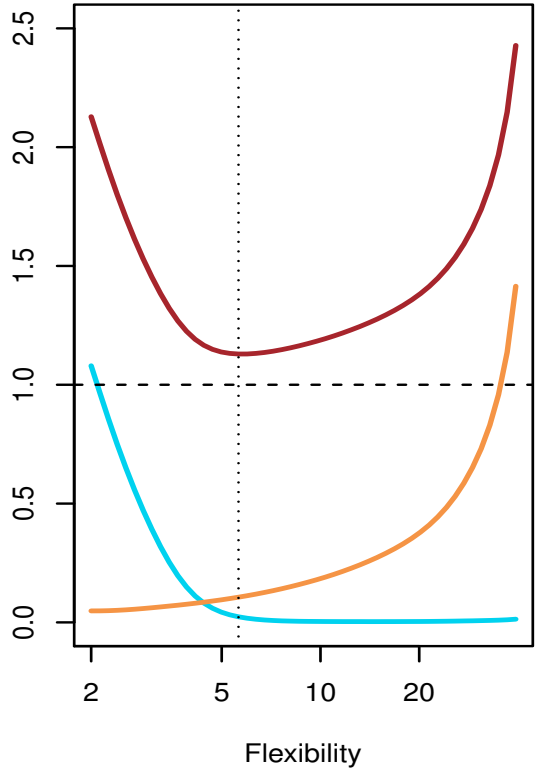


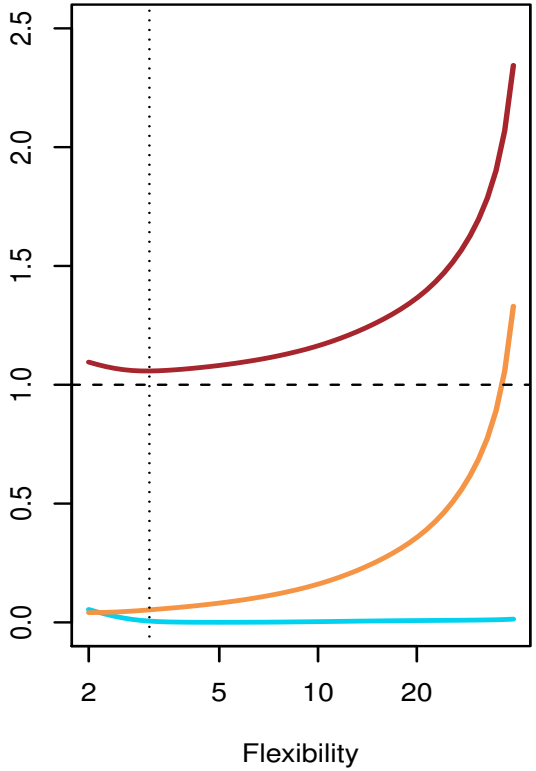
Sesgo al cuadrado, varianza, Var(ε) (línea discontinua) y MSE de prueba para los tres conjuntos de datos de las Figuras 2.9 a 2.11. La línea de puntos vertical indica el nivel de flexibilidad correspondiente al MSE de prueba más pequeño.

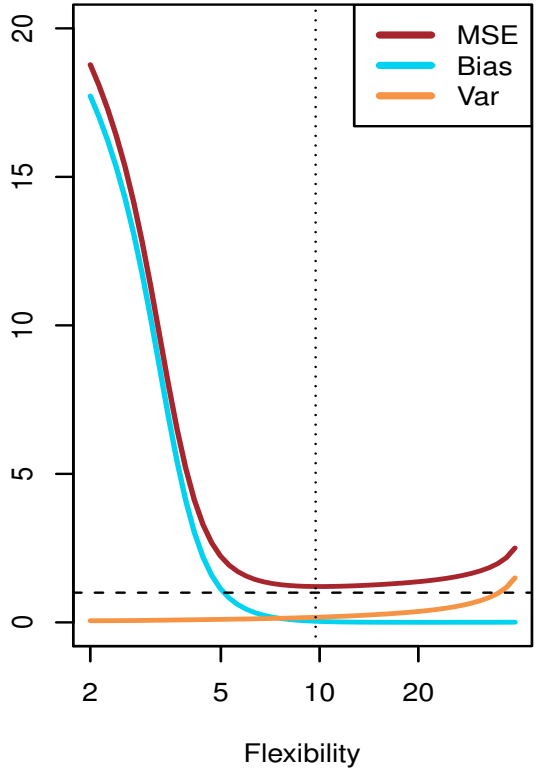

Y
X
0
20
40
60
80
100
2
4
6
8
10
12


Y
X
0
20
40
60
80
100
2
4
6
8
10
12


Y
X
0
20
40
60
80
100
−10
0
10
20


Flexibility
2
5
10
20
0.0
0.5
1.0
1.5
2.0
2.5


Flexibility
2
5
10
20
0.0
0.5
1.0
1.5
2.0
2.5


MSE
Bias
Var
Flexibility
2
5
10
20
0
5
10
15
20

Canje Sesgo -Varianza (3)

- Como regla general, a medida que utilizamos métodos más flexibles, la varianza aumentará y el sesgo disminuirá.
- Tasa de cambio relativa de estas dos cantidades determina si el MSE de prueba aumenta o disminuye
- A medida que aumentamos la flexibilidad de una clase de métodos, el sesgo tiende inicialmente a disminuir más rápido de lo que aumenta la varianza. En consecuencia, el MSE de prueba esperado disminuye
- Sin embargo, en algún momento, aumentar la flexibilidad tiene poco impacto en el sesgo, pero comienza a aumentar significativamente la varianza. Cuando esto sucede, el MSE de prueba aumenta

2.2.3

Entorno de Clasificación

- Hasta ahora, la discusión sobre la precisión del modelo se ha centrado en el escenario de la *regresión*
- Muchos de los conceptos, como el canje entre sesgo y varianza, se transfieren al entorno de *clasificación* con algunas modificaciones debido al hecho de que y_i no es cuantitativo

Clasificación

- Supongamos que buscamos estimar f sobre la base de observaciones de entrenamiento
  - $(x_1, y_1), \ldots, (x_n, y_n)$
  - $y_1, \ldots, y_n$ son cualitativos
- El enfoque más común para cuantificar la precisión de nuestra estimación $\hat{f}$ es la tasa de error de entrenamiento
  - La proporción de errores que se cometen si aplicamos nuestra estimación $\hat{f}$ a las observaciones de entrenamiento

Tasa de Error de Entrenamiento

- Calcula la fracción de clasificaciones incorrectas

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

  - y_i etiqueta de clase real
  - $\hat{y}_i$ etiqueta de clase predicha
  - $I(y_i \neq \hat{y}_i)$
    - 1 si $y_i \neq \hat{y}_i$ clasificación errónea
    - 0 si $y_i = \hat{y}_i$ clasificación correcta

Clasificación (2)

- Similar a regresión, lo que más interesa son las *tasas de error de prueba*, que resultan de aplicar nuestro clasificador para probar observaciones que no se utilizaron en el entrenamiento.
- La tasa de error de prueba asociada con un conjunto de observaciones de prueba de la forma (x^p, y^p) está dada por

$$Ave\ I(y^p \neq \hat{y}^p)$$

  - $\hat{y}^p$ etiqueta de clase predicha que resulta de aplicar el clasificador a la observación de prueba con el predictor x^p
- Un buen clasificador es aquel para el cual el error de prueba es el más pequeño

El Clasificador de Bayes

- Es posible demostrar que la tasa de error de prueba dada se minimiza, en promedio, mediante un clasificador muy simple que asigna cada observación a la clase más probable, dados sus valores predictores.

- En otras palabras, simplemente deberíamos asignar una observación de prueba con el vector predictor x^p a la clase j para la cual la probabilidad es mayor

$$Pr(Y = j \,|\, X = x^p)$$

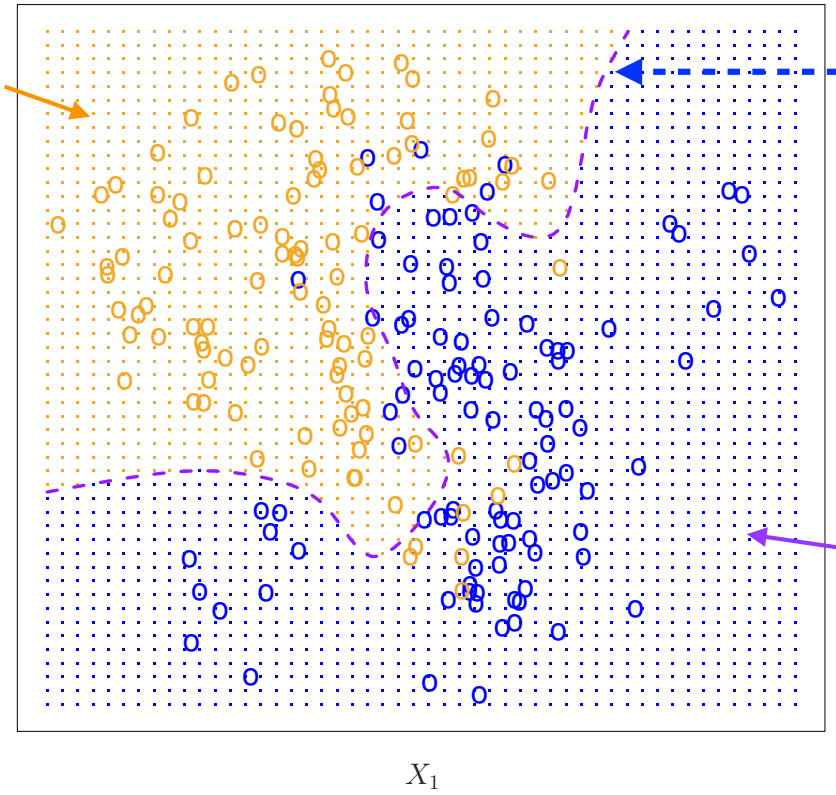
El Clasificador de Bayes (2)

$$Pr(Y = j \,|\, X = x^p)$$

- Probabilidad condicional: la probabilidad de que $Y = j$, dado el vector predictor observado $X = x^p$
- Este clasificador muy simple se llama clasificador de Bayes.
- Ej., problema de dos clases donde, el clasificador de Bayes corresponde a predecir
  - *clase uno* si $Pr(Y = 1 \,|\, X = x^p) > 0.5$
  - *clase dos* en caso contrario

El Clasificador de Bayes (3)

$Pr(Y = naranja \mid X) > 50\,\%$



$Pr(Y = naranja \mid X) = 50\,\%$

Límite de decisión de Bayes

$Pr(Y = naranja \mid X) < 50\,\%$

- Para cada valor de X_1 y X_2, existe una probabilidad diferente de que la respuesta sea naranja o violeta.
- Podemos calcular las probabilidades condicionales

Ejemplo utilizando un conjunto de **datos simulados** en un espacio bidimensional que consta de predictores X_1 y X_2. Los círculos naranja y violeta corresponden a observaciones de entrenamiento que pertenecen a dos clases diferentes.

El Clasificador de Bayes (3)

- El clasificador Bayes produce la tasa de error de prueba más baja posible, análoga al error irreducible, llamada la tasa de error de Bayes.
- El clasificador de Bayes siempre elegirá la clase para la cual la probabilidad es mayor
- Tasa de error

$$1 - \max_j \; Pr(Y = j \,|\, X = x^p)$$

$$1 - E\Big(\max_j \; Pr(Y = j \,|\, X)\Big)$$

- En el ejemplo, la tasa de error de Bayes es 0,133 > 0, porque las clases se superponen en la población verdadera

El Clasificador de Bayes (4)

- Para cada valor de X_1y X_2, existe una probabilidad diferente de que la respuesta sea naranja o azul.
- Dado que se trata de datos simulados, sabemos cómo se generaron los datos y podemos calcular las probabilidades condicionales para cada valor de X_1 y X_2.

El Clasificador de Bayes (5)

- En teoría, siempre nos gustaría predecir respuestas cualitativas utilizando el clasificador de Bayes
- Pero para *datos reales, no conocemos la distribución condicional de Y dado X*, por lo que calcular el clasificador de Bayes es imposible
- Por lo tanto, el clasificador de Bayes sirve como un estándar de oro inalcanzable con el que comparar otros métodos

K-Vecinos más Cercanos

- Muchos enfoques intentan *estimar la distribución condicional de Y dado X* y
- Clasificar una observación en la clase con la probabilidad estimada más alta.
- Uno de esos métodos es el clasificador *K-vecinos más cercanos* (KNN).

K-Vecinos más Cercanos (2)

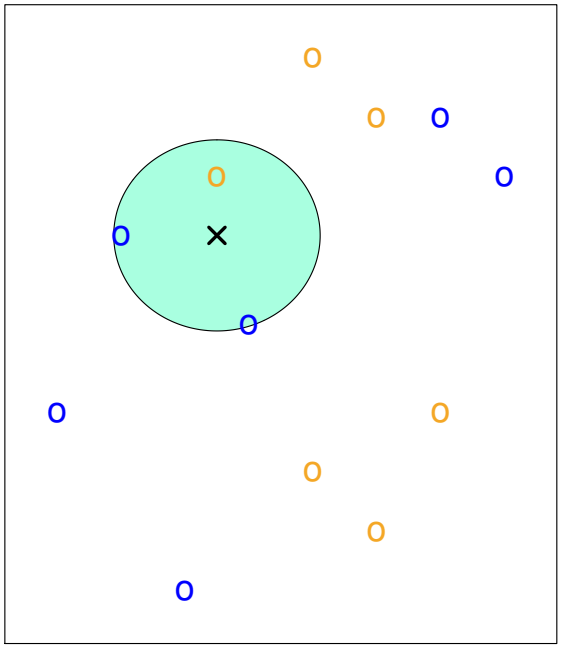
- Dado un entero positivo K y una observación de prueba x^p, el *clasificador KNN*
  - Identifica los K puntos en los datos de entrenamiento más cercanos a x^p, representados por N^P
  - Estima la probabilidad condicional para la clase j como la fracción de puntos en N^P cuyos valores de respuesta son iguales a j

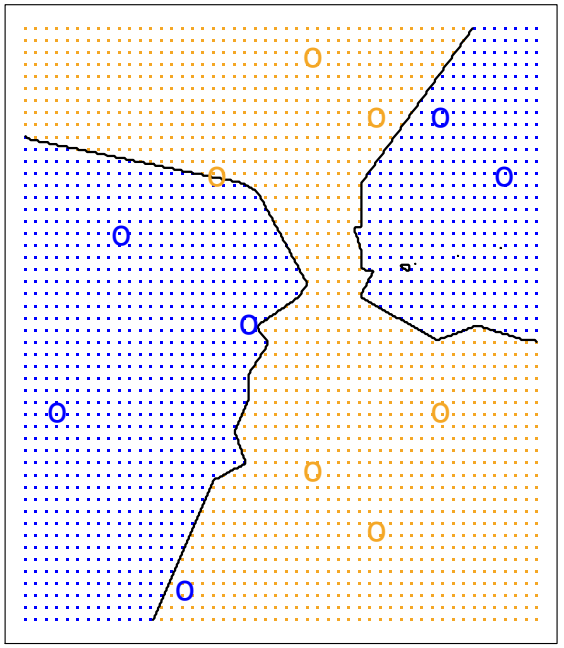
$$Pr(Y = j \,|\, X = x^p) = \frac{1}{K} \sum_{i \in N^p} I(y_i = j)$$

  - KNN clasifica la observación de prueba x^p en la clase que tiene la mayor probabilidad

KNN : Ejemplo

- Datos de entrenamiento, 6 observaciones azules y 6 naranjas
- Hacer una predicción para el punto marcado con la cruz negra



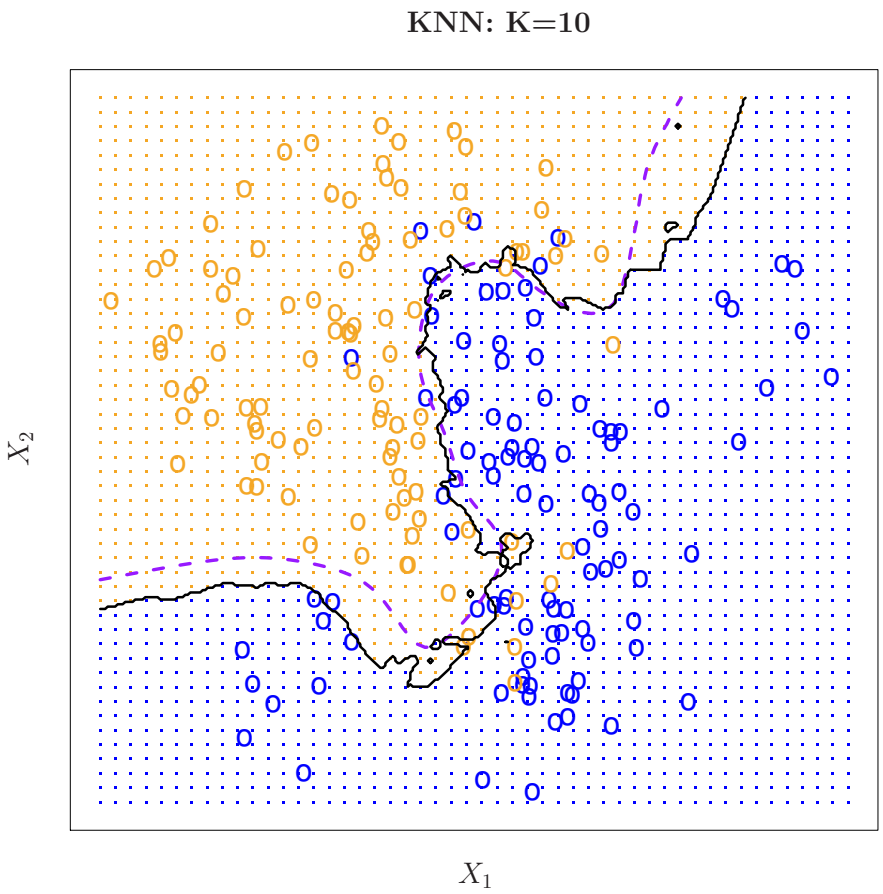


- K = 3, KNN identificará primero las 3 observaciones más cercanas a la cruz
- Vecindad: 2 puntos azules y 1 punto naranja: probabilidades estimadas de 2/3 y 1/3 para las clases
- KNN predecirá que la cruz negra pertenece a la clase azul

- Aplicamos KNN con K = 3 en todos los valores posibles para X1 y X2
- Dibujamos el límite de decisión KNN correspondiente

KNN y Bayes Optimo

- Enfoque simple
- KNN → clasificadores cercanos del Bayes óptimo
- Límite de decisión de KNN es muy cercano al del clasificador Bayes
- KNN no conoce la distribución verdadera

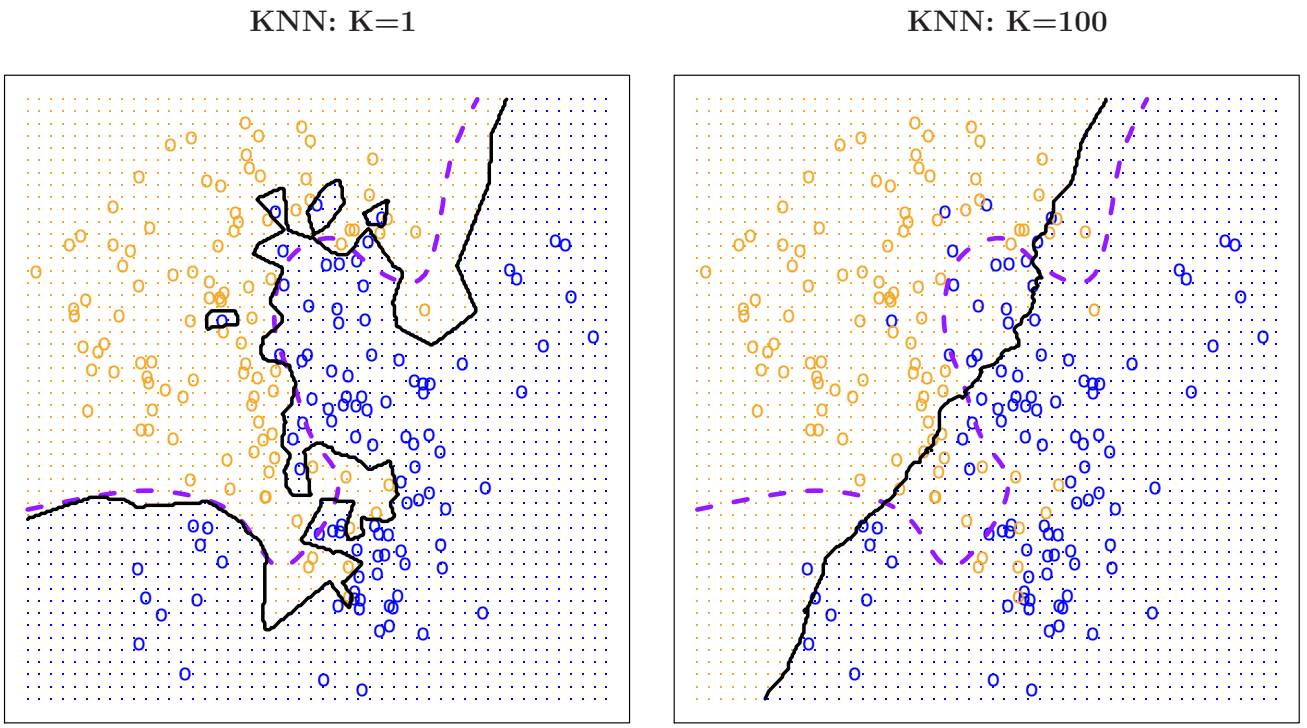


Tasa de Error de Prueba		
Bayes	0,1304	- - - - -
KNN	0,1363	———

Límite de decisión de KNN, K = 10
Conjunto de datos simulados

Número de Vecinos K

- K=1, límite de decisión es demasiado flexible
- Encuentra patrones en los datos que no corresponden al límite de decisión de Bayes
- Clasificador con sesgo bajo y varianza muy alta



- A medida que K crece, el clasificador se vuelve menos flexible
- K=100, el límite de decisión casi lineal
- Clasificador de baja varianza pero alto sesgo

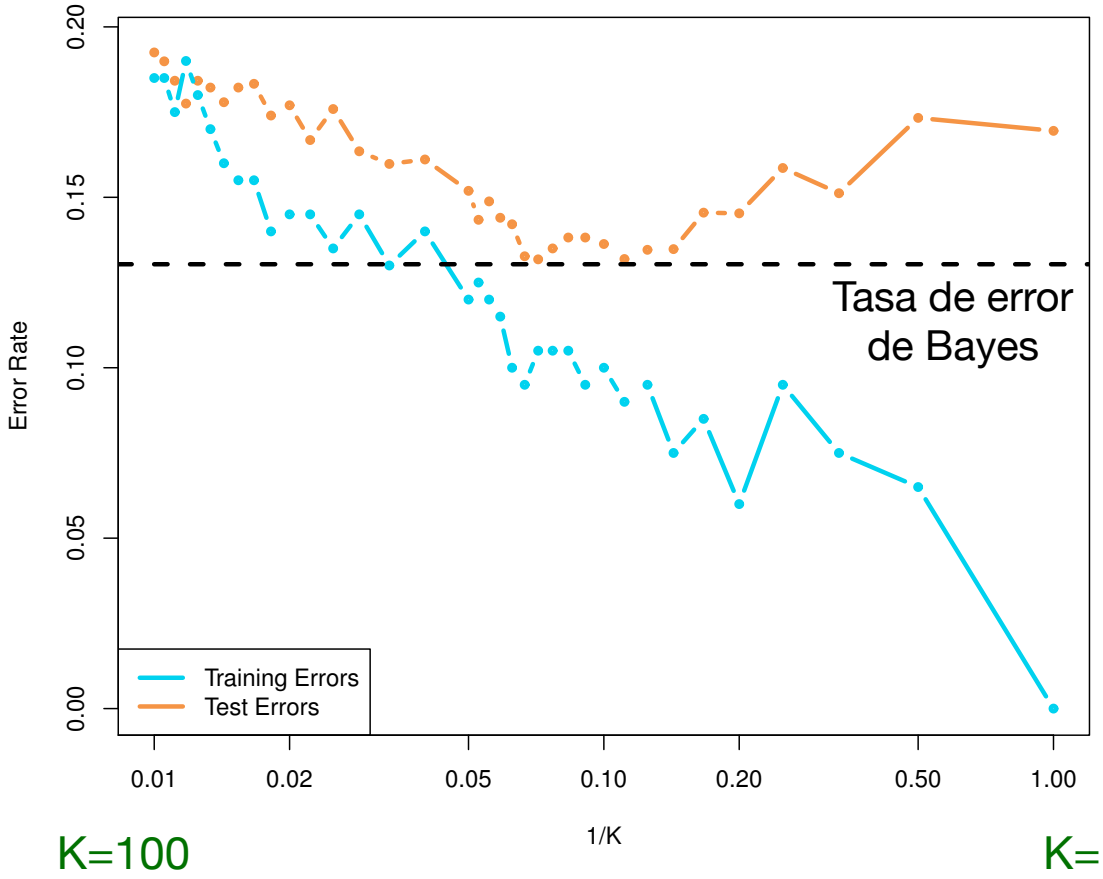
Ajustes de KNN a datos simulados, K = 1 y K = 100

La elección de K tiene un efecto drástico en el clasificador KNN obtenido

Tasa de Error de Prueba	
Bayes	0,1304
K = 1	0,1695
K = 10	0,1363
K = 100	0,1925

Entrenamiento y Prueba

- La tasa de error de entrenamiento disminuye constantemente a medida que aumenta la flexibilidad
- El error de prueba presenta una forma de U característica, que disminuye al principio antes de aumentar nuevamente cuando el método se vuelve excesivamente flexible y se sobreajusta



- K = 1, la tasa de error de entrenamiento de KNN es 0, pero la tasa de error de prueba puede ser bastante alta.
- En general, a medida que utilizamos métodos de clasificación más flexibles, la tasa de error de entrenamiento disminuirá, pero es posible que la tasa de error de prueba no

Errores de entrenamiento y prueba con KNN en función de 1/K
A medida que aumenta 1/K, el método se vuelve más flexible

No existe una relación fuerte entre la tasas de error de entrenamiento y prueba
(al igual que en en el entorno de regresión)

2.3

Laboratorio: Introducción a Python

1. Para empezar
2. Comandos básicos
3. Introducción a Python numérico
4. Gráficos
5. Notación de secuencias y cortes
6. Indexando datos
7. Cargando datos
8. Bucles *for*
9. Resúmenes gráficos y numéricos adicionales

2.4

Ejercicios