Introducción al Aprendizaje automático

Dra Ana Georgina Flesia

Optativa Ciencias de la Computación
FaMAF-UNC
Oficina 370
georgina.flesia@unc.edu.ar

2020

Objetivo central del aprendizaje supervisado:

Predicción, no causalidad

El foco está puesto en poder realizar predicciones de interés bajo condiciones complejas, no en estudiar el mecanismo causal que rige bajo esas condiciones

Se necesita un modelo que generalize bien con datos no vistos antes.

Suposiciones fundamentales del aprendizaje supervisado

- suposición i.i.d.: los ejemplos de entrenamiento y prueba son muestras independientes provenientes de la misma distribución de probabilidad.

Suposiciones fundamentales del aprendizaje supervisado

- suposición i.i.d.: los ejemplos de entrenamiento y prueba son muestras independientes provenientes de la misma distribución de probabilidad.
- El error del modelo sobre la muestra de entrenamiento y sobre la de test se suponen iguales, aun cuando solo se entrenó el modelo sobre la muestra de entrenamiento.

Suposiciones fundamentales del aprendizaje supervisado

- suposición i.i.d.: los ejemplos de entrenamiento y prueba son muestras independientes provenientes de la misma distribución de probabilidad.
- El error del modelo sobre la muestra de entrenamiento y sobre la de test se suponen iguales, aun cuando solo se entrenó el modelo sobre la muestra de entrenamiento.
- El error de entrenamiento o exactitud provee un estimador con un desvío optimista del desempeño al generalizar

Capacidad del Modelo

- Underfitting: los errores de entrenamiento y testeo son ambos grandes. (high Bias)

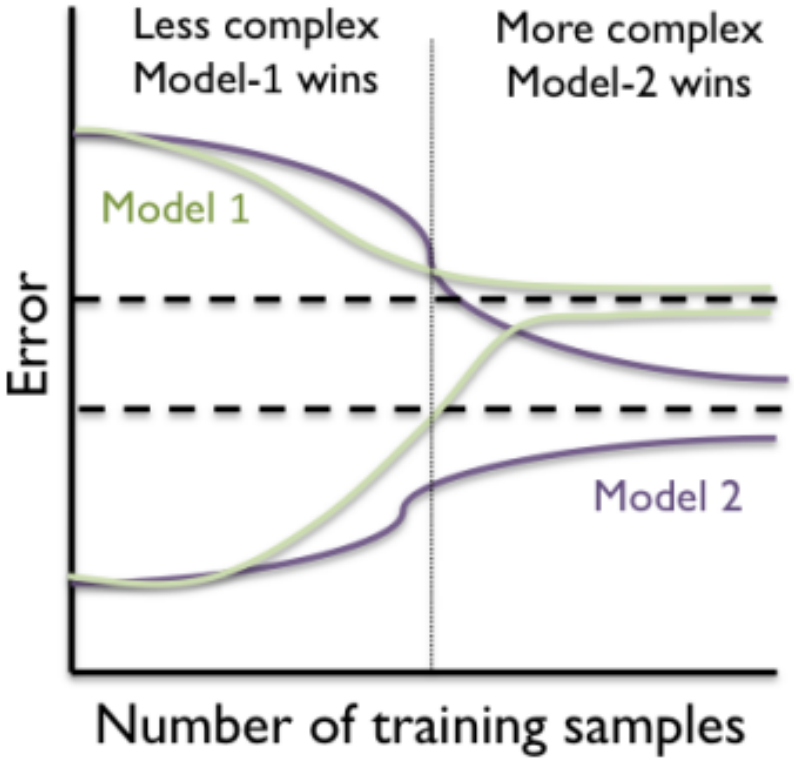
Capacidad del Modelo

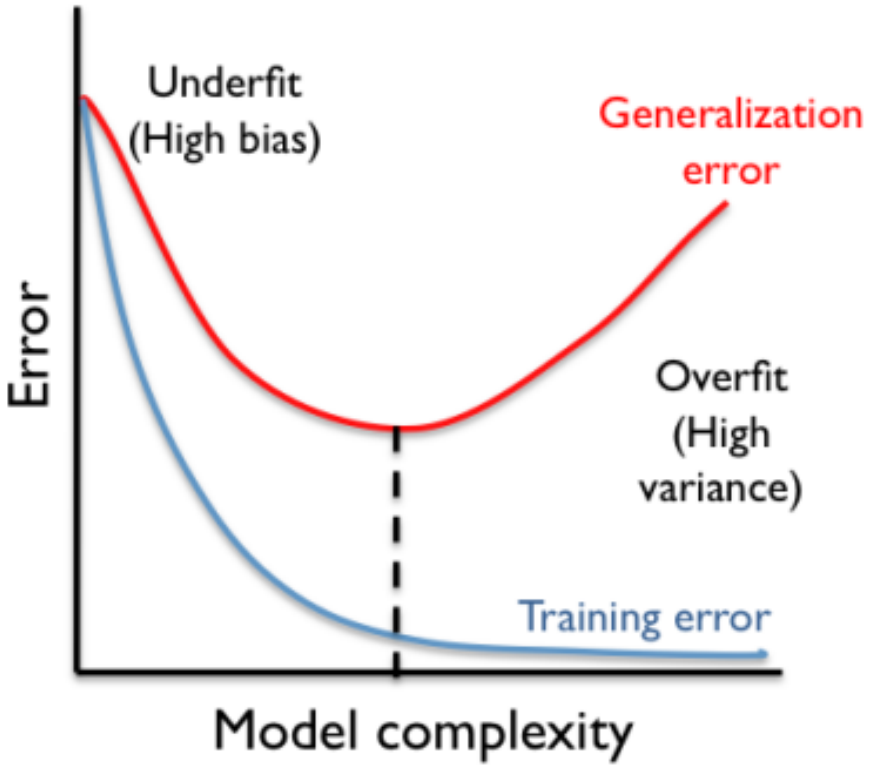
- Underfitting: los errores de entrenamiento y testeo son ambos grandes. (high Bias)
- Overfitting: error al entrenar es pequeño, pero al testear aumenta. (High Variance)

Capacidad del Modelo

- Underfitting: los errores de entrenamiento y testeo son ambos grandes. (high Bias)
- Overfitting: error al entrenar es pequeño, pero al testear aumenta. (High Variance)
- Si el espacio de hipótesis que se estudia es grande hay mas tendencia a sobreajustar.

Sobre-entrenamiento y Sub-entrenamiento





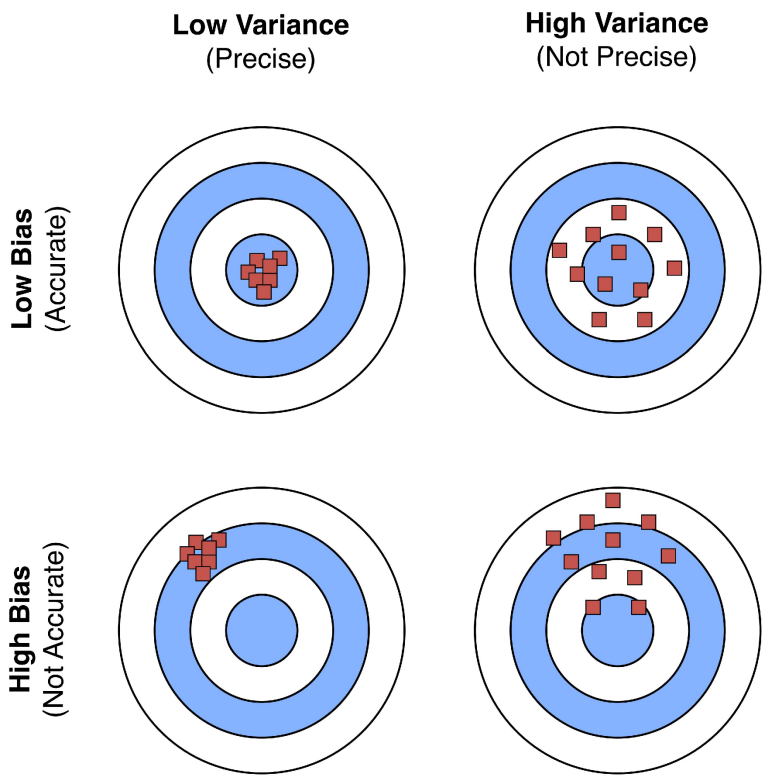
Descomposición en sesgo y varianza

- La descomposición de la pérdida en sesgo y varianza nos ayuda a comprender algoritmos de aprendizaje, los conceptos se correlacionan con el ajuste y el sobreajuste

Descomposición en sesgo y varianza

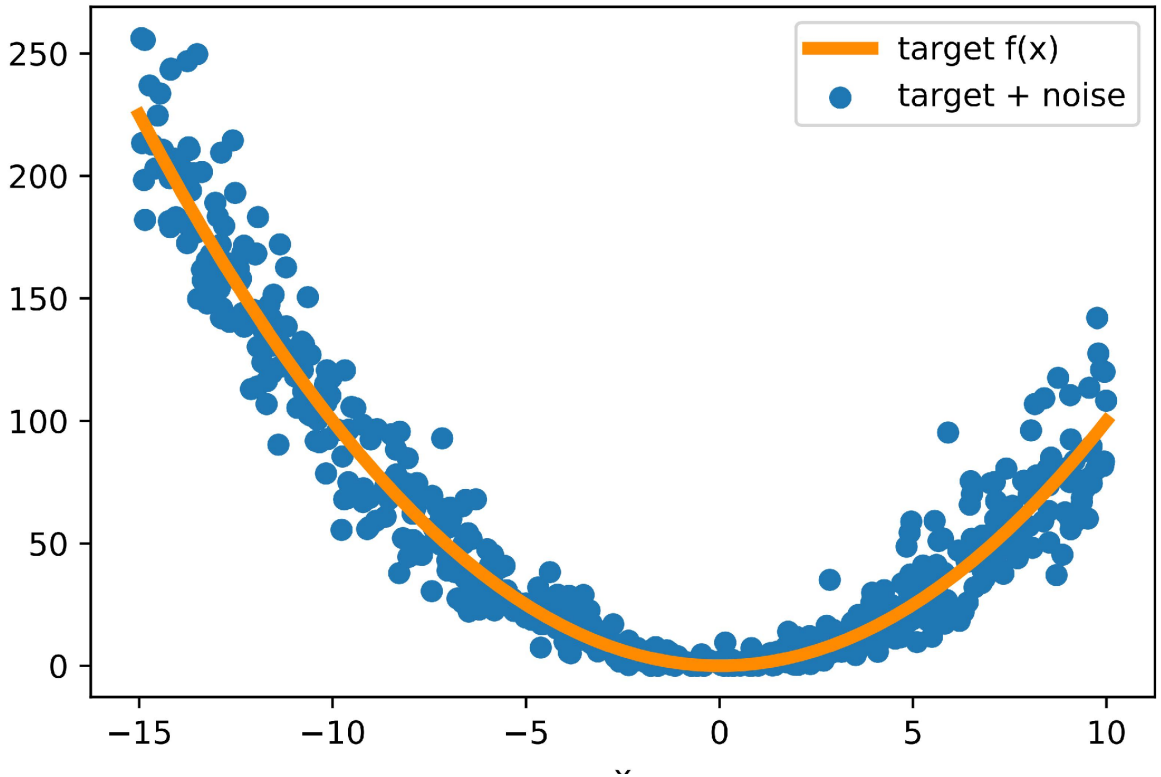
- La descomposición de la pérdida en sesgo y varianza nos ayuda a comprender algoritmos de aprendizaje, los conceptos se correlacionan con el ajuste y el sobreajuste
- Ayuda a explicar por qué los métodos de ensemble como random forests podrían funcionar mejor que los modelos individuales como los árboles de decisión

Intuición sobre sesgo y varianza



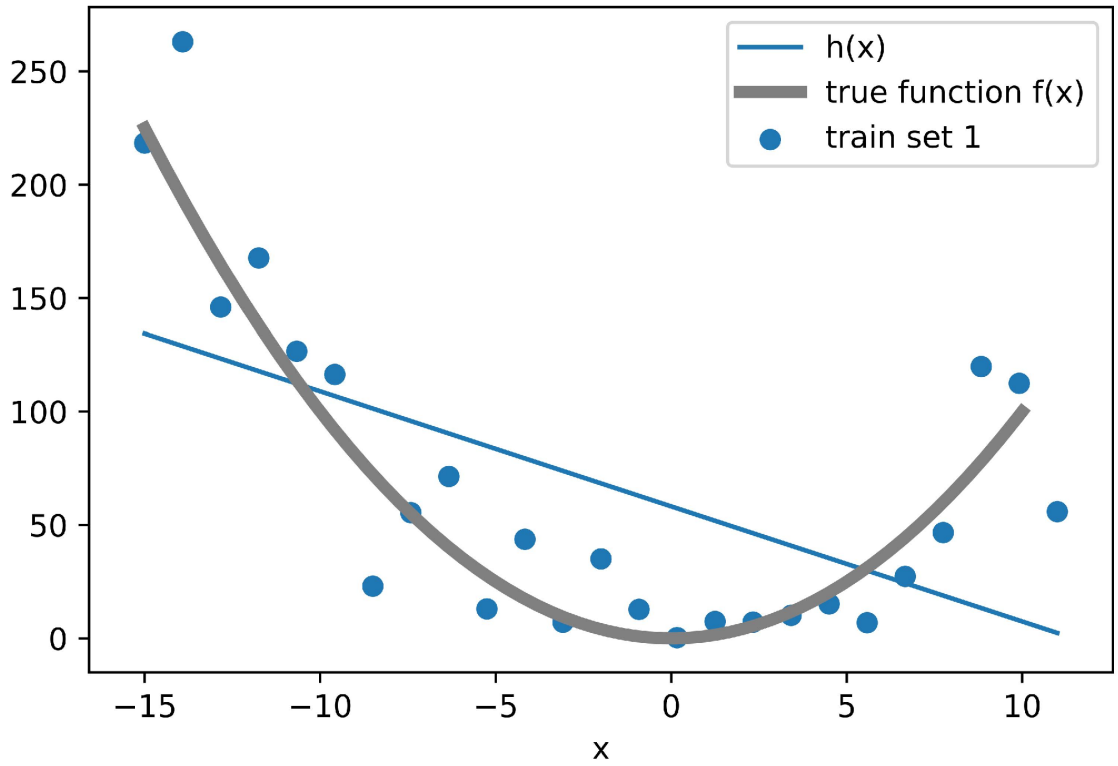
Ejemplo de regresión

Supongamos tener una función de este tipo que gobierna un fenómeno



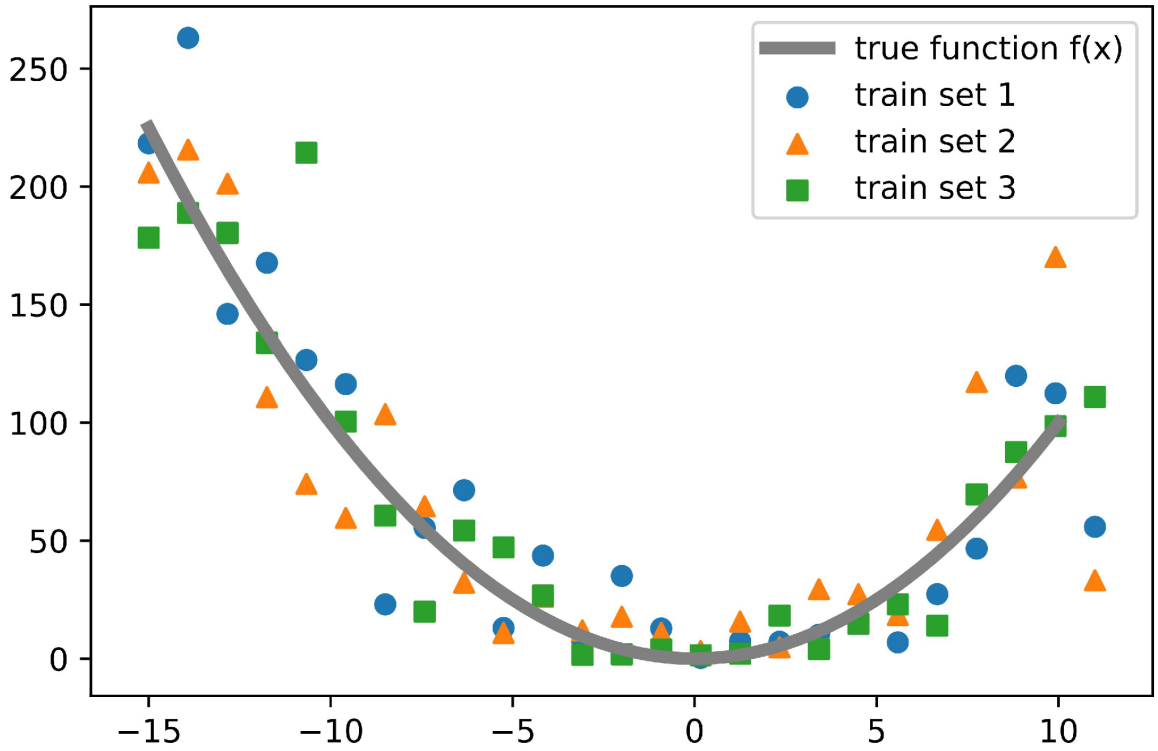
Ejemplo de regresión

Se muestran algunos datos del fenómeno en forma ruidosa



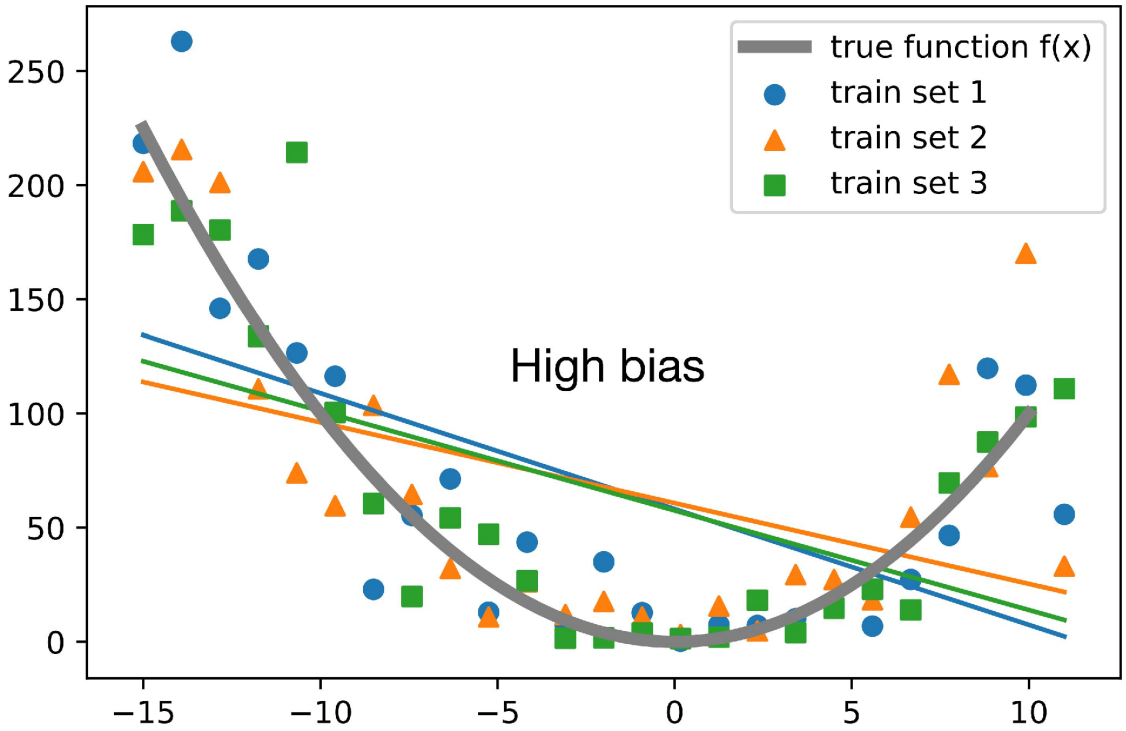
Ejemplo de regresión

Si los datos son pocos, uno puede ajustar una función lineal



Ejemplo de regresión

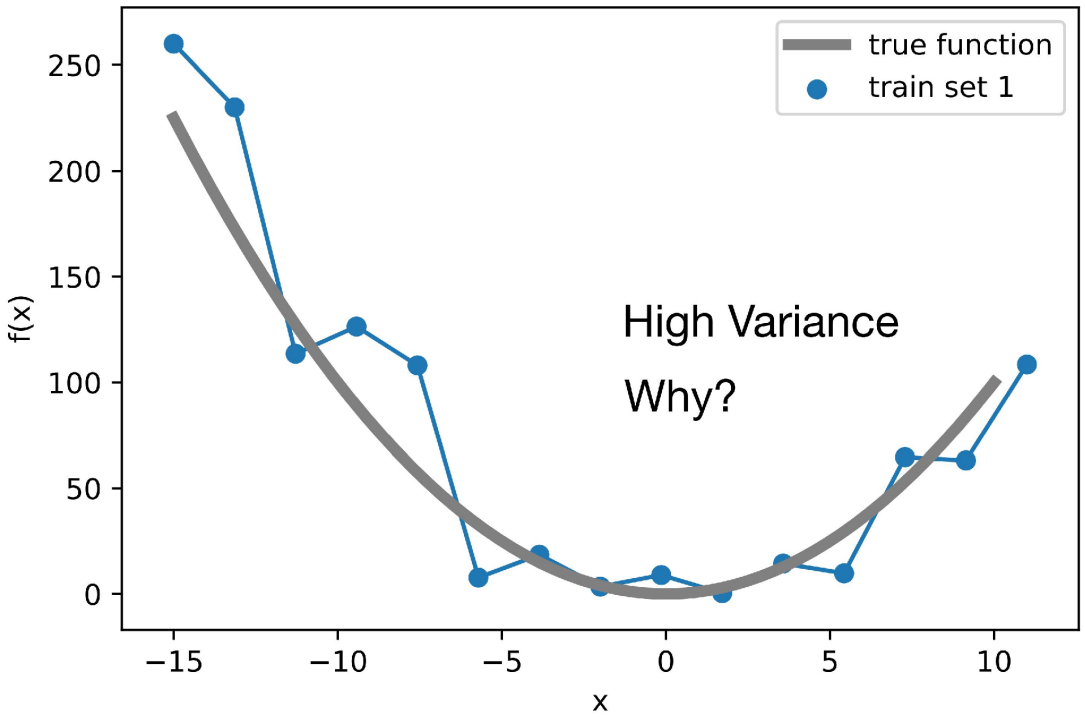
Si uno muestra en forma independiente y ajusta rectas a cada grupo



Hay dos puntos donde el sesgo es nulo

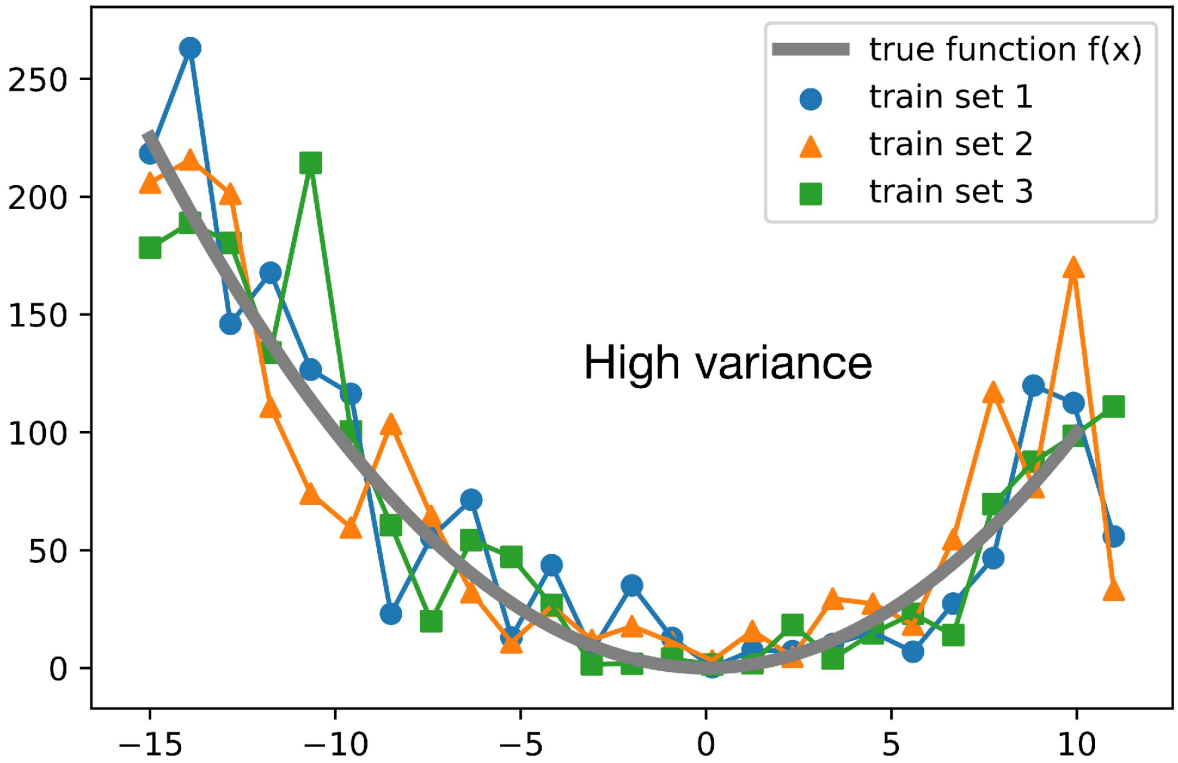
Ejemplo de regresión

Si se ajusta en árbol de decisión, el sesgo es menor



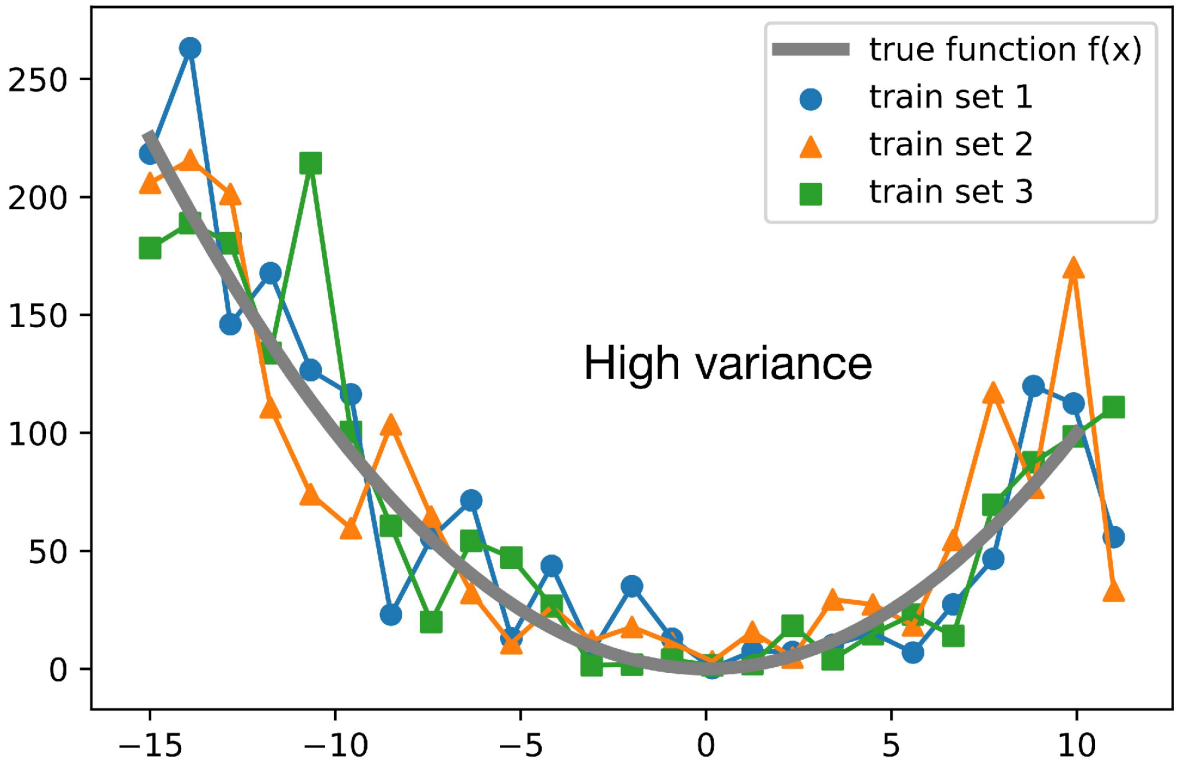
Ejemplo de regresión

Si se ajustan árboles a cada muestra en forma independiente



Ejemplo de regresión

Que pasa si se toma el promedio de estos árboles ajustados? eso es lo que ensamble hace para reducir la varianza.



Métodos de Evaluación

- En estadística, definimos como estimador $\hat{\theta}$ a cualquier función de los datos de una muestra aleatoria que tiene una distribución de probabilidad caracterizada por el parámetro θ

Métodos de Evaluación

- En estadística, definimos como estimador $\hat{\theta}$ a cualquier función de los datos de una muestra aleatoria que tiene una distribución de probabilidad caracterizada por el parámetro θ
- Hay estimadores buenos y malos..

Métodos de Evaluación

- En estadística, definimos como estimador $\hat{\theta}$ a cualquier función de los datos de una muestra aleatoria que tiene una distribución de probabilidad caracterizada por el parámetro θ
- Hay estimadores buenos y malos..
- Suficiencia, consistencia y otras propiedades nos ayudan a decidir cuando un estimador es bueno.

Métodos de Evaluación

- En estadística, definimos como estimador $\hat{\theta}$ a cualquier función de los datos de una muestra aleatoria que tiene una distribución de probabilidad caracterizada por el parámetro θ
- Hay estimadores buenos y malos..
- Suficiencia, consistencia y otras propiedades nos ayudan a decidir cuando un estimador es bueno.
- Sesgo nulo y varianza convergente a cero implican que el error cuadrático medio del estimador tiende a cero.

Terminología

- Parámetro θ, Estimador $\hat{\theta}$

Terminología

- Parámetro θ, Estimador $\hat{\theta}$
- Sesgo

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Terminología

- Parámetro θ, Estimador $\hat{\theta}$
- Sesgo

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta$$

- Varianza

$$var(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2] = E[\hat{\theta}^2] - [E\hat{\theta}]^2$$

Terminología

- Parámetro θ, Estimador $\hat{\theta}$
- Sesgo

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta$$

- Varianza

$$var(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2] = E[\hat{\theta}^2] - [E\hat{\theta}]^2$$

- Como medimos el error? con una función de pérdida.

Terminología

- Parámetro θ, Estimador $\hat{\theta}$
- Sesgo

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta$$

- Varianza

$$var(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2] = E[\hat{\theta}^2] - [E\hat{\theta}]^2$$

- Como medimos el error? con una función de pérdida.
  - Pérdida cuadrática $S = (\hat{\theta} - \theta)^2$

Terminología

- Parámetro θ, Estimador $\hat{\theta}$
- Sesgo
$$B(\hat{\theta}) = E[\hat{\theta}] - \theta$$
- Varianza
$$var(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2] = E[\hat{\theta}^2] - [E\hat{\theta}]^2$$
- Como medimos el error? con una función de pérdida.
  - Pérdida cuadrática $S = (\hat{\theta} - \theta)^2$
  - Pérdida cero uno $L = I_{[\hat{\theta} - \theta = 0]}$

Descomposición de la pérdida cuadrática

En el contexto de machine learning

- $\theta = y = f(x)$ la función objetivo, ya sea para regresión o para clasificación

Descomposición de la pérdida cuadrática

En el contexto de machine learning

- $\theta = y = f(x)$ la función objetivo, ya sea para regresión o para clasificación
- $\hat{\theta} = \hat{y}$ es la predicción de la función objetivo

Descomposición de la pérdida cuadrática

En el contexto de machine learning

- $\theta = y = f(x)$ la función objetivo, ya sea para regresión o para clasificación
- $\hat{\theta} = \hat{y}$ es la predicción de la función objetivo
- Pérdida cuadrática $S = (\hat{y} - y)^2$

$$
\begin{aligned}
S &= (y - \hat{y})^2 \\
&= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\
&= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})
\end{aligned}
$$

Error cuadrático medio

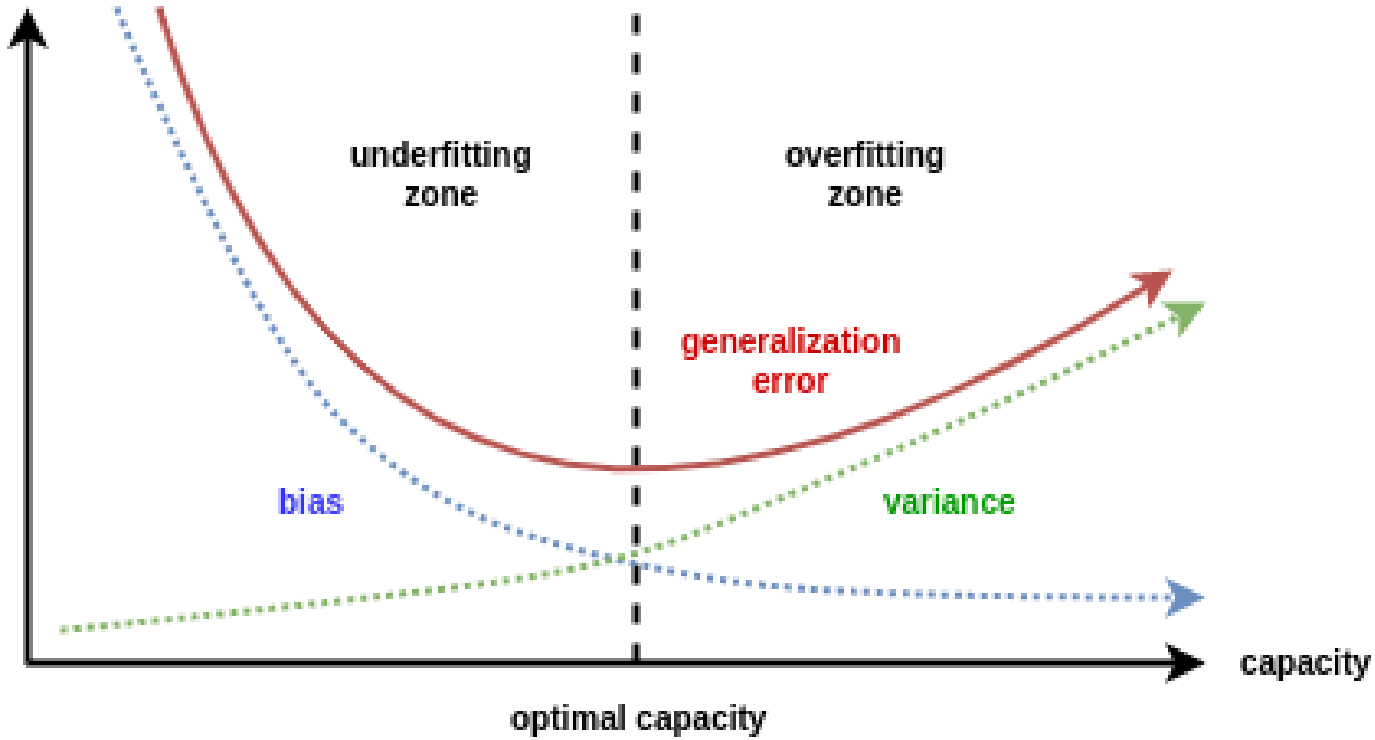
tomando esperanza a ambos lados

$$
\begin{aligned}
E[S] &= E\left[(y-\hat{y})^2\right] \\
E\left[(y-\hat{y})^2\right] &= (y-E[\hat{y}])^2 + E\left[(E[\hat{y}]-\hat{y})^2\right] \\
&= [\text{Bias}]^2 + \text{ Variance}
\end{aligned}
$$

dado que

$$
\begin{aligned}
E[2(y-E[\hat{y}])(E[\hat{y}]-\hat{y})] &= 2E[(y-E[\hat{y}])(E[\hat{y}]-\hat{y})] \\
&= 2(y-E[\hat{y}])E[(E[\hat{y}]-\hat{y})] \\
&= 2(y-E[\hat{y}])(E[E[\hat{y}]]-E[\hat{y}]) \\
&= 2(y-E[\hat{y}])(E[\hat{y}]-E[\hat{y}]) \\
&= 0
\end{aligned}
$$

Sobre-entrenamiento y Sub-entrenamiento



Descomposición de la pérdida cero-uno

- La pérdida cero-uno toma el valor cero cuando se clasifica la etiqueta correctamente, y toma el valor uno cuando no lo hace.

$$\text{Loss} = \begin{cases} 1 \text{ if } y \neq \hat{y} \\ 0 \text{ otherwise} \end{cases}$$

Descomposición de la pérdida cero-uno

- La pérdida cero-uno toma el valor cero cuando se clasifica la etiqueta correctamente, y toma el valor uno cuando no lo hace.

$$\text{Loss} = \begin{cases} 1 \text{ if } y \neq \hat{y} \\ 0 \text{ otherwise} \end{cases}$$

- La predicción principal es la predicción que minimiza el promedio de la perdida

$$\overline{\hat{y}} = \underset{\hat{y}'}{\operatorname{argmin}} E[L(\hat{y}, \hat{y})]$$

Descomposición de la pérdida cero-uno

- La pérdida cero-uno toma el valor cero cuando se clasifica la etiqueta correctamente, y toma el valor uno cuando no lo hace.

$$\text{Loss} = \begin{cases} 1 \text{ if } y \neq \hat{y} \\ 0 \text{ otherwise} \end{cases}$$

- La predicción principal es la predicción que minimiza el promedio de la perdida

$$\overline{\hat{y}} = \underset{\hat{y}'}{\operatorname{argmin}} E[L(\hat{y}, \hat{y})]$$

  - Para la pérdida cuadrática -> Mean

Descomposición de la pérdida cero-uno

- La pérdida cero-uno toma el valor cero cuando se clasifica la etiqueta correctamente, y toma el valor uno cuando no lo hace.

$$\text{Loss} = \begin{cases} 1 \text{ if } y \neq \hat{y} \\ 0 \text{ otherwise} \end{cases}$$

- La predicción principal es la predicción que minimiza el promedio de la perdida

$$\overline{\hat{y}} = \underset{\hat{y}'}{\operatorname{argmin}} E[L(\hat{y}, \hat{y})]$$

  - Para la pérdida cuadrática -> Mean
  - Para la pérdida $0-1->$ Mode

Descomposición de la pérdida cero-uno

	Squared Loss	$0-1$ Loss
single loss	$(y-\hat{y})^2$	$L(y,\hat{y})$
Expected loss	$E\left[(y-\hat{y})^2\right]$	$E[L(y,\hat{y})]$
Main prediction $E[\hat{y}]$	mean (average)	mode
Bias ^{2}	$(y-E[\hat{y}])^2$	$L(y,E[\hat{y}])$
Variance	$E\left[(E[\hat{y}]-\hat{y})^2\right]$	$E[L(\hat{y},E[\hat{y}])]$

Descomposición de la pérdida cero-uno

- Al utilizar la moda para definir la predicción principal, el sesgo es 1 si la predicción principal no coincide con la verdadera etiqueta y y 0 en otro caso.

$$\text{Bias} = \begin{cases} 1 \text{ if } y \neq E[\hat{y}] \\ 0 \text{ otherwise} \end{cases}$$

Descomposición de la pérdida cero-uno

- Al utilizar la moda para definir la predicción principal, el sesgo es 1 si la predicción principal no coincide con la verdadera etiqueta y y 0 en otro caso.

$$\text{Bias} = \begin{cases} 1 \text{ if } y \neq E[\hat{y}] \\ 0 \text{ otherwise} \end{cases}$$

- La varianza de la pérdida cero-uno es la probabilidad de que la etiqueta predicha no coincida con la predicción principal.

$$\underline{\text{Variance} = P(\hat{y} \neq E[\hat{y}])}$$

Descomposición de la pérdida cero-uno

- Si consideramos a la perdida esperada como sesgo mas varianza, si el sesgo es 0 entonces :

$$\text{E(Loss)} = P(y \neq \hat{y}) = 0 + \text{Variance} = P(\hat{y} \neq E[\hat{y}])$$

Descomposición de la pérdida cero-uno

- Si consideramos a la perdida esperada como sesgo mas varianza, si el sesgo es 0 entonces :

$$\text{E(Loss)} = P(y \neq \hat{y}) = 0 + \text{Variance} = P(\hat{y} \neq E[\hat{y}])$$

- En otras palabras, si el modelo es insesgado, la pérdida está caracterizada por la varianza el cual es proporcional al sobreajuste.

Descomposición de la pérdida cero-uno

- Si el sesgo es 1 entonces $(y \neq E(\hat{y}))$ por lo cual

$$\text{E(Loss)} = P(\hat{y} \neq y) = 1 - P(\hat{y} = y) = 1 - P(\hat{y} \neq E[\hat{y}])$$

esto es, la pérdida esperada es 1 menos la varianza.

Descomposición de la pérdida cero-uno

- Si el sesgo es 1 entonces $(y \neq E(\hat{y}))$ por lo cual

$$\text{E(Loss)} = P(\hat{y} \neq y) = 1 - P(\hat{y} = y) = 1 - P(\hat{y} \neq E[\hat{y}])$$

  esto es, la pérdida esperada es 1 menos la varianza.
- En este caso, si el sesgo es tan grande que la predicción principal es siempre mal, aumentar la varianza reduce la pérdida, dado que algunos puntos van a estar mas cerca, de pura suerte , de los valores reales.