Tema 2 – Optimización y Regularización (Parte 4)

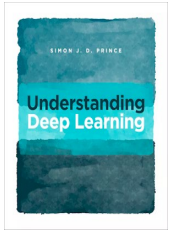
Aprendizaje Automático II - Grado en Inteligencia Artificial
Universidad Rey Juan Carlos

Iván Ramírez Díaz
ivan.ramirez@urjc.es

José Miguel Buenaposada Biencinto
josemiguel.buenaposada@urjc.es

Bibliografía

- Understanding Deep Learning. Capítulo 8.



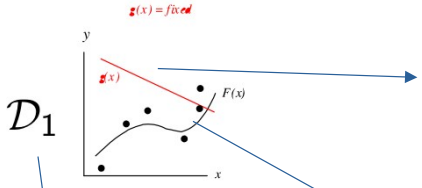
- Deep Learning: CS 182 2021. Lecture 3, Part 3.
  Sergey Levine. UC Berkeley.
  Curso en youtube.

2.5 Estimación del rendimiento

- Ruido, sesgo y varianza
- Reducir la varianza
- Reducir el sesgo y el relación sesgo-varianza
- Doble descenso
- Elegir los hiperparámetros

2.5 Estimación del rendimiento

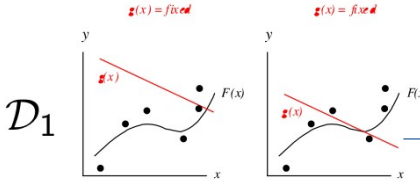
- Ruido, sesgo y varianza
- Reducir la varianza
- Reducir el sesgo y el relación sesgo-varianza
- Doble descenso
- Elegir los hiperparámetros



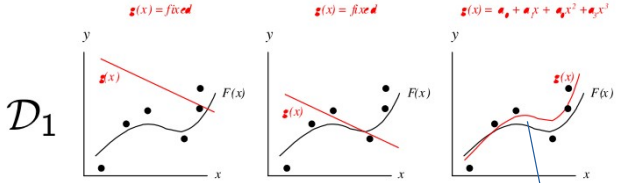
Estimación de la curva (en rojo) con datos de entrenamiento (f[x;Φ]). En este caso recta de pendiente fija.

Verdadera curva (en negro) a estimar con los datos de entrenamiento (μ[x])

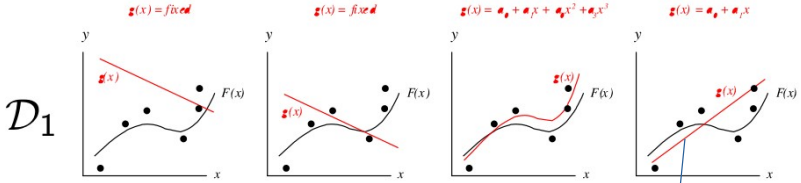
Muestra de 6 datos para entrenar un modelo de regresión
(y[x] = μ[x] + ruido)



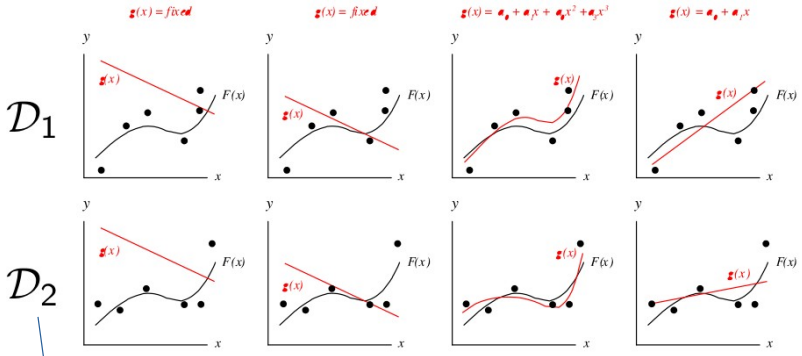
Otro modelo que estima la curva (en rojo) con datos de entrenamiento (f[x;Φ]). En este caso otra recta fija.



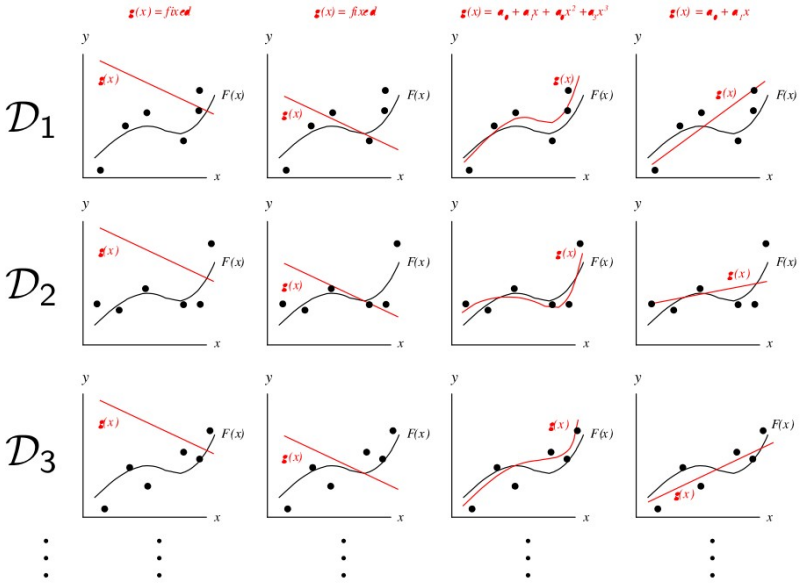
Polinomio cúbico (en rojo) que estima la curva con datos de entrenamiento (f[x;Φ])

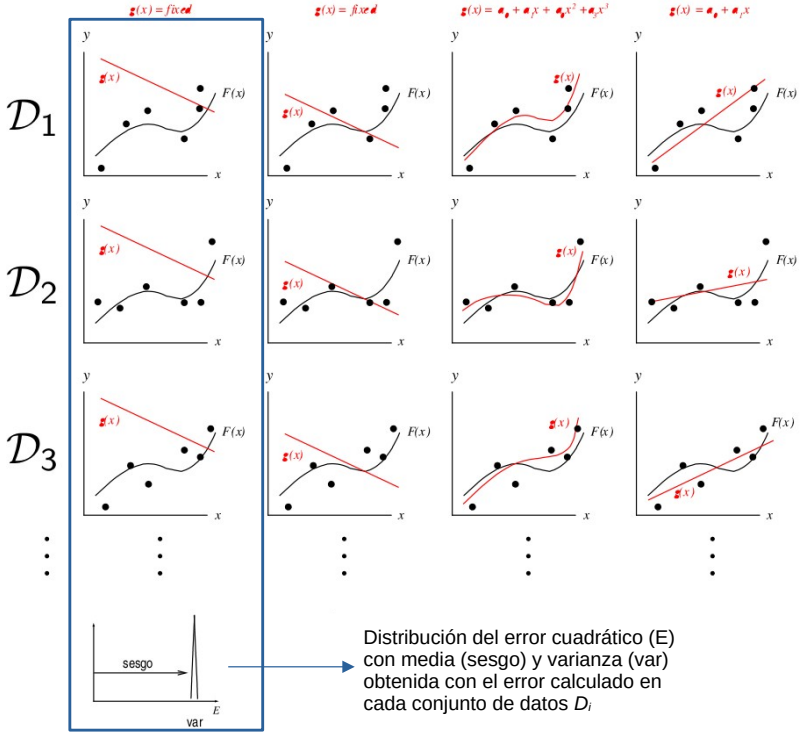


Recta (en rojo) que estima la curva con datos de entrenamiento (f[x;Φ])

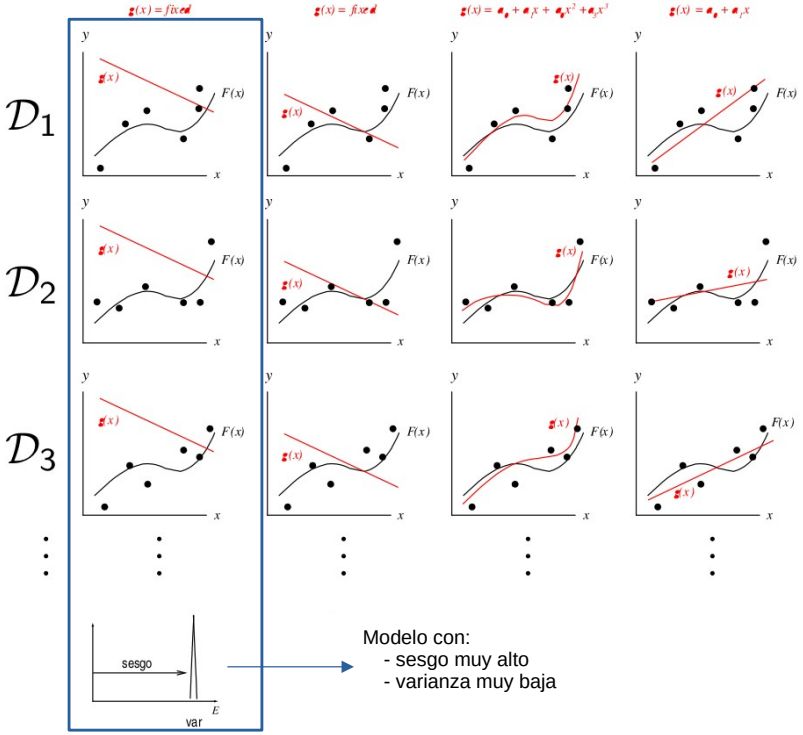


Otra muestra diferente
de 6 datos para entrenar
un modelo de regresión

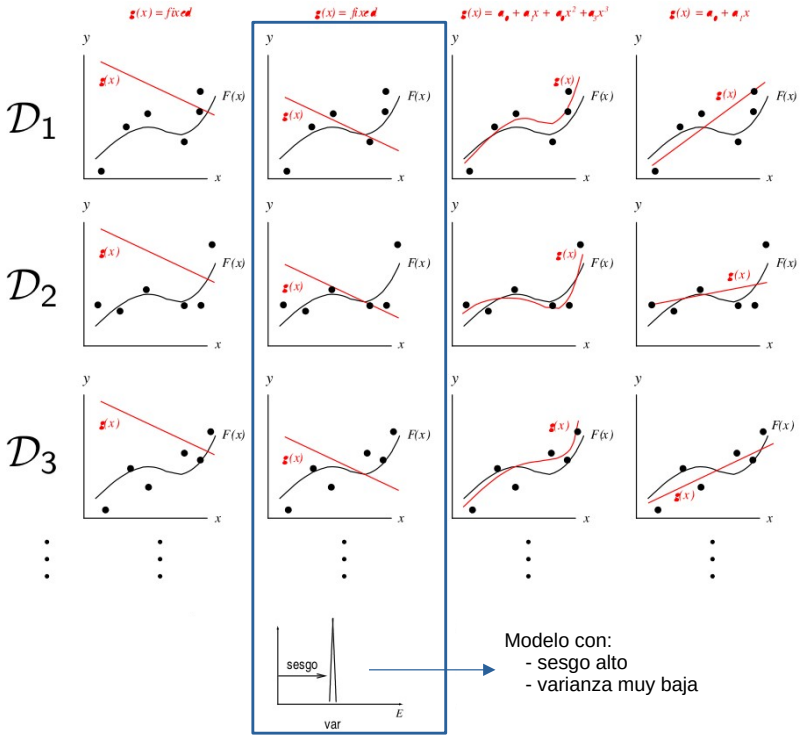




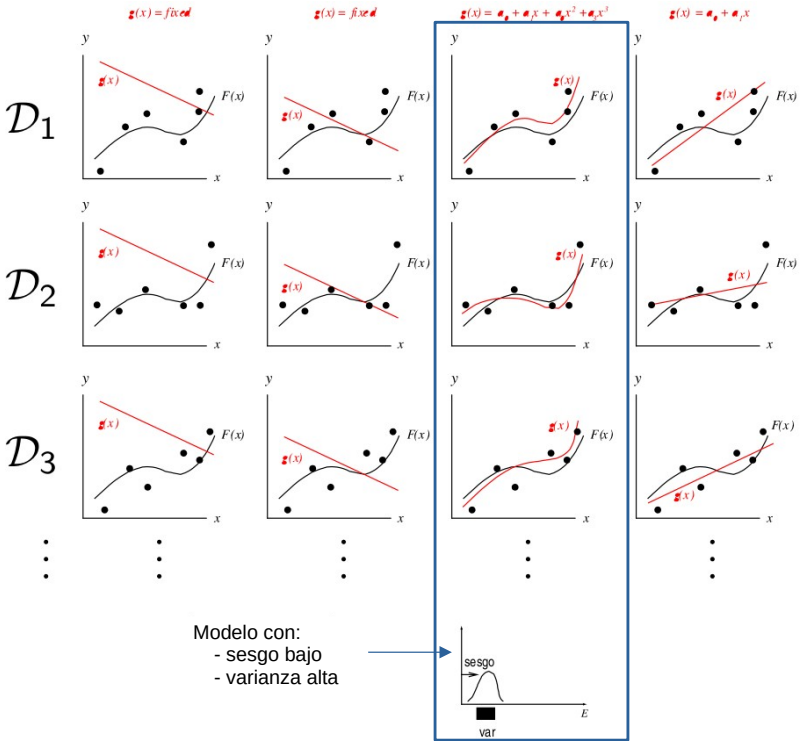
Distribución del error cuadrático (E)
con media (sesgo) y varianza (var)
obtenida con el error calculado en
cada conjunto de datos D_i



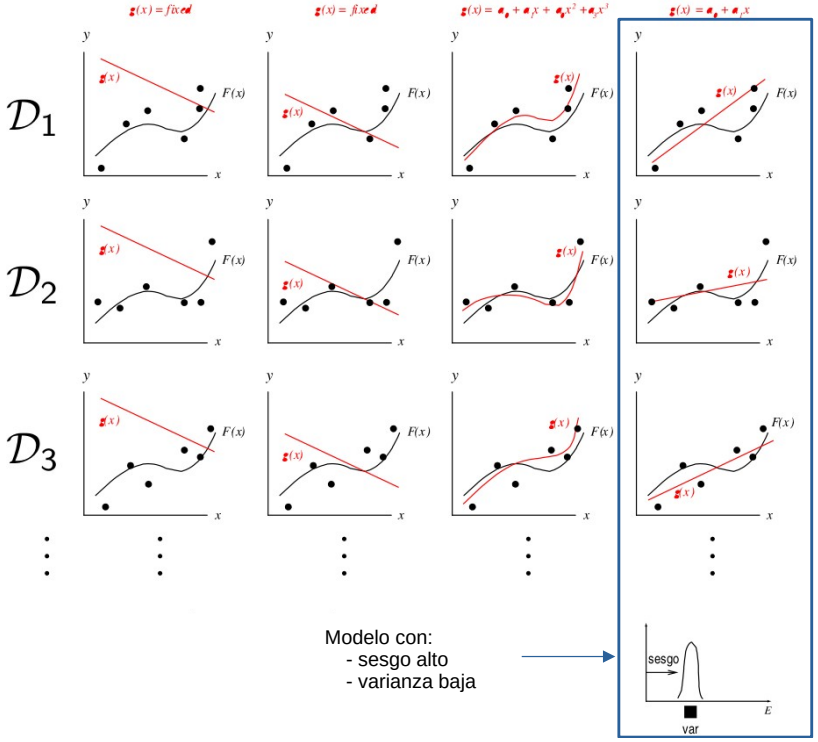
Modelo con:
- sesgo muy alto
- varianza muy baja



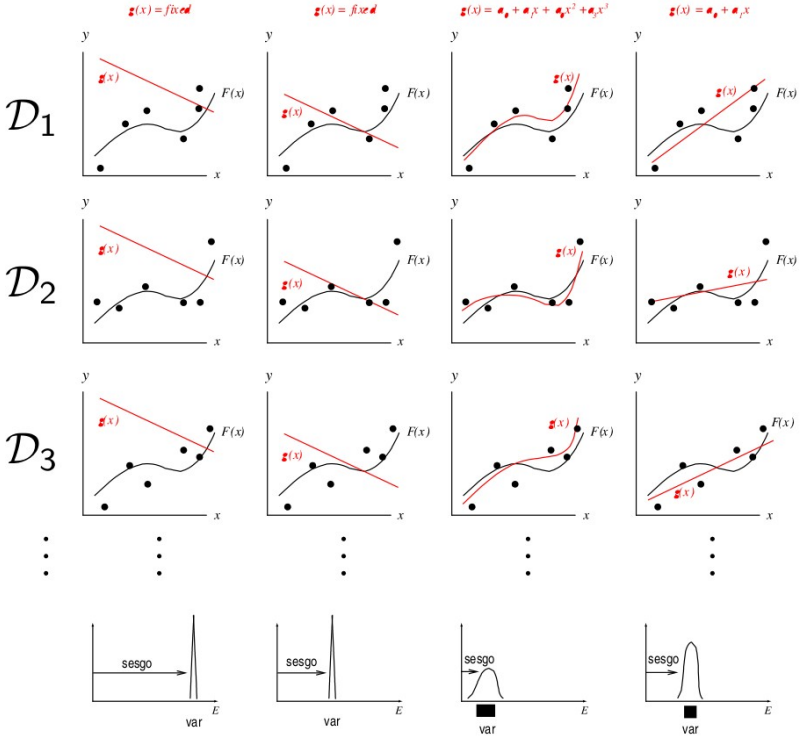
Modelo con:
- sesgo alto
- varianza muy baja



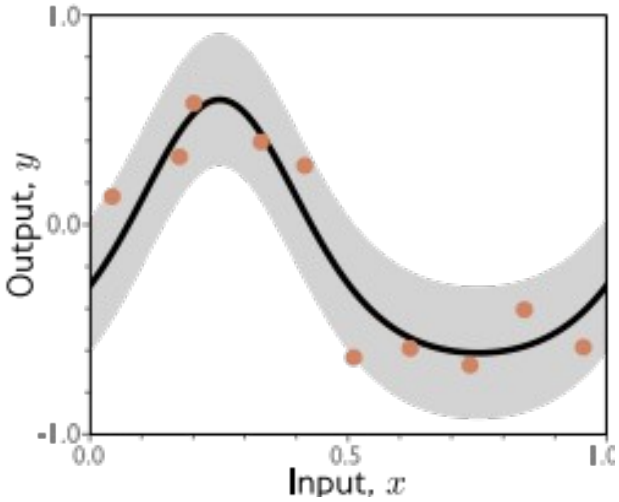
Modelo con:
- sesgo bajo
- varianza alta



Modelo con:
- sesgo alto
- varianza baja

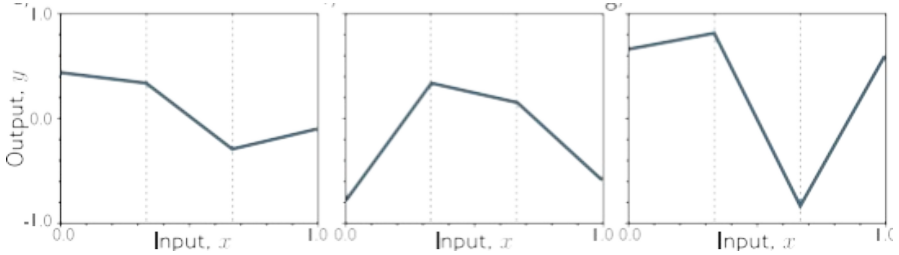


Ejemplo de regresión

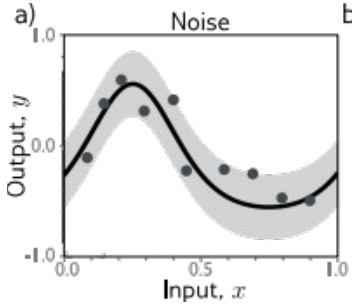


Modelo básico para la estimación

- Modelo que estima tres modelos lineales en el intervalo [0, 1] en tramos de igual longitud.

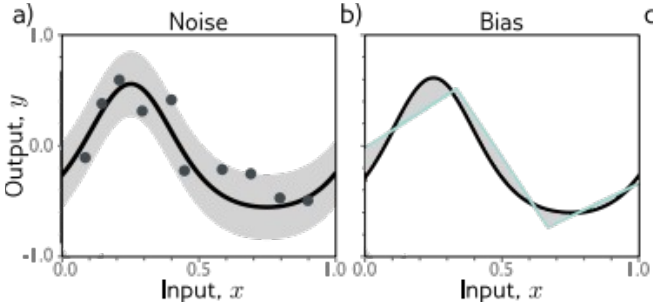


Ruido, sesgo (bias) y varianza (variance)

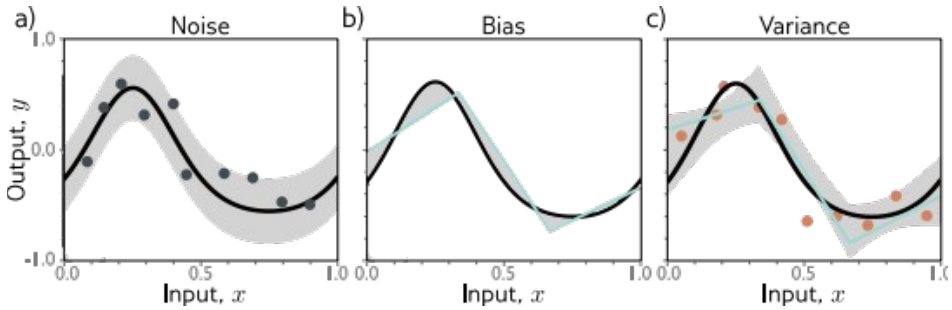


- Posibles causas:
  - Ruido en las medidas
  - Algunas variables del modelo son ocultas
  - Datos mal etiquetados

Ruido, sesgo (bias) y varianza (variance)



Ruido, sesgo (bias) y varianza (variance)



Ruido, sesgo (bias) y varianza (variance)

- La varianza es la incertidumbre en el modelo entrenado debido a la elección del conjunto de datos.
- El sesgo es la desviación sistemática de la media de la verdadera función que estamos estimando debida a las limitaciones de nuestro modelo.
- El ruido es la incertidumbre inherente en la función verdadera que lleva un dato x a una salida y.

Para un problema de mínimos cuadrados ...

$$L[x] = \left(\mathrm{f}[x, \boldsymbol{\phi}] - y[x]\right)^2$$

Regresor con parámetros $\boldsymbol{\Phi}$ que estima y dado un x.

Verdadera función que genera los datos de entrenamiento con un cierto ruido de dist. Normal (y[x] = μ[x] + ruido)

Para un problema de mínimos cuadrados ...

$$L[x] = \left(\mathrm{f}[x, \boldsymbol{\phi}] - y[x]\right)^2$$

Se puede demostrar que:

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_y[L[x]]\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(\mathrm{f}[x, \boldsymbol{\phi}[\mathcal{D}]] - f_\mu[x]\right)^2\right]}_{\text{variance}} + \underbrace{\left(f_\mu[x] - \mu[x]\right)^2}_{\text{bias}} + \underbrace{\sigma^2}_{\text{noise}}$$

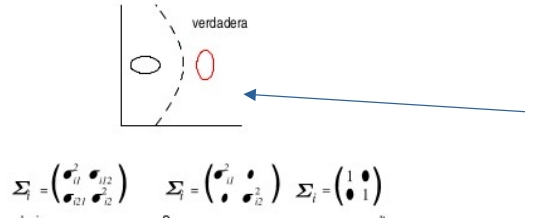
Esperanza sobre las muestras de datos

Esperanza sobre el ruido en el conjunto de test

Media de las estimaciones de f para infinitos conjuntos de datos de entrenamiento *D*

Verdadera función sin ruido

Para un problema de clasificación ...



Dos clases de distribución normal en 2D (roja y negra). La frontera es la parábola con línea discontinua.

$\Sigma_i = \begin{pmatrix} \sigma_{i1}^2 & \sigma_{i12} \\ \sigma_{i21} & \sigma_{i2}^2 \end{pmatrix}$ $\Sigma_i = \begin{pmatrix} \sigma_{i1}^2 & 0 \\ 0 & \sigma_{i2}^2 \end{pmatrix}$ $\Sigma_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

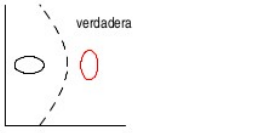
G_1 G_2 G_3

Clasif. estadístico bayesiano con gaussianas de matriz de covarianza completa.

Clasif. estadístico bayesiano con gaussianas de matriz de covarianza diagonal.

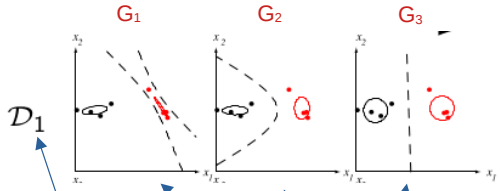
Clasif. estadístico bayesiano con gaussianas de matriz de covarianza identidad.

Para un problema de clasificación ...



$\Sigma_i = \begin{pmatrix} \sigma_{i1}^2 & \sigma_{i12} \\ \sigma_{i21} & \sigma_{i2}^2 \end{pmatrix}$ $\Sigma_i = \begin{pmatrix} \sigma_{i1}^2 & 0 \\ 0 & \sigma_{i2}^2 \end{pmatrix}$ $\Sigma_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

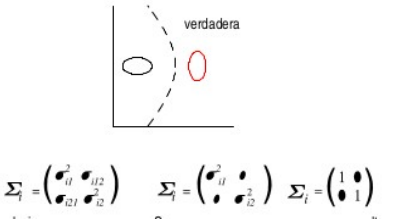
G_1 G_2 G_3



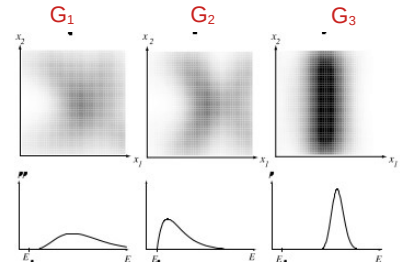
Muestra de 4 datos por clase para entrenar (sacados de las distribuciones verdaderas)

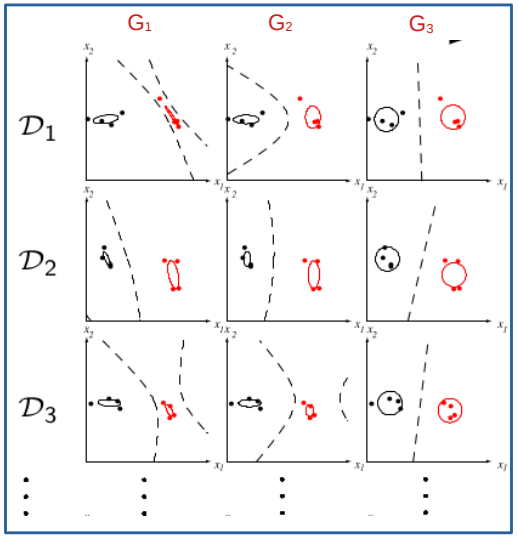
Ajuste de las dos dist. gaussianas a cada clase con los datos de entrenamiento para G_1, G_2 y G_3.

Para un problema de clasificación ...



$\Sigma_i = \begin{pmatrix} \sigma_{i1}^2 & \sigma_{i12} \\ \sigma_{i21} & \sigma_{i2}^2 \end{pmatrix}$ $\Sigma_i = \begin{pmatrix} \sigma_{i1}^2 & 0 \\ 0 & \sigma_{i2}^2 \end{pmatrix}$ $\Sigma_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$





sesgo

bajo → alto

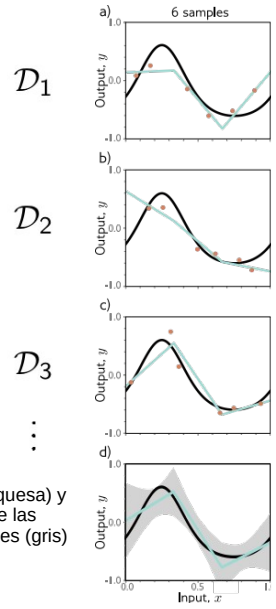
alta ← baja

varianza

2.5 Estimación del rendimiento

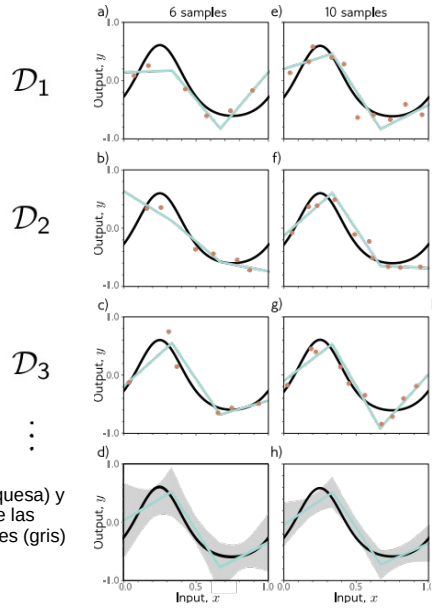
- Ruido, sesgo y varianza
- Reducir la varianza
- Reducir el sesgo y el compromiso sesgo-varianza
- Doble descenso
- Elegir los hiperparámetros

Reducir la varianza



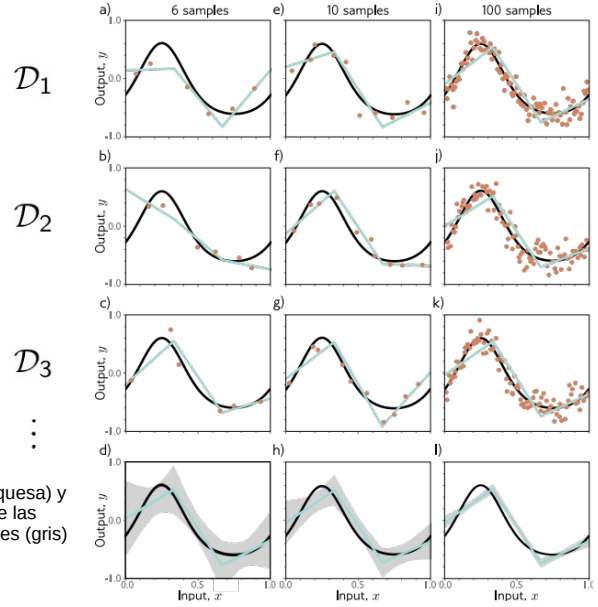
Media (turquesa) y varianza de las estimaciones (gris)

Reducir la varianza



Media (turquesa) y varianza de las estimaciones (gris)

Reducir la varianza



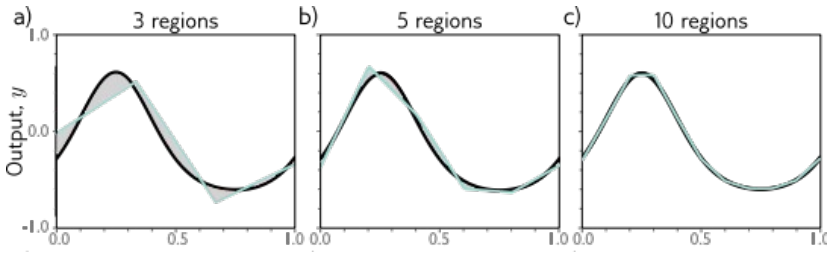
Media (turquesa) y varianza de las estimaciones (gris)

Aumentando el n.º de datos de entrenamiento reducimos la varianza.

2.5 Estimación del rendimiento

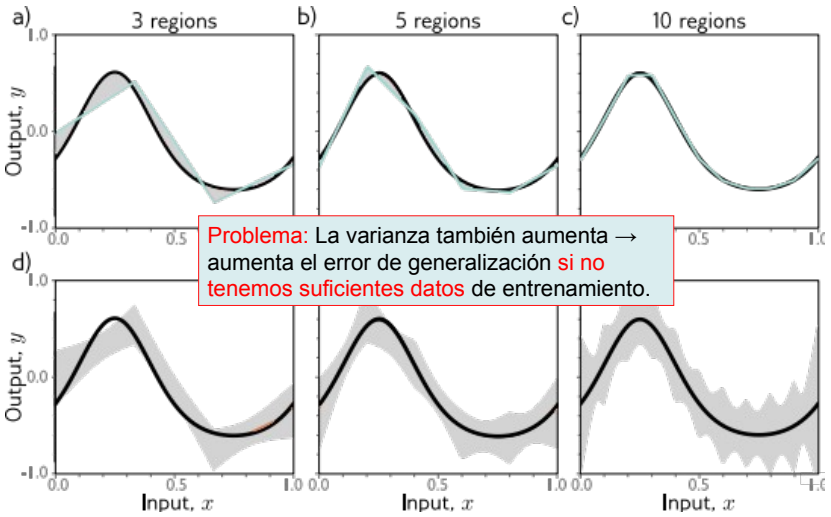
- Ruido, sesgo y varianza
- Reducir la varianza
- Reducir el sesgo y la relación sesgo-varianza
- Doble descenso
- Elegir los hiperparámetros

Reducir el sesgo



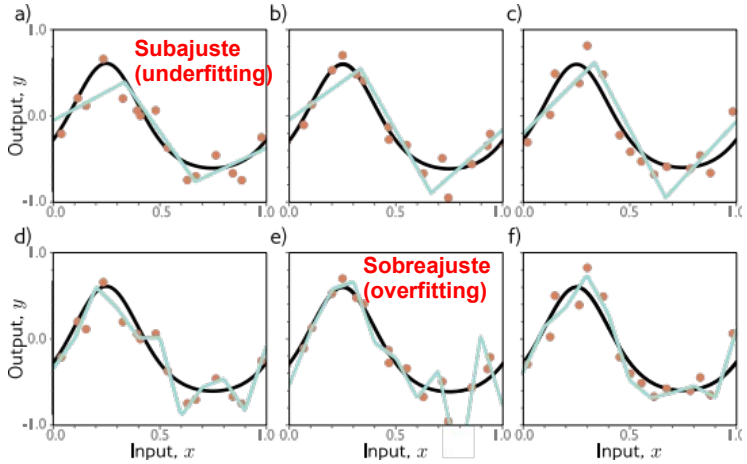
Aumentamos el número de parámetros del estimador (= flexibilidad para ajustarse a los datos) → $f_\mu[x]$ se acerca más a $\mu[x]$

Reducir el sesgo



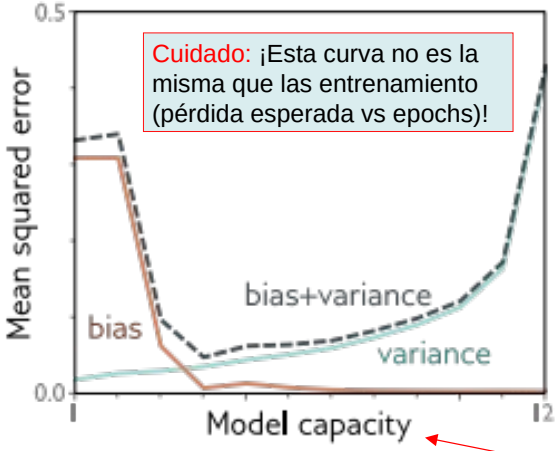
Problema: La varianza también aumenta → aumenta el error de generalización si no tenemos suficientes datos de entrenamiento.

¿Por qué aumenta la varianza?



$f[x, \Phi]$ (turquesa) describe mejor los datos de entrenamiento, pero no la verdadera curva (en negro), $\mu[x]$ → **Sobreajuste (overfitting)**

Relación sesgo-varianza (bias-variance trade-off)



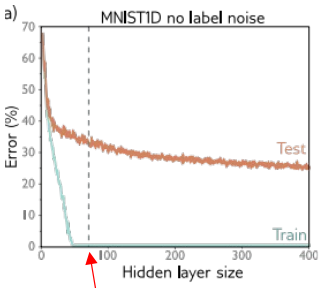
Cuidado: ¡Esta curva no es la misma que las entrenamiento (pérdida esperada vs epochs)!

Relacionado con:
- número de datos de entrenamiento que el modelo puede ajustar sin error
- número de parámetros del modelo

2.5 Estimación del rendimiento

- Ruido, sesgo y varianza
- Reducir la varianza
- Reducir el sesgo y el relación sesgo-varianza
- Doble descenso
- Elegir los hiperparámetros

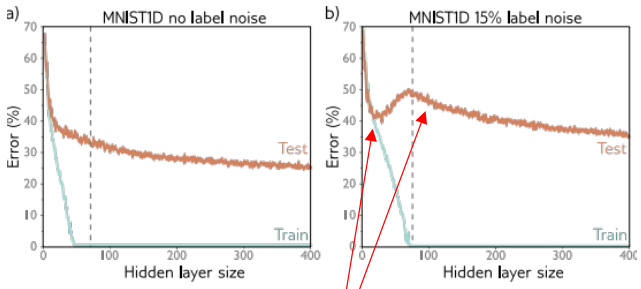
Doble descenso



Cuidado: ¡Esta curva no es la misma que las entrenamiento (pérdida esperada vs epochs)!

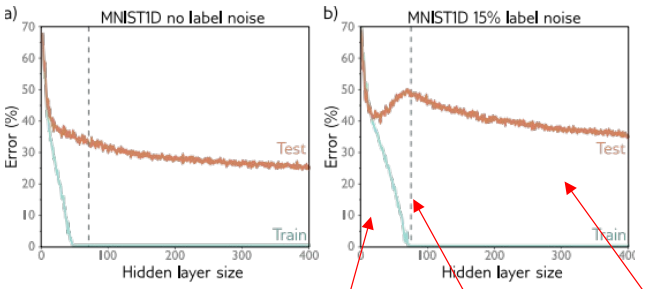
Nº parámetros del modelo = Nº datos entrenamiento

Doble descenso



Aparece el **Doble Descenso**

Doble descenso

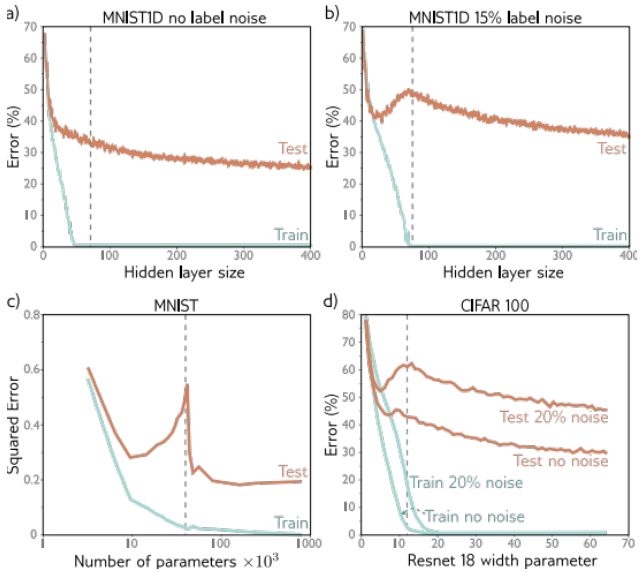


Régimen subparametrizado (o clásico)

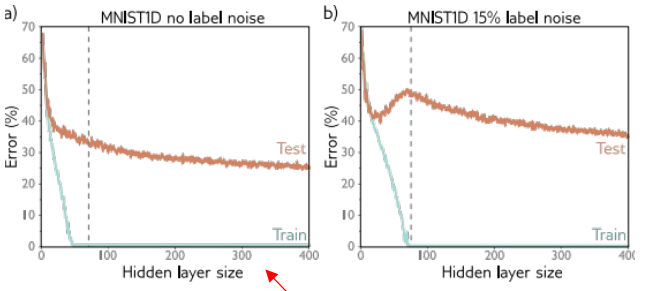
Régimen crítico

Régimen sobreparametrizado (o moderno)

Doble descenso

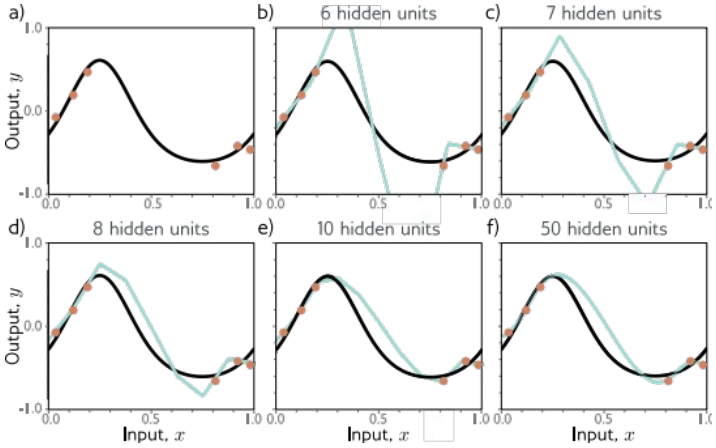


Doble descenso



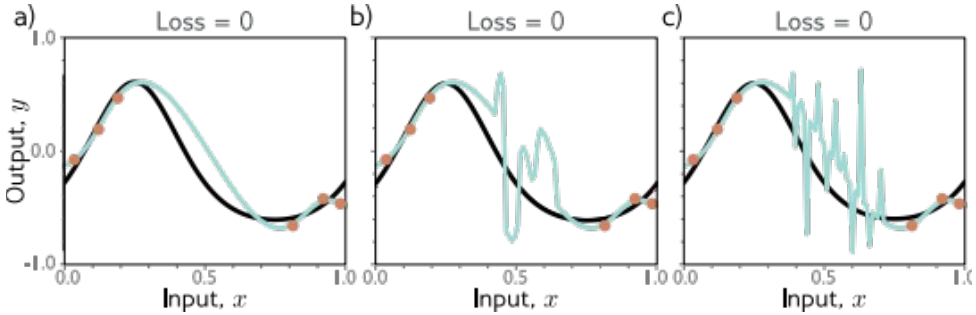
- El error de entrenamiento es cercano a 0
- Lo que esté ocurriendo no está afectando a los datos de entrenamiento
- ¿Debería de ocurrir fuera de los puntos usados en entrenamiento?

En puntos distintos a los de entrenamiento ...



Explicación potencial:
- Con más parámetros estimamos funciones más suaves
- Suavidad fuera de los datos de entrenamiento es algo razonable

¿Explicación del doble descenso?



Todas estas soluciones equivalentes en términos de pérdida esperada.

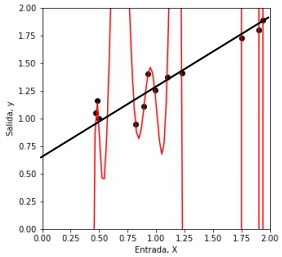
- ¿Por qué el modelo debería elegir la función más suave en a)?
- Explicaciones potenciales:
  - La inicialización elige funciones suaves y la optimización no se sale de ellas
  - El algoritmo de entrenamiento “prefiere” converger a funciones suaves.

2.5 Estimación del rendimiento

- Ruido, sesgo y varianza
- Reducir la varianza
- Reducir el sesgo y el compromiso sesgo-varianza
- Doble descenso
- Elegir los hiperparámetros

Algunas preguntas …

- ¿Cómo saber si estamos sobreajustando o subajustando?
- ¿Cómo elegir el algoritmo de optimización / modelo?
- ¿Cómo elegir los hiperparámetros?
- Idea: elegir lo que haga la pérdida esperada muy baja



$$J(\phi) = 0$$

¡No se puede detectar **sobreajuste** con la pérdida esperada en el entrenamiento!

Método de trabajo en Aprendizaje Automático

Datos disponibles

Método de trabajo en Aprendizaje Automático

Datos disponibles

Usar para entrenar

$$J(\boldsymbol{\phi}, D_{train})$$

1. Entrenar $\boldsymbol{\Phi}$ con $J(\boldsymbol{\Phi}, D_{train})$
   Si $J(\boldsymbol{\Phi}, D_{train})$ no pequeña
   Estás **subajustando**
   bajar la regularización
   mejorar el optimizador
2. Mirar $J(\boldsymbol{\Phi}, D_{val})$
   Si $J(\boldsymbol{\Phi}, D_{val}) >> J(\boldsymbol{\Phi}, D_{train})$
   Estás **sobreajustando**
   incrementar regularización

Reservar para:

- elegir hiperparámetros
- añadir/eliminar características
- elegir el tipo de modelo

$$J(\boldsymbol{\phi}, D_{val})$$

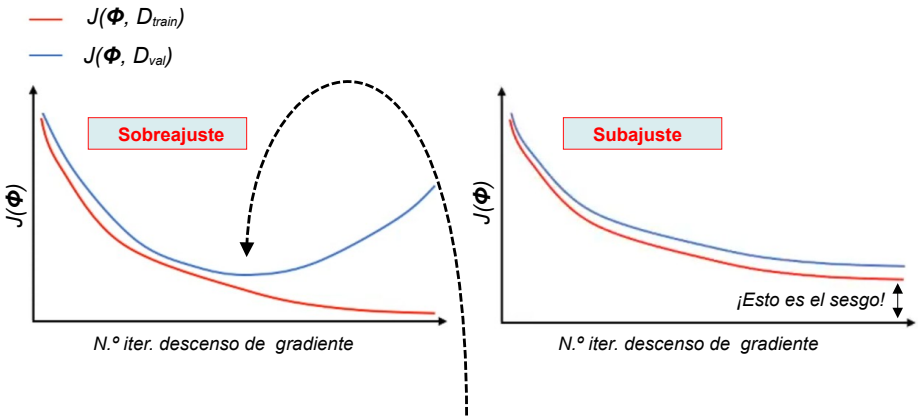
Método de trabajo adecuado

Datos disponibles

Usar para seleccionar:
- Φ *(vía optimización)*
- Hiperparámetros optimización (p.ej. learning rate)

Usar para seleccionar:
- Tipo de modelo (p.ej. Logistic regresion vs otro)
- Hiperparámetros regularización
- Selección características

Curvas de aprendizaje



Pregunta: ¿Podemos parar aquí?

Early stopping

El conjunto de test final

Datos disponibles

1. Entrenar Φ con $J(\Phi, D_{train})$
   - Si $J(\Phi, D_{train})$ no pequeña
     - Estás **subajustando**
     - bajar la regularización
     - mejorar el optimizador
2. Mirar $J(\Phi, D_{val})$
   - Si $J(\Phi, D_{val}) >> J(\Phi, D_{train})$
     - Estás **sobreajustando**
     - incrementar regularización

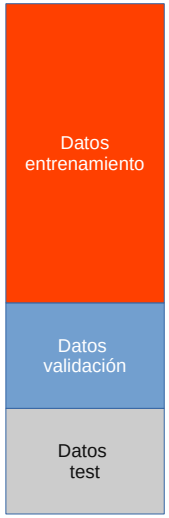
¿Hemos terminado?

¿Nos quedamos con el clasificador que minimiza $J(\Phi, D_{val})$?

No es buena idea - hemos usado el conjunto de validación para elegir hiperparámetros

Método de trabajo en Aprendizaje Automático

Datos disponibles



Usar para seleccionar:
- Φ *(vía optimización)*
- Hiperparámetros optimización (p.ej. learning rate)

Usar para seleccionar:
- Tipo de modelo (p.ej. Logistic regresion vs otro)
- Hiperparámetros regularización
- Selección características

Usar únicamente para reportar rendimiento final

Resumen

- ¿De dónde vienen los errores?
  - Varianza: demasiada capacidad, información insuficiente para encontrar los parámetros adecuados
  - Sesgo: poca capacidad, no se puede representar la función correcta
  - Error = Variance + Bias² (en regresión)
  - Sobreajuste (overfitting): demasiada varianza
  - Subajuste (underfitting): demasiado sesgo

Resumen

- ¿Cómo equilibrar sesgo y varianza?
  - Seleccionar el tipo de modelo cuidadosamente
  - Seleccionar las características cuidadosamente
  - Regularización: añadida a la función de coste para reducir la varianza

Resumen

- ¿Cómo seleccionar hiperparámetros?
  - Separación entrenamiento/validación
  - Datos de entrenamiento para optimización (aprendizaje)
  - Datos de validación para seleccionar hiperparámetros
  - ¡Datos de test para obtener el resultado final y nada más!