Tema 1: Introducción a la clasificación convencional.

Machine Learning: Conceptos:

- Que es aprender:
  - El aprendizaje es cualquier proceso mediante el cual un sistema mejora el rendimiento a partir de la experiencia (Herbert Simon)
  - Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T con una medida de rendimiento P, si su rendimiento en las tareas en T, medido por P, mejora con la experiencia E. (Tom Mitchell).

Ejemplo:

- El filtro de spam es un programa de aprendizaje automático:
  - Puede aprender a marcar el correo no deseado utilizando ejemplos de correos electrónicos no deseados y ejemplos de correos electrónicos que no son correo no deseado, etiquetados por los usuarios.
  - Los ejemplos que utiliza el sistema para aprender se denominan conjunto de entrenamiento.
  - En este caso, la tarea T es marcar el spam para nuevos correos electrónicos, la experiencia E son los datos de entrenamiento y la medida de rendimiento P podría ser la proporción de correos electrónicos clasificados correctamente.

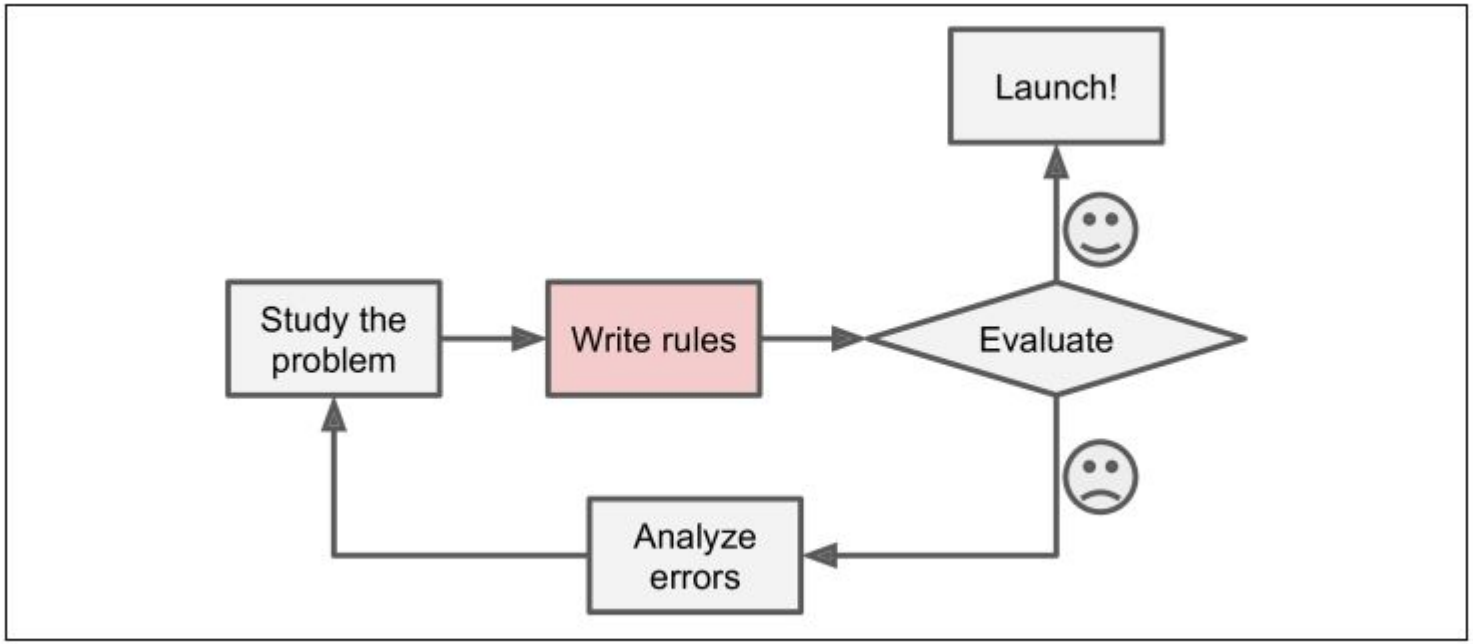


Figure 1-1. The traditional approach

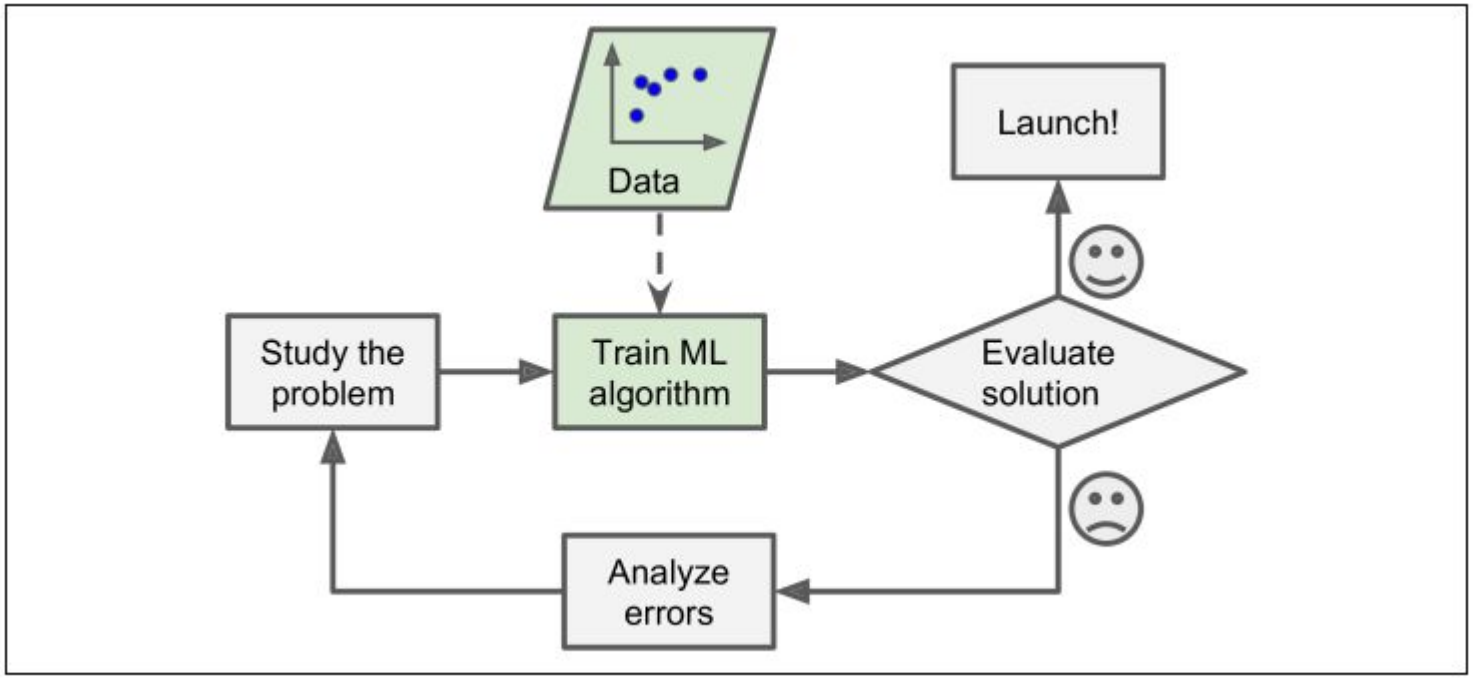


Figure 1-2. Machine Learning approach

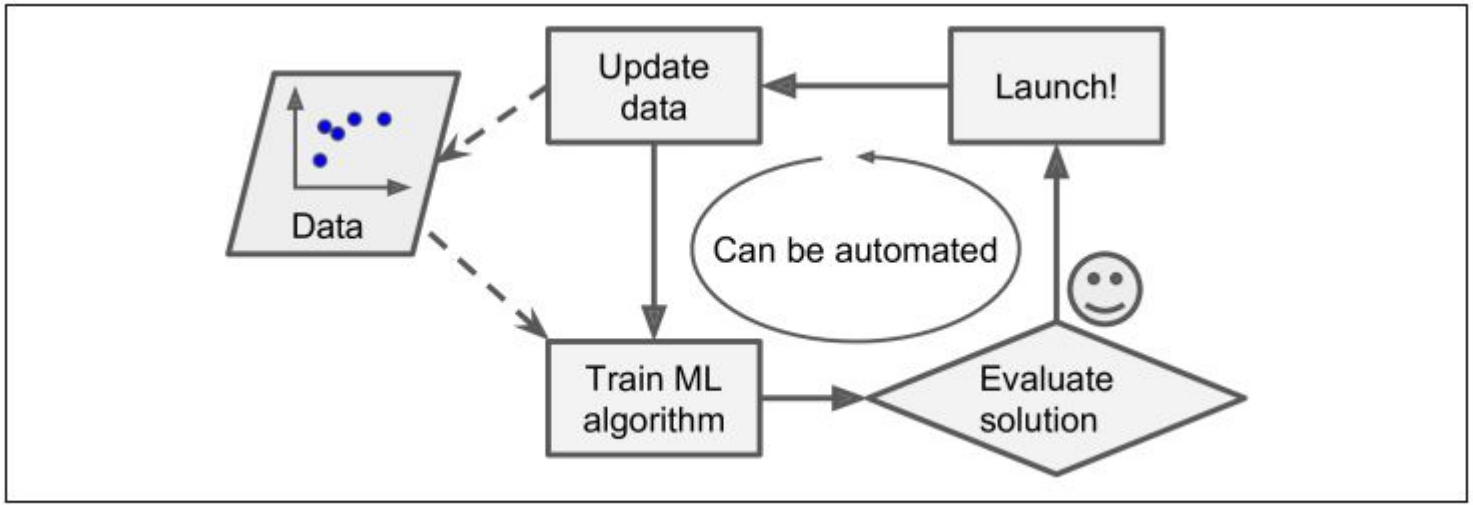


Figure 1-3. Automatically adapting to change

¿Qué son los datos?

- Colección de objetos de datos y sus atributos.
- Un atributo es una propiedad o característica de un objeto. También se conoce como variable, campo, característica, rasgo, entrada. Por ejemplo color de ojos, altura, peso, edad, temperatura.
- Una colección de atributos describe un objeto. También se le conoce como registro, punto, caso, muestra, entidad, instancia.

Tipos de Atributos

- Nominal: color de ojos, códigos postales, número de teléfono.
- Ordinal: clasificaciones, calificaciones, altura en (alto, mediano, bajo), peso en (gordo, mediano, flaco)
- Intervalo: Calendario o fecha, temperaturas en Celsius o Fahrenheit. La ratio no permanece cte o no tiene sentido
- Ratio: longitud, tiempo, altura, peso, conteo, edad..

Atributos y metodos

- La mayoría de los métodos solo pueden tratar con atributos de valor real.
- Otros atributos se transforman.
  - Nominal: generalmente se convierten usando codificación 1-out N (sklearn.preprocessing.OneHotEncoder)
  - Ordinales: pueden tratarse como nominales o convertirse a una escala numérica significativa.
  - Intervalo: la mejor manera es convertir el intervalo en ratio. Muchas veces simplemente se ignoran.
  - Ratio: Se pueden utilizar sin modificación.

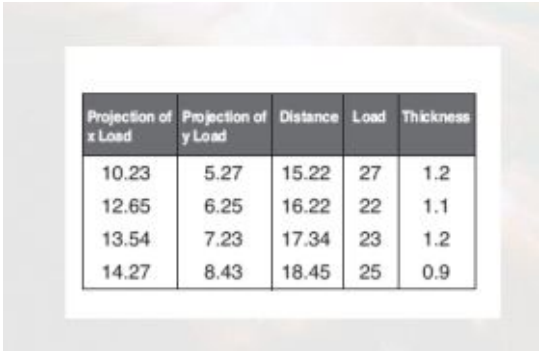
Tipos de conjuntos de datos: Registros de Datos (1)

Registros de Datos: datos que consisten en una colección de registros, cada uno de los cuales consta de un conjunto fijo de atributos. Para el aprendizaje supervisado se incluye la clase como atributo especial. Hay tres tipos.

- Matriz de datos
- Datos del documento
- Datos de la transacción

Tipos de conjuntos de datos: Registros de Datos (2)

Matriz de datos: los objetos de datos tienen el mismo conjunto fijo de atributos numéricos y pueden considerarse como puntos en un espacio multidimensional, donde cada dimensión representa un atributo. Se pueden representar mediante una matriz mxn, donde hay m filas, una para cada objeto, y n columnas, una para cada atributo.

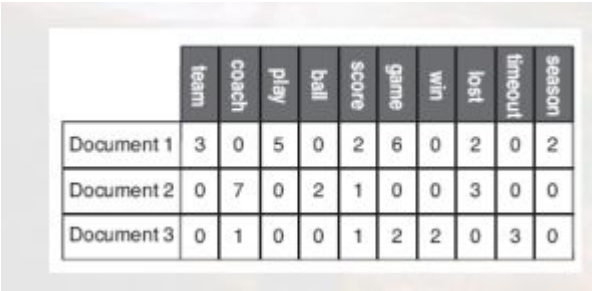


Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

Tipos de conjuntos de datos: Registros de Datos (3)

Datos del documento: cada documento se convierte en un vector de términos. Bolsa de palabras.

- Cada término es un componente del vector
- El valor de cada componente es la frecuencia del término en el documento.



	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Tipos de conjuntos de datos: Registros de Datos (4)

Datos de la transacción

- cada registro (transacción) involucra un conjunto de términos. Por ejemplo, el conjunto de productos comprados por un cliente durante un viaje de compras es una transacción y los términos son los productos.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Áreas y tareas de aprendizaje automático (1): aprendizaje supervisado

Los datos y las etiquetas correspondientes son conocidos.

- Clasificación: predecir valores categóricos, es decir, etiquetas. El filtro de spam está entrenado con muchos correos electrónicos de ejemplo, junto con su clase (spam o ham), y debe aprender a clasificar nuevos correos electrónicos.
- Regresión: predecir valores numéricos. Por ejemplo, predecir el precio de un automóvil, dado un conjunto de características como el kilometraje, la antigüedad, la marca, etc.

Hay que tener en cuenta que algunos algoritmos de regresión también se pueden usar para la clasificación y viceversa.

Áreas y tareas de aprendizaje automático (2): aprendizaje no supervisado

Solo se dan datos, no se proporcionan etiquetas.

- Agrupamiento (clustering): agrupa los datos según la "distancia". Por ejemplo, en la agrupación de clientes, puede agrupar a sus clientes en función de sus compras, su actividad en su sitio web, etc.
- Detección de anomalías: por ejemplo, detección de transacciones inusuales con tarjetas de crédito para evitar fraudes.
- Aprendizaje de reglas de asociación: en el que el objetivo es profundizar en grandes cantidades de datos y descubrir relaciones interesantes entre atributos.
- Reducción de datos: reducir características/atributos. Por ejemplo, el kilometraje de un automóvil puede estar muy relacionado con su edad, por lo que el algoritmo de reducción de dimensionalidad los fusionará en una sola característica. (PCA)

Áreas y tareas de aprendizaje automático (3): aprendizaje semisupervisado

Solo algunas de las etiquetas están presentes. Se utiliza en tareas de clasificación. Google Fotos, es buen ejemplo de ello. Una vez que carga todas las fotos de su familia en el servicio, reconoce automáticamente a las personas de su familia en otras fotos.

La mayoría de los algoritmos de aprendizaje semisupervisados son combinaciones de algoritmos supervisados y no supervisados.

Áreas y tareas de aprendizaje automático (4): aprendizaje por refuerzo

Un agente que interactúa con el mundo hace observaciones, realiza acciones y es recompensado o castigado; debe aprender a elegir acciones de tal manera que obtenga una gran recompensa. Luego debe aprender por sí mismo cuál es la mejor estrategia, llamada política, para obtener la mayor recompensa con el tiempo.

El programa AlphaGo de DeepMind es un buen ejemplo de aprendizaje por refuerzo: apareció en los titulares en mayo de 2017 cuando venció al campeón mundial Ke Jie en el juego de Go. Aprendió su política ganadora analizando millones de juegos y luego jugando muchos juegos contra sí mismo. Tenga en cuenta que el aprendizaje se apagó durante los juegos contra el campeón; AlphaGo solo estaba aplicando la política que había aprendido. Ejemplo de Alpha zero con el ajedrez.

Aprendizaje por lotes y en línea

Otro criterio utilizado para clasificar los sistemas de aprendizaje automático es si el sistema puede o no aprender de forma incremental a partir de un flujo de datos entrantes.

- Aprendizaje por lotes: el sistema es incapaz de aprender de forma incremental: debe entrenarse utilizando todos los datos disponibles. Por lo general, esto requerirá mucho tiempo y recursos informáticos.
- Aprendizaje en línea: el sistema se entrena de forma incremental alimentándose con instancias de datos de forma secuencial, ya sea individualmente o en pequeños grupos llamados mini lotes. Cada paso de aprendizaje es rápido y económico, por lo que el sistema puede aprender sobre nuevos datos sobre la marcha, a medida que llegan.

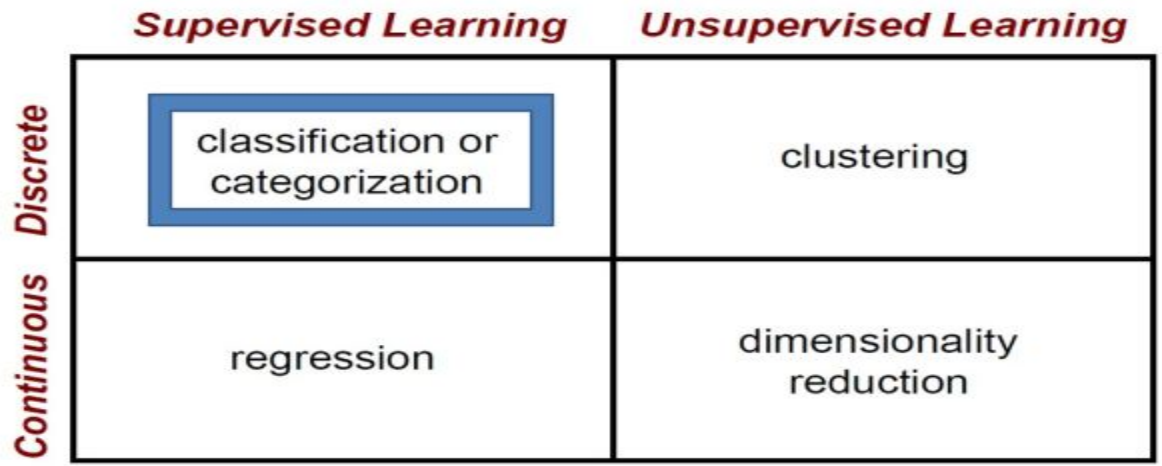
Aprendizaje basado en instancias versus aprendizaje basado en modelos

Una forma más de categorizar los sistemas de Machine Learning es por cómo generalizan.

Método basado en instancias: el sistema aprende los ejemplos de memoria, luego los generaliza a nuevos casos comparándolos con los ejemplos aprendidos (o un subconjunto de ellos), utilizando una medida de similitud. Por ejemplo, KNN.

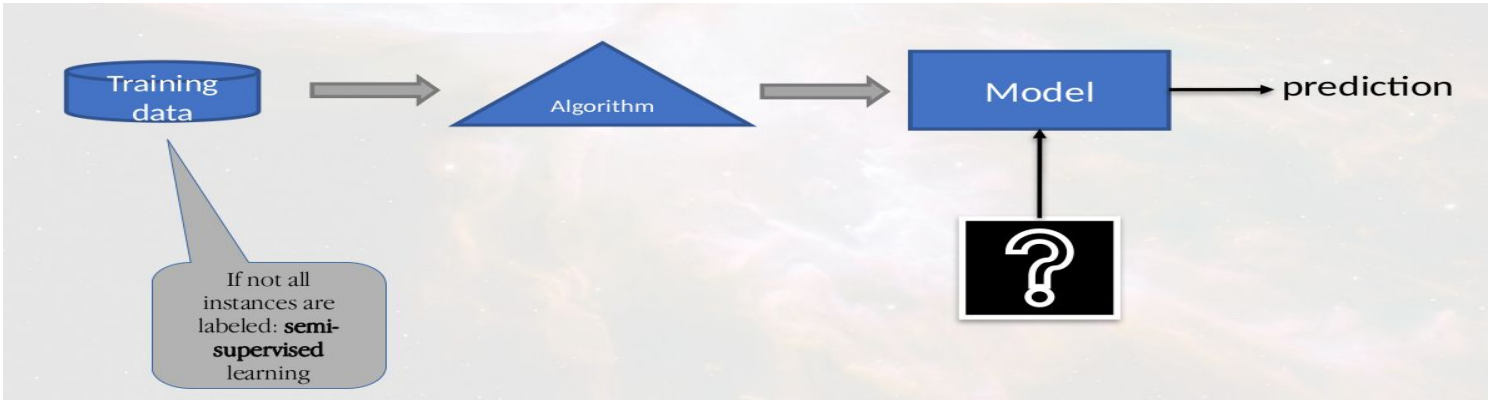
Aprendizaje basado en modelos: el sistema construye un modelo a partir de los ejemplos y luego usa el modelo para hacer predicciones. Por ejemplo, un modelo de regresión lineal.

Problemas de aprendizaje automático.



Aprendizaje supervisado

En el aprendizaje supervisado, a los algoritmos se les presenta un conjunto de instancias clasificadas de las que aprenden una forma de clasificar las instancias no vistas. Cuando el atributo a predecir es numérico en lugar de nominal, se denomina regresión.

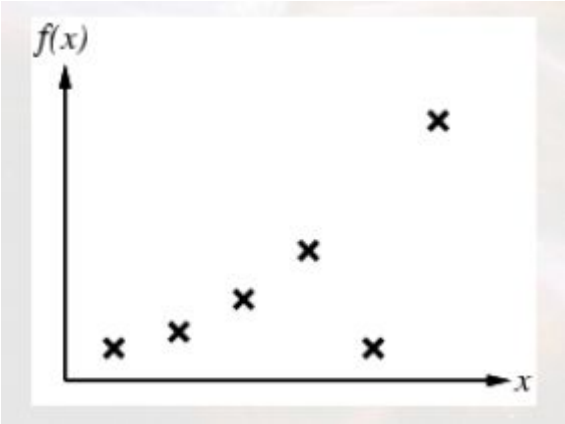


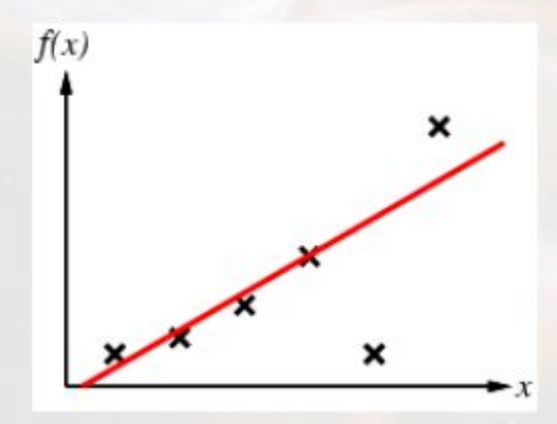
Método de aprendizaje inductivo (1)

- Forma más simple: aprende una función a partir de ejemplos
- f es la función objetivo
- Un ejemplo es un par (x, f(x))
- Tarea de inducción pura:
  - Dada una colección de ejemplos de f, obtener una función que se aproxime a f
  - Encontrar una hipótesis h, tal que h ≈ f, dado un conjunto de ejemplos de entrenamiento
- Este es un modelo muy simplificado de aprendizaje real:
  - Ignora conocimientos previos
  - Se supone que se dan ejemplos

Método de aprendizaje inductivo (2)

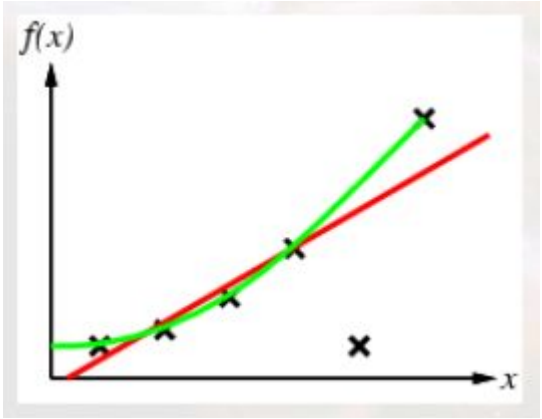
- Construya/ajuste h para que coincida con f en el conjunto de entrenamiento.
- h es consistente si concuerda con f en todos los ejemplos.
- p.ej. ajuste de curvas.

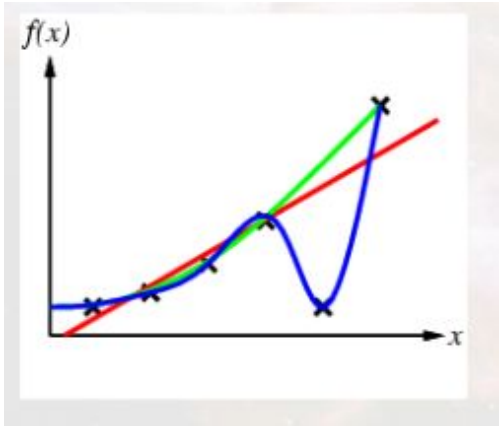




Método de aprendizaje inductivo (3)

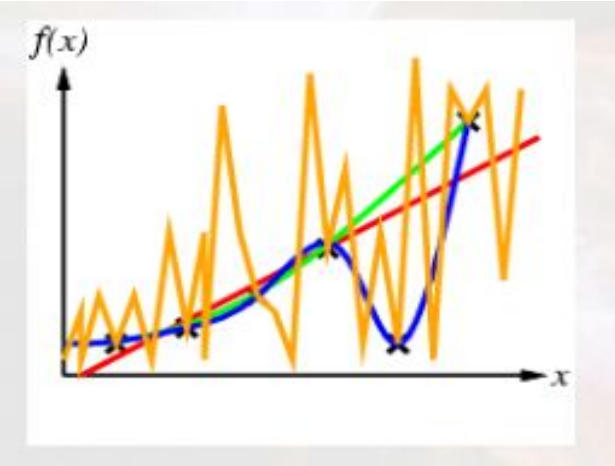
- Construya/ajuste h para que coincida con f en el conjunto de entrenamiento.
- h es consistente si concuerda con f en todos los ejemplos.
- p.ej. ajuste de curvas





Método de aprendizaje inductivo (4)

- La navaja de Occam: seleccione la hipótesis más simple consistente con los datos.



Clasificación

- **Idea: construir un modelo basado en datos anteriores para predecir la clase de los nuevos datos.**
- Dada una colección de registros **(conjunto de entrenamiento).**
  - Cada registro contiene un conjunto de atributos, uno de los atributos es la clase.
- Encontrar un modelo para el atributo de clase como una función de los valores de otros atributos.
- Objetivo: se debe asignar una clase a los nuevos registros con la mayor precisión posible.

Ejemplo: filtro de Spam

Input: email
Output: spam/ham
Setup:
- Get a large collection of example emails, each labeled "spam" or "ham"
- Note: someone has to hand label all this data!
- Want to learn to predict labels of new, future emails

Features: The attributes used to make the ham / spam decision
- Words: FREE!
- Text Patterns: $dd, CAPS
- Non-text: SenderInContacts
- …

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Ejemplo: reconocimiento de dígitos manuscritos

Input: images / pixel grids

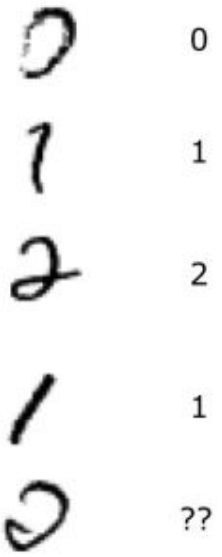
Output: a digit 0-9

Setup:

- Get a large collection of example images, each labeled with a digit
- Note: someone has to hand label all this data!
- Want to learn to predict labels of new, future digit images

Features: The attributes used to make the digit decision

- Pixels: (6,8)=ON
- Shape Patterns: NumComponents, AspectRatio, NumLoops



Ejemplos de clasificación

En la clasificación, las etiquetas se predicen a partir de las entradas x

Ejemplos:

- OCR (Entrada: imágenes, clases: caracteres)
- Diagnóstico médico (entrada: síntomas, clases: enfermedades)
- Clasificador automático de trabajos (entrada: documento, clases: calificaciones)
- Detección de fraude (entrada: actividad de la cuenta, clases: fraude/sin fraude)
- Artículos recomendados en un periódico, libros recomendados
- Identificación de secuencias de ADN y proteínas
- Categorización e identificación de imágenes astronómicas
- Detección de peatones
- ……

Aprendizaje supervisado: conceptos importantes

- Datos: instancias etiquetadas <Xi, y>, p. correo electrónico marcado como spam/no spam
  - Conjunto de entrenamiento, Conjunto de prueba.
- Características: pares de valores de atributo que caracterizan cada X
- Ciclo de Experimentación
  - Aprender parámetros en el conjunto de entrenamiento
  - Calcular la precisión del conjunto de prueba.
  - Muy importante: ¡nunca “mire” el conjunto de prueba!
- Evaluación
  - Precisión: fracción de instancias predichas correctamente.
- Sobreajuste y generalización
  - Obtener un clasificador que funcione bien con los datos de prueba.
  - Sobreajuste: ajustar los datos de entrenamiento en exceso. No generaliza bien.

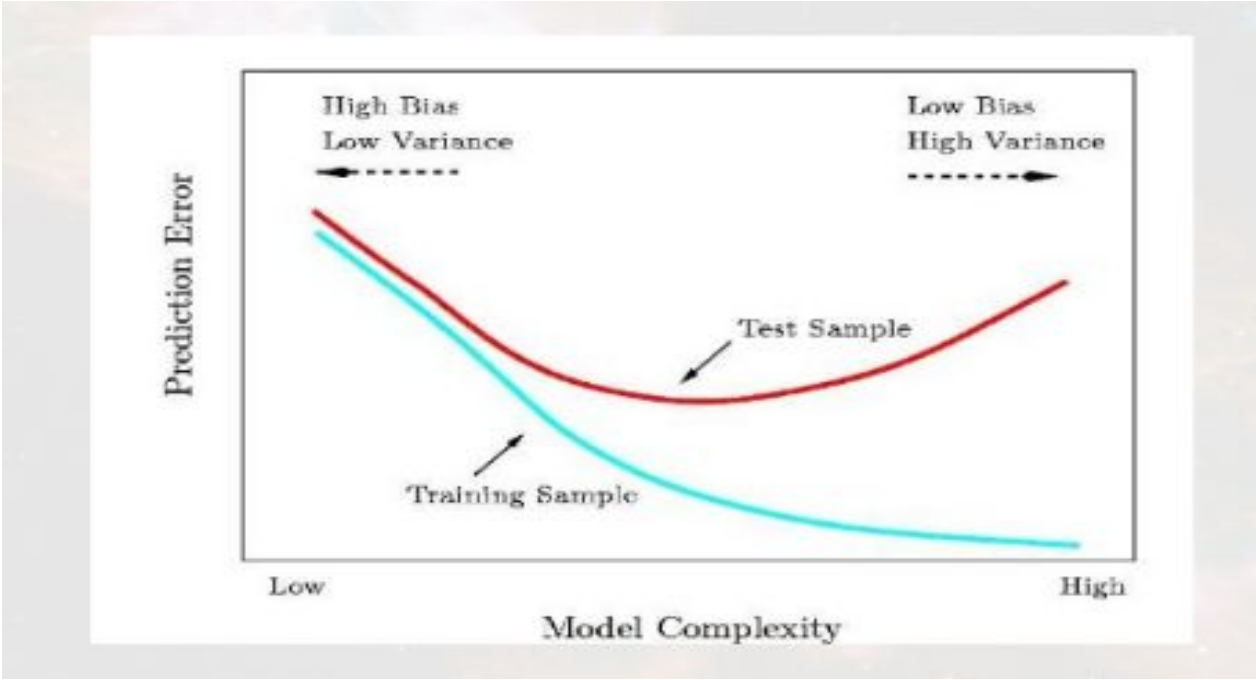
Generalización

Las hipótesis deben generalizar para clasificar correctamente las instancias que no están en los datos de entrenamiento.

Simplemente memorizar ejemplos de entrenamiento es una hipótesis consistente que no generaliza bien.

Navaja de Occam: Encontrar hipótesis simples ayuda a asegurar una correcta generalización.

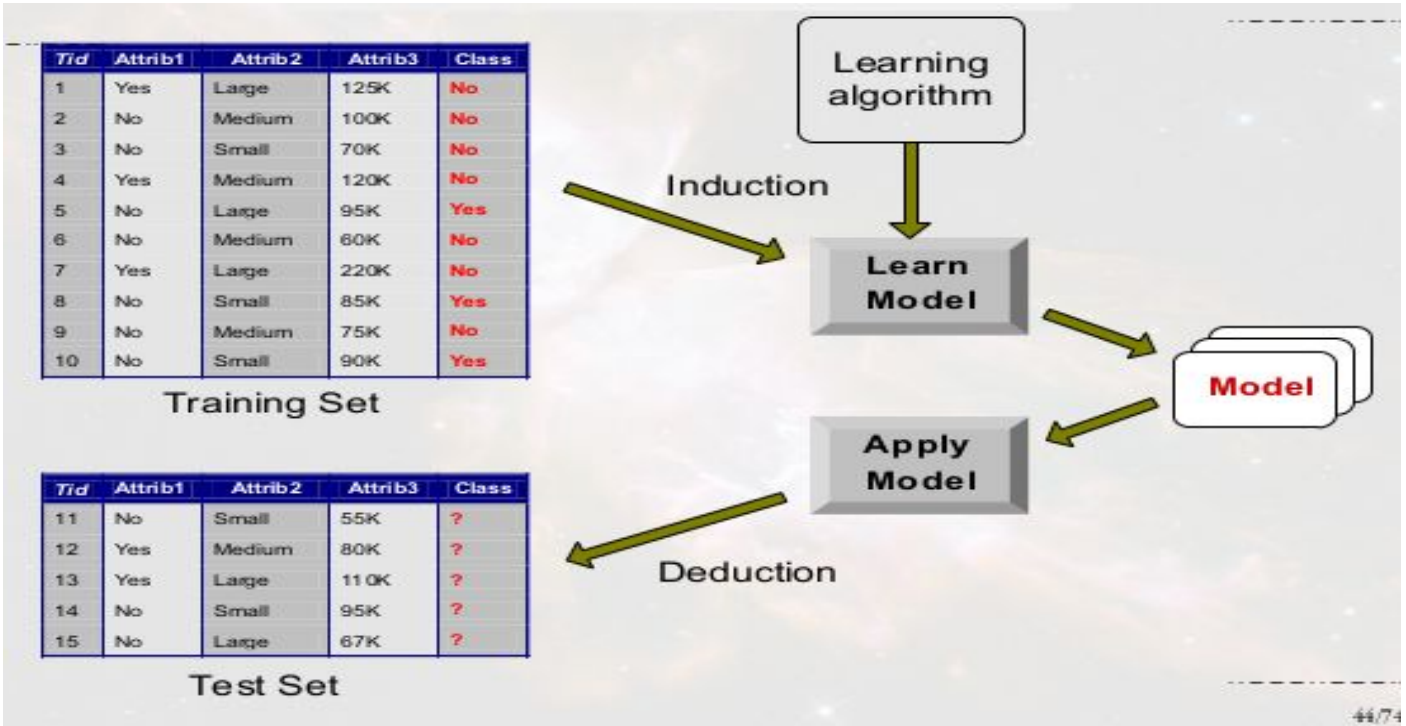
Error de entrenamiento vs error de prueba



Clasificación: un proceso de dos pasos

- Construcción de modelos: descripción de un conjunto de clases predeterminadas
  - Se supone que cada tupla/muestra pertenece a una clase predeterminada (etiqueta de clase).
  - El conjunto de tuplas utilizado para la construcción del modelo es el conjunto de entrenamiento.
  - El modelo se representa como reglas de clasificación, árboles de decisión o fórmulas matemáticas.
- Uso del modelo: clasificar futuros objetos desconocidos.
  - Estimar la precisión del modelo
    - La etiqueta conocida de la muestra de prueba se compara con los resultados del modelo.
    - El conjunto de prueba es independiente del conjunto de entrenamiento; de lo contrario, se producirá un ajuste excesivo
  - Si la precisión es aceptable, utilice el modelo para clasificar las tuplas de datos cuyas etiquetas de clase no se conocen.

Ilustración de la tarea de clasificación

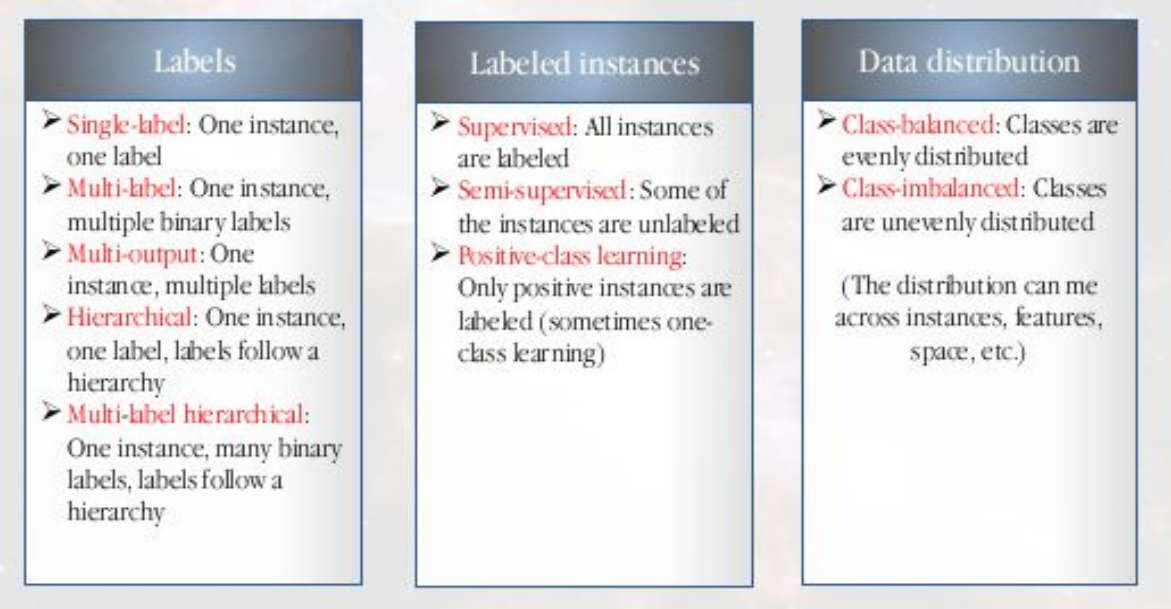


Problemas: preparación de datos

- Limpieza de datos
  - Preprocesar los datos para reducir el ruido y manejar los valores faltantes.
- Análisis de relevancia
  - Selección de características: elimine los atributos irrelevantes o redundantes
  - Selección de instancias: elimine las instancias irrelevantes o redundantes
- Transformación de datos
  - Generalizar datos (discretización)
  - Normalizar valores de atributo

Escenarios de clasificación

Dependiendo de las características de los conjuntos de datos.

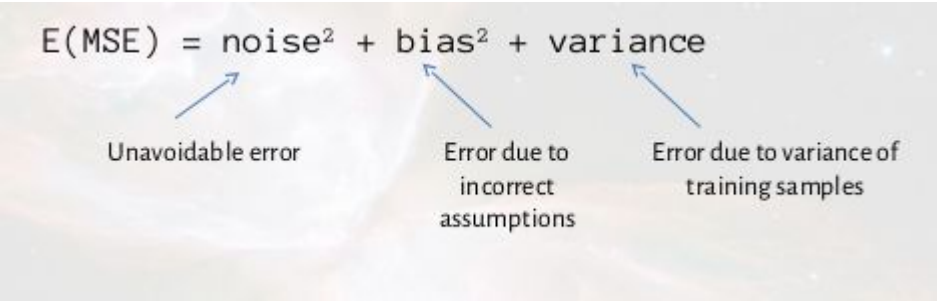


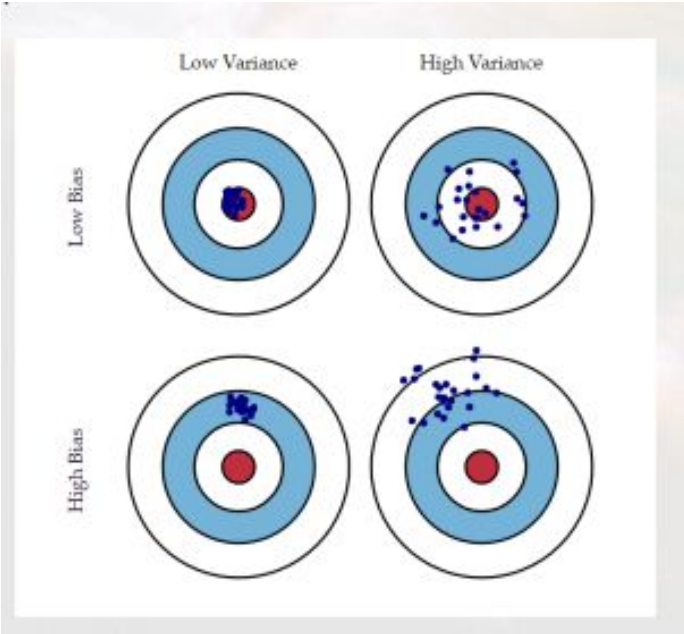
Generalización

- Componentes del error de generalización
  - Sesgo: ¿cuánto difiere el modelo promedio de todos los conjuntos de entrenamiento del modelo verdadero?
    - Error debido a supuestos inexactos/simplificaciones realizadas por el modelo
  - Varianza: cuánto difieren entre sí los modelos estimados a partir de diferentes conjuntos de entrenamiento
- Desajuste: el modelo es demasiado "simple" para representar todas las características de clase relevantes
  - Alto sesgo y baja varianza
  - Alto error de entrenamiento y alto error de prueba
- Sobreajuste: el modelo es demasiado "complejo" y se ajusta a características irrelevantes (ruido) en los datos
  - Bajo sesgo y alta varianza
  - Bajo error de entrenamiento y alto error de prueba

Sesgo/varianza de error

- Error = error irreducible + sesgo + varianza
- Sesgo
  - Error de tendencia central del modelo
- Varianza
  - Error de separación de la tendencia central del modelo.

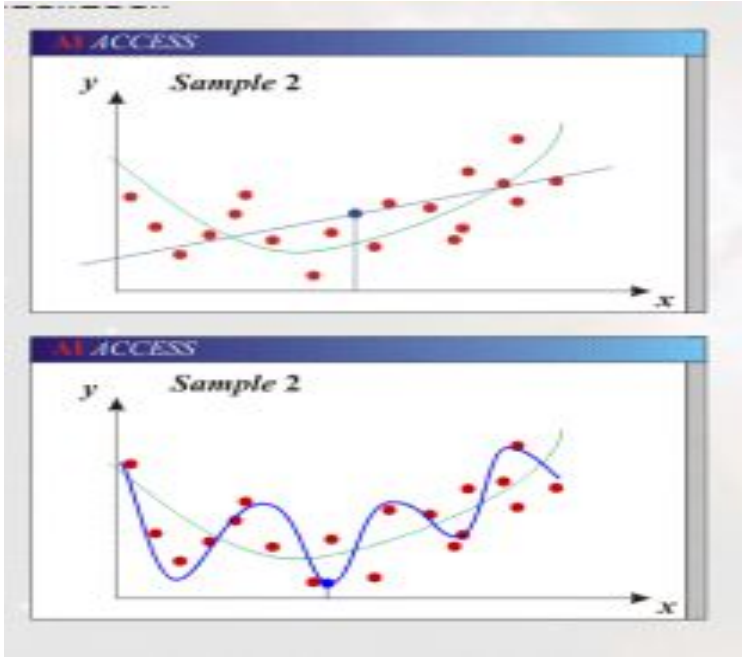




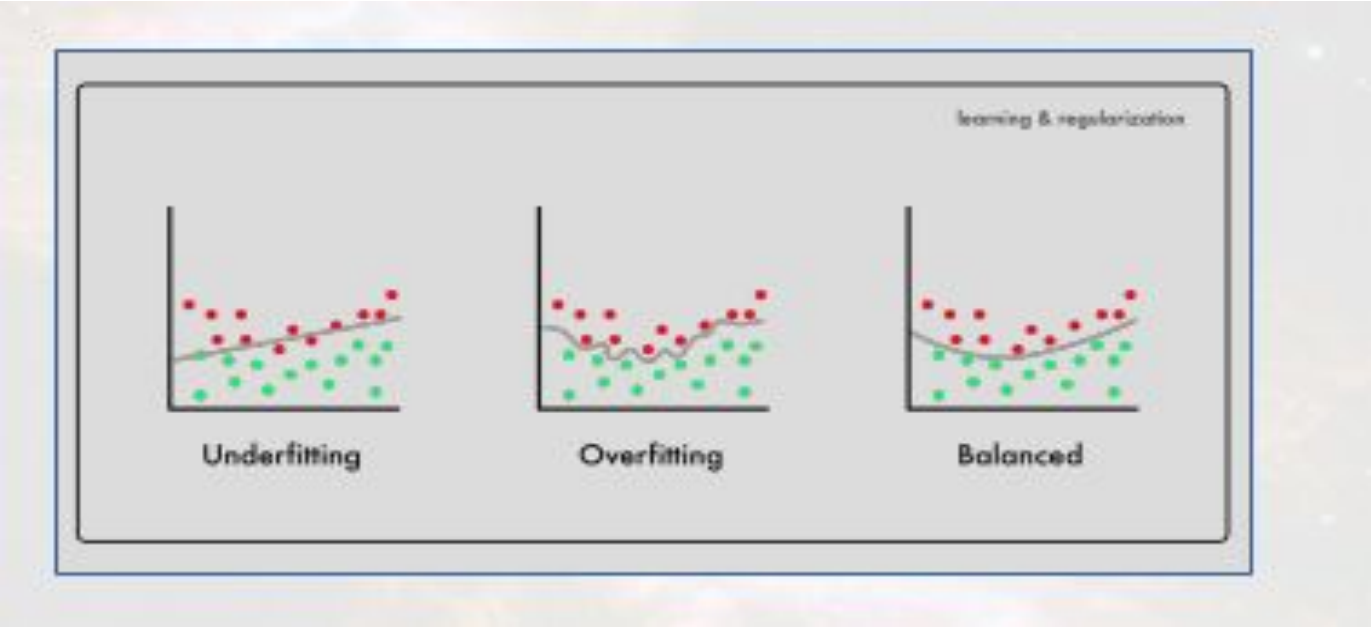
Compensación de sesgo-varianza

Los modelos con muy pocos parámetros son inexactos debido a un gran sesgo (falta de flexibilidad).

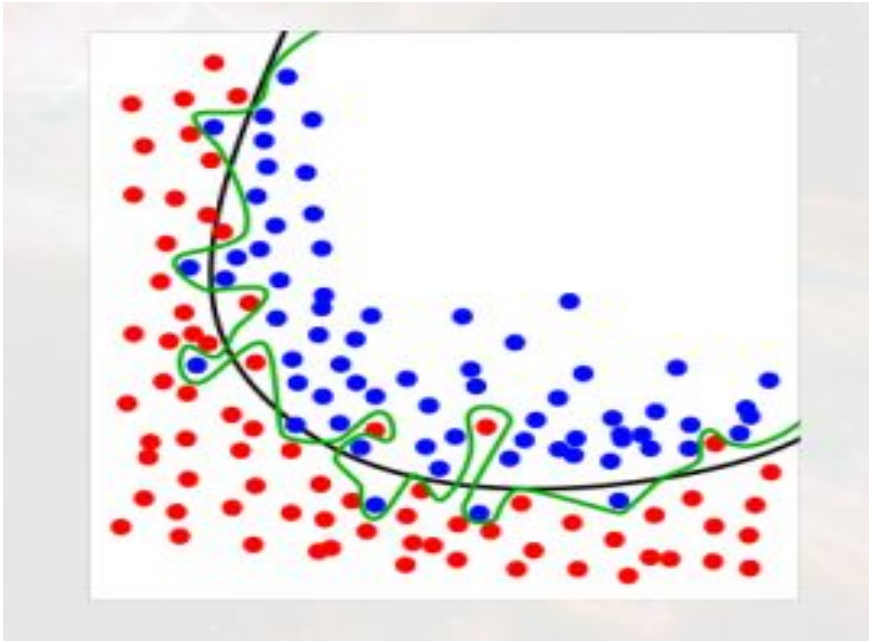
Los modelos con demasiados parámetros son inexactos debido a una gran variación (demasiada sensibilidad a la muestra).



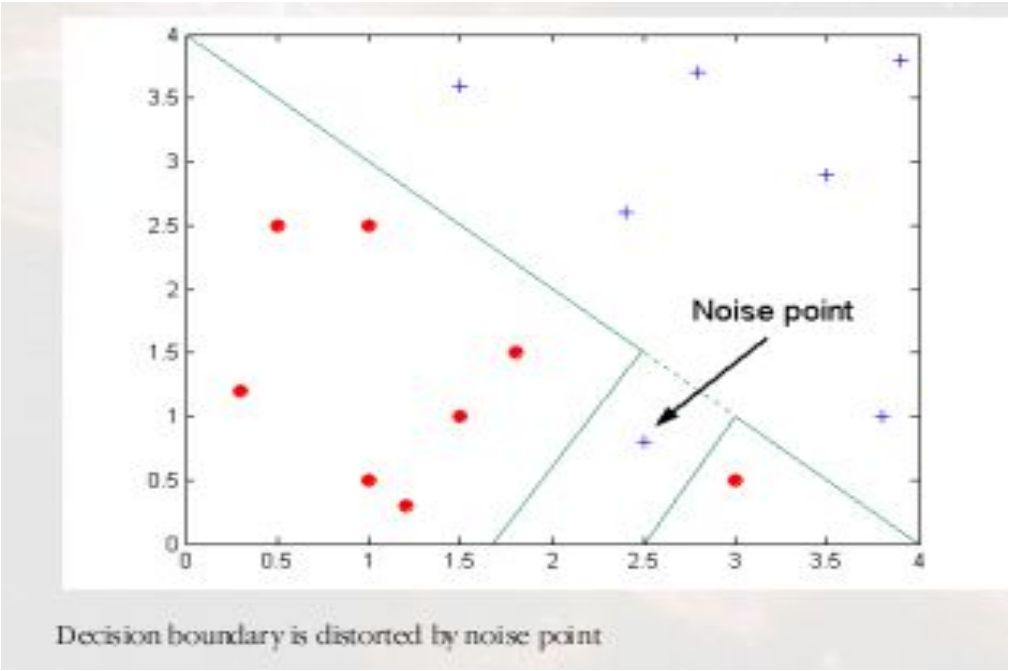
Underfitting and Overfitting



Underfitting and Overfitting



Sobreajuste debido al ruido



Decision boundary is distorted by noise point

Overfitting

Maldición del sobreajuste:

- Relacionado con el aprendizaje
- Peor cuando tu algoritmo de aprendizaje es mejor
- No importa cómo de duro lo intentes, es peor

La navaja de Occam

- Dados dos modelos de errores de generalización similares, uno debería preferir el modelo más simple sobre el modelo más complejo
- Para modelos complejos, existe una mayor posibilidad de que se haya ajustado accidentalmente por errores en los datos.
- Por lo tanto, se debe incluir la complejidad del modelo al evaluar un modelo.
- Problema: No es fácil saber cuál es el modelo más simple