Aprendizaje Estadístico y Computacional

Clase en vivo 1

Jonathan Acosta

Magíster en Ciencia de Datos
Pontificia Universidad Católica de Chile

Quinto Bimestre

Esquema

1. Conceptos Fundamentales semana 1
   - Tipos de aprendizajes: Supervisado y No supervisado
   - Los Problemas de Regresión y Clasificación
   - Etapas de modelamiento en Proyectos de Ciencia de Datos

2. Conceptos Fundamentales Semana 2
   - Ejemplo Regresión Lineal
   - Ejemplo Regresión Logística

Esquema

1. Conceptos Fundamentales semana 1
   - Tipos de aprendizajes: Supervisado y No supervisado
   - Los Problemas de Regresión y Clasificación
   - Etapas de modelamiento en Proyectos de Ciencia de Datos

2. Conceptos Fundamentales Semana 2
   - Ejemplo Regresión Lineal
   - Ejemplo Regresión Logística

Tipos de aprendizajes: Supervisado y No supervisado

¿Qué es el Aprendizaje Supervisado?

Tipos de aprendizajes: Supervisado y No supervisado

¿Qué es el Aprendizaje Supervisado?

Se llama **supervisado** por la presencia de la variable de resultado para guiar el proceso de aprendizaje.

Tipos de aprendizajes: Supervisado y No supervisado

¿Qué es el Aprendizaje Supervisado?

Se llama **supervisado** por la presencia de la variable de resultado para guiar el proceso de aprendizaje.

¿Qué es el Aprendizaje No Supervisado?

Tipos de aprendizajes: Supervisado y No supervisado

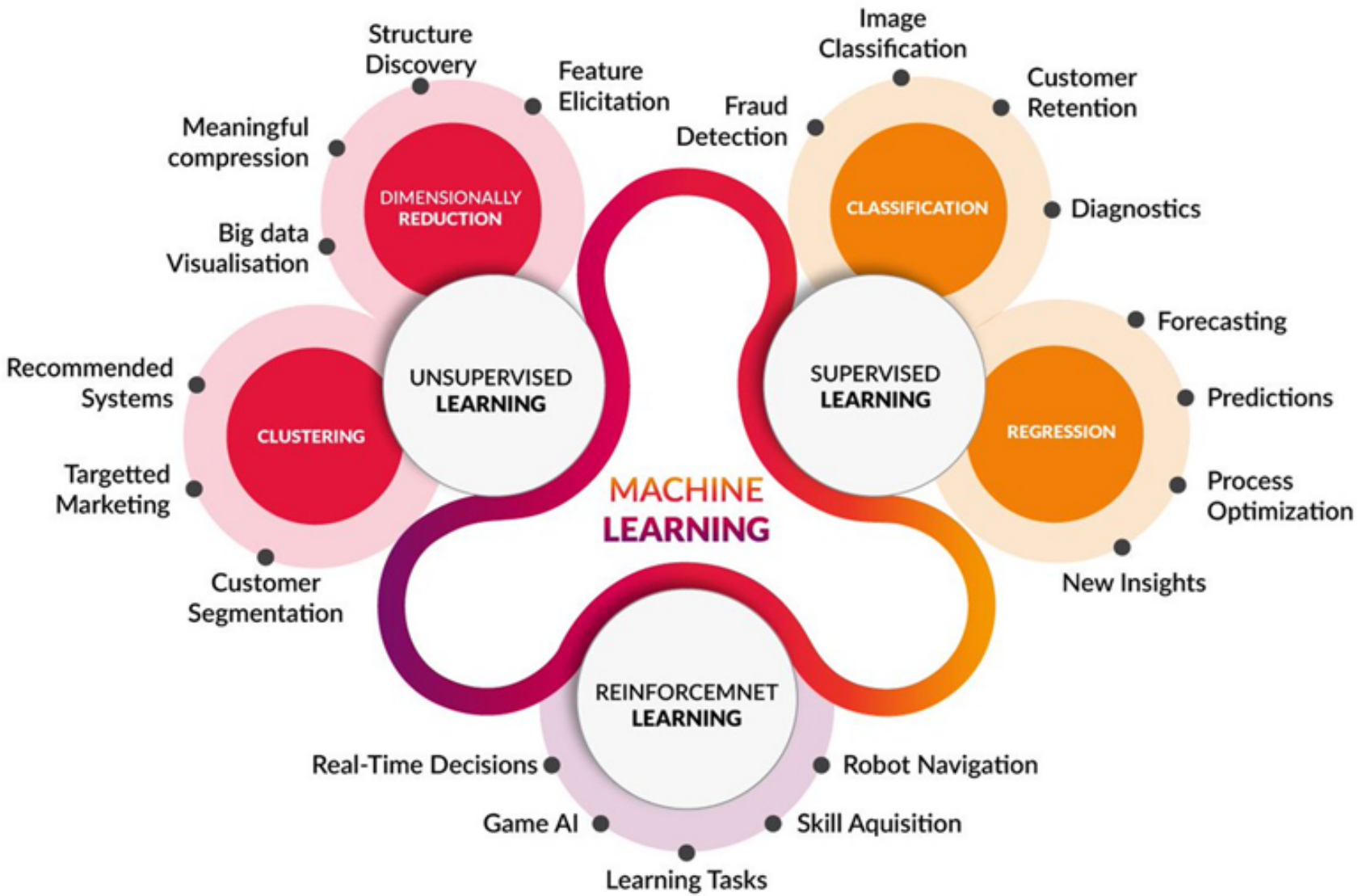
¿Qué es el Aprendizaje Supervisado?

Se llama **supervisado** por la presencia de la variable de resultado para guiar el proceso de aprendizaje.

¿Qué es el Aprendizaje No Supervisado?

Como contra-parte del Aprendizaje Supervisado, el Aprendizaje No Supervisado es aquel donde no se conoce la respuesta. Corresponde al aprendizaje "sin profesor".

Tipos de aprendizajes: Supervisado y No supervisado



Conceptos Fundamentales del Aprendizaje Supervisado

- La ciencia del aprendizaje desempeña un papel fundamental en los campos de la estadística, la minería de datos y la inteligencia artificial, y se cruza con áreas de la ingeniería y otras disciplinas.

Conceptos Fundamentales del Aprendizaje Supervisado

- La ciencia del aprendizaje desempeña un papel fundamental en los campos de la estadística, la minería de datos y la inteligencia artificial, y se cruza con áreas de la ingeniería y otras disciplinas.
- La mezcla de áreas provoca que varias palabras que signifiquen lo mismo:

  $\boldsymbol{X}$: Entrada = Caracteristicas = Regresores = Predictores = V. Independientes.

  Y: Salida = Respuesta = V. Dependiente.

Conceptos Fundamentales del Aprendizaje Supervisado

- La ciencia del aprendizaje desempeña un papel fundamental en los campos de la estadística, la minería de datos y la inteligencia artificial, y se cruza con áreas de la ingeniería y otras disciplinas.
- La mezcla de áreas provoca que varias palabras que signifiquen lo mismo:

  $\boldsymbol{X}$: Entrada = Caracteristicas = Regresores = Predictores = V. Independientes.

  Y: Salida = Respuesta = V. Dependiente.
- Utilizaremos las anteriores indistintamente.

Los Problemas de Regresión y Clasificación

- En el aprendizaje supervisado es natural pensar que la naturaleza de la variable respuesta influye en las técnicas de aprendizajes específicas.

Los Problemas de Regresión y Clasificación

- En el aprendizaje supervisado es natural pensar que la naturaleza de la variable respuesta influye en las técnicas de aprendizajes específicas.
- El problema de regresión surge cuando la variable de interés es numérica (continua o discreta).

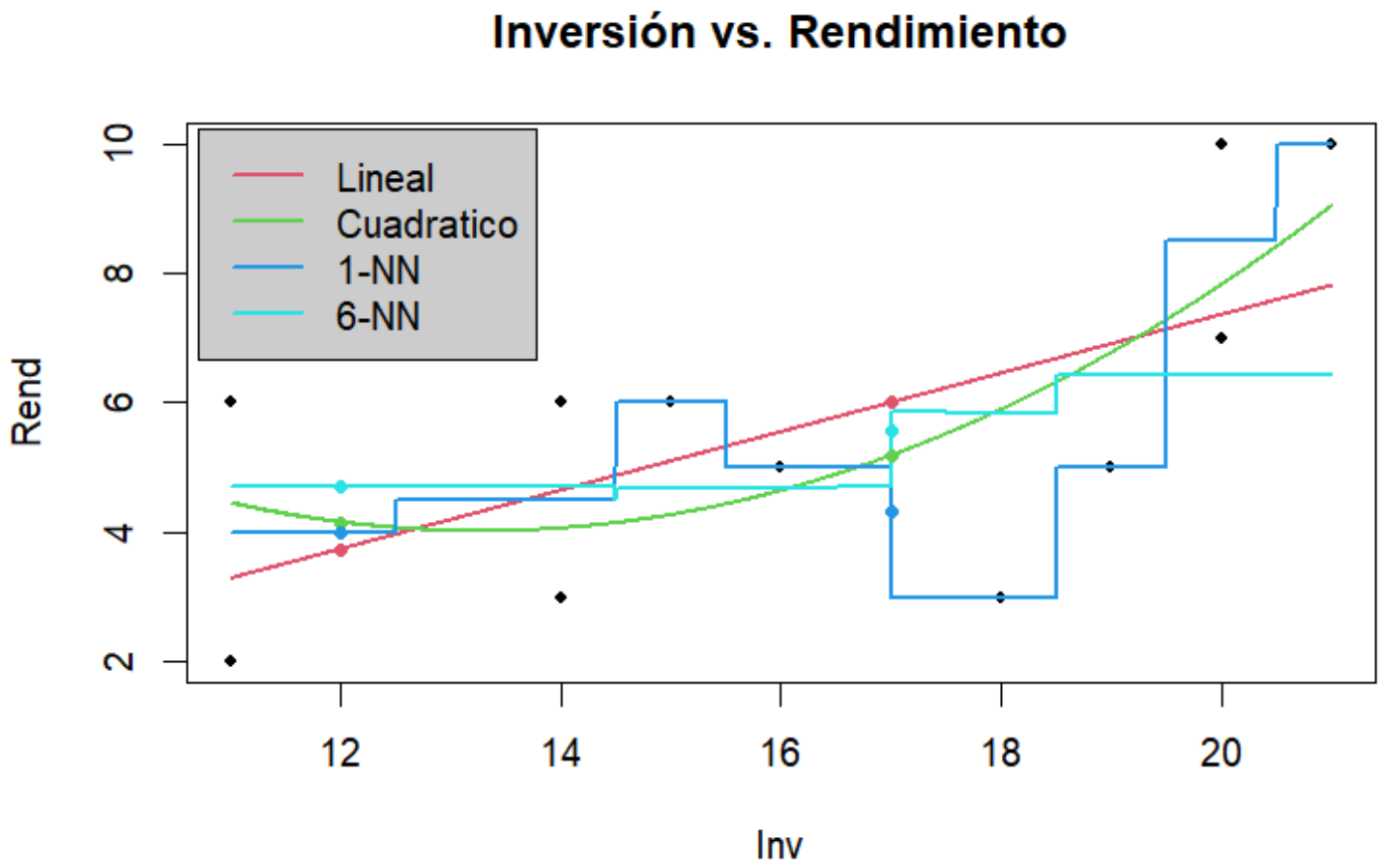
Los Problemas de Regresión y Clasificación

- En el aprendizaje supervisado es natural pensar que la naturaleza de la variable respuesta influye en las técnicas de aprendizajes específicas.
- El problema de regresión surge cuando la variable de interés es numérica (continua o discreta).
- El problema de clasificación surge cuando la variable de interés es categórica (nominal u ordinal).

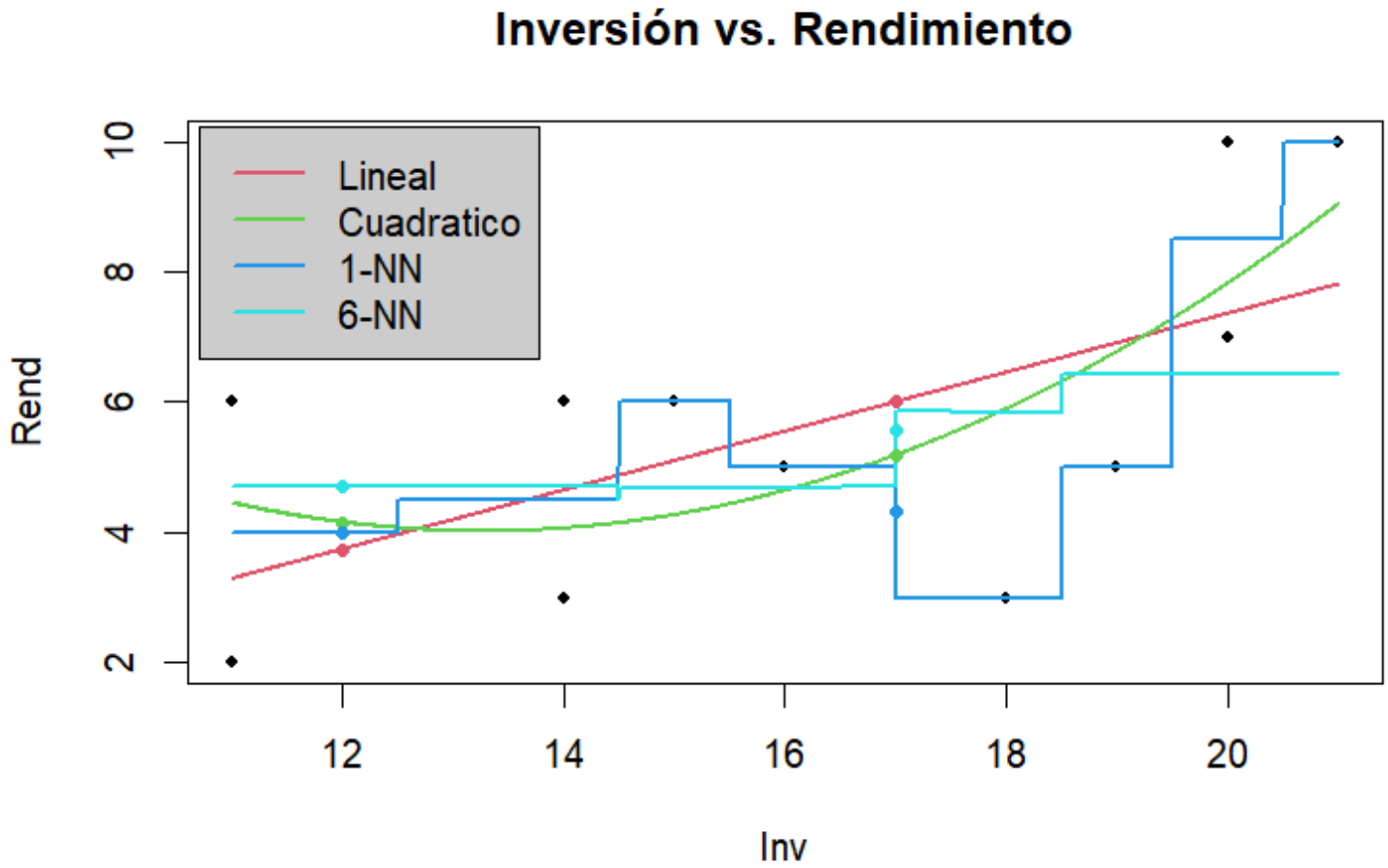
Los Problemas de Regresión y Clasificación

- En el aprendizaje supervisado es natural pensar que la naturaleza de la variable respuesta influye en las técnicas de aprendizajes específicas.
- El problema de regresión surge cuando la variable de interés es numérica (continua o discreta).
- El problema de clasificación surge cuando la variable de interés es categórica (nominal u ordinal).

El problema de Regresión



El problema de Regresión



Dentro de los modelos de regresión lineal, ¿cuál es más apropiado?, y en k-NN, ¿cómo elegir k?

El problema de Clasificación

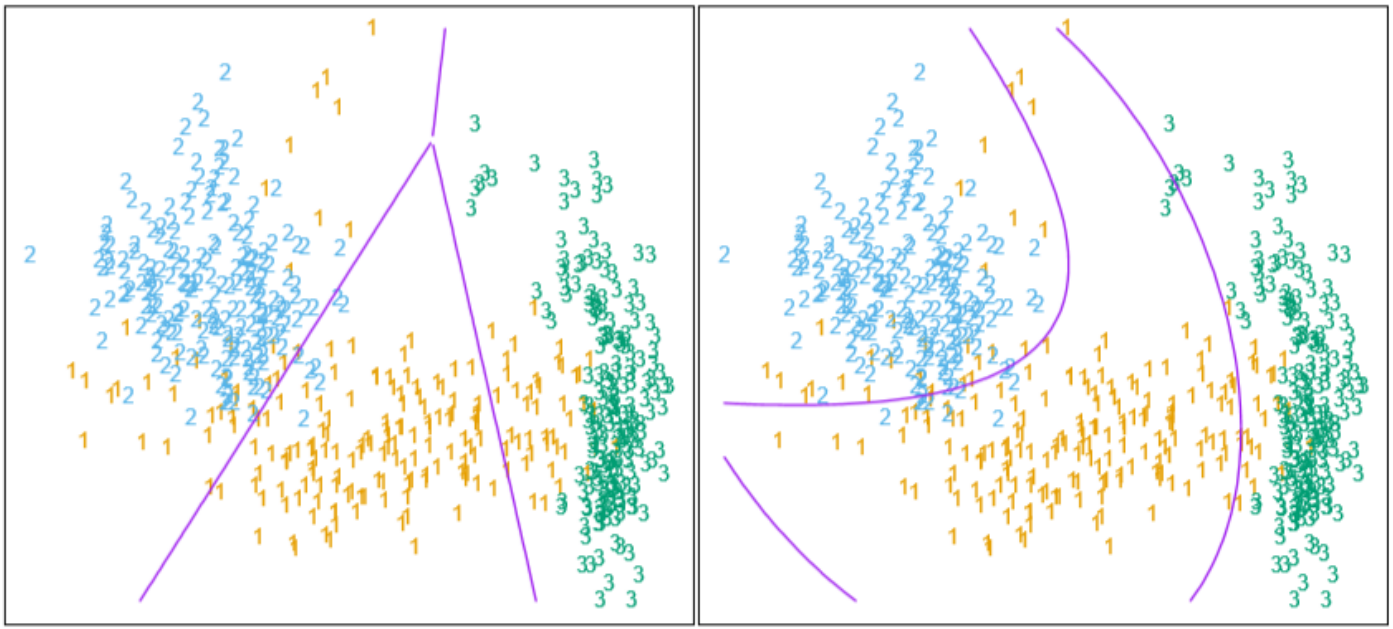


Figura 1: A la izquierda el clasificador $\boldsymbol{x} = (X_1, X_2)$ y la derecha el clasificador $\widetilde{\boldsymbol{x}} = (X_1, X_2, X_1^2, X_2^2, X_1X_2)$. Fuente **?**

El problema de Clasificación

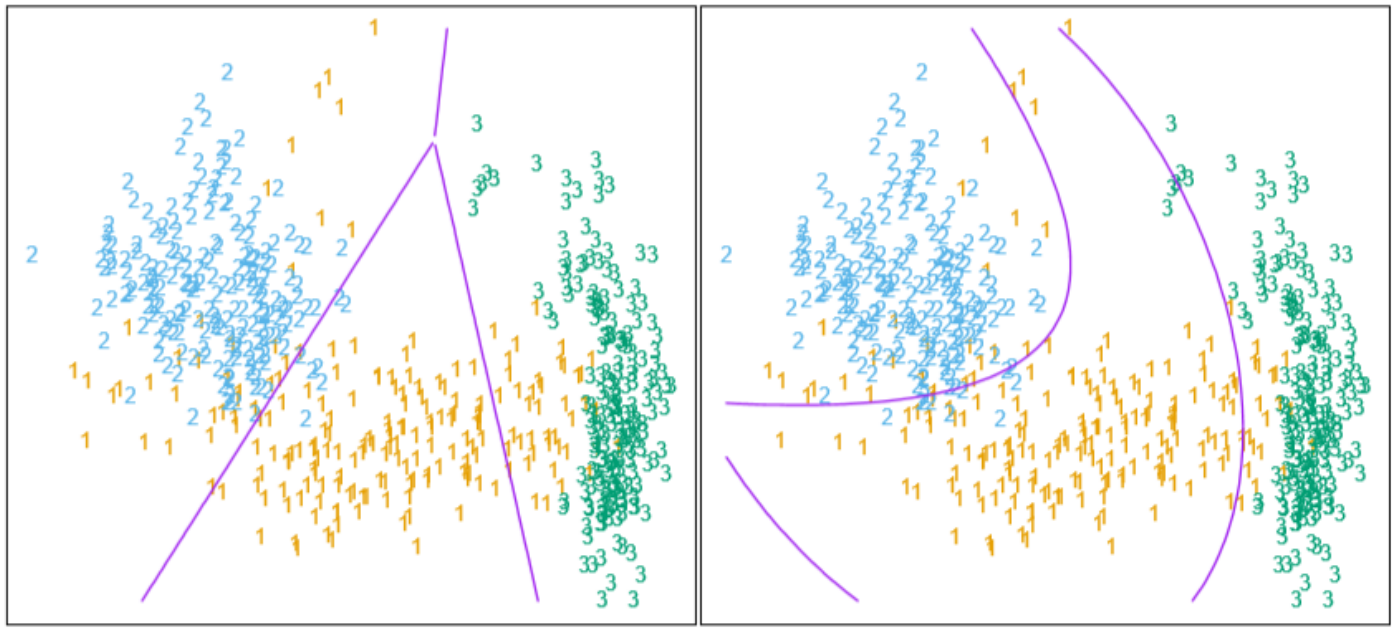


Figura 1: A la izquierda el clasificador $\boldsymbol{x} = (X_1, X_2)$ y la derecha el clasificador $\widetilde{\boldsymbol{x}} = (X_1, X_2, X_1^2, X_2^2, X_1X_2)$. Fuente **?**

Como definir decidir si es más apropiado las fronteras lineales o curvas?

overfitting/underfitting

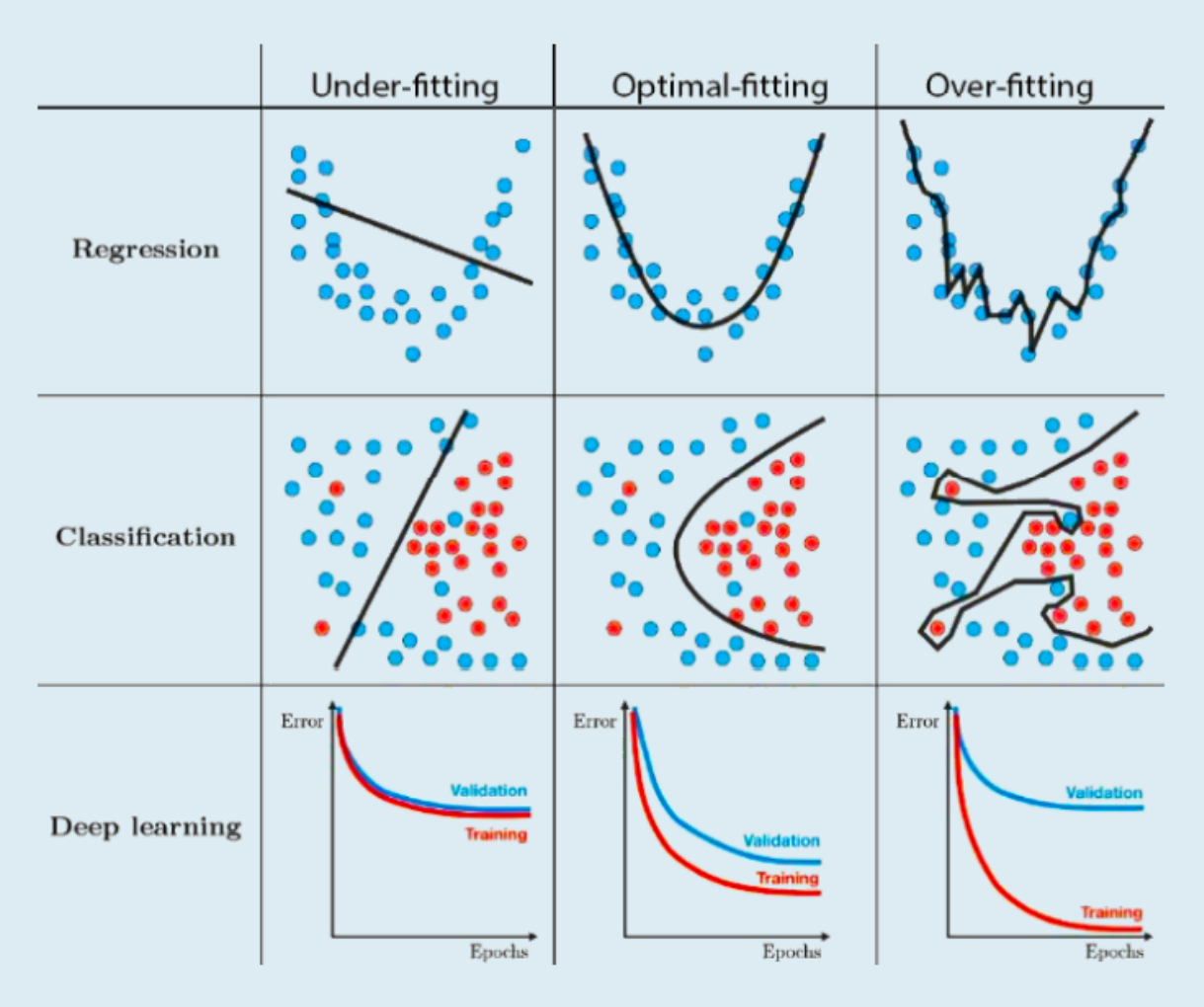


Figura 2: overfitting/underfitting: Creterios a tener en cuenta tanto para el problema de clasificación como regresión

Ciclo de un Proyecto de Ciencia de Datos

Figura 3: Ciclo de vida de un proyecto de Ciencia de Datos

Esquema

1. Conceptos Fundamentales semana 1
   - Tipos de aprendizajes: Supervisado y No supervisado
   - Los Problemas de Regresión y Clasificación
   - Etapas de modelamiento en Proyectos de Ciencia de Datos

2. Conceptos Fundamentales Semana 2
   - Ejemplo Regresión Lineal
   - Ejemplo Regresión Logística

Ejemplo Regresión Lineal

Se dispone de datos anuales durante un período de 14 años para las variables:

- Consumo (C),
- Exportaciones (Ex),
- Oferta Monetaria (OM),

de la economía de un determinado país. Al ajustar un modelo de regresión a las variables Consumo y Oferta Monetaria, se obtuvo:

$$\begin{array}{rcccc} C_i & = & 851{,}3 & + & 0{,}7945 \cdot OM_i \\ & & (15{,}52) & & (11{,}60) \end{array}$$

Además, se obtuvo $R^2 = 0{,}9182$ y $\text{SCError}_1 = 143868{,}69$.

Posteriormente, se ajustó el modelo de regresión con todas las variables, obteniendo:

$$\begin{array}{rcccccc} C_i & = & 655{,}8 & + & 3{,}359 \cdot Ex_i & + & 0{,}0556 \cdot OM_i \\ & & (11{,}39) & & (4{,}282) & & (0{,}312) \end{array}$$

Además, se obtuvo $R^2 = 0{,}969$ y $\text{SCError}_2 = 53938{,}8$.

En ambos modelos las cifras entre paréntesis corresponden a los valores de los estadísticos t-Student.

Ejemplo Regresión Lineal

Si, además, se sabe que la inversa de la matriz $\boldsymbol{X}^\top \boldsymbol{X}$, donde el orden corresponde al modelo 2, es

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{pmatrix} 0{,}6758016 & -0{,}007303 & 0{,}001341 \\ -0{,}007303 & 0{,}000125 & -0{,}0000276 \\ 0{,}001341 & -0{,}0000276 & 0{,}0000064 \end{pmatrix}$$

a) Realice un análisis de significancia, para determinar si se justifica la incorporación de la variable Exportaciones.

b) En ambos modelos estime σ^2.

c) Suponga que $Ex = 275$, y $OM = 1000$. Utilizando ambos modelos y ésta información, estime el consumo C. Compare los intervalos de confianza de las predicciones realizadas por ambos modelos.

d) ¿Usted propondría alguno de estos dos modelos? Justifique su respuesta.

Ejemplo Regresión Logística

En un estudio se examinaron las relaciones entre las condiciones meteorológicas durante los primeros 21 días después de la eclosión de las crías de codorniz escalada y su supervivencia hasta los 21 días de edad. Sea p la probabilidad de que las crías sobrevivan más de 21 días. Se utilizó un total de 54 crías en el estudio donde se ajustó un modelo logístico, cuyos resultados se muestran en la siguiente Tabla, donde $\widehat{\beta}_j$ son estimaciones de los parámetros, $se(\widehat{\beta}_j)$ sus respectivos errores estándar, mientras, que C el estadístico de prueba de razón de verosimilitud (delta de Deviance), cuando la variable indicada se excluyó del modelo completo que contenía las tres variables explicativas.

Variable explicativa	$\widehat{\beta}_j$	$se(\widehat{\beta}_j)$	C
Temperatura mínima durante los primeros 12 días	0,143	0,19	0,602
Temperatura máxima durante los primeros 7 días	1,247	0,45	14,83
Número de días con precipitaciones durante los primeros 7 días	-0,706	0,45	2,83

Ejemplo Regresión Logística

1. Con la información entregada, proponga un modelo basado en los estadísticos de la razón de verosimilitud (delta de Deviance).

2. Utilice las pruebas de Wald para determinar qué variables explicativas son significativas.

3. Suponga que el estimador del intercepto es $\widehat{\beta}_0 = -20{,}45$. Obtenga la estimación de la probabilidad que una cría sobreviva más de 21 días si se tiene una temperatura mínima durante los primeros 12 días fue de 2^oC, una temperatura máxima de 18^oC durante los primeros 7 días y solo un día de precipitaciones durante los primeros 7 días.

¿Alguna consulta?