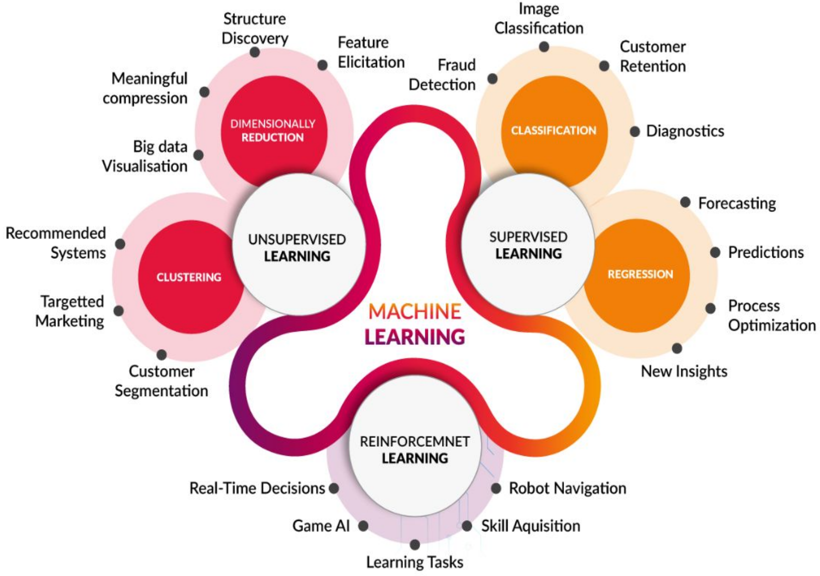
Machine Learning

Teoría Bayesiana de la Decisión

Christian Oliva Moya
Luis Fernando Lago Fernández

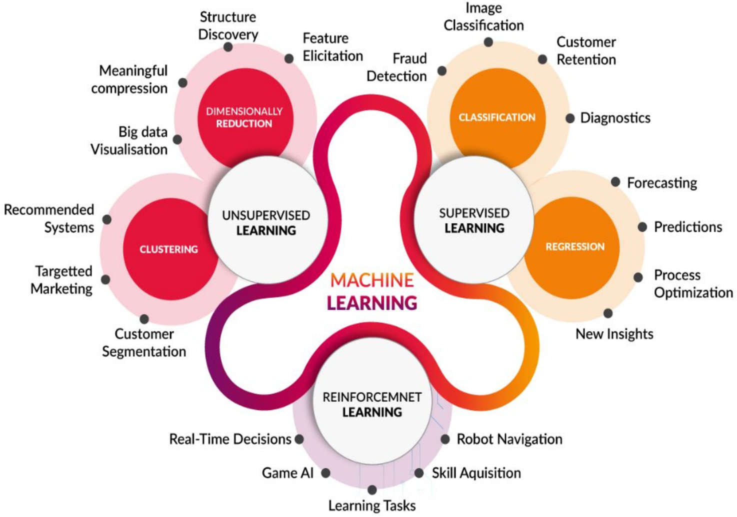
Repaso

- Aprendizaje Automático
  - Supervisado
    - Clasificación
    - Regresión
  - No supervisado
    - Clustering
  - Otros paradigmas



Repaso

- Aprendizaje Supervisado
  - Problema: $\{(\boldsymbol{x}_1, t_1), (\boldsymbol{x}_2, t_2), \ldots, (\boldsymbol{x}_N, t_N)\}$
    - $\boldsymbol{x}_i$ ≡ vector de atributos
    - t_i ≡ variable objetivo (target)
    - N ≡ número de datos/patrones
  - Objetivo: predecir t_i a partir de $\boldsymbol{x}_i$ → $y_i = f(\boldsymbol{x}_i) \approx t_i$



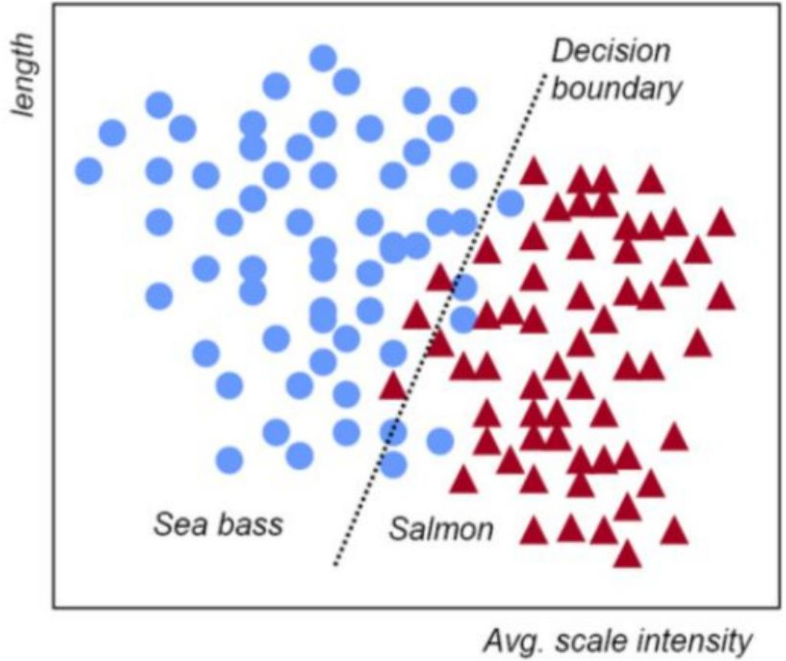
Repaso

- Ejemplo: salmones vs lubinas

- Atributos:
  - Longitud
  - Brillo
  - Posición de las aletas
  - Color
  - …

Repaso

- Ejemplo: salmones vs lubinas

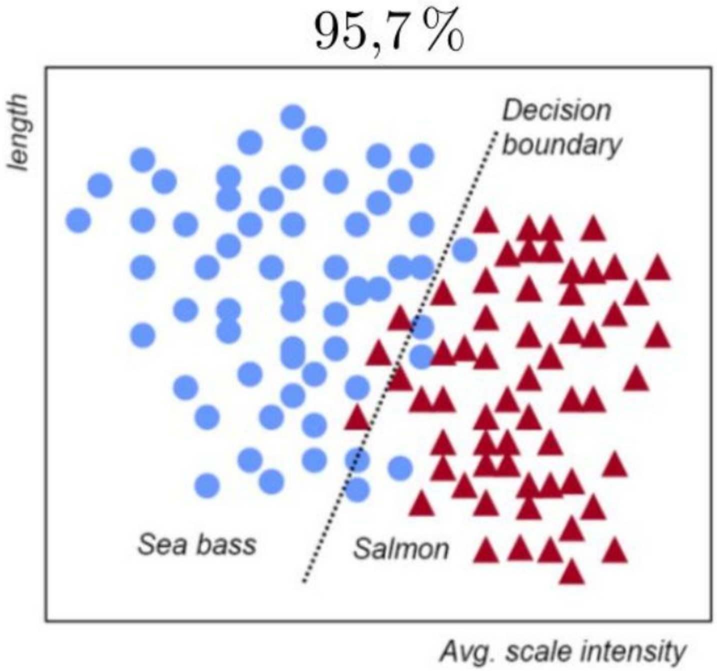


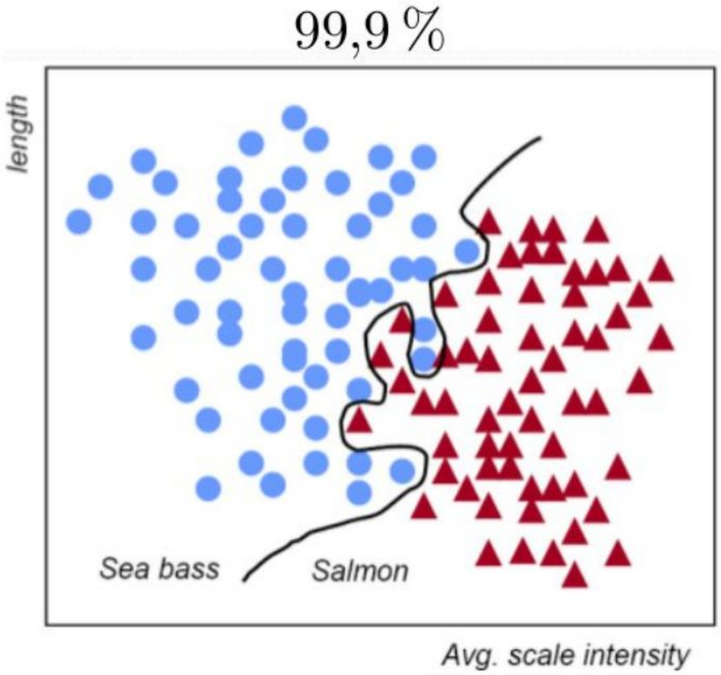
(From Duda, Hart and Stork, *Pattern Classification*, 2001)

Accuracy: 95%

Repaso

¿Cuál es el mejor modelo?

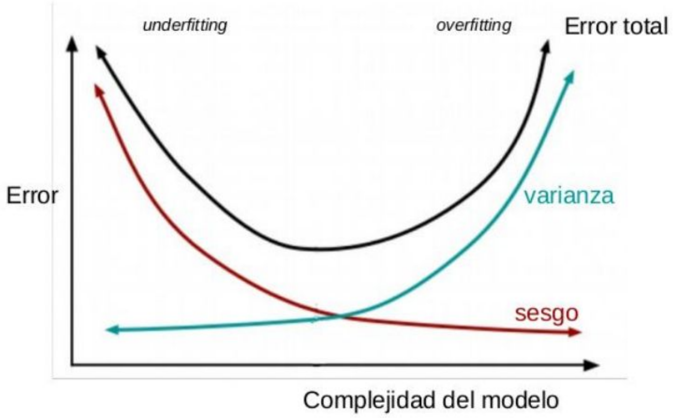




(From Duda, Hart and Stork, *Pattern Classification*, 2001)

Repaso

- Complejidad, generalización y sobreajuste
  - Modelos más complejos se adaptan mejor a los datos
  - Modelos más complejos generalizan peor (sobreajuste / overfitting)



Teoría Bayesiana de la Decisión

- Planteamiento estadístico del problema, asumiendo que:
  - El problema se puede plantear en términos de probabilidades y costes
  - Todas las probabilidades relevantes son conocidas
- Problema de ejemplo:

	L < 50 cm	**L > 50 cm**	**Total**
Salmón	20	40	60
Lubina	30	10	40

Teoría Bayesiana de la Decisión

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40

- Probabilidad a priori:
  - Probabilidad de observar un salmón/lubina en la cinta
    - P(salmón) = 60 / 100 = 60% = 0.6
    - P(lubina) = 40 / 100 = 40% = 0.4
    - P(salmón) + P(lubina) = 0.6 + 0.4 = 1.0
  - En general: P(C1) + P(C2) + … + P(Cn) = 1.0

Teoría Bayesiana de la Decisión

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40

- Regla de decisión:
  - Define qué acción a tomar basándose en la observación
- Supongamos que:
  - Solo conocemos el prior, es decir, no hay datos de entrada
- Lo mejor que podemos hacer es:

$$\text{decisión} = \begin{cases} \textit{lubina si } p(\textit{lubina}) > p(\textit{salmón}) \\ \textit{salmón si } p(\textit{lubina}) < p(\textit{salmón}) \end{cases}$$

Teoría Bayesiana de la Decisión

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40

- Regla de decisión en base al prior:
  - Elegimos siempre la clase más probable
  - Es la regla óptima en ausencia de más información
  - Siempre se asigna la misma clase a todos los patrones

En el ejemplo, P(salmón) > P(lubina) → todos los datos son salmones

Teoría Bayesiana de la Decisión

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40

- Regla de decisión:
  - Define qué acción a tomar basándose en la observación
- Supongamos que:
  - Conocemos un conjunto de variables que describen a cada patrón
- Probabilidad a posteriori:
  - P(salmón | L < 50 cm) = 20 / 50 = 0.4 P(lubina | L < 50 cm) = 30 / 50 = 0.6
  - P(salmón | L > 50 cm) = 40 / 50 = 0.8 P(lubina | L > 50 cm) = 10 / 50 = 0.2

Teoría Bayesiana de la Decisión

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40

- Regla de decisión:
  - Define qué acción a tomar basándose en la observación
- Supongamos que:
  - Conocemos un conjunto de variables que describen a cada patrón
- Probabilidad a posteriori:
  - P(salmón | L < 50 cm) = 20 / 50 = 0.4   P(lubina | L < 50 cm) = 30 / 50 = 0.6
  - P(salmón | L > 50 cm) = 40 / 50 = 0.8   P(lubina | L > 50 cm) = 10 / 50 = 0.2

Teoría Bayesiana de la Decisión

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40

- Regla de decisión MAP:
  - Elegimos siempre la clase más probable dados los atributos observados
  - Estimación MAP (máximo a posteriori)

En el ejemplo:

Si L < 50 cm → P(salmón | L < 50 cm) = 20 / 50 = 0.4 P(lubina | L < 50 cm) = 30 / 50 = 0.6

Si L > 50 cm → P(salmón | L > 50 cm) = 40 / 50 = 0.8 P(lubina | L > 50 cm) = 10 / 50 = 0.2

Teoría Bayesiana de la Decisión

- Teorema de Bayes:

$$P(C_j|x_i) = \frac{P(x_i|C_j)P(C_j)}{P(x_i)}$$

Teoría Bayesiana de la Decisión

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40
Total	50	50	100

- Teorema de Bayes:

$$p(\text{salmón} \mid L < 50) = \frac{p(L < 50 \mid \text{salmón})\, p(\text{salmón})}{p(L < 50)}$$

Teoría Bayesiana de la Decisión

- Teorema de Bayes:

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40
Total	50	50	100

$$p(salmón \mid L < 50) = \frac{p(L < 50 \mid salmón)\, p(salmón)}{p(L < 50)}$$

20/50 = 0.4

Teoría Bayesiana de la Decisión

- Teorema de Bayes:

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40
Total	50	50	100

$$\underbrace{p(\text{salmón} \mid L < 50)}_{20/50\,=\,0.4} = \frac{\overbrace{p(L < 50 \mid \text{salmón})}^{20/60\,=\,0.33}\; p(\text{salmón})}{p(L < 50)}$$

Teoría Bayesiana de la Decisión

- Teorema de Bayes:

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40
Total	50	50	100

20/60 = 0.33 60/100 = 0.6

$$p(salmón \mid L < 50) = \frac{p(L < 50 \mid salmón)\; p(salmón)}{p(L < 50)}$$

20/50 = 0.4

Teoría Bayesiana de la Decisión

- Teorema de Bayes:

	L < 50 cm	L > 50 cm	Total
Salmón	20	40	60
Lubina	30	10	40
Total	50	50	100

$$\underbrace{p(salmón \mid L < 50)}_{20/50 = 0.4} = \frac{\overbrace{p(L < 50 \mid salmón)}^{20/60 = 0.33}\ \overbrace{p(salmón)}^{60/100 = 0.6}}{\underbrace{p(L < 50)}_{50/100 = 0.5}}$$

Teoría Bayesiana de la Decisión

$$P(C_j|x_i) = \frac{P(x_i|C_j)P(C_j)}{P(x_i)}$$

- Regla de decisión con Teorema de Bayes:
  - La evidencia $P(x_i)$ no depende de C
  - Por tanto:
    - Decisión $C^* = \text{argmax } P(x \mid C_i)\, P(C_i)$
  - La decisión depende solo de los prioris y las verosimilitudes.

¿Cómo podemos estimarlos a partir de una muestra finita?

Teoría Bayesiana de la Decisión

- Estimación paramétrica:
    - Suponemos una forma funcional conocida f(x, w)
    - Ajustamos los parámetros w para maximizar la probabilidad de las observaciones
    - Algoritmo EM para clasificación suponiendo distribuciones Gaussianas

Teoría Bayesiana de la Decisión

- Clasificador Naive Bayes:
  - Combina el teorema de Bayes suponiendo que los atributos son independientes dada la clase

$$\boxed{P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n \mid y)}{P(x_1, \dots, x_n)}} \Longrightarrow \boxed{P(y \mid x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)}}$$

$$\boxed{\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y)}$$

Teoría Bayesiana de la Decisión

1.9.1. Gaussian Naive Bayes

`GaussianNB` implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

Presupone independencia entre los atributos