2016-2017

Aprendizaje Automático

1. Introducción al Aprendizaje Automático

Enrique Vidal Ruiz
(evidal@dsic.upv.es)

Francisco Casacuberta Nolla
(fcn@dsic.upv.es)

Departament de Sistemas Informàtics i Computació (DSIC)

Universitat Politècnica de València (UPV)

Septiembre, 2016

Index

Introducción

Aprendizaje automático (AA), *aprendizaje computacional* o *"machine learning"* (ML):

- Tecnologías desarrollados en el marco de varias disciplinas relacionadas con los *sistemas inteligentes*: *reconocimiento de formas*, *cibernética*, *inteligencia artificial*, *estadística aplicada*, entre otras.
- Modernamente se suele considerar como un planteamiento *integrador* de todas estas disciplinas
- *No* solo se interesa en el "aprendizaje de modelos" propiamente dicho, sino en todo el proceso de resolución de problemas, basado más o menos explícitamente en una aplicación rigurosa de la *teoría de la decisión estadística*.

Index

Aprendizaje automático: predicción y generalización

Aprendizaje:

- Se asume la existencia de *datos de aprendizaje o entrenamiento*; típicamente datos de *entrada* $x \in \mathcal{X}$ y *salida* $y \in \mathcal{Y}$ de un sistema o proceso
- El objetivo es obtener un modelo (o función $f : \mathcal{X} \to \mathcal{Y}$) que *generalice* estos datos adecuadamente
- "Generalizar" frecuentemente consiste en *predecir* la salida a partir de nuevos datos de entrada diferentes de los de entrenamiento

Regresión y clasificación

- *Regresión*: Tanto los datos de entrada como los de salida pertenecen a dominios ($\mathcal{X}$, $\mathcal{Y}$) arbitarios

  Ejemplos:

  - $\mathcal{X} \subset \mathbb{R}^d$ (vectores de d componentes reales), $\mathcal{Y} \subset \Sigma^\star$ (cadenas de símbolos)
  - Un caso simple: $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ y el modelo predictor es una función $f : \mathbb{R} \to \mathbb{R}$

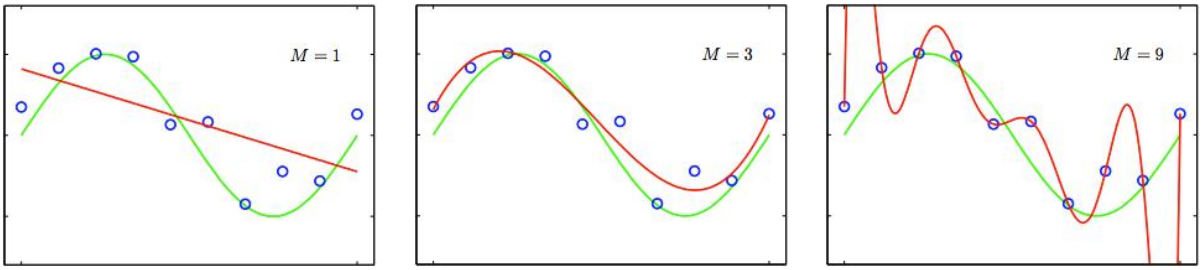
- *Clasificación*: $\mathcal{X}$ es arbritario, pero $\mathcal{Y}$ es un conjunto finito (y generalmente pequeño) de C elementos llamados *clases*. Sin pérdida de generalidad, se puede asumir que $\mathcal{Y} = \{1, 2, \ldots, C\} \subset \mathbb{N}$.

  Ejemplos:

  - Reconocimiento de imágenes de dígitos: $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} = \{1, 2, \ldots, 10\}$
  - Detección de *"spam"*: $\mathcal{X} \subset \Sigma^\star$, $\mathcal{Y} = \{1, 2\}$, donde Σ es el alfabeto ASCII (o UTF) y las etiquetas $\{1, 2\}$ corresponden a *spam* y *no-spam*

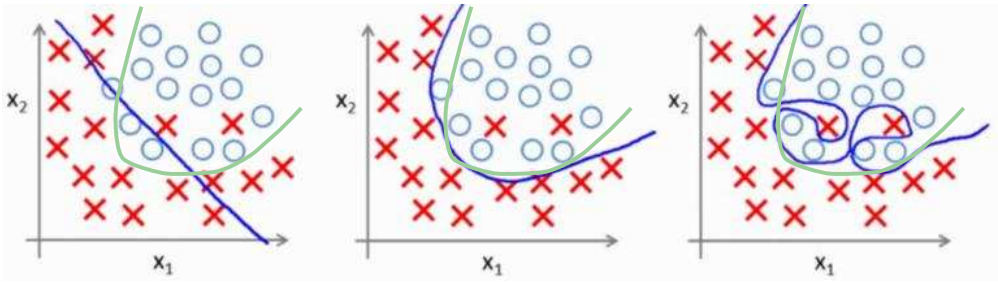
Sobregeneralización y sobreajuste: ejemplos

Modelos de regresión f (en rojo) que aproximan a $g : \mathbb{R} \to \mathbb{R}$ (en verde)



sobregeneralización O.K. sobreajuste

Front. de decisión (azul) que aproximan a las de un clasificador $g : \mathbb{R}^2 \to \{\times, \circ\}$ (verde)



sobregeneralización O.K. sobreajuste

La amenaza de la dimensionalidad

- Si $\mathcal{X} \equiv \mathbb{R}^d$, cuando d es muy grande, aparecen diversos fenómenos adversos que se conocen comunmente como la *"amenaza de la dimensionalidad"*
- La causa común de estos problemas es que, cuando aumenta d, el volumen del espacio (por ej., de un hipercubo) aumenta exponencialmente y los datos aparecen muy dispersos
- Ej.: bastan $10^2 = 100$ puntos para muestrear un intervalo unidad (un hipercubo en $\mathbb{R}^1$) para que los puntos no disten más de $10^{-2} = 0.01$ entre sí. Pero en $\mathbb{R}^{10}$ harían falta 10^{20} puntos
- *Curiosidad* relacionada con lo anterior:
  si $d>>>$, *¡los puntos de un hipercubo tienden a concentrarse "cerca de sus vértices"!*.

  Si $d>>>$, el volumen de un hipercubo de lado $2r$ es $(2r)^d$, mientras que el de una hiperesfera de radio r (contenida en él) es *mucho* menor: $2r^d\pi^{d/2}/d\,\Gamma(d/2)$.

  Al aumentar d, el volumen de la hiperesfera resulta insignificante con respecto al del hipercubo:

$$d \to \infty \;\Rightarrow\; \frac{2r^d\pi^{d/2}}{d\,\Gamma(d/2)}\frac{1}{(2r)^d} \to 0$$

- Con frecuencia, estos fenómenos causan problemas de *sobregeneralización* y *sobreajuste*
- Soluciones: determinación de la *"dimensionalidad intrínseca"*, técnicas de *reducción de la dimensionalidad*, etc.

Evaluación: Esperanza de error

Sea $f : \mathcal{X} \to \mathcal{Y}$ la función obtenida mediante un sistema de AA. La *probabilidad de error* de f es la esperanza estadística de que la salida de f sea incorrecta.

Supongamos que $\mathcal{X}$ es un dominio discreto y sea $P(x)$ la (verdadera) distribución de probabilidad incondicional de las entradas $x \in \mathcal{X}$.

$$E_x[\text{error}(f(x))] \;=\; \sum_{x \in \mathcal{X}} \text{error}(f(x))P(x)$$

Si $\mathcal{X}$ es continuo (por ejemplo, $\mathcal{X} = \mathbb{R}^d$):

$$E_{\boldsymbol{x}}[\text{error}(f(\boldsymbol{x})] \;=\; \int_{\boldsymbol{x} \in \mathcal{X}} \text{error}(f(\boldsymbol{x}))\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

donde $p(\boldsymbol{x})$ es ahora la *densidad de probabilidad* incondicional.

Esta es la *"verdadera"* probabilidad de error, pero normalmente no se conoce $P(x)$, o solo se conocen aproximaciones que no permiten calcular $E[\text{error}(f)]$.

En la práctica, $E[\text{error}(f)]$ se suele *estimar* mediante datos de *"test etiquetados"*; es decir, datos similares a los de entrenamiento, que contienen información de *entrada* y la correspondiente "etiqueta" (información de *salida correcta*).

Estimación de la probabilidad de error

Sea $p = E[\text{error}(f)]$ la *verdadera esperanza* (probabilidad) de error de un sistema basado en f. Una estimación empírica ($\hat{p}$) de p puede obtenerse contabilizando el número de errores de decisión, N_e, que se producen en una *muestra de test* con N datos:

$$\hat{p} = \frac{N_e}{N}$$

Si $N >>$, podemos asumir que $\hat{p}$ se distribuye normalmente como:

$$\hat{p} \sim \mathcal{N}\left(p,\, \frac{p(1-p)}{N}\right)$$

Intervalo de confianza al 95%:

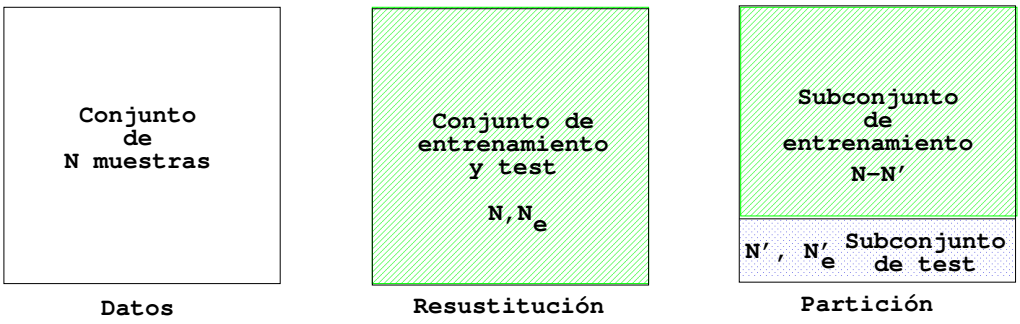
$$P(\hat{p} - \epsilon \leq p \leq \hat{p} + \epsilon) \;=\; 0.95; \qquad \epsilon \;=\; 1.96\,\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Métodos de partición de datos

Para evaluar un sistema de *Aprendizaje Automático*, se necesitan datos etiquetados, no solo para estimar el error, sino para aprender los modelos de decisión. Dado un conjunto de datos, este se puede dividir de diversas formas en subconjuntos de *entrenamiento* y de *test*:

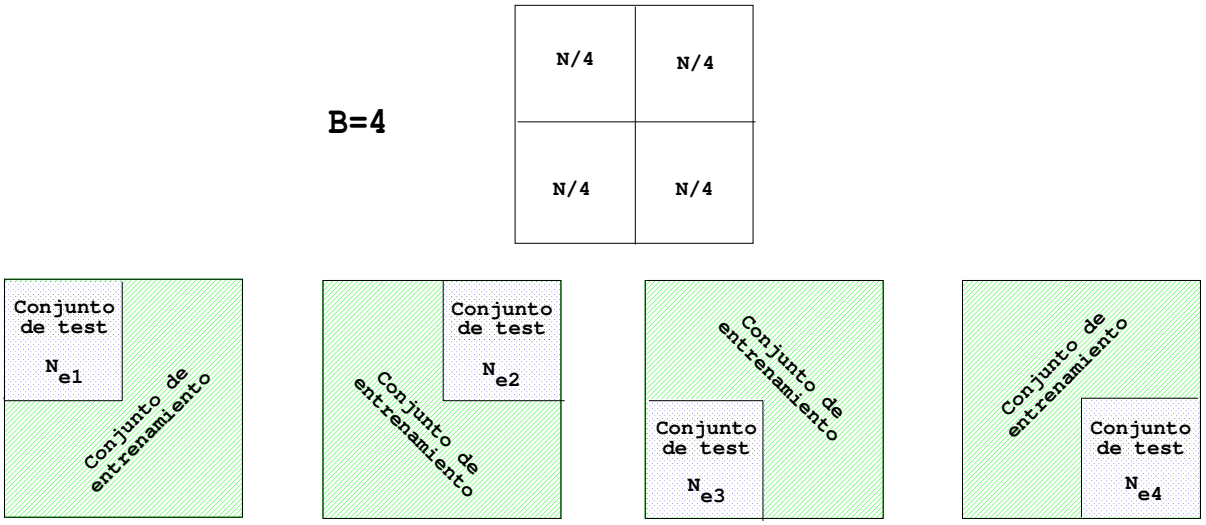
- *Resustitución (Resubstitution):* Todos los datos disponibles se utilizan tanto para para entrenamiento como para test. Inconveniente: es *(muy) optimista*
- *Partición (Hold Out):* Los datos se dividen en un subconjunto para entrenamiento y otro para test. Inconveniente: desaprovechamiento de datos
- *Validación Cruzada en B bloques (B-fold Cross Validation):* Los datos se dividen aleatoriamente en B bloques. Cada bloque se utiliza como test para un sistema entrenado con el resto de bloques. Inconvenientes: Reduce el número de datos de entrenamiento (sobre todo cuando B es pequeño) y el coste computacional se incrementa con B
- *Exclusión individual (Leaving One Out):* Cada dato individual se utiliza como dato único de test de un sistema entrenado con los $N-1$ datos restantes. Equivale a Validación Cruzada en N bloques. Inconveniente: máximo coste computacional

Resustitución y partición



- Resustitución. Error: $\frac{N_e}{N}$. Talla de entrenamiento: N
- Partición: Error: $\frac{N'_e}{N'}$. Talla de entrenamiento: $N - N'$

Validación cruzada



- Error: $\frac{N_{e1}+N_{e2}+N_{e3}+N_{e4}}{N}$
- Talla de entrenamiento efectiva: $\frac{3N}{4}$

Index

Aprendizaje deductivo e inductivo

- *Aprendizaje Deductivo* (o "por instrucción"):
  Se asume que existe un agente (humano) que posee el conocimiento necesario, el cual se transfiere de alguna forma al sistema.

  En el contexto de AA, no se considera que esto sea propiamente "aprendizaje", sino más bien se trataría de un modo de "enseñanza" en el que el sistema es "programado" para resolver cierta tarea.

- *Aprendizaje Inductivo* (o "a partir de ejemplos"):
  Es el planteamiento propio de AA.

  El sistema posee escaso conocimiento a-priori sobre la tarea a resolver y debe construir su(s) modelo(s) principalmente mediante la observación de *ejemplos* o *muestras de aprendizaje* de *entrada/salida* de dicha tarea.

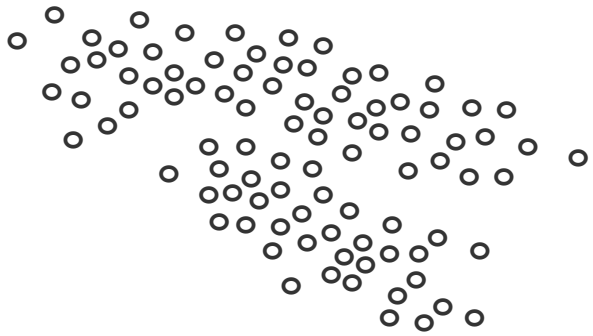
Aprendizaje supervisado y no supervisado

Aprendizaje supervisado: Información (completa) de *entrada* y *salida* en los datos de entrenamiento

Aprendizaje no supervisado:

- Los datos de entrenamiento solo contienen información de la *entrada* $x \in \mathcal{X}$
- El objetivo es obtener información sobre la estructura del dominio de *salida*, $\mathcal{Y}$
- En problemas de *clasificación*, esta información se refiere a la (posible) estructura en clases de los datos $x \in \mathcal{X}$. En este caso, el problema se conoce como *agrupamiento* o *"clustering"*

Ejemplo: Datos de entrenamiento en $\mathcal{X} = \mathbb{R}^2$:



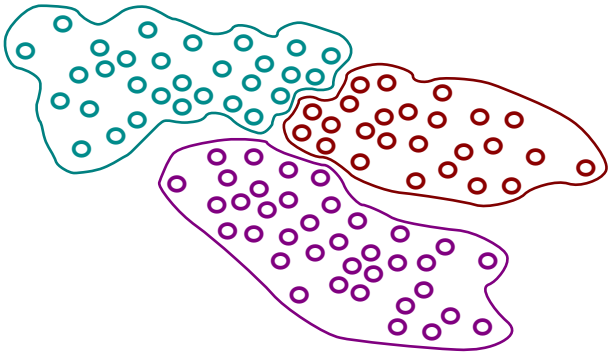
Aprendizaje supervisado y no supervisado

Aprendizaje supervisado: Información (completa) de *entrada* y *salida* en los datos de entrenamiento

Aprendizaje no supervisado:

- Los datos de entrenamiento solo contienen información de la *entrada* $x \in \mathcal{X}$
- El objetivo es obtener información sobre la estructura del dominio de *salida*, $\mathcal{Y}$
- En problemas de *clasificación*, esta información se refiere a la (posible) estructura en clases de los datos $x \in \mathcal{X}$. En este caso, el problema se conoce como *agrupamiento* o *"clustering"*

Ejemplo: Datos de entrenamiento en $\mathcal{X} = \mathbb{R}^2$, agrupados en tres clases:



Otros modos de aprendizaje automático

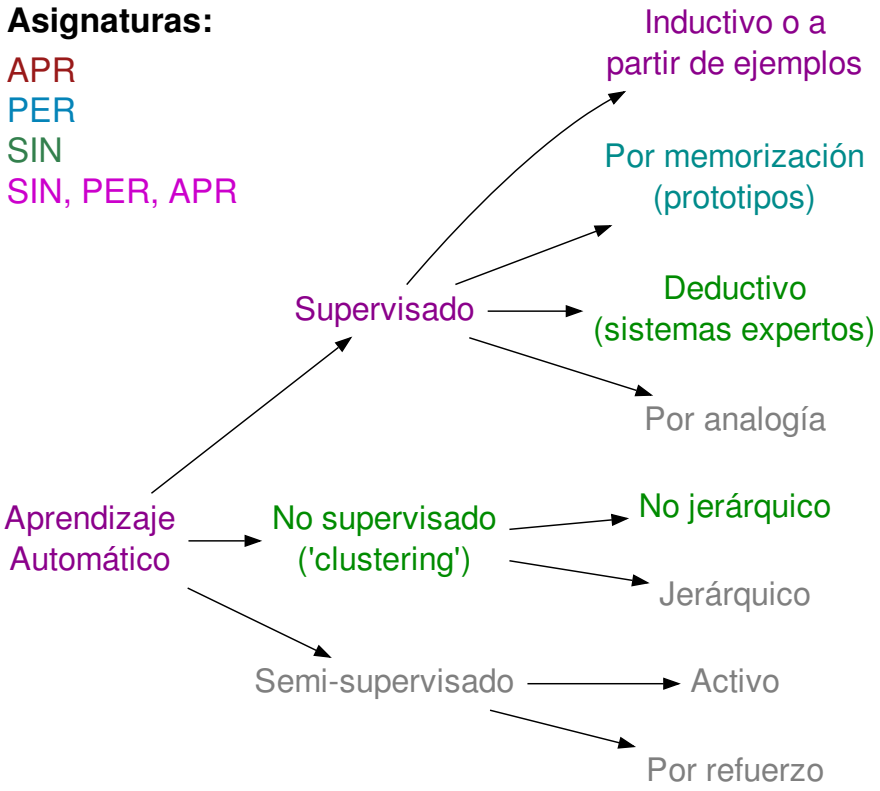
- *Aprendizaje "semi-supervisado" (ASS):* se refiere a planteamientos de AA situados entre el aprendizaje totalmente supervisado y totalmente no-supervisado
- *Aprendizaje adaptativo (AAD):* se parte de un modelo previo, cuyos parámetros se modifican ("adaptan") usando los (nuevos) datos de entrenamiento
- *Aprendizaje "on-line" (AOL):* no hay distinción explícita entre las fases de "entrenamiento" y "test"; el sistema aprende (posiblemente partiendo de cero) mediante el propio proceso de predicción, con supervisión humana.

  Para cada *entrada* $x \in \mathcal{X}$, la supervisión consiste en la validación o corrección de la salida $y = f(x) \in \mathcal{Y}$ predicha por el sistema
- *Aprendizaje activo (AAC):* no se dispone de la salida, y, de cada dato (x) de entrenamiento y el sistema escoje las muestras x más adecuadas para que un agente externo (humano) las etiquete con su y correcta
- *Aprendizaje por refuerzo (AR):* Puede considerarse como un caso de AOL y ASS en el que la supervisión es "incompleta"; típicamente una información (booleana) de *premio* o *castigo* con respecto a la salida predicha por el sistema

Una taxonomía de técnicas de AA

Asignaturas:

APR
PER
SIN
SIN, PER, APR



Index

Orígenes y evolución histórica del AA

Desde los años 40 del pasado siglo, se han venido desarrollando de forma más o menos paralela dos enfoques principales para la disciplina que modernamente se conoce como *sistemas inteligentes* (SI):

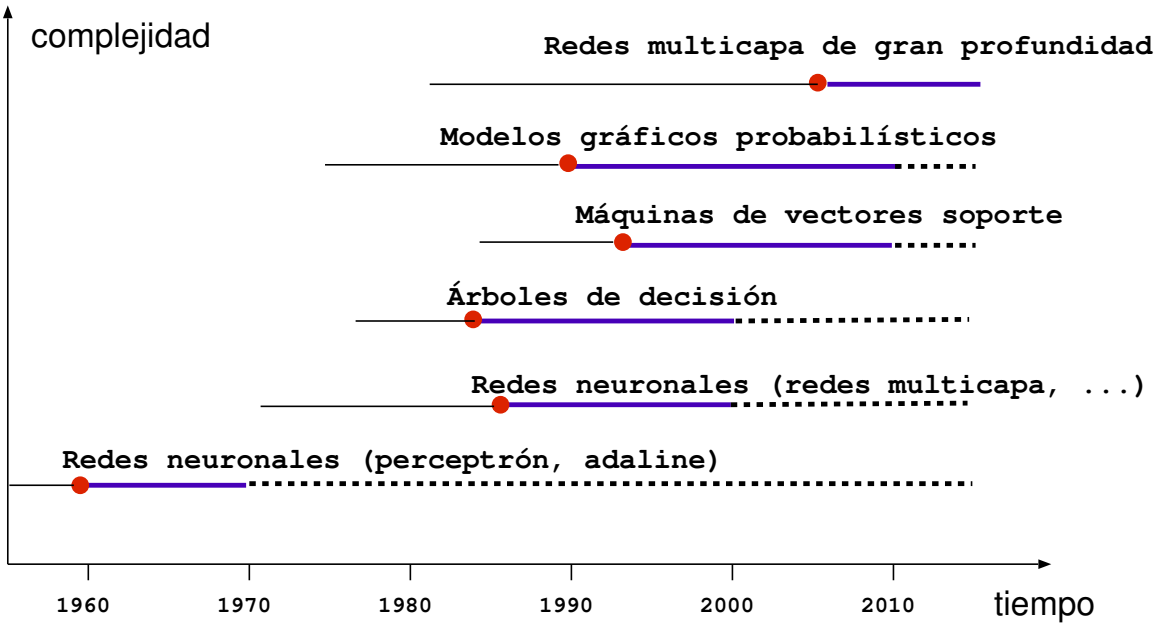
- *Inteligencia artificial* (propiamente dicha, o "clásica" – IA), que se ocupa principalmente de los aspectos mas cognitivos, con claras relaciones con la lógica, el conocimiento y su procesamiento

- *Reconocimiento de formas* (RF – también "reconocimiento de patrones" o, en inglés, "pattern recognition"), que se ocupa de aspectos más "perceptivos", relacionados con la visión, el habla, etc.

El *aprendizaje automático* surge en los años 80-90 como planteamiento integrador de los enfoques IA y RF, entre otros.

Grandes avances y espectaculares resultados prácticos en los últimos 20 años.

Evolución de algunas tecnologías importantes de AA



Para cada tecnología, la línea continua indica el periodo de desarrollo teórico-experimental y la de puntos el periodo de vigencia como tecnología consolidada.

Index

Áreas y aplicaciones

- *Reconocimiento de Imágenes*
  Reconocimiento de caracteres, de texto manuscrito, firmas, análisis de documentos, identificación de placas de matrícula y tipos de vehículos, piezas industriales, reconocimiento de texturas y detección de defectos para control de calidad, etc.
- *Visión Artificial y Robótica*
  Reconocimiento de rostros y expresiones faciales, visión para navegación de robots, vehiculos autónomos y exploración espacial, etc.
- *Teledetección (imágenes aéreas o de satélite)*
  Exploración de recursos naturales, predicción de cosechas y explotaciones forestales, localización de posibles yacimientos minerales, etc.
- *Análisis de Señales Sísmicas*
  Señales naturales: predicción de terremotos. Señales artificiales: localización de yacimientos minerales y petróleo, etc.
- *Reconocimiento del Habla y Procesado del Lenguaje*
  Reconocimiento de palabras aisladas, habla contínua, comprensión, traducción automática, etc.
- *Aplicaciones en Biomedecina y Genómica*
  Detección de tumores y tejidos cancerosos, análisis de electro cardio/encefalo–gramas, detección de situaciones críticas en UVI, diagnóstico a partir de síntomas, reconocimiento de cromosomas para detección de malformaciones congénitas, análisis de secuencias genómicas, etc.
- *Aplicaciones Biométricas*
  Reconocimiento de huellas dactilares, de rostros, iris, análisis de voz para identificación del locutor, etc.
- *Aplicaciones Agrícolas*
  Visión artificial para recolección automática, localización de “malas hierbas” para su eliminación selectiva, detección de puntos de injerto para su automatización, detección de defectos y selección de frutos para su envasado, etc.
- *Protección Civil*
  Predicción del clima, predicción de terremotos, control incendios forestales, detección de situaciones de alerta en sistemas hidrológicos, etc.
- *Economía*
  Segmentación de mercados, predicción de tendencias, detección de patrones de fraude, minería y analítica de datos, etc.
- *Ayudas Discapacitados*
  Ayudas para la visión, control del entorno mediante reconocimiento del habla, ayudas al aprendizaje del habla, etc.
- *Análisis de Datos (Analytics)*
  Datos masivos (big data analytics), web (web analytics), internet de las cosas (IoT analytics))

Index

Notación

- $\mathbb{R}$, $\mathbb{N}$ y $\mathbb{B}$; espacios de los reales, de los naturales y de los booleanos, respectivamente
- $\mathbb{R}^d$: espacio vectorial de d dimensiones
- Σ^*: espacio de cadenas de longitud finita de símbolos
- $\mathcal{X}, \mathcal{Y}$: espacios de datos de entrada y de salida, respectivamente
- x, y: un dato de entrada y un dato de salida, respectivamente
- $f, g : \mathcal{X} \to \mathcal{Y}$: funciones entre el espacio de entrada y el de salida
- C: número de clases en un problema de clasificación
- Γ: distribución o densidad de probabilidad gamma
- N: número de muestras
- $\mathcal{N}$: distribución o densidad de probabilidad gaussiana o normal

Conceptos básicos de Estadística y Probabilidad

Probabilidad, densidad $P(x),\ p(x):\ \sum_x P(x) = 1\ ,\quad \int_x p(x)d(x) = 1$

Probabilidad conjunta $P(x,y):\ \sum_x \sum_y P(x,y) = 1$

Probabilidad condicional $P(x \mid y):\ \sum_x P(x \mid y) = 1 \quad \forall y$

Marginales $P(x) = \sum_y P(x,y), \qquad P(y) = \sum_x P(x,y)$

Regla de la probabilidad conjunta $P(x,y) = P(x)\,P(y \mid x)$

Regla de la cadena $P(x_1, x_2, \ldots, x_N) = P(x_1) \prod_{i=2}^{N} P(x_i \mid x_1, \ldots, x_{i-1})$

Regla de Bayes $P(y \mid x)\ P(x) = P(y)\,P(x \mid y)$

Esperanza $E_P(f(x)) \equiv E_x(f(x)) = \sum_x f(x)\ P(x)$