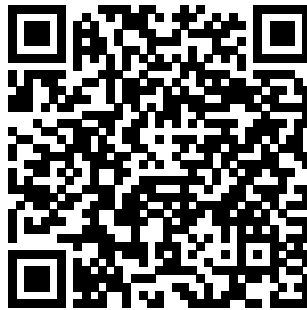


# El Diccionario de Aprendizaje Automático de **A'**alto

Alexander Jung, Konstantina Olioumtsevits, Juliette Gronier y Salvatore Rastelli

Traducido al español por Tommi Flores

July 16, 2025



por favor cite como: A. Jung, K. Olioumtsevits and J. Gronier,  
*The Aalto Dictionary of Machine Learning*. Espoo, Finland:  
Aalto University, 2025.

## Agradecimientos

Este diccionario de aprendizaje automático evolucionó a través del desarrollo y la enseñanza de varios cursos, incluyendo CS-E3210 Machine Learning: Basic Principles, CS-C3240 Machine Learning, CS-E4800 Artificial Intelligence, CS-EJ3211 Machine Learning with Python, CS-EJ3311 Deep Learning with Python, CS-E4740 Federated Learning, y CS-E407507 Human-Centered Machine Learning. Estos cursos se ofrecieron en Aalto University <https://www.aalto.fi/en>, a estudiantes adultos a travez de The Finnish Institute of Technology (FITech) <https://fitech.io/en/>, y a estudiantes internacionales a través de European University Alliance Unite! <https://www.aalto.fi/en/unite>.

Agradecemos a los estudiantes que brindaron valiosos comentarios que ayudaron a dar forma a este diccionario. Un agradecimiento especial a Mikko Seesto por su meticulosa corrección.

## Lists of Symbols

### Conjuntos y Funciones

$a \in \mathcal{A}$  El objeto  $a$  es un elemento del conjunto  $\mathcal{A}$ .

---

$a := b$  Utilizamos  $a$  como una abreviatura para  $b$ .

---

$|\mathcal{A}|$  La cardinalidad (es decir, el número de elementos) de un conjunto finito  $\mathcal{A}$ .

---

$\mathcal{A} \subseteq \mathcal{B}$   $\mathcal{A}$  es un subconjunto de  $\mathcal{B}$ .

---

$\mathcal{A} \subset \mathcal{B}$   $\mathcal{A}$  es un subconjunto propio de  $\mathcal{B}$ .

---

$\mathbb{N}$  Los números naturales  $1, 2, \dots$

---

$\mathbb{R}$  Los números reales  $x$  [1].

---

$\mathbb{R}_+$  Los números reales no negativos  $x \geq 0$ .

---

$\mathbb{R}_{++}$  Los números reales positivos  $x > 0$ .

$\{0, 1\}$	El conjunto que consta de los dos números reales 0 y 1.
$[0, 1]$	El intervalo cerrado de números reales $x$ con $0 \leq x \leq 1$ .
$\operatorname{argmin}_{\mathbf{w}} f(\mathbf{w})$	El conjunto de minimizadores para una función de valor real $f(\mathbf{w})$ .
$\mathbb{S}^{(n)}$	El conjunto de vectores de norma unitaria en $\mathbb{R}^{n+1}$ .
$\log a$	El logaritmo del número positivo $a \in \mathbb{R}_{++}$ .
$h(\cdot) : \mathcal{A} \rightarrow \mathcal{B} : a \mapsto h(a)$	Una función (aplicación) que acepta cualquier elemento $a \in \mathcal{A}$ de un conjunto $\mathcal{A}$ como entrada y entrega un elemento bien definido $h(a) \in \mathcal{B}$ de un conjunto $\mathcal{B}$ . El conjunto $\mathcal{A}$ es el dominio de la función $h$ y el conjunto $\mathcal{B}$ es el codominio de $h$ . El aprendizaje automático (aprendizaje automático) tiene como objetivo encontrar (o aprender) una función $h$ (es decir, una hipótesis) que lea las atributos $\mathbf{x}$ de un punto de datos y entregue una predicción $h(\mathbf{x})$ para su etiqueta $y$ .
$\nabla f(\mathbf{w})$	El gradiente de una función diferenciable de valor real $f : \mathbb{R}^d \rightarrow \mathbb{R}$ es el vector $\nabla f(\mathbf{w}) = \left( \frac{\partial f}{\partial w_1}, \dots, \frac{\partial f}{\partial w_d} \right)^T \in \mathbb{R}^d$ [2, Ch. 9].

## Matrices y Vectores

$\mathbf{x} = (x_1, \dots, x_d)^T$	Un vector de longitud $d$ , con su $j$ -ésima entrada siendo $x_j$ .
$\mathbb{R}^d$	El conjunto de vectores $\mathbf{x} = (x_1, \dots, x_d)^T$ que consiste en $d$ entradas de valor real $x_1, \dots, x_d \in \mathbb{R}$ .
$\mathbf{I}_{l \times d}$	Una matriz identidad generalizada con $l$ filas y $d$ columnas. Los elementos de $\mathbf{I}_{l \times d} \in \mathbb{R}^{l \times d}$ son iguales a 1 en la diagonal principal y iguales a 0 en los demás casos.
$\mathbf{I}_d, \mathbf{I}$	Una matriz identidad cuadrada de tamaño $d \times d$ . Si el tamaño es claro por el contexto, omitimos el subíndice.
$\ \mathbf{x}\ _2$	La norma euclidiana (o $\ell_2$ ) del vector $\mathbf{x} = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ definida como $\ \mathbf{x}\ _2 := \sqrt{\sum_{j=1}^d x_j^2}$ .
$\ \mathbf{x}\ $	Alguna norma del vector $\mathbf{x} \in \mathbb{R}^d$ [3]. A menos que se especifique lo contrario, nos referimos a la norma euclidiana $\ \mathbf{x}\ _2$ .
$\mathbf{x}^T$	La traspuesta de una matriz que tiene el vector $\mathbf{x} \in \mathbb{R}^d$ como su única columna.
$\mathbf{X}^T$	La traspuesta de una matriz $\mathbf{X} \in \mathbb{R}^{m \times d}$ . Una matriz cuadrada de valores reales $\mathbf{X} \in \mathbb{R}^{m \times m}$ se denomina simétrica si $\mathbf{X} = \mathbf{X}^T$ .
$\mathbf{0} = (0, \dots, 0)^T$	El vector en $\mathbb{R}^d$ con cada entrada igual a cero.
$\mathbf{1} = (1, \dots, 1)^T$	El vector en $\mathbb{R}^d$ con cada entrada igual a uno.

$(\mathbf{v}^T, \mathbf{w}^T)^T$	El vector de longitud $d + d'$ obtenido al concatenar las entradas del vector $\mathbf{v} \in \mathbb{R}^d$ con las entradas de $\mathbf{w} \in \mathbb{R}^{d'}$ .
$\text{span}\{\mathbf{B}\}$	El espacio generado (span) por una matriz $\mathbf{B} \in \mathbb{R}^{a \times b}$ , que es el subespacio de todas las combinaciones lineales de las columnas de $\mathbf{B}$ , $\text{span}\{\mathbf{B}\} = \{\mathbf{B}\mathbf{a} : \mathbf{a} \in \mathbb{R}^b\} \subseteq \mathbb{R}^a$ .
$\det(\mathbf{C})$	El determinante de la matriz $\mathbf{C}$ .
$\mathbf{A} \otimes \mathbf{B}$	El producto de Kronecker de $\mathbf{A}$ y $\mathbf{B}$ [4].

## Teoría de la Probabilidad

$\mathbb{E}_p\{f(\mathbf{z})\}$	La expectation de una función $f(\mathbf{z})$ de una variable aleatoria (RV) $\mathbf{z}$ cuya distribución de probabilidad es $\mathbb{P}(\mathbf{z})$ . Si la distribución de probabilidad es clara por el contexto, simplemente escribimos $\mathbb{E}\{f(\mathbf{z})\}$ .
$\mathbb{P}(\mathbf{x}, y)$	Una distribución de probabilidad conjunta de una RV cuyas realizaciones son punto de datos con atributos $\mathbf{x}$ y etiqueta $y$ .
$\mathbb{P}(\mathbf{x} y)$	Una distribución de probabilidad condicional de una RV $\mathbf{x}$ dado el valor de otra RV $y$ [5, Sec. 3.5].
$\mathbb{P}(\mathbf{x}; \mathbf{w})$	Una distribución de probabilidad parametrizada de una RV $\mathbf{x}$ . La distribución de probabilidad depende de un vector de parámetros $\mathbf{w}$ . Por ejemplo, $\mathbb{P}(\mathbf{x}; \mathbf{w})$ podría ser una distribución normal multivariante con el vector de parámetros $\mathbf{w}$ dado por las entradas del vector de media $\mathbb{E}\{\mathbf{x}\}$ y la matriz de covarianza $\mathbb{E}\left\{(\mathbf{x} - \mathbb{E}\{\mathbf{x}\})(\mathbf{x} - \mathbb{E}\{\mathbf{x}\})^T\right\}$ .
$\mathcal{N}(\mu, \sigma^2)$	La distribución de probabilidad de una variable aleatoria gaussiana (VA gaussiana) $x \in \mathbb{R}$ con media (o expectation) $\mu = \mathbb{E}\{x\}$ y varianza $\sigma^2 = \mathbb{E}\{(x - \mu)^2\}$ .
$\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$	La distribución normal multivariante de un vector de valores VA gaussiana $\mathbf{x} \in \mathbb{R}^d$ con media (o expectation) $\boldsymbol{\mu} = \mathbb{E}\{\mathbf{x}\}$ y matriz de covarianza $\mathbf{C} = \mathbb{E}\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\}$ .

## Aprendizaje Automático

$r$	Un índice $r = 1, 2, \dots$ que enumeran los punto de datoss.
$m$	El número de punto de datoss en (es decir, el tamaño de) un conjunto de datos.
$\mathcal{D}$	Un conjunto de datos $\mathcal{D} = \{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}\}$ es una lista de punto de datoss individuales $\mathbf{z}^{(r)}$ , para $r = 1, \dots, m$ .
$d$	El número de atributos que caracterizan un punto de datos.
$x_j$	La $j$ -ésima característica (feature) de un punto de datos. La primera atributo se denota como $x_1$ , la segunda como $x_2$ , y así sucesivamente.
$\mathbf{x}$	El vector de atributos (feature vector) $\mathbf{x} = (x_1, \dots, x_d)^T$ de un punto de datos cuyas entradas son los atributos individuales de un punto de datos.
$\mathcal{X}$	El espacio de atributos $\mathcal{X}$ es el conjunto de todos los valores posibles que las atributos $\mathbf{x}$ de un punto de datos pueden tomar.
$\mathbf{z}$	En lugar del símbolo $\mathbf{x}$ , a veces usamos $\mathbf{z}$ como otro símbolo para denotar un vector cuyas entradas son las atributos individuales de un punto de datos. Necesitamos dos símbolos diferentes para distinguir entre atributos (features) crudas y atributos aprendidos [6, Ch. 9].
$\mathbf{x}^{(r)}$	El vector de atributos de el $r$ -ésimo punto de datos dentro de un conjunto de datos. <sup>8</sup>
$x_j^{(r)}$	La $j$ -ésima atributo del $r$ -ésimo punto de datos dentro de un conjunto de datos.



$\mathcal{B}$	Un mini-lote (o subconjunto) de punto de datos seleccionados aleatoriamente.
$B$	El tamaño de (es decir, el número de punto de datos en) un mini-lote.
$y$	La etiqueta (o cantidad de interés) de un punto de datos.
$y^{(r)}$	La etiqueta del $r$ -ésimo punto de datos.
$(\mathbf{x}^{(r)}, y^{(r)})$	Las atributos y la etiqueta del $r$ -ésimo punto de datos.
$\mathcal{Y}$	El espacio de etiquetas $\mathcal{Y}$ de un método de ML consiste en todos los valores potenciales de etiqueta que un punto de datos puede tener. El espacio de etiquetas nominal puede ser más grande que el conjunto de diferentes valores de etiqueta que surgen en un conjunto de datos dado (por ejemplo, un conjunto de entrenamiento). Los problemas (o métodos) de ML que utilizan un espacio de etiquetas numérico, como $\mathcal{Y} = \mathbb{R}$ o $\mathcal{Y} = \mathbb{R}^3$ , se conocen como problemas de regresión. Los problemas (o métodos) de ML que utilizan un espacio de etiquetas discreto, como $\mathcal{Y} = \{0, 1\}$ o $\mathcal{Y} = \{gato, perro, ratón\}$ , se conocen como problemas de clasificación.
$\eta$	La tasa de aprendizaje (o tamaño de paso) utilizada por los métodos de métodos de gradiente.
$h(\cdot)$	Un mapa de hipótesis que toma como entrada las atributos $\mathbf{x}$ de un punto de datos y entrega una predicción $\hat{y} = h(\mathbf{x})$ para su etiqueta $y$ .
$\mathcal{Y}^{\mathcal{X}}$	Dado dos conjuntos $\mathcal{X}$ y $\mathcal{Y}$ , denotamos por $\mathcal{Y}^{\mathcal{X}}$ el conjunto de todos los posibles mapas de hipótesis $h : \mathcal{X} \rightarrow \mathcal{Y}$ .

$\mathcal{H}$	Un espacio de hipótesis o modelo utilizado por un método de ML. El espacio de hipótesis consiste en diferentes mapas de hipótesis $h : \mathcal{X} \rightarrow \mathcal{Y}$ , entre los cuales el método de ML debe elegir.
$d_{\text{eff}}(\mathcal{H})$	La dimensión efectiva de un espacio de hipótesis $\mathcal{H}$ .
$B^2$	El sesgo cuadrado de una hipótesis aprendida $\hat{h}$ producida por un método de ML. El método se entrena con punto de datos que se modelan como las realizaciones de RVs. Dado que los datos son una realización de RVs, la hipótesis aprendida $\hat{h}$ también es una realización de una RV.
$V$	La varianza de los (parámetros de la) hipótesis aprendida producida por un método de ML. El método se entrena con punto de datos que se modelan como las realizaciones de RVs. Dado que los datos son una realización de RVs, la hipótesis aprendida $\hat{h}$ también es una realización de una RV.
$L((\mathbf{x}, y), h)$	La pérdida incurrida al predecir la etiqueta $y$ de un punto de datos utilizando la predicción $\hat{y} = h(\mathbf{x})$ . La predicción $\hat{y}$ se obtiene al evaluar la hipótesis $h \in \mathcal{H}$ para la vector de atributos $\mathbf{x}$ del punto de datos.
$E_v$	El error de validación de una hipótesis $h$ , que es su pérdida promedio incurrida en un conjunto de validación.
$\hat{L}(h \mathcal{D})$	El riesgo empírico o pérdida promedio incurrido por la hipótesis $h$ en un conjunto de datos $\mathcal{D}$ .

$E_t$	El error de entrenamiento de una hipótesis $h$ , que es su pérdida promedio incurrido en un conjunto de entrenamiento.
$t$	Un índice de tiempo discreto $t = 0, 1, \dots$ utilizado para enumerar eventos secuenciales (o instantes de tiempo).
$t$	Un índice que enumera las tarea de aprendizajes dentro de un problema de aprendizaje multitarea.
$\alpha$	Un parámetro de regularización que controla la cantidad de regularización.
$\lambda_j(\mathbf{Q})$	El $j$ -ésimo valor propio (eigenvalue) (ordenado en forma ascendente o descendente) de una matriz semi-definida positiva (psd) $\mathbf{Q}$ . También usamos la abreviatura $\lambda_j$ si la matriz correspondiente es clara por el contexto.
$\sigma(\cdot)$	La función de activación utilizada por una neurona artificial dentro de una red neuronal artificial (RNA).
$\mathcal{R}_{\hat{y}}$	Una región de decisión dentro de un espacio de atributos.
$\mathbf{w}$	Un vector de parámetros $\mathbf{w} = (w_1, \dots, w_d)^T$ de un modelo, por ejemplo, los pesos de un modelo lineal o en una ANN.
$h^{(\mathbf{w})}(\cdot)$	Un mapa de hipótesis que involucra parámetros del modelo ajustables $w_1, \dots, w_d$ apilados en el vector $\mathbf{w} = (w_1, \dots, w_d)^T$ .
$\phi(\cdot)$	Un mapa de atributos $\phi : \mathcal{X} \rightarrow \mathcal{X}' : \mathbf{x} \mapsto \mathbf{x}' := \phi(\mathbf{x}) \in \mathcal{X}'$ .
$K(\cdot, \cdot)$	Dado un espacio de atributos $\mathcal{X}$ , un kernel es un mapa $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ que es psd.

## Aprendizaje Federado

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Un grafo no dirigido cuyos nodos $i \in \mathcal{V}$ representan dispositivos dentro de un red de aprendizaje federado (red FL). Los bordes ponderados no dirigidos $\mathcal{E}$ representan conectividad entre dispositivos y similitudes estadísticas entre sus conjunto de datoss y tarea de aprendizajes.
$i \in \mathcal{V}$	Un nodo que representa un dispositivo dentro de un red FL. El dispositivo puede acceder a un conjunto de datos local y entrenar un modelo local.
$\mathcal{G}^{(\mathcal{C})}$	El subgrafo inducido de $\mathcal{G}$ utilizando los nodos en $\mathcal{C} \subseteq \mathcal{V}$ .
$\mathbf{L}^{(\mathcal{G})}$	La matriz laplaciana de un grafo $\mathcal{G}$ .
$\mathbf{L}^{(\mathcal{C})}$	La matriz laplaciana del grafo inducido $\mathcal{G}^{(\mathcal{C})}$ .
$\mathcal{N}^{(i)}$	La entorno de un nodo $i$ en un grafo $\mathcal{G}$ .
$d^{(i)}$	El grado ponderado $d^{(i)} := \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'}$ de un nodo $i$ en un grafo $\mathcal{G}$ .
$d_{\max}^{(\mathcal{G})}$	El máximo grado ponderado de nodo en un grafo $\mathcal{G}$ .
$\mathcal{D}^{(i)}$	El conjunto de datos local $\mathcal{D}^{(i)}$ que posee el nodo $i \in \mathcal{V}$ de un red FL.
$m_i$	El número de punto de datoss (es decir, el tamaño de la muestra) contenidos en el conjunto de datos local $\mathcal{D}^{(i)}$ en el nodo $i \in \mathcal{V}$ .

$\mathbf{x}^{(i,r)}$	Las atributos del $r$ -ésimo punto de datos en el conjunto de datos local $\mathcal{D}^{(i)}$ .
$y^{(i,r)}$	La etiqueta del $r$ -ésimo punto de datos en el conjunto de datos local $\mathcal{D}^{(i)}$ .
$\mathbf{w}^{(i)}$	Los parámetros del modelo locales del dispositivo $i$ dentro de un red FL.
$L_i(\mathbf{w})$	La función de pérdida local utilizada por el dispositivo $i$ para medir la utilidad de alguna elección $\mathbf{w}$ para los parámetros del modelo locales.
$L^{(d)}(\mathbf{x}, h(\mathbf{x}), h'(\mathbf{x}))$	La pérdida incurrida por una hipótesis $h'$ en un punto de datos con atributos $\mathbf{x}$ y etiqueta $h(\mathbf{x})$ obtenida de otra hipótesis.
$\text{stack}\{\mathbf{w}^{(i)}\}_{i=1}^n$	El vector $\left( (\mathbf{w}^{(1)})^T, \dots, (\mathbf{w}^{(n)})^T \right)^T \in \mathbb{R}^{dn}$ que se obtiene apilando verticalmente los parámetros del modelo locales $\mathbf{w}^{(i)} \in \mathbb{R}^d$ .

## Machine Learning Concepts

**$k$ -means** El algoritmo  $k$ -means es un método de agrupamiento rígido que asigna cada punto de datos de un conjunto de datos a precisamente uno de  $k$  clusters diferentes. El método alterna entre actualizar las asignaciones de clúster (asignando al clúster con el medio mas cercano) y, dado estas asignaciones actualizadas recalculan los medios de los clúster [6, Ch. 8].

Vea también: media, algoritmo, agrupamiento rígido, punto de datos, conjunto de datos, cluster.

**agrupamiento** Los métodos de agrupamiento descomponen un conjunto dado de puntos de datos en pocos subconjuntos, que se denominan clústers. Cada clúster está compuesto por puntos de datos que son más similares entre sí que con respecto a los puntos de datos fuera del clúster. Los diferentes métodos de agrupamiento utilizan distintas medidas de similitud entre puntos de datos y diferentes formas de representación de los clústers. El método de agrupamiento  $k$ -means utiliza el vector de atributo promedio (media del clúster) de un clúster como su representante. Un método popular de agrupamiento suave basado en el modelo de mezcla gaussiana representa un clúster mediante una distribución normal multivariante.

Vea también: punto de datos, cluster,  $k$ -means, atributo, media, agrupamiento suave, modelo de mezcla gaussiana (GMM), distribución

normal multivariante.

### **agrupamiento basado en densidad para aplicaciones espaciales con ruido (DBSCAN)**

DBSCAN es un algoritmo de agrupamiento para puntos de datos caracterizados por vectores de atributos numéricos. Al igual que  $k$ -means y agrupamiento suave mediante modelo de mezcla gaussiana, DBSCAN utiliza las distancias euclidianas entre los vectores de atributos para determinar los clusters. Sin embargo, a diferencia de  $k$ -means y modelo de mezcla gaussiana, DBSCAN emplea una noción diferente de similitud entre puntos de datos. DBSCAN considera que dos puntos de datos son similares si están conectados a través de una secuencia (camino) de puntos de datos intermedios cercanos. Por lo tanto, DBSCAN puede considerar que dos puntos de datos son similares (y, por lo tanto, pertenecen al mismo cluster) incluso si sus vectores de atributos tienen una gran distancia euclidiana.

Vea también: agrupamiento, algoritmo, punto de datos, vector de atributos,  $k$ -means, agrupamiento suave, GMM, cluster.

**agrupamiento basado en flujo** El agrupamiento basado en flujo agrupa los nodos de un grafo no dirigido aplicando el algoritmo de  $k$ -means sobre vectores de atributos específicos para cada nodo. Estos vectores de atributos construyen a partir de flujos de red entre nodos fuente y destino seleccionados cuidadosamente [7].

Vea también: agrupamiento, grafo,  $k$ -means, vector de atributos

**agrupamiento convexo** Considere un conjunto de datos  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)} \in \mathbb{R}^d$ . El agrupamiento convexo aprende vectores  $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(m)}$  mini-

mizando

$$\sum_{r=1}^m \|\mathbf{x}^{(r)} - \mathbf{w}^{(r)}\|_2^2 + \alpha \sum_{i,i' \in \mathcal{V}} \|\mathbf{w}^{(i)} - \mathbf{w}^{(i')}\|_p.$$

Aquí,  $\|\mathbf{u}\|_p := (\sum_{j=1}^d |u_j|^p)^{1/p}$  denota la norma- $p$  (para  $p \geq 1$ ). Resulta que muchos de los vectores óptimos  $\hat{\mathbf{w}}^{(1)}, \dots, \hat{\mathbf{w}}^{(m)}$  coinciden. Un clúster consiste entonces en aquellos puntos de datos  $r \in \{1, \dots, m\}$  con valores idénticos de  $\hat{\mathbf{w}}^{(r)}$  [8, 9].

Vea también: conjunto de datos, convexo, agrupamiento, norma, cluster, punto de datos.

**agrupamiento en grafos** El agrupamiento en grafos tiene como objetivo agrupar puntos de datos que están representados como nodos de un grafo  $\mathcal{G}$ . Las aristas del  $\mathcal{G}$  representan similitudes por pares entre los puntos de datos. En algunos casos, es posible cuantificar el grado de estas similitudes mediante un peso de arista [7, 10].

Vea también: agrupamiento, grafo, punto de datos, peso de arista

**agrupamiento espectral** El agrupamiento espectral es una instancia particular del agrupamiento en grafos, es decir, agrupa puntos de datos representados como los nodos  $i = 1, \dots, n$  de un grafo  $\mathcal{G}$ . El agrupamiento espectral utiliza los vectores propios de la matriz laplaciana  $\mathbf{L}^{(\mathcal{G})}$  para construir vectores de atributos  $\mathbf{x}^{(i)} \in \mathbb{R}^d$  para cada nodo (es decir, para cada punto de datos)  $i = 1, \dots, n$ . Podemos utilizar estos vectores de atributos como entrada para métodos de agrupamiento en el espacio euclidiano, como k-means o agrupamiento suave mediante modelo de mezcla gaussiana. Mas o menos, los vectores de atributos de los nodos que pertenecen a un subconjunto bien conectado (o agru-



pamiento) de nodos en  $\mathcal{G}$  están ubicados cerca en el espacio euclidiano  $\mathbb{R}^d$  (Vea la Figura 1).

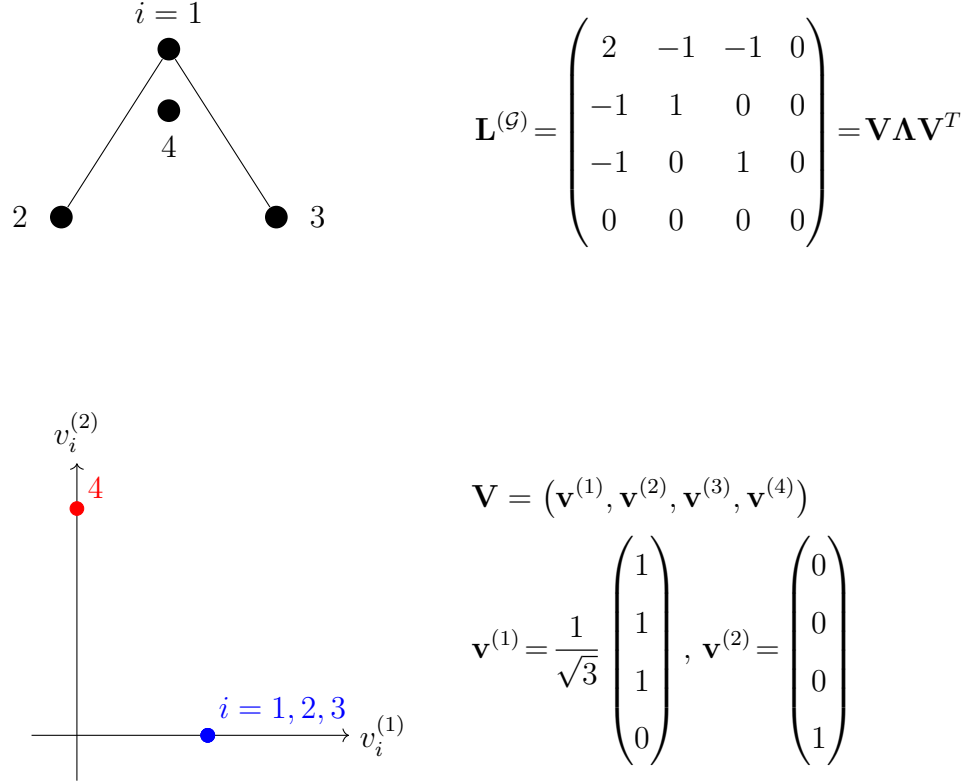


Figure 1: **Arriba.** Izquierda: Un grafo no dirigido  $\mathcal{G}$  con cuatro nodos  $i = 1, 2, 3, 4$ , donde cada nodo representa un punto de datos. Derecha: La matriz laplaciana  $\mathbf{L}^{(\mathcal{G})} \in \mathbb{R}^{4 \times 4}$  y su descomposición en valores propios **Abajo.** Izquierda: Un diagrama de dispersión de los puntos de datos usando los vectores de atributos  $\mathbf{x}^{(i)} = (v_i^{(1)}, v_i^{(2)})^T$ . Derecha: Dos vectores propios  $\mathbf{v}^{(1)}, \mathbf{v}^{(2)} \in \mathbb{R}^d$  correspondientes al valor propio  $\lambda = 0$  de la matriz laplaciana  $\mathbf{L}^{(\mathcal{G})}$ .

.  
Vea también: agrupamiento, agrupamiento en grafos, punto de datos, grafo, vector propio, matriz laplaciana, vector de atributos, espacio euclidiano,  $k$ -means, agrupamiento suave, GMM, descomposición en valores propios (EVD), diagrama de dispersión, valor propio,

**agrupamiento rígido** El agrupamiento rígido se refiere a la tarea de dividir un conjunto dado de puntos de datos en (pocos) clusters no superpuestos. El método de agrupamiento rígido más utilizado es  $k$ -means. Vea también: agrupamiento, punto de datos, cluster,  $k$ -means.

**agrupamiento suave** El agrupamiento suave se refiere a la tarea de dividir un conjunto dado de puntos de datos en (pocos) clusters superpuestos. Cada punto de datos se asigna a varios clusters diferentes con diversos grados de pertenencia. Los métodos de agrupamiento suave determinan el grado de pertenencia (o asignación suave a un clúster) para cada punto de datos y cada clúster. Un enfoque fundamentado para el agrupamiento suave consiste en interpretar los puntos de datos como realizaciones iid de un modelo de mezcla gaussiana. De este modo, se obtiene una elección natural para el grado de pertenencia como la probabilidad condicional de que un punto de datos pertenezca a un componente específico de la mezcla.

Vea también: agrupamiento, punto de datos, cluster, grado de pertenencia, independientes e idénticamente distribuidos (i.i.d.), realización, GMM, probabilidad.

**algoritmo** Un algoritmo es una especificación precisa y paso a paso de cómo

producir una salida a partir de una entrada dada en un número finito de pasos computacionales [11]. Por ejemplo, un algoritmo para entrenar un modelo lineal describe explícitamente cómo transformar un conjunto de entrenamiento dado en parámetros del modelo a través de una secuencia de pasos de gradiente. Esta caracterización informal puede formalizarse rigurosamente mediante diferentes modelos matemáticos [12]. Un modelo simple de un algoritmo es una colección de ejecuciones posibles. Cada ejecución es una secuencia:

$$\text{input}, s_1, s_2, \dots, s_T, \text{output}$$

que respeta las restricciones inherentes al ordenador que ejecuta el algoritmo. Los algoritmos pueden ser deterministas, donde cada entrada resulta únicamente en una sola ejecución, o aleatorios, donde las ejecuciones pueden variar probabilísticamente. Los algoritmos aleatorios pueden analizarse modelando las secuencias de ejecución como resultados de experimentos aleatorios, considerando el algoritmo como un proceso estocástico [5, 13, 14]. Crucialmente, un algoritmo abarca más que solo un mapeo de entrada a salida; también incluye los pasos computacionales intermedios  $s_1, \dots, s_T$ .

Vea también: modelo lineal, conjunto de entrenamiento, parámetros del modelo, paso de gradiente, modelo

**algoritmo distribuido** Un algoritmo distribuido distributed es un algoritmo diseñado para un tipo especial de computadora: una colección de dispositivos de cómputo interconectados (o nodos). Estos dispositivos se comunican y coordinan sus cálculos locales intercambiando mensajes

a través de una red [15, 16]. A diferencia de un algoritmo clásico, que se ejecuta en un solo dispositivo, un algoritmo distribuido se ejecuta de forma concurrente en múltiples dispositivos con capacidades de cómputo. Cada ejecución involucra tanto cálculos locales como eventos de intercambio de mensajes. Una ejecución genérica podría verse así:

$$\begin{aligned} \text{Node 1: } & \text{input}_1, s_1^{(1)}, s_2^{(1)}, \dots, s_{T_1}^{(1)}, \text{output}_1; \\ \text{Node 2: } & \text{input}_2, s_1^{(2)}, s_2^{(2)}, \dots, s_{T_2}^{(2)}, \text{output}_2; \\ & \vdots \\ \text{Node N: } & \text{input}_N, s_1^{(N)}, s_2^{(N)}, \dots, s_{T_N}^{(N)}, \text{output}_N. \end{aligned}$$

Cada dispositivo  $i$  inicia con su entrada local y ejecuta una secuencia de cálculos intermedios  $s_k^{(i)}$  en instantes de tiempo discretos  $k = 1, \dots, T_i$ . Estos cálculos pueden depender tanto de cálculos previos locales como de mensajes recibidos de otros dispositivos. Uno de los usos clave de los algoritmos distribuidos es en aprendizaje federado, donde una red de dispositivos colaboran para entrenar un modelo personalizado por dispositivo..

Vea también: algoritmo, dispositivo, aprendizaje federado (FL)

**algoritmo en línea** Un algoritmo en línea es un algoritmo que procesa datos de forma incremental, recibiendo elementos de datos uno por uno y tomando decisiones o generando salidas inmediatamente, sin tener acceso a toda la entrada desde el inicio [17, 18]. A diferencia de un algoritmo fuera de línea, que dispone de toda la entrada desde el comienzo, un algoritmo en línea debe lidiar con la incertidumbre del futuro y no puede cambiar decisiones pasadas. Puede modelarse como

una ejecución del tipo:

$$\text{init}, s_1, \text{out}_1, \text{in}_2, s_2, \text{out}_2, \dots, \text{in}_T, s_T, \text{out}_T.$$

Cada ejecución comienza en un estado inicial y alterna entre cálculos, salidas y nuevas entradas. Un ejemplo importante en aprendizaje automático es el descenso por gradiente en línea, que actualiza los modelos de parametro conforme llegan nuevos puntos de datos.

Vea también: algoritmo, datos, ML, descenso por gradiente en línea (online GD), parámetros del modelo, punto de datos

**algoritmo estocástico** Un algoritmo estocástico utiliza un mecanismo aleatorio durante su ejecución. Por ejemplo, descenso por gradiente estocástico utiliza un subconjunto seleccionado aleatoriamente de puntos de datos para calcular una aproximación del gradiente de una funcion objetivo. Podemos representar un algoritmo estocástico mediante un proceso estocástico, es decir, la secuencia de ejecución posible es el conjunto de resultados posibles de un experimento aleatorio [5], [13], [14].

Véa también: estocástico, algoritmo, descenso de gradiente estocástico (SGD), punto de datos, gradiente, función objetivo, método de optimización, métodos de gradiente.

**análisis de componentes principales (PCA)** El PCA determina una mapa de atributos lineal tales que las nuevos atributos permiten reconstruir los atributos originales con el minimo error de reconstrucción [6].

Vea también: mapa de atributos, atributo, mínimo.

**análisis de componentes principales probabilístico (PPCA)** El análisis de componentes principales probabilístico (PPCA) extiende el pca

básico mediante el uso de un modelo de probabilidad para los puntos de datos. El modelo de probabilidad de PPCA reduce la tarea de reducción de dimensionalidad a un problema de estimación que se puede resolver utilizando métodos de esperanza-maximización.

Vea también: análisis de componentes principales (PCA), modelo probabilístico, punto de datos, EM.

**aplicación** Utilizamos el término aplicación como sinónimo de función.

Véase también: función.

**aplicación lineal** Una aplicación lineal  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  es una función que satisface las propiedades de aditividad:  $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$ , y homogeneidad:  $f(c\mathbf{x}) = cf(\mathbf{x})$  para todos los vectores  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  y escalares  $c \in \mathbb{R}$ . En particular,  $f(\mathbf{0}) = \mathbf{0}$ . Toda aplicación lineal puede representarse como una multiplicación matricial  $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$  para alguna matriz  $\mathbf{A} \in \mathbb{R}^{m \times n}$ . El conjunto de aplicaciones lineales con valores reales para una dimensión  $n$  dada constituye un modelo lineal, el cual se utiliza en muchos métodos de aprendizaje automático.

Vea también: modelo lineal, regresión lineal, PCA, vector de atributos.

**aprendizaje automático (ML)** El aprendizaje automático tiene como objetivo predecir una etiqueta a partir de los atributos de un punto de datos. Los métodos de ML logran esto aprendiendo una hipótesis de un espacio de hipótesis (o modelo) mediante la minimización de una función de pérdida [6, 19]. Una formulación precisa de este principio es la minimización empírica del riesgo (ERM). Diferentes métodos de ML se obtienen de distintas elecciones de diseño para los puntos de datos

(sus atributos y etiquetas), el modelo, y la función de pérdida [6, Cap. 3].

Vea también: etiqueta, atributos, punto de datos, hipótesis, espacio de hipótesis, modelo, función de pérdida, minimización empírica del riesgo (ERM)

**aprendizaje automático explicable (explainable ML)** El aprendizaje automático explicable (explainable ML) consiste en métodos de aprendizaje automático que tienen como objetivo complementar cada predicción con una explicación explícita de cómo se ha obtenido dicha predicción. La construcción de una explicación explícita puede no ser necesaria si el método de aprendizaje automático utiliza un modelo lo suficientemente simple (o interpretable) [20].

Vea también: predicción, explicación, ML, modelo.

**aprendizaje de atributos** Consideremos una aplicación de aprendizaje automático con puntos de datos caracterizados por atributos crudos  $\mathbf{x} \in \mathcal{X}$ . El aprendizaje de atributos se refiere a la tarea de aprender un mapeo

$$\Phi : \mathcal{X} \rightarrow \mathcal{X}' : \mathbf{x} \mapsto \mathbf{x}',$$

que recibe como entrada el atributo  $\mathbf{x} \in \mathcal{X}$  de un punto de datos y entrega nuevos atributos  $\mathbf{x}' \in \mathcal{X}'$  de un nuevo espacio de atributos  $\mathcal{X}'$ . Se obtienen diferentes métodos de aprendizaje de atributos a partir de diferentes elecciones de  $\mathcal{X}, \mathcal{X}'$ , de un espacio de hipótesis  $\mathcal{H}$  de posibles mapeos  $\Phi$ , y de una medida cuantitativa de la utilidad de un mapeo específico  $\Phi \in \mathcal{H}$ . Por ejemplo, pca utiliza  $\mathcal{X} := \mathbb{R}^d$ ,  $\mathcal{X}' := \mathbb{R}^{d'}$  con

$d' < d$ , y un espacio de hipótesis

$$\mathcal{H} := \{ \Phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'} : \mathbf{x}' := \mathbf{F}\mathbf{x} \text{ con alguna } \mathbf{F} \in \mathbb{R}^{d' \times d} \}.$$

PCA mide la utilidad de un mapeo específico  $\Phi(\mathbf{x}) = \mathbf{F}\mathbf{x}$  por el mínimo error de reconstrucción lineal incurrido sobre un conjunto de datos,

$$\min_{\mathbf{G} \in \mathbb{R}^{d' \times d}} \sum_{r=1}^m \left\| \mathbf{G}\mathbf{F}\mathbf{x}^{(r)} - \mathbf{x}^{(r)} \right\|_2^2.$$

Vea también: ML, punto de datos, atributo, espacio de atributos, PCA, mínimo

**aprendizaje en línea** Algunos métodos de aprendizaje automático están diseñados para procesar datos de forma secuencial, actualizando sus parámetros del modelo a medida que nuevos puntos de datos se vuelven disponibles—uno a la vez. Un ejemplo típico son los datos de series temporales, como las temperaturas mínimas y máximas diarias registradas por una estación meteorológica de FMI. Estos valores forman una secuencia cronológica de observaciones. En el aprendizaje en línea, la hipótesis (o sus parámetros del modelo) se refina incrementalmente con cada nuevo punto de datos observado, sin volver a visitar los datos anteriores.

Vea también: ML, datos, parámetros del modelo, punto de datos, Instituto Meteorológico de Finlandia (FMI), hipótesis, GD en línea, algoritmo en línea.

**aprendizaje federado agrupado (CFL)** El aprendizaje federado agrupado asume que los conjuntos de datos locales están naturalmente organi-



zados en clusteres. Los conjuntos de datos locales que pertenecen al mismo cluster comparten propiedades estadísticas similares. El aprendizaje federado agrupado agrega los conjuntos de datos locales dentro del mismo cluster para formar un conjunto de entrenamiento para el entrenamiento de un modelo específico del cluster. La Minimización de variación total generalizada facilita esta agrupación de manera implícita al imponer una similitud aproximada de los parámetros del modelo a través de subconjuntos bien conectados del red de aprendizaje federado. Vea también: modelo local, dispositivo, FL, suposición de agrupamiento, red FL, cluster, conjunto de datos local, conjunto de entrenamiento, modelo, minimización de variación total generalizada (GTVMin), parámetros del modelo.

**aprendizaje federado en red (NFL)** El aprendizaje federado en red (NFL) se refiere a métodos que aprenden modelos personalizados de manera distribuida. Estos métodos aprenden a partir de conjuntos de datos locales que están relacionados por una estructura de red intrínseca. Vea también: modelo, conjunto de datos local

**aprendizaje federado horizontal (horizontal FL)** El aprendizaje federado horizontal utiliza conjunto de datos locales constituidos por diferentes puntos de datos, pero emplea los mismos atributos para caracterizarlos [21]. Por ejemplo, la predicción meteorológica utiliza una red de estaciones meteorológicas (observación) distribuidas espacialmente. Cada estación mide las mismas cantidades, como la temperatura diaria, la presión atmosférica y la precipitación. Sin embargo, distintas

estaciones miden las características o atributos de diferentes regiones espaciotemporales. Cada región espaciotemporal representa un punto de datos individual, caracterizado por los los mismos atributos (por ejemplo, temperatura diaria o presión atmosférica).

Vea también: conjunto de datos local, punto de datos, atributos

**aprendizaje federado vertical (VFL)** El aprendizaje federado vertical (VFL) se refiere a aplicaciones de aprendizaje federado en las que los distintos dispositivos tienen acceso a diferentes atributos del mismo conjunto de puntos de datos [22]. Formalmente, el conjunto de datos global subyacente es

$$\mathcal{D}^{(\text{global})} := \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(m)}, y^{(m)})\}.$$

Denotamos por  $\mathbf{x}^{(r)} = (x_1^{(r)}, \dots, x_{d'}^{(r)})^T$ , para  $r = 1, \dots, m$ , los vectores de atributos completos de los puntos de datos. Cada dispositivo  $i \in \mathcal{V}$  observa solo un subconjunto  $\mathcal{F}^{(i)} \subseteq \{1, \dots, d'\}$  de atributos, lo que da lugar a un conjunto de datos local  $\mathcal{D}^{(i)}$  con vectores de atributos

$$\mathbf{x}^{(i,r)} = (x_{j_1}^{(r)}, \dots, x_{j_d}^{(r)})^T.$$

Algunos de los dispositivos también pueden tener acceso a las etiquetas  $y^{(r)}$ , para  $r = 1, \dots, m$ , del conjunto de datos global. Una posible aplicación del VFL es permitir la colaboración entre distintos proveedores de atención médica. Cada proveedor recopila distintos tipos de mediciones, como análisis de sangre, electrocardiogramas y radiografías de tórax, para los mismos pacientes. Otra aplicación es un sistema nacional de seguro social, donde los registros de salud, los indicadores

financieros, el comportamiento del consumidor y los datos de movilidad son recolectados por diferentes instituciones. El VFL permite el aprendizaje conjunto entre estas entidades mientras se mantienen niveles bien definidos de protección de la privacidad.

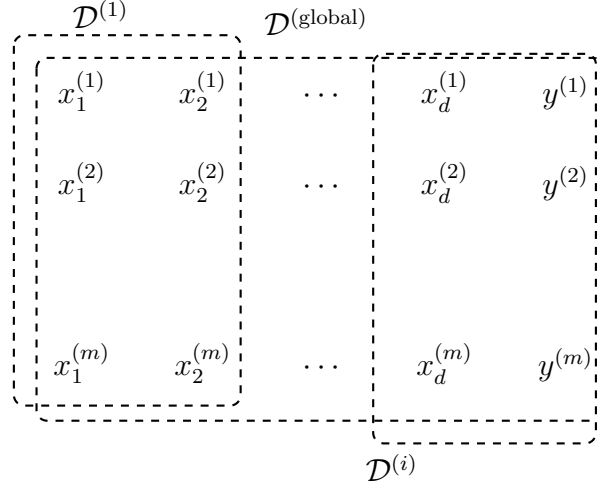


Figure 2: El VFL utiliza conjuntos de datos locales que se derivan de los puntos de datos de un conjunto de datos global común. Los conjuntos de datos locales difieren en la selección de atributos utilizadas para caracterizar los puntos de datos.

Vea tambien: FL, dispositivo, atributo, punto de datos, conjunto de datos, vector de atributos, conjunto de datos local, etiqueta, datos, protección de la privacidad.

**aprendizaje federado(FL)** FL es un término general para los métodos de aprendizaje automático que entrenan modelos de manera colaborativa,

utilizando datos y computación descentralizadas.

Vea también: ML, modelo, datos.

**aprendizaje multitarea** El aprendizaje multitarea tiene como objetivo aprovechar las relaciones entre diferentes tareas de aprendizaje. Considere dos tareas de aprendizaje obtenidas del mismo conjunto de datos de capturas de webcam. La primera tarea consiste en predecir la presencia de un ser humano, mientras que la segunda consiste en predecir la presencia de un automóvil. Podría ser útil utilizar la misma estructura de red profunda para ambas tareas y permitir que solo los pesos de la capa de salida final sean diferentes.

Vea también: tarea de aprendizaje, conjunto de datos, red profunda, pesos

**aprendizaje semi-supervisado (SSL)** El aprendizaje semi-supervisado (SSL) utiliza puntos de datos no etiquetados para apoyar el aprendizaje de una hipótesis a partir de puntos de datos etiquetados [23]. Este enfoque es particularmente útil para aplicaciones de aprendizaje automático que ofrecen una gran cantidad de puntos de datos no etiquetados, pero solo un número limitado de puntos de datos etiquetados.

Vea también: punto de datos, hipótesis, punto de dato etiquetado, ML, punto de datos

**arrepentimiento (regret)** El arrepentimiento de una hipótesis  $h$  en relación con otra hipótesis  $h'$ , que sirve como referencia, es la diferencia entre la pérdida incurrida por  $h$  y la pérdida incurrida por  $h'$  [18]. La hipótesis de referencia  $h'$  también se denomina experto.

Vea también: hipótesis, referencia (baseline), pérdida, experto

**aspectos computacionales** Por aspectos computacionales de un método de aprendizaje automático, nos referimos principalmente a los recursos computacionales requeridos para su implementación. Por ejemplo, si un método de aprendizaje automático utiliza técnicas de optimización iterativas para resolver ERM, sus aspectos computacionales incluyen: 1) cuántas operaciones aritméticas se necesitan para implementar una sola iteración (paso de gradiente); y 2) cuántas iteraciones se requieren para obtener parámetros del modelo útiles. Un ejemplo importante de técnica de optimización iterativa es el descenso por gradiente.

Vea también: ML, ERM, paso de gradiente, parámetros del modelo, descenso por gradiente (GD)

**aspectos estadísticos** Por aspectos estadísticos de un método de aprendizaje automático, nos referimos a las propiedades de la distribución de probabilidad de su salida bajo un modelo de probabilidad para los datos introducidos en el método.

Vea también: ML, distribución de probabilidad, modelo probabilístico, datos

**ataque** Un ataque a un sistema de aprendizaje automático se refiere a una acción intencional—ya sea activa o pasiva—que compromete la integridad, disponibilidad o confidencialidad del sistema. Los ataques activos implican la alteración de componentes como los conjuntos de datos (por medio de envenenamiento de datos) o de los enlaces de comunicación entre dispositivos en un entorno de aprendizaje federado. Los ataques

pasivos, como los ataques a la privacidad, buscan inferir atributos sensibles sin modificar el sistema. Según su objetivo, distinguimos entre ataques de denegación de servicio, ataques de puerta trasera y ataques a la privacidad.

Véa también: envenenamiento de datos, ataque a la privacidad, atributo sensible, ataque de denegación de servicio (DoS), puerta trasera (backdoor).

**ataque a la privacidad** Un ataque a la privacidad en un sistema de aprendizaje automático tiene como objetivo inferir atributos sensibles de individuos mediante el aprovechamiento del acceso parcial a un modelo de aprendizaje automático ya entrenado. Una forma de ataque a la privacidad es la inversión de modelo.

Véa también: ataque, ML, atributo sensible, model inversion, inteligencia artificial confiable (IA confiable), reglamento general de protección de datos (RGPD).

**ataque de denegación de servicio (DoS)** Un ataque de denegación de servicio (DoS) tiene como objetivo (por ejemplo, mediante envenenamiento de datos) dirigir el entrenamiento de un modelo para que tenga un rendimiento deficiente con puntos de datos típicos.

Vea también: envenenamiento de datos, modelo, punto de datos

**atributo** Un atributo de un punto de datos es una de sus propiedades que se puede medir o calcular fácilmente sin la necesidad de supervisión humana. Por ejemplo, si un punto de datos es una imagen digital (por ejemplo, almacenada como un archivo `.jpeg`), entonces podríamos

usar las intensidades rojo-verde-azul de sus píxeles como atributos. Sinónimos específicos del dominio para el término atributo incluyen "covariable", "variable explicativa", "variable independiente", "variable de entrada", "predictor (variable)" o "regresor" [24], [25], [26].

Vea también: punto de datos

**atributo sensible** El aprendizaje automático busca aprender una hipótesis que prediga la etiqueta de un punto de datos a partir de sus atributos. En algunas aplicaciones, es crucial garantizar que la salida del sistema no permita inferir atributos sensibles de los puntos de datos. Qué se considera atributo sensible depende del dominio de aplicación y debe definirse explícitamente.

Vea también: ML, hipótesis, etiqueta, punto de datos, atributo,

**aumentación de datos** Los métodos de aumentación de datos añaden puntos de datos sintéticos a un conjunto existente de punto de datos. Estos puntos de datos sintéticos se obtienen mediante perturbaciones (por ejemplo, añadir ruido a mediciones físicas) o transformaciones (por ejemplo, rotaciones de imágenes) de los puntos de datos originales. Estas perturbaciones y transformaciones son tales que los puntos de datos sintéticos resultantes deben tener la misma etiqueta. Como ejemplo, una imagen de un gato rotada sigue siendo una imagen de un gato, aunque sus vectores de atributos (obtenidos al apilar las intensidades de color de los píxeles) sean muy diferentes (ver Figura 3). La aumentación de datos puede ser una forma eficiente de regulación.

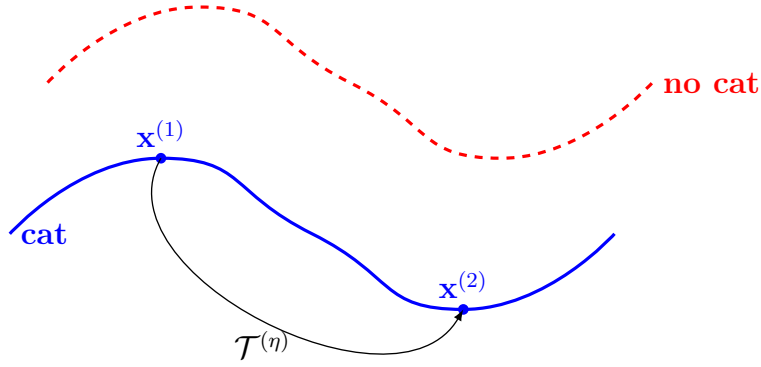


Figure 3: La aumentación de datos aprovecha las simetrías intrínsecas de los puntos de datos en algún espacio de atributos  $\mathcal{X}$ . Podemos representar una simetría mediante un operador  $\mathcal{T}^{(\eta)} : \mathcal{X} \rightarrow \mathcal{X}$ , parametrizado por algún número  $\eta \in \mathbb{R}$ . Por ejemplo,  $\mathcal{T}^{(\eta)}$  podría representar el efecto de rotar una imagen de un gato  $\eta$  grados. Un punto de datos con vector de atributos  $\mathbf{x}^{(2)} = \mathcal{T}^{(\eta)}(\mathbf{x}^{(1)})$  debete tener la misma etiqueta  $y^{(2)} = y^{(1)}$  que un punto de datos con vector de atributos  $\mathbf{x}^{(1)}$ .

Vea también: datos, punto de datos, etiqueta, vector de atributos, regularización, espacio de atributos.

**autoencoder** Un autoencoder es un método de aprendizaje automático que aprende simultáneamente un mapeo codificador  $h(\cdot) \in \mathcal{H}$  y un mapeo decodificador  $h^*(\cdot) \in \mathcal{H}^*$ . Es una instancia de minimización empírica del riesgo (ERM) que utiliza una perdida calculada a partir del error de reconstrucción  $\mathbf{x} - h^*(h(\mathbf{x}))$ .

Vea también: ML, ERM, pérdida

**bagging** El bagging (o agregación por bootstrap) es una técnica genérica



para mejorar (la robustez de) un método de aprendizaje automático dado. La idea es utilizar el bootstrap para generar copias perturbadas de un conjunto de datos dado y luego aprender una hipótesis separada para cada copia. Posteriormente, se predice la etiqueta de un punto de datos combinando o agregando las predicciones individuales de cada hipótesis separada. Para mapas de hipótesis que entregan valores numéricos de etiqueta, esta agregación podría implementarse calculando el promedio de las predicciones individuales.

Vea también: ML, bootstrap, conjunto de datos, hipótesis, etiqueta, punto de datos, predicción.

**bandido multi-brazo (MAB)** Un problema de bandido multi-brazo (MAB) modela un escenario de toma de decisiones repetida en el que, en cada paso temporal  $k$ , un aprendiz debe elegir una de varias acciones posibles, a menudo denominadas brazos, de un conjunto finito  $\mathcal{A}$ . Cada brazo  $a \in \mathcal{A}$  produce una recompensa estocástica  $r^{(a)}$  extraída de una distribución de probabilidad desconocida con media  $\mu^{(a)}$ . El objetivo del aprendiz es maximizar la recompensa acumulada a lo largo del tiempo mediante un equilibrio estratégico entre exploración (recopilar información sobre brazos inciertos) y explotación (seleccionar brazos que se sabe que funcionan bien). Este equilibrio se cuantifica mediante la noción de arrepentimiento, que mide la brecha de rendimiento entre la estrategia del aprendiz y la estrategia óptima que siempre selecciona el mejor brazo. Los problemas MAB forman un modelo fundamental en aprendizaje en línea, aprendizaje por refuerzo y diseño experimental secuencial [27].

Vea también: recompensa, distribución de probabilidad, media, arrepentimiento (regret)

**boosting** El boosting es un metodo de optimización iterativo para aprender una función de hipótesis precisa (o strong learner) combinando secuencialmente funciones de hipótesis menos precisas (conocidas como weak learners) [28, Cap. 10]. Por ejemplo, los weak learners suelen ser árboles de decisión poco profundos que se combinan para obtener un árbol de decisión profundo. El boosting puede entenderse como una generalización de los metodos de descenso por gradiente para ERM utilizando modelos paramétricos y funciones de perdida suaves [29]. Así como el descenso por gradiente actualiza iterativamente los parametros del modelo para reducir el rieso empírico, el boosting combina iterativamente (ej., por suma) funciones de hipótesis para reducir dicho riesgo empírico. Una variante ampliamente utilizada del enfoque general de boosting es el boosting gradiente, que emplea el gradiente de la función de perdida para combinar los weak learners [29].

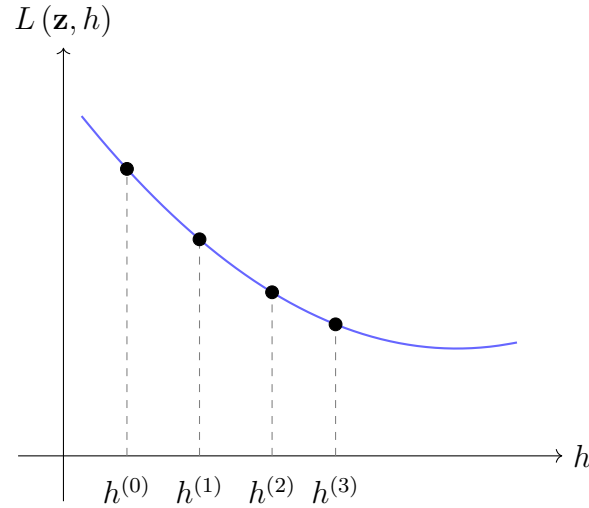


Figure 4: Los métodos de boosting construyen una secuencia de funciones de hipótesis  $h^{(0)}, h^{(1)}, \dots$  que son cada vez mejores (es decir, incurren en menor pérdida).

Véa también: hipótesis, aplicación, árbol de decisión, generalización, métodos de gradiente, ERM, modelo, suave, función de pérdida, GD, parámetros del modelo, riesgo empírico, gradiente, pérdida, paso de gradiente.

**bootstrap** Para el análisis de métodos de aprendizaje automático, es a menudo útil interpretar un conjunto dado de puntos de datos  $\mathcal{D} = \{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}\}$  como realizaciones de variables aleatorias iid con una distribución de probabilidad común  $p(\mathbf{z})$ . En general, no conocemos  $p(\mathbf{z})$  exactamente, por lo que necesitamos estimarla. El método bootstrap utiliza el histograma de  $\mathcal{D}$  como un estimador para la distribución de probabilidad subyacente  $p(\mathbf{z})$ .

Vea también: ML, punto de datos, realización, RV, i.i.d., distribución de probabilidad

**bosque aleatorio** Un bosque aleatorio es un conjunto (o ensamblaje) de diferentes árboles de decisiones. Cada uno de estos árboles de decisiones se obtiene al ajustar una copia perturbada del conjunto de datos original. Vea también: árbol de decisión, conjunto de datos.

**brecha de generalización** La diferencia entre el rendimiento de un modelo entrenado sobre el conjunto de entrenamiento y su rendimiento sobre otros puntos de datos (como los del conjunto de validación). Véa también: modelo, conjunto de entrenamiento, punto de datos, conjunto de validación, hipótesis, árbol de decisión, generalización, métodos de gradiente, ERM, suave, función de pérdida, GD, parámetros del modelo, riesgo empírico, gradiente, pérdida, paso de gradiente.

**caracterización min-max de Courant–Fischer–Weyl** Considere una matriz semi-definida positiva  $\mathbf{Q} \in \mathbb{R}^{d \times d}$  con descomposición en valores propios (o descomposición espectral),

$$\mathbf{Q} = \sum_{j=1}^d \lambda_j \mathbf{u}^{(j)} (\mathbf{u}^{(j)})^T.$$

Aquí, utilizamos los valores propios ordenados (de forma creciente)

$$\lambda_1 \leq \dots \leq \lambda_n.$$

La caracterización min-max de Courant–Fischer–Weyl [3, Th. 8.1.2] representa los valores propios de  $\mathbf{Q}$  como las soluciones a ciertos proble-

mas de optimización.

Vea también: psd, EVD, valor propio.

**clasificación** La clasificación es la tarea de determinar una etiqueta  $y$  con valor discreto para un punto de datos, basado únicamente en sus atributos  $\mathbf{x}$ . La etiqueta  $y$  pertenece a un conjunto finito, como por ejemplo  $y \in \{-1, 1\}$  o  $y \in \{1, \dots, 19\}$ , y representa la categoría a la que pertenece el punto de datos.

Vea también: punto de datos, atributo

**clasificación multi-etiqueta** Los problemas y métodos de clasificación multi-etiqueta utilizan puntos de datos caracterizados por varias etiquetas. Por ejemplo, considere un punto de datos que representa una imagen con dos etiquetas. Una etiqueta indica la presencia de un ser humano en la imagen y otra etiqueta indica la presencia de un automóvil. Vea también: clasificación, punto de datos, etiqueta

**clasificador** Un clasificador es una hipótesis (función)  $h(\mathbf{x})$  usada para predecir una etiqueta que toma valores de un espacio de etiquetas finito. Podemos usar directamente el valor  $h(\mathbf{x})$  como predicción  $\hat{y}$  para la etiqueta. Pero normalmente se usa una función  $h(\cdot)$  que entrega una cantidad numérica. La predicción es obtenida a travez de un paso de umbral. Por ejemplo, en un problema de clasificación binaria con  $\mathcal{Y} \in \{-1, 1\}$ , podríamos usar una hipótesis de valores reales  $h(\mathbf{x}) \in \mathbb{R}$  como clasificador. Una predicción  $\hat{y}$  puede obtenerse mediante:

$$\hat{y} = 1 \text{ for } h(\mathbf{x}) \geq 0 \text{ and } \hat{y} = -1 \text{ otherwise.} \quad (1)$$

Podemos caracterizar un clasificador mediante sus regiones de decisión  $\mathcal{R}_a$ , para cada valor posible de etiqueta  $a \in \mathcal{Y}$ .

Vea también: hipótesis, etiqueta, espacio de etiquetas, predicción, clasificación, región de decisión

**clasificador lineal** Consideremos puntos de datos caracterizados por atributos numéricos  $\mathbf{x} \in \mathbb{R}^d$  y una etiqueta  $y \in \mathcal{Y}$  de algún espacio de etiqueta finito  $\mathcal{Y}$ . Un clasificador lineal se caracteriza por tener regiones de decisión que están separadas por hiperplanos en  $\mathbb{R}^d$  [6, Ch. 2].

Vea también: punto de datos, atributo, etiqueta, espacio de etiquetas, región de decisión

**clúster** Un clúster es un subconjunto de puntos de datos que son más similares entre sí que a los puntos de datos fuera del clúster. La medida cuantitativa de similitud entre los puntos de datos es una decisión de diseño. Si los puntos de datos se caracterizan por vectores de atributos euclidianos  $\mathbf{x} \in \mathbb{R}^d$ , podemos definir la similitud entre dos puntos de datos a través de la distancia euclidiana entre sus vectores de atributos. Un ejemplo de tal clúster es. 5.

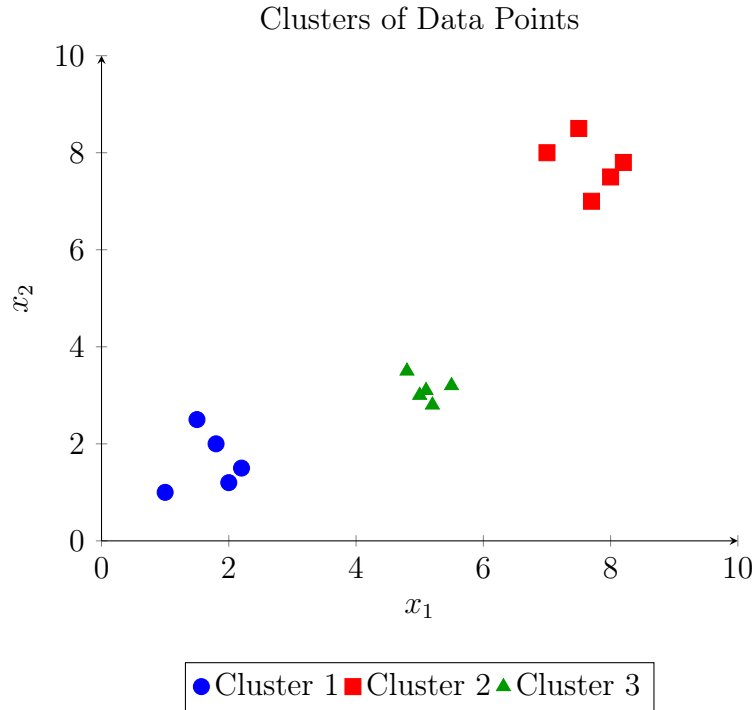


Figure 5: Ilustración de tres clústers en un espacio de atributos bidimensional. Cada clúster agrupa puntos de datos que son más similares entre sí que con aquellos en otros clusters, basándose en la distancia euclidiana.

Vea también: punto de datos, vector de atributos, espacio de atributos.

**condición de gradiente cero** Considera el problema de optimización sin restricciones  $\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$  con una función de objetivo suave y convexa  $f(\mathbf{w})$ . Una condición necesaria y suficiente para que un vector  $\hat{\mathbf{w}} \in \mathbb{R}^d$  resuelva este problema es que el gradiente  $\nabla f(\hat{\mathbf{w}})$  sea el vector cero,

$$\nabla f(\hat{\mathbf{w}}) = \mathbf{0} \Leftrightarrow f(\hat{\mathbf{w}}) = \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}).$$

Vea también: suave, convexo, función objetivo, gradiente.

**conectividad algebraica** La conectividad algebraica de un grafo no dirigido es el segundo valor propio más pequeño  $\lambda_2$  de su matriz laplaciana. Un grafo es conexo si y solo si  $\lambda_2 > 0$ .

Véa también: grafo, valor propio, matriz laplaciana.

**conjunto de datos** Un conjunto de datos se refiere a una colección de puntos de datos. Estos puntos de datos contienen información sobre alguna cantidad de interés (o etiqueta) dentro de una aplicación de aprendizaje automático. Los métodos de aprendizaje automático utilizan conjuntos de datos para el entrenamiento de un modelo (por ejemplo, a través de ERM) y para la validación de modelos. Nuestra noción de conjunto de datos es muy flexible, ya que permite diferentes tipos de puntos de datos. De hecho, los puntos de datos pueden ser objetos físicos concretos (como humanos o animales) o objetos abstractos (como números). Como ejemplo, la Figura 6 muestra un conjunto de datos que consiste en vacas como punto de datos.



Figure 6: A cowherd.

Con frecuencia, un ingeniero de aprendizaje automático no tiene acceso



directo a un conjunto de datos. De hecho, acceder al conjunto de datos en la Figura requeriría visitar el rebaño de vacas en los Alpes. En su lugar, necesitamos utilizar una aproximación (o representación) del conjunto de datos que sea más conveniente para trabajar. Se han desarrollado diferentes modelos matemáticos para la representación (o aproximación) de conjuntos de datos [30], [31], [32], [33]. Uno de los modelos de datos más adoptados es el modelo relacional, que organiza los datos en una tabla (o relación) [34], [30]. Una tabla se compone de filas y columnas:

- Cada fila de la tabla representa un solo punto de datos.
- Cada columna de la tabla corresponde a un atributo específico del punto de datos. Los métodos de aprendizaje automático pueden utilizar estos atributos como atributos y etiquetas del punto de datos.

Por ejemplo, la Tabla 1 muestra una representación del conjunto de datos en la Figura 6. En el modelo relacional, el orden de las filas es irrelevante, y cada atributo (es decir, columna) debe estar definido de manera precisa con un dominio, que especifica el conjunto de valores posibles. En las aplicaciones de aprendizaje automático, estos dominios de atributos se convierten en el espacio de atributos y el espacio de etiquetas.

Name	Weight	Age	Height	Stomach temp
Zenzi	100	4	100	25
Berta	140	3	130	23
Resi	120	4	120	31

Table 1: Una relación (o tabla) que representa el conjunto de datos en la Figura .

Si bien el modelo relacional es útil para el estudio de muchas aplicaciones de aprendizaje automático, puede ser insuficiente para cumplir con los requisitos de IA confiable. Enfoques modernos como las hojas de datos para conjuntos de datos proporcionan una documentación más completa, incluyendo detalles sobre el proceso de recolección del conjunto de datos, su uso previsto y otra información contextual [35].

Vea también: punto de datos, etiqueta, ML, modelo, ERM, validación, datos, espacio de atributos, espacio de etiquetas, trustworthy AI

**conjunto de datos local** El concepto de un conjunto de datos local se encuentra entre el concepto de un punto de datos y un conjunto de datos. Un conjunto de datos local consiste en varios puntos de datos individuales, que se caracterizan por atributos y etiquetas. A diferencia de un único conjunto de datos utilizado en métodos de aprendizaje automático básicos, un conjunto de datos local también está relacionado con otros conjuntos de datos locales mediante diferentes nociones de similitud. Estas similitudes pueden surgir de modelos de probabilidad o de la infraestructura de comunicación y están codificadas en las aristas de una red de aprendizaje federado.

Vea también: conjunto de datos, punto de datos, atributo, etiqueta, ML, modelo probabilístico, red FL.

**conjunto de entrenamiento** Un conjunto de entrenamiento es un conjunto de datos  $\mathcal{D}$  que consiste en algunos puntos de datos usados en ERM para aprender una hipótesis  $\hat{h}$ . La pérdida promedio de  $\hat{h}$  en el conjunto de entrenamiento se denomina error de entrenamiento. La comparación del error de entrenamiento con el error de validación de  $\hat{h}$  nos permite diagnosticar el método de aprendizaje automático e informa cómo mejorar el error de validación (por ejemplo, utilizando un espacio de hipótesis diferente o recopilando más puntos de datos) [6, Sec. 6.6].  
Vea también: conjunto de datos, punto de datos, ERM, hipótesis, pérdida, error de entrenamiento, error de validación

**conjunto de prueba** Un conjunto de puntos de datos que no ha sido utilizado ni para entrenar un modelo (por ejemplo, mediante ERM) ni en un conjunto de validación para elegir entre diferentes modelos.  
Vea también: punto de datos, modelo, ERM, conjunto de validación, modelo

**conjunto de validación** Un conjunto de puntos de datos usado para estimar el riesgo de una hipótesis  $\hat{h}$  que ha sido aprendida mediante algún método de aprendizaje automático (por ejemplo, resolviendo ERM). La pérdida promedio de  $\hat{h}$  en el conjunto de validación se denomina error de validación y puede utilizarse para diagnosticar un método de aprendizaje automático (vea [6, Sec. 6.6]). La comparación entre el error de entrenamiento y el error de validación puede proporcionar

información sobre cómo mejorar el método de aprendizaje automático (como usar un espacio de hipótesis diferente).

Vea también: punto de datos, riesgo, hipótesis, ML, ERM, pérdida, validación, error de validación, error de entrenamiento, espacio de hipótesis

**contraction operator** An operator  $\mathcal{F} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is a contraction if, for some  $\kappa \in [0, 1)$ ,

$$\|\mathcal{F}\mathbf{w} - \mathcal{F}\mathbf{w}'\|_2 \leq \kappa \|\mathbf{w} - \mathbf{w}'\|_2 \text{ holds for any } \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d.$$

**convexo** Un subconjunto  $\mathcal{C} \subseteq \mathbb{R}^d$  del espacio euclidiano  $\mathbb{R}^d$  se denomina convexo si contiene el segmento de recta entre cualesquiera dos puntos  $\mathbf{x}, \mathbf{y} \in \mathcal{C}$  en ese conjunto. Una función  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  es convexa si su epígrafo  $\{(\mathbf{w}^T, t)^T \in \mathbb{R}^{d+1} : t \geq f(\mathbf{w})\}$  es un conjunto convexo [36]. Ilustramos un ejemplo de un conjunto convexo y una función convexa en la Figura 7.



Figure 7: Izquierda: Un conjunto convexo  $\mathcal{C} \subseteq \mathbb{R}^d$ . Derecha: Una función convexa  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ .

Vea también: espacio euclidiano.

**cota superior de confianza (UCB)** Considera una aplicación de aprendizaje automático que requiere seleccionar, en cada paso temporal  $k$ , una acción  $a_k$  de un conjunto finito de alternativas  $\mathcal{A}$ . La utilidad de

seleccionar la acción  $a_k$  se cuantifica mediante una señal numérica de recompensa  $r^{(a_k)}$ . Un modelo de probabilidad ampliamente utilizado para este tipo de problema de toma de decisiones secuencial es el entorno de bandidos multi-brazo estocásticos [27]. En este modelo, la recompensa  $r^{(a)}$  se considera como la realización de una variable aleatoria con media desconocida  $\mu^{(a)}$ . Idealmente, siempre elegiríamos la acción con la mayor recompensa esperada  $\mu^{(a)}$ , pero estas medias son desconocidas y deben estimarse a partir de datos observados. Simplemente elegir la acción con la mayor estimación  $\hat{\mu}^{(a)}$  puede conducir a resultados subóptimos debido a la incertidumbre en la estimación. La estrategia UCB aborda esto seleccionando acciones no solo basándose en sus medias estimadas, sino también incorporando un término que refleja la incertidumbre en estas estimaciones, favoreciendo acciones con alto potencial de recompensa y alta incertidumbre. Las garantías teóricas para el rendimiento de las estrategias UCB, incluyendo cotas de arrepentimiento logarítmicas, se establecen en [27].

Vea también: ML, recompensa, modelo probabilístico, realización, RV, media, datos

**covariance** La covarianza entre dos RVs reales  $x$  y  $y$ , definidas sobre un mismo espacio de probabilidad, mide su dependencia lineal. Se define como

$$\text{cov}(x, y) = \mathbb{E}\{(x - \mathbb{E}\{x\})(y - \mathbb{E}\{y\})\}.$$

Una covarianza positiva indica que  $x$  y  $y$  tienden a aumentar juntos, mientras que una covarianza negativa sugiere que uno tiende a aumentar cuando el otro disminuye. Si  $\text{cov}(x, y) = 0$ , se dice que las RVs son

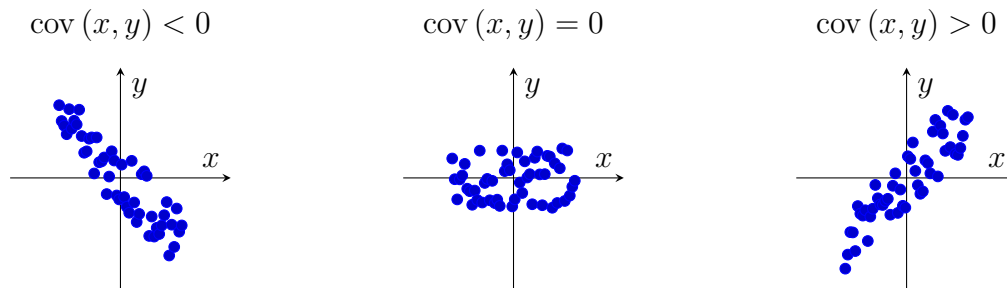


Figure 8: Diagrama de dispersións illustrating realizaci3n from three different modelo probabilísticos for two RVs with different covariance values: negative (left), zero (center), and positive (right).

no correlacionadas, aunque esto no implica necesariamente que sean estadísticamente independientes. Véase la Figura 8 para ilustraciones visuales.

**criterio de parada** Muchos métodos de aprendizaje automático utilizan algoritmos iterativos que construyen una secuencia de parametros del modelo (como los pesos de un mapa lineal o los pesos de una red neuronal artificial). Estos parámetros (idealmente) convergen a una elecci3n óptima para los parametros del modelo. En la práctica, dados recursos computacionales finitos, necesitamos detener la iteraci3n después de un número finito de repeticiones. Un criterio de parada es cualquier condici3n bien definida requerida para detener la iteraci3n.

Vea tambien: ML, algoritmo, parámetros del modelo, pesos, ANN.

**datos** Los datos se refieren a objetos que llevan informaci3n. Estos objetos pueden ser tanto objetos físicos concretos (como personas o animales) como conceptos abstractos (como números). A menudo, utilizamos

representaciones (o aproximaciones) de los datos originales que son más convenientes para su procesamiento. Estas aproximaciones se basan en diferentes modelos de datos, siendo el modelo de datos relacional uno de los más utilizados [34].

**datos en red** Los datos en red consisten en conjuntos de datos locales relacionados por alguna noción de similitud por pares. Podemos representar los datos en red utilizando un grafo cuyos nodos contienen conjuntos de datos locales y cuyas aristas codifican similitudes por pares. Un ejemplo de datos en red surge en las aplicaciones de aprendizaje federado donde los conjuntos de datos locales son generados por dispositivos distribuidos espacialmente.

Vea también: datos, conjunto de datos local, grafo, conjunto de datos local, FL, dispositivo

**datos faltantes** Considere un conjunto de datos constituido por puntos de datos recopilados a través de algún dispositivo físico. Debido a imperfecciones y fallas, algunos de los valores de atributos o etiquetas de los puntos de datos podrían estar corruptos o simplemente faltar. La imputación de datos tiene como objetivo estimar estos valores faltantes [37]. Podemos interpretar la imputación de datos como un problema de aprendizaje automático donde la etiqueta de un punto de datos es el valor del atributo corrupto.

Vea también: conjunto de datos, punto de datos, dispositivo, atributo, etiqueta, punto de datos, Datos

**descenso de gradiente estocástico (SGD)** El descenso de gradiente es-

tocástico es una variante del descenso por gradiente en la que se reemplaza el gradiente de la función objetivo por una aproximación estocástica. Una aplicación principal del descenso por gradiente estocástico es entrenar un modelo parametrizado mediante ERM sobre un conjunto de entrenamiento  $\mathcal{D}$  que es muy grande o no está fácilmente disponible (por ejemplo, cuando los puntos de datos están almacenados en bases de datos distribuidas por todo el mundo). Para evaluar el gradiente del riesgo empírico (como función de los parámetros del modelo  $\mathbf{w}$ ), se requiere calcular una suma  $\sum_{r=1}^m \nabla_{\mathbf{w}} L(\mathbf{z}^{(r)}, \mathbf{w})$  sobre todos los puntos de datos del conjunto de entrenamiento. Una aproximación estocástica se obtiene al reemplazar esta suma  $\sum_{r=1}^m \nabla_{\mathbf{w}} L(\mathbf{z}^{(r)}, \mathbf{w})$  por una suma parcial  $\sum_{r \in \mathcal{B}} \nabla_{\mathbf{w}} L(\mathbf{z}^{(r)}, \mathbf{w})$  considerando un subconjunto aleatorio  $\mathcal{B} \subseteq \{1, \dots, m\}$  (see Figure 9). A estos puntos de datos seleccionados aleatoriamente se les denomina a menudo lote. El tamaño del lote, denotado  $|\mathcal{B}|$  es un parámetro importante del descenso por gradiente estocástico. El descenso por gradiente estocástico con  $|\mathcal{B}| > 1$  se conoce como mini-lote SGD [38].



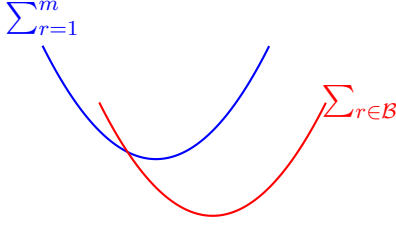


Figure 9: El descenso por gradiente estocástico (descenso por gradiente) para ERM aproxima el gradiente  $\sum_{r=1}^m \nabla_{\mathbf{w}} L(\mathbf{z}^{(r)}, \mathbf{w})$  reemplazando la suma sobre todos los puntos de datos del conjunto de entrenamiento (indexados por  $r = 1, \dots, m$ ) con una suma sobre un subconjunto aleatorio  $\mathcal{B} \subseteq \{1, \dots, m\}$ .

Vea tambien: GD, gradiente, función objetivo, modelo, ERM, conjunto de entrenamiento, punto de datos, riesgo empírico, parámetros del modelo, lote.

**descenso por gradiente (GD)** El descenso por gradiente es un método iterativo para encontrar el mínimo de una función diferenciable  $f(\mathbf{w})$  con un argumento vectorial  $\mathbf{w} \in \mathbb{R}^d$ . Considera una estimación o aproximación actual  $\mathbf{w}^{(k)}$  para el mínimo de la función  $f(\mathbf{w})$ . Queremos encontrar un nuevo vector (mejor)  $\mathbf{w}^{(k+1)}$  que tenga un valor objetivo menor  $f(\mathbf{w}^{(k+1)}) < f(\mathbf{w}^{(k)})$  que la estimación actual  $\mathbf{w}^{(k)}$ . Esto se puede lograr típicamente utilizando un paso de gradiente.

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \eta \nabla f(\mathbf{w}^{(k)}) \quad (2)$$

tamaño de paso suficientemente pequeño  $\eta > 0$ . La Figura 10 ilustra el efecto de un solo paso de gradiente (2).

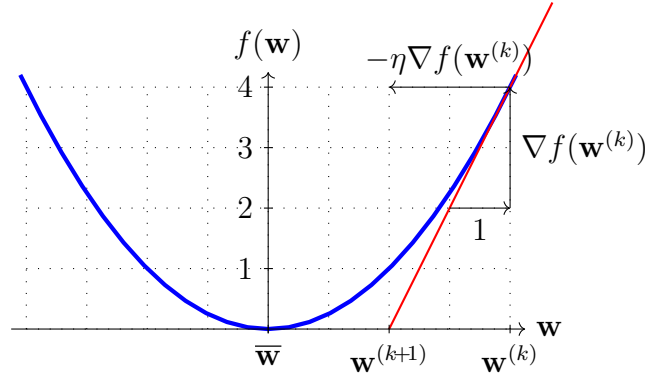


Figure 10: Un paso de gradiente singular (2) hacia el minimizador  $\bar{\mathbf{w}}$  de  $f(\mathbf{w})$ .

Vea también: mínimo, diferenciable, paso de gradiente, tamaño de paso, gradiente.

**descenso por gradiente en línea (online GD)** Considere un método de aprendizaje automático que aprende los parámetros del modelo  $\mathbf{w}$  a partir de un espacio de parámetros  $\mathcal{W} \subseteq \mathbb{R}^d$ . El proceso de aprendizaje utiliza puntos de datos  $\mathbf{z}^{(t)}$  que llegan en instantes de tiempo consecutivos  $t = 1, 2, \dots$ . Interpretamos los puntos de datos  $\mathbf{z}^{(t)}$  como copias iid de una variable aleatoria  $\mathbf{z}$ . El riesgo  $\mathbb{E}\{L(\mathbf{z}, \mathbf{w})\}$  de una hipótesis  $h^{(\mathbf{w})}$  puede entonces (bajo condiciones suaves) obtenerse como el límite  $\lim_{T \rightarrow \infty} (1/T) \sum_{t=1}^T L(\mathbf{z}^{(t)}, \mathbf{w})$ . Podríamos usar este límite como la función objetivo para aprender los parámetros del modelo  $\mathbf{w}$ . Sin embargo, este límite solo puede evaluarse si esperamos un tiempo infinito para recolectar todos los puntos de datos. Algunas aplicaciones de aprendizaje automático requieren métodos que aprendan en línea: tan pronto como llega un nuevo punto de datos  $\mathbf{z}^{(t)}$  en el tiempo  $t$ ,

actualizamos los parametros del modelo actuales  $\mathbf{w}^{(t)}$ . Nótese que el nuevo punto de datos  $\mathbf{z}^{(t)}$  contribuye con el componente  $L(\mathbf{z}^{(t)}, \mathbf{w})$  al riesgo. Como sugiere el nombre, el descenso por gradiente en línea actualiza  $\mathbf{w}^{(t)}$  mediante un (proyectado) paso de gradiente:

$$\mathbf{w}^{(t+1)} := P_{\mathcal{W}}(\mathbf{w}^{(t)} - \eta_t \nabla_{\mathbf{w}} L(\mathbf{z}^{(t)}, \mathbf{w})). \quad (3)$$

Nótese que (3) es un paso de gradiente para el componente actual  $L(\mathbf{z}^{(t)}, \cdot)$  del riesgo. La actualización (3) ignora todos los componentes anteriores  $L(\mathbf{z}^{(t')}, \cdot)$  para  $t' < t$ . Por tanto, podría suceder que, comparado con  $\mathbf{w}^{(t)}$ , los parametros del modelo actualizados  $\mathbf{w}^{(t+1)}$  aumenten la perdida promedio retrospectivo  $\sum_{t'=1}^{t-1} L(\mathbf{z}^{(t')}, \cdot)$ . Sin embargo, para un tasa de aprendizaje  $\eta_t$  elegido apropiadamente, se puede demostrar que el descenso por gradiente en línea es óptimo en escenarios de interés práctico. Por óptimo, entendemos que los parametros del modelo  $\mathbf{w}^{(T+1)}$  entregados por el descenso por gradiente en línea tras observar  $T$  puntos de datos  $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(T)}$  son al menos tan buenos como los entregados por cualquier otro método de aprendizaje [17, 39].

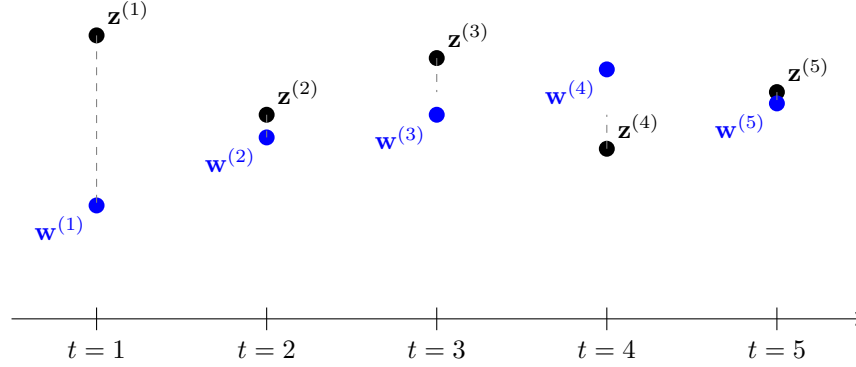


Figure 11: Una instancia de descenso por gradiente en línea que actualiza los parámetros del modelo  $\mathbf{w}^{(t)}$  usando el punto de datos  $\mathbf{z}^{(t)} = x^{(t)}$  que llega en el tiempo  $t$ . Esta instancia emplea la pérdida de error cuadrático  $L(\mathbf{z}^{(t)}, w) = (x^{(t)} - w)^2$ .

Vea también: ML, parámetros del modelo, espacio de parámetros, punto de datos, i.i.d., RV, riesgo, hipótesis, función objetivo, GD, paso de gradiente, pérdida, tasa de aprendizaje, pérdida de error cuadrático.

**descenso por gradiente proyectado (GD proyectado)** Consideremos un método basado en ERM que utiliza un modelo parametrizado con un espacio de parámetros  $\mathcal{W} \subseteq \mathbb{R}^d$ . Aun si la función objetivo de ERM es suave, no podemos usar el descenso por gradiente básico, ya que este no toma en cuenta las restricciones sobre la variable de optimización (es decir, los parámetros del modelo). El descenso por gradiente proyectado extiende el descenso por gradiente básico para controlar restricciones sobre la variable de optimización (es decir, los parámetros del modelo). Una sola iteración del descenso por gradiente proyectado consiste

primero en realizar un paso de gradiente y luego proyectar el resultado sobre el espacio de parámetros

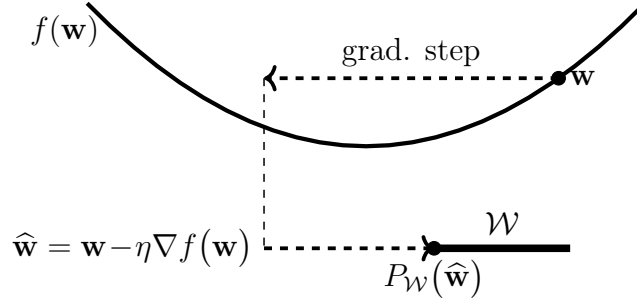


Figure 12: GD proyectado amplía un paso de gradiente básico con una proyección de regreso al conjunto de restricciones  $\mathcal{W}$ .

Vea también: ERM, modelo, espacio de parámetros, función objetivo, suave, GD, parámetros del modelo, paso de gradiente

**descenso por subgradiente** El descenso por subgradiente es una generalización del descenso por gradiente que no requiere la diferenciabilidad de la función a minimizar. Esta generalización se obtiene al reemplazar el concepto de gradiente por el de subgradiente. Similar a los gradientes, los subgradientes permiten construir aproximaciones locales de una función objetivo. La función objetivo podría ser el riesgo empírico  $\widehat{L}(h^{(\mathbf{w})}|\mathcal{D})$  visto como una función de los parámetros del modelo  $\mathbf{w}$  que seleccionan una hipótesis  $h^{(\mathbf{w})} \in \mathcal{H}$ .

Vea también: subgradiente, generalización, GD, gradiente, función objetivo, riesgo empírico, parámetros del modelo, hipótesis.

**descomposición en valores propios (EVD)** La descomposición en valores propios para una matriz cuadrada  $\mathbf{A} \in \mathbb{R}^{d \times d}$  es una factorización de la forma

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}.$$

Las columnas de la matriz  $\mathbf{V} = (\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(d)})$  son los vectores propios de la matriz  $\mathbf{A}$ . La matriz diagonal  $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \dots, \lambda_d\}$  contiene los valores propios  $\lambda_j$  correspondientes a los vectores propios  $\mathbf{v}^{(j)}$ . Es importante notar que esta descomposición solo existe si la matriz  $\mathbf{A}$  es diagonalizable.

Vea también: vector propio, valor propio.

**descomposición en valores singulares (SVD)** La descomposición en valores singulares (SVD) para una matriz  $\mathbf{A} \in \mathbb{R}^{m \times d}$  es una factorización de la forma

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^T,$$

con matrices ortonormales  $\mathbf{V} \in \mathbb{R}^{m \times m}$  y  $\mathbf{U} \in \mathbb{R}^{d \times d}$  [3]. La matriz  $\mathbf{\Lambda} \in \mathbb{R}^{m \times d}$  es no nula solo en la diagonal principal, cuyos elementos  $\Lambda_{j,j}$  son no negativos y se denominan valores singulares.

**desigualdad de concentración** Una cota superior para la probabilidad de que una RV se desvíe más de una cantidad prescrita respecto a su expectation [40].

Véase también: probabilidad, RV, expectation.

**determinante** El determinante  $\det(\mathbf{A})$  de una matriz cuadrada  $\mathbf{A} = (\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(d)}) \in \mathbb{R}^{d \times d}$  es una función de sus columnas  $\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(d)} \in \mathbb{R}^d$ , que satisface [41]

- Normalización:

$$\det(\mathbf{I}) = 1$$

- Multilinealidad:

$$\begin{aligned} \det(\mathbf{a}^{(1)}, \dots, \alpha \mathbf{u} + \beta \mathbf{v}, \dots, \mathbf{a}^{(d)}) &= \alpha \det(\mathbf{a}^{(1)}, \dots, \mathbf{u}, \dots, \mathbf{a}^{(d)}) \\ &\quad + \beta \det(\mathbf{a}^{(1)}, \dots, \mathbf{v}, \dots, \mathbf{a}^{(d)}) \end{aligned}$$

- Antisimetría:

$$\det(\dots, \mathbf{a}^{(j)}, \dots, \mathbf{a}^{(j')}, \dots) = -\det(\dots, \mathbf{a}^{(j')}, \dots, \mathbf{a}^{(j)}, \dots).$$

Podemos interpretar una matriz  $\mathbf{A}$  como una transformación lineal en  $\mathbb{R}^d$ . El determinante  $\det(\mathbf{A})$  caracteriza cómo (y en qué orientación) se alteran los volúmenes en  $\mathbb{R}^d$  por esta transformación [3], [42]. En particular,  $\det(\mathbf{A}) > 0$  preserva la orientación,  $\det(\mathbf{A}) < 0$  la invierte, y  $\det(\mathbf{A}) = 0$  colapsa el volumen completamente, lo que indica que  $\mathbf{A}$  no es invertible. El determinante también satisface  $\det(\mathbf{AB}) = \det(\mathbf{A}) \cdot \det(\mathbf{B})$ , y si  $\mathbf{A}$  es diagonalizable con valor propio  $\lambda_1, \dots, \lambda_d$ , entonces  $\det(\mathbf{A}) = \prod_{j=1}^d \lambda_j$  [43]. Para los casos especiales  $d = 2$  (2D) y  $d = 3$  (3D), el determinante puede interpretarse como un área o volumen orientado definido por los vectores columna de  $\mathbf{A}$ .

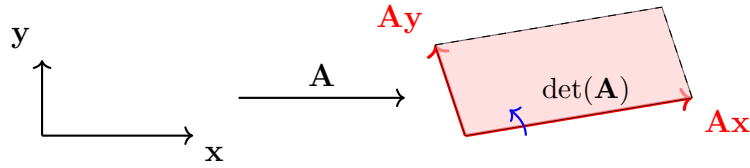


Figure 13: Podemos interpretar una matriz cuadrada  $\mathbf{A}$  como una transformación lineal de  $\mathbb{R}^d$  en sí misma. El determinante  $\det(\mathbf{A})$  caracteriza cómo esta transformación altera un volumen orientado.

Vea tambien: valor propio, inverse matrix.

**diagrama de dispersión** Un método de visualización que representa puntos de datos mediante marcadores en un plano bidimensional. La Figura 14 depicts un ejemplo de un diagrama de dispersión.

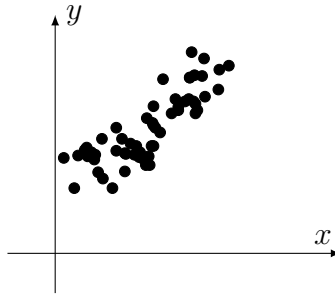


Figure 14: Un diagrama de dispersión de algunos puntos de datos que representan las condiciones climáticas diarias en Finlandia. Cada punto de datos se caracteriza por su temperatura mínima diaria  $x$  como atributo y su temperatura máxima diaria  $y$  como etiqueta. Las temperaturas se han medido en la estación meteorológica FMI de Helsinki Kaisaniemi durante el período 1.9.2024 - 28.10.2024.



Vea también: punto de datos, mínimo, atributo, máximo, etiqueta, FMI

**diferenciable** Una función real  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  es diferenciable si, en cualquier punto, puede aproximarse localmente mediante una función lineal. La aproximación lineal local en el punto  $\mathbf{x}$  es determinada por el gradiente  $\nabla f(\mathbf{x})$  [2].

Vea también: gradiente

**dimensión de Vapnik–Chervonenkis (dimensión VC)** La dimensión VC es una medida ampliamente utilizada para cuantificar el tamaño de un espacio de hipótesis infinito. Remitimos a la literatura (véase [44]) para una definición precisa de la dimensión VC, así como una discusión sobre sus propiedades básicas y su uso en ML.

Vea también: espacio de hipótesis, ML, dimensión efectiva.

**dimensión efectiva** La dimensión efectiva  $d_{\text{eff}}(\mathcal{H})$  de un espacio de hipótesis infinito  $\mathcal{H}$  es una medida de su tamaño. A grandes rasgos, la dimensión efectiva es igual al número efectivo de parámetros del modelo independientes y ajustables. Estos parámetros pueden ser los coeficientes utilizados en un mapa lineal o los pesos y términos de sesgo de una red neuronal artificial.

Vea también: espacio de hipótesis, parámetros del modelo, parámetros, pesos, ANN

**discrepancia** Considere una aplicación de aprendizaje federado con datos en red representada por un red FL. Los métodos de aprendizaje federado utilizan una medida de discrepancia para comparar los mapas de hipótesis generados por los modelos locales en los nodos  $i, i'$  conectados

por una arista en la red FL.

Vea también: FL, datos en red, red FL, hipótesis, modelo local

**dispositivo** Cualquier sistema físico que pueda usarse para almacenar y procesar datos. En el contexto de aprendizaje automático, generalmente nos referimos a un ordenador que puede leer puntos de datos de diferentes fuentes y, a su vez, entrenar un modelo de aprendizaje automático utilizando estos puntos de datos.

Vea también: datos, ML, punto de datos, modelo.

**distribución de probabilidad** Para analizar métodos de aprendizaje automático, puede ser útil interpretar los puntos de datos como realizaciones iid de una variable aleatoria. Las propiedades típicas de tales puntos de datos están gobernadas por la distribución de probabilidad de esta variable aleatoria. La distribución de probabilidad de una variable aleatoria binaria  $y \in \{0, 1\}$  se especifica completamente mediante las probabilidades  $\mathbb{P}(y = 0)$  y  $\mathbb{P}(y = 1) = 1 - \mathbb{P}(y = 0)$ . La distribución de probabilidad de una variable aleatoria con valores reales  $x \in \mathbb{R}$  puede especificarse mediante una función de densidad de probabilidad  $p(x)$  tal que  $\mathbb{P}(x \in [a, b]) \approx p(a)|b - a|$ . En el caso más general, una distribución de probabilidad se define mediante una medida de probabilidad [45, 46]. Vea también: ML, punto de datos, i.i.d., realización, RV, probabilidad, función de densidad de probabilidad (pdf)

**distribución normal multivariante** La distribución normal multivariante, denotada como  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , es un modelo de probabilidad fundamental para vectores de atributos numéricos de dimensión fija  $d$ . Define una

familia de distribución de probabilidad sobre variables aleatorias vectoriales  $\mathbf{x} \in \mathbb{R}^d$  [5], [45], [47]. Cada distribución de esta familia está completamente determinada por su vector de media  $\boldsymbol{\mu} \in \mathbb{R}^d$  y su matriz de covarianza  $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ . Cuando la matriz de covarianza  $\boldsymbol{\Sigma}$  es invertible, la correspondiente distribución de probabilidad se caracteriza por la siguiente pdf:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right].$$

Nótese que esta pdf solo está definida cuando  $\boldsymbol{\Sigma}$  es invertible. Más generalmente, cualquier variable aleatoria  $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  admite la siguiente representación:

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$$

donde  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  es un vector normal estandar, y  $\mathbf{A} \in \mathbb{R}^{d \times d}$  satisface  $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$ . Esta representación sigue siendo válida incluso si  $\boldsymbol{\Sigma}$  es singular, en cuyo caso  $\mathbf{A}$  no es de rango completo [48, Ch. 23]. La familia de distribuciones normales multivariantes es excepcional entre los modelos de probabilidad para cantidades numéricas, al menos por las siguientes razones. Primero, esta familia es cerrada bajo transformaciones afines, es decir,

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \mathbf{B}\mathbf{x} + \mathbf{c} \sim \mathcal{N}(\mathbf{B}\boldsymbol{\mu} + \mathbf{c}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T).$$

Segundo, la distribución de probabilidad  $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$  maximiza la entropía diferencial entre todas las distribuciones con la misma matriz de covarianza  $\boldsymbol{\Sigma}$  [49].

Véa también: modelo probabilístico, distribución de probabilidad, vector normal estándar, entropía diferencial, VA gaussiana.

**divergencia de Kullback-Leibler (divergencia KL)** La divergencia KL es una medida cuantitativa de cuánto se diferencia una distribución de probabilidad de otra distribución de probabilidad [49].  
Vea también: distribución de probabilidad

**embudo de privacidad** El embudo de privacidad es un método para aprender atributos amigables con la privacidad de los puntos de datos [50].  
Vea también: atributo, punto de datos

**entorno** El entorno de un nodo  $i \in \mathcal{V}$  es el subconjunto de nodos constituido por los vecinos de  $i$ .  
Vea también: vecinos

**entropía** La entropía cuantifica la incertidumbre o imprevisibilidad asociada a una variable aleatoria [49]. Para una variable aleatoria discreta  $x$  que toma valores en un conjunto finito  $\mathcal{S} = \{x_1, \dots, x_n\}$  con función de masa de probabilidad  $p_i := \mathbb{P}(x = x_i)$ , la entropía se define como

$$H(x) := - \sum_{i=1}^n p_i \log p_i.$$

La entropía se maximiza cuando todos los resultados son igualmente probables, y se minimiza (es decir, es cero) cuando el resultado es determinista. Una generalización del concepto de entropía para variables aleatorias continuas es la entropía diferencial.

Vea también: incertidumbre, modelo probabilístico.

**entropía diferencial** Para una variable aleatoria con valores reales  $\mathbf{x} \in \mathbb{R}^d$  con función de densidad de probabilidad  $p(x)$ , la entropía diferencial se

define como [49]

$$h(\mathbf{x}) := - \int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}.$$

La entropía diferencial puede ser negativa y carece de algunas propiedades de la entropía para variables aleatorias discretas, como la invariancia ante cambios de variable [49]. Entre todas las variables aleatorias con media  $\boldsymbol{\mu}$  y matriz de covarianza  $\boldsymbol{\Sigma}$  dadas,  $h(\mathbf{x})$  se maximiza cuando  $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .

Véa también: RV, pdf, media, matriz de covarianza, incertidumbre, modelo probabilístico.

**envenenamiento de datos** El envenenamiento de datos se refiere a la manipulación intencional (o fabricación) de puntos de datos para influir en el entrenamiento de un modelo de aprendizaje automático [51, 52]. La protección contra el envenenamiento de datos es especialmente importante en aplicaciones distribuidas de aprendizaje automático donde los conjuntos de datos están descentralizados.

Vea también: punto de datos, modelo, ML, conjunto de datos

**epígrafo** El epígrafo de una función real  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  es el conjunto de puntos que se encuentran sobre o en la grafo de la función:

$$\text{epi}(f) = \{(\mathbf{x}, t) \in \mathbb{R}^n \times \mathbb{R} \mid f(\mathbf{x}) \leq t\}.$$

Una función es convexo si y solo si su epígrafo es un conjunto convexo [36], [53].

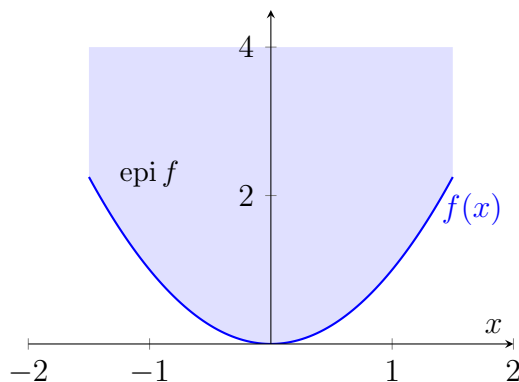


Figure 15: Epígrafo de la función  $f(x) = x^2$  (área sombreada).

Vea también: función, grafo, convexo.

**error cuadrático medio de estimación (MSEE)** Consideremos un método de aprendizaje automático que aprende parámetros del modelo  $\hat{\mathbf{w}}$  a partir de un conjunto de datos  $\mathcal{D}$ . Si interpretamos los puntos de datos en  $\mathcal{D}$  como realizaciones iid de una variable aleatoria  $\mathbf{z}$ , definimos el error de estimación como  $\Delta \mathbf{w} := \hat{\mathbf{w}} - \bar{\mathbf{w}}$ . Aquí,  $\bar{\mathbf{w}}$  representa los verdaderos parámetros del modelo de la distribución de probabilidad de  $\mathbf{z}$ . El error cuadrático medio de estimación se define como la expectativa  $\mathbb{E}\{\|\Delta \mathbf{w}\|^2\}$  del cuadrado de la norma euclidiana del error de estimación [54, 55].

Vea también: ML, parámetros del modelo, conjunto de datos, punto de datos, realización, error de estimación, distribución de probabilidad, expectation, norma

**error de entrenamiento** La pérdida promedio de una hipótesis al predecir las etiquetas de los puntos de datos en un conjunto de entrenamiento. A

veces también nos referimos al error de entrenamiento como la pérdida promedio mínimo que se logra mediante una solución de ERM.

Vea también: pérdida, hipótesis, etiqueta, punto de datos, conjunto de entrenamiento, ERM

**error de estimación** Consideremos varios puntos de datos, cada uno con un vector de atributos  $\mathbf{x}$  y una etiqueta  $y$ . En algunas aplicaciones, podemos modelar la relación entre el vector de atributos y la etiqueta de un punto de datos como  $y = \bar{h}(\mathbf{x}) + \varepsilon$ . Aquí  $\bar{h}$  representa la hipótesis verdadera subyacente y  $\varepsilon$  es un término de ruido que resume errores de modelado o etiquetado. El error de estimación incurrido por un método de aprendizaje que aprende una hipótesis  $\hat{h}$ , por ejemplo usando ERM, se define como  $\hat{h}(\mathbf{x}) - \bar{h}(\mathbf{x})$ , para algún vector de atributos dado. En un espacio de hipótesis paramétrico, donde las hipótesis se determinan mediante parámetros del modelo  $\mathbf{w}$ , podemos definir el error de estimación como  $\Delta\mathbf{w} = \hat{\mathbf{w}} - \bar{\mathbf{w}}$  [28, 55].

Vea también: punto de datos, vector de atributos, etiqueta, hipótesis, ML, ERM, espacio de hipótesis, parámetros del modelo

**error de validación** Consideremos una hipótesis  $\hat{h}$  obtenida por algún método de aprendizaje automático, por ejemplo, utilizando ERM en un conjunto de entrenamiento. La pérdida promedio de  $\hat{h}$  en un conjunto de validación, que es diferente del conjunto de entrenamiento, se denomina error de validación.

Vea también: hipótesis, ML, ERM, conjunto de entrenamiento, pérdida, conjunto de validación

**espacio de atributos** El espacio de atributos de una aplicación o método de aprendizaje automático está constituido por todos los valores potenciales que el vector de atributos de un punto de datos puede asumir. Una elección común para el espacio de atributos es el espacio euclidiano  $\mathbb{R}^d$ , donde la dimensión  $d$  es el número de atributos individuales de un punto de datos.

Vea también: ML, vector de atributos, punto de datos, espacio euclidiano, atributo

**espacio de etiquetas** Consideremos una aplicación de aprendizaje automático que involucra puntos de datos caracterizados por atributos y etiquetas. El espacio de etiquetas está constituido por todos los valores potenciales que una etiqueta de un punto de datos puede asumir. Los métodos de regresión, que buscan predecir etiquetas numéricas a menudo utilizan el espacio de etiquetas  $\mathcal{Y} = \mathbb{R}$ . Los métodos de clasificación binaria utilizan un espacio de etiquetas que consiste de dos elementos diferentes, por ejemplo,  $\mathcal{Y} = \{-1, 1\}$ ,  $\mathcal{Y} = \{0, 1\}$ , o  $\mathcal{Y} = \{\text{“imagen de gato”}, \text{“sin imagen de gato”}\}$ .

Vea también: ML, punto de datos, atributo, etiqueta, Regresión, clasificación

**espacio de Hilbert** Un espacio de Hilbert es un espacio vectorial lineal equipado con un producto interno entre pares de vectores. Un ejemplo importante de espacio de Hilbert es el espacio euclidiano  $\mathbb{R}^d$ , para alguna dimensión  $d$ , que consiste en vectores euclidianos  $\mathbf{u} = (u_1, \dots, u_d)^T$  junto con el producto interno  $\mathbf{u}^T \mathbf{v}$ .



Vea también: espacio euclidiano.

**espacio de hipótesis** Cada método práctico de aprendizaje automático utiliza un espacio de hipótesis (o modelo)  $\mathcal{H}$ . El espacio de hipótesis de un método de aprendizaje automático es un subconjunto de todos los posibles mapeos del vector de atributos al espacio de etiquetas. La elección del diseño del espacio de hipótesis debe considerar los recursos computacionales disponibles y los aspectos estadísticos. Si la infraestructura computacional permite operaciones matriciales eficientes y existe una relación (aproximadamente) lineal entre un conjunto de atributos y una etiqueta, una elección útil para el espacio de hipótesis podría ser el modelo lineal.

Vea también: ML, hipótesis, modelo, espacio de atributos, espacio de etiquetas, aspectos estadísticos, atributo, etiqueta, modelo lineal.

**espacio de parámetros** El espacio de parámetros  $\mathcal{W}$  de un modelo de aprendizaje automático  $\mathcal{H}$  es el conjunto de todas las elecciones factibles para los parámetros del modelo (vea la Figura 16). Muchos métodos de aprendizaje automático importantes usan un modelo que está parametrizado por vectores del espacio euclidiano  $\mathbb{R}^d$ . Dos ejemplos ampliamente utilizados de modelos parametrizados son los modelos lineales y las redes profundas. El espacio de parámetros es entonces a menudo un subconjunto  $\mathcal{W} \subseteq \mathbb{R}^d$ , por ejemplo, todos los vectores  $\mathbf{w} \in \mathbb{R}^d$  con una norma menor a uno.

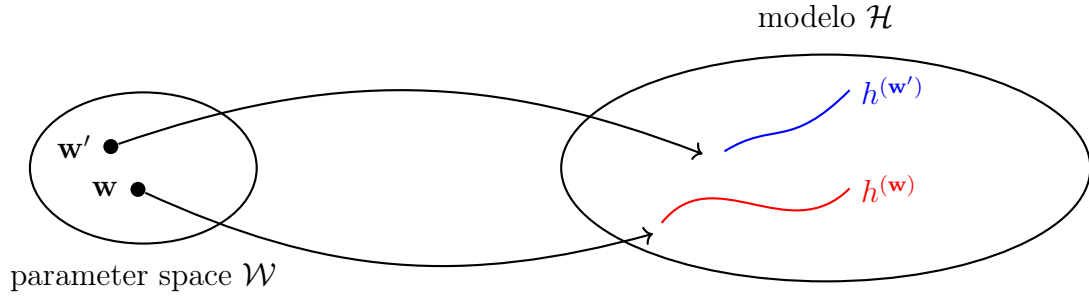


Figure 16: El espacio de parámetros  $\mathcal{W}$  de un modelo de aprendizaje automático  $\mathcal{H}$  consiste en todas las elecciones factibles para los parámetros del modelo. Cada elección  $\mathbf{w}$  para los parámetros del modelo seleccionan un mapa de hipótesis  $h^{(\mathbf{w})} \in \mathcal{H}$ .

Vea también: ML, modelo, parámetros del modelo, espacio euclidiano, modelo lineal, red profunda, norma, hipótesis.

**espacio de probabilidad** Un espacio de probabilidad es un modelo matemático de un proceso físico (un experimento aleatorio) con un resultado incierto. Formalmente, un espacio de probabilidad  $\mathcal{P}$  es una tripleta  $(\Omega, \mathcal{F}, P)$  donde:

- $\Omega$  es el espacio muestral que contiene todos los posibles resultados elementales del experimento aleatorio;
- $\mathcal{F}$  es una sigma-álgebra, es decir, una colección de subconjuntos de  $\Omega$  (llamados eventos) que satisface ciertas propiedades de clausura bajo operaciones de conjuntos;
- $P$  es una medida de probabilidad, una función que asigna una probabilidad  $P(\mathcal{A}) \in [0, 1]$  a cada evento  $\mathcal{A} \in \mathcal{F}$ . Esta función

debe satisfacer  $P(\Omega) = 1$  y

$$P\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) = \sum_{i=1}^{\infty} P(\mathcal{A}_i)$$

para cualquier secuencia numerable de eventos mutuamente disjuntos  $\mathcal{A}_1, \mathcal{A}_2, \dots$  en  $\mathcal{F}$ .

Los espacios de probabilidad constituyen la base para definir variables aleatorias y razonar sobre la incertidumbre en aplicaciones de aprendizaje automático [45, 46, 56].

Véa también: probabilidad, modelo, RV, ML

**espacio euclidiano** El espacio euclidiano  $\mathbb{R}^d$  de dimensión  $d \in \mathbb{N}$  consiste en vectores  $\mathbf{x} = (x_1, \dots, x_d)$ , con  $d$  entradas de valores reales  $x_1, \dots, x_d \in \mathbb{R}$ . Dicho espacio euclidiano está equipado con una estructura geométrica definida por el producto interno  $\mathbf{x}^T \mathbf{x}' = \sum_{j=1}^d x_j x'_j$  entre dos vectores cualesquiera  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$  [2].

**espacio vectorial** Un espacio vectorial (también llamado espacio lineal) es un conjunto de elementos (llamados vectores) cerrado bajo la suma de vectores y la multiplicación por escalares, es decir:

- Si  $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ , entonces  $\mathbf{x} + \mathbf{y} \in \mathcal{V}$ .
- Si  $\mathbf{x} \in \mathcal{V}$  y  $c \in \mathbb{R}$ , entonces  $c\mathbf{x} \in \mathcal{V}$ .
- En particular,  $\mathbf{0} \in \mathcal{V}$ .

El espacio euclidiano  $\mathbb{R}^n$  es un espacio vectorial. Los modelos lineales y las aplicaciones lineales operan dentro de tales espacios.

Véa también: espacio euclidiano, modelo lineal, aplicación lineal.

**espectrograma** Un espectrograma representa la distribución tiempo-frecuencia de la energía de una señal temporal  $x(t)$ . Intuitivamente, cuantifica la cantidad de energía de la señal presentedentro de un segmento de tiempo específico  $[t_1, t_2] \subseteq \mathbb{R}$  y en un intervalo de frecuencia  $[f_1, f_2] \subseteq \mathbb{R}$ . Formalmente, el espectrograma de una señal se define como el módulo al cuadrado de su transformada de Fourier de ventana corta (STFT, en inglés) [57]. La Figure 17 muestra una señal temporal junto con su espectrograma.

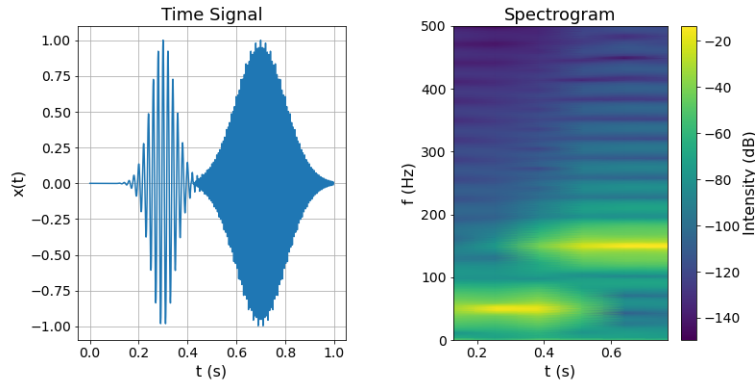


Figure 17: Izquierda: una señal temporal compuesta por dos pulsos gaussianos modulados. Derecha: representación de intensidad de su espectrograma.

La representación de intensidad del espectrograma puede considerarse como una imagen de la señal. Una estrategia sencilla para la clasificación de señales de audio consiste en introducir esta imagen en una red profunda desarrollada originalmente para tareas de clasificación de imágenes y detección de objetos [58]. Conviene señalar que, además del espectrograma, existen otras representaciones alternativas para describir

la distribución tiempo-frecuencia de la energía de una señal [59, 60].

Vea también: clasificación, red profunda

**esperanza** Consideremos un vector de atributos numérico  $\mathbf{x} \in \mathbb{R}^d$  que interpretamos como la realización de una variable aleatoria con una distribución de probabilidad  $p(\mathbf{x})$ . La esperanza de  $\mathbf{x}$  se define como la integral  $\mathbb{E}\{\mathbf{x}\} := \int \mathbf{x}p(\mathbf{x})$  [2, 46, 61]. Nótese que la esperanza solo está definida si esta integral existe, es decir, si la variable aleatoria es integrable. [2], [46], [61]. Fig. 18 ilustra la esperanza de una variable aleatoria discreta  $x$  que toma valores solo de un conjunto finito.

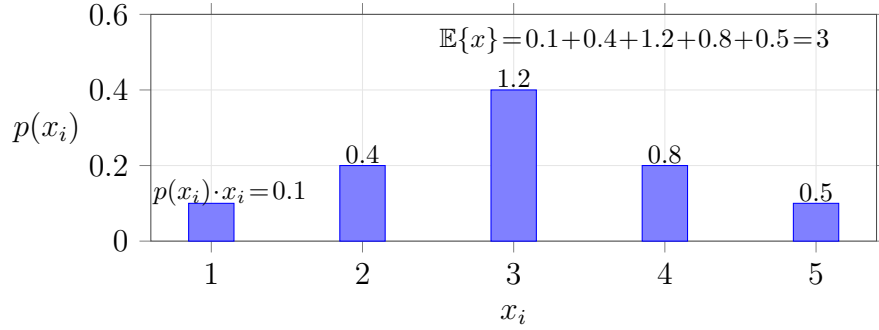


Figure 18: La esperanza de una variable aleatoria discreta RV  $x$  se obtiene sumando sus posibles valores  $x_i$ , ponderados por la correspondiente probabilidad  $p(x_i) = \mathbb{P}(x = x_i)$ .

Vea también: vector de atributos, realización, RV, distribución de probabilidad, probabilidad.

**esperanza-maximización (EM)** Considera un modelo de probabilidad  $\mathbb{P}(\mathbf{z}; \mathbf{w})$  para los puntos de datos  $\mathcal{D}$  generados en alguna aplicación de aprendizaje Automático. El estimador de maxima verosimilitud

para los parametros del modelo  $\mathbf{w}$  se obtienen maximizando  $\mathbb{P}(\mathcal{D}; \mathbf{w})$ . Sin embargo, el problema de optimización resultante puede ser computacionalmente desafiante. El algoritmo de expectativa-maximización (EM) aproxima el estimador de maxima verosimilitud introduciendo una variable aleatoria latente  $\mathbf{z}$  de modo que maximizar  $\mathbb{P}(\mathcal{D}, \mathbf{z}; \mathbf{w})$  sea más fácil [28, 62, 63]. Dado que no observamos  $\mathbf{z}$ , necesitamos estimarlo a partir del conjunto de datos observado  $\mathcal{D}$  utilizando una esperanza condicional. La estimación resultante  $\hat{\mathbf{z}}$  se utiliza para calcular una nueva estimación  $\hat{\mathbf{w}}$  resolviendo  $\max_{\mathbf{w}} \mathbb{P}(\mathcal{D}, \hat{\mathbf{z}}; \mathbf{w})$ . El punto clave es que la esperanza condicional  $\hat{\mathbf{z}}$  depende de los parametros del modelo  $\hat{\mathbf{w}}$ , que hemos actualizado en función de  $\hat{\mathbf{z}}$ . Por lo tanto, debemos volver a calcular  $\hat{\mathbf{z}}$ , lo que a su vez resulta en una nueva elección  $\hat{\mathbf{w}}$  para los parametros del modelo. En la práctica, repetimos el cálculo de la esperanza condicional (es decir, el E-step) y la actualización de los parametros del modelo (es decir, the M-step) hasta que se cumpla algún criterio de parada.

Vea también: modelo probabilístico, punto de datos, ML, máxima verosimilitud, parámetros del modelo, RV, conjunto de datos, expectation, criterio de parada.

**estimador de Bayes** Considera un modelo de probabilidad con una distribución de probabilidad conjunta  $p(\mathbf{x}, y)$  para los atributos  $\mathbf{x}$  y la etiqueta  $y$  de un punto de datos. Para una función de pérdida dada  $L(\cdot, \cdot)$ , denominamos a una hipótesis  $h$  como un estimador de Bayes si su riesgo  $\mathbb{E}\{L((\mathbf{x}, y), h)\}$  es el mínimo [54]. Nótese que la propiedad de una hipótesis de ser un estimador de Bayes depende de la distribución

de probabilidad subyacente y de la elección de la función de pérdida  $L(\cdot, \cdot)$ .

Vea también: modelo probabilístico, distribución de probabilidad, atributo, etiqueta, punto de datos, función de pérdida, hipótesis, riesgo, mínimo.

**estocástico** Un proceso o método se denomina estocástico si implica un componente aleatorio o está regido por leyes probabilísticas. En aprendizaje automático, los métodos estocásticos a menudo incorporan aleatoriedad por razones como la optimización (por ejemplo, descenso por gradiente estocástico) o la modelización de la incertidumbre (por ejemplo, mediante un modelo de probabilidad). Un proceso estocástico es una colección de variables aleatorias indexadas por tiempo o espacio, que se utilizan para modelar fenómenos aleatorios que evolucionan en el tiempo (p. ej., ruido en sensores o series temporales financieras).

Vea también: ML, SGD, incertidumbre, modelo probabilístico, RV, modelo estocástico de bloques (SBM).

**etiqueta** Una es un hecho de nivel superior o una cantidad de interés asociada a un punto de datos. Por ejemplo, si el punto de datos es una imagen, la etiqueta podría indicar si la imagen contiene un gato o no. Los sinónimos de etiqueta, comúnmente utilizados en dominios específicos, incluyen "variable de respuesta", "variable de salida" y "objetivo" [24], [25], [26].

Vea también: punto de datos

**experto** El aprendizaje automático tiene como objetivo aprender una hipótesis  $h$  que prediga con precisión la etiqueta de un punto de datos basado

en sus atributos Medimos el error de prediccion utilizando una funcion de perdida. Idealmente, buscamos una hipótesis que incurra en la perdida mínima para cualquier punto de datos. Podemos hacer este objetivo más preciso mediante la suposición de independencia e idéntica distribución y utilizando el riesgo de bayes como referencia (baseline) para la perdida promedio de una hipótesis. Una manera alternativa de obtener una referencia es utilizar la hipótesis  $h'$  aprendida por un método de aprendizaje automático existente. A esta hipótesis  $h'$  la denominamos experto [18]. Los métodos de minimización de arrepentimiento aprenden una hipótesis que incurre en una perdida comparable a la del mejor experto [17, 18].

Vea también: ML, hipótesis, etiqueta, glsdatapoint, atributo, predicción, función de pérdida, pérdida, punto de datos, suposición de independencia e idéntica distribución (suposición i.i.d.), riesgo de Bayes, referencia (baseline), Arrepentimiento (regret)

**explicabilidad** Definimos la (subjetiva) explicabilidad de un método de aprendizaje automático como el nivel de simulabilidad [64] de las predicciones entregadas por un sistema de aprendizaje automático a un usuario humano. Se pueden construir medidas cuantitativas para la explicabilidad (subjetiva) de un modelo entrenado comparando sus predicciones con las predicciones proporcionadas por un usuario en un conjunto de prueba [64, 65]. Alternativamente, podemos usar modelos probabilísticos para los datos y medir la explicabilidad de un modelo de aprendizaje automático entrenado mediante la entropía condicional (diferencial) de sus predicciones, dadas las predicciones del usuario [66, 67].



Vea también: ML, predicción, modelo, conjunto de prueba, modelo probabilístico, datos

### **Explicaciones Locales Interpretables e Independientes del Modelo (LIME)**

Consideremos un modelo entrenado (o una hipótesis aprendida)  $\hat{h} \in \mathcal{H}$ , que asigna el vector de atributos de un punto de datos a la predicción  $\hat{y} = \hat{h}$ . Las explicaciones Locales Interpretables e Independientes del Modelo (LIME) son una técnica para explicar el comportamiento de  $\hat{h}$ , localmente, alrededor de un punto de datos con vector de atributos  $\mathbf{x}^{(0)}$  [68]. La explicación se da en forma de una aproximación local  $g \in \mathcal{H}'$  de  $\hat{h}$  (véa Fig. ). Esta aproximación puede obtenerse mediante una instancia de ERM con un conjunto de entrenamiento diseñado cuidadosamente. En particular, el conjunto de entrenamiento consiste en puntos de datos con vector de atributos  $\mathbf{x}$  cercana a  $\mathbf{x}^{(0)}$  y la (pseudo-)etiqueta  $\hat{h}(\mathbf{x})$ . Nótese que podemos utilizar un modelo  $\mathcal{H}'$  diferente para la aproximación que el modelo original  $\mathcal{H}$ . Por ejemplo, podemos usar un árbol de decisión para aproximar (localmente) una red profunda. Otra elección ampliamente utilizada para  $\mathcal{H}'$  es el modelo lineal.

Vea también: modelo, hipótesis, vector de atributos, punto de datos, predicción, ERM, conjunto de entrenamiento, árbol de decisión, red profunda, modelo lineal

**explicación** Una manera para hacer que los métodos de aprendizaje automático sean transparentes consiste en proporcionar una explicación junto con la predicción generada por el método aprendizaje automático. Las explicaciones pueden adoptar muchas formas diferentes. Pueden ser

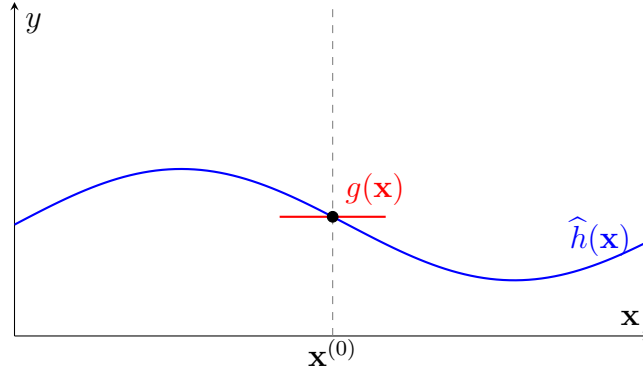


Figure 19: Para explicar un modelo  $\hat{h} \in \mathcal{H}$  entrenado, alrededor de un vector de atributos  $\mathbf{x}^{(0)}$ , podemos usar una aproximación local  $g \in \mathcal{H}'$ .

un texto natural o una medida cuantitativa que indique la importancia de atributos individuales de un punto de datos [69]. También podemos usar formas visuales, como los mapas de intensidad usados en tareas de clasificación de imágenes [70].

Vea también: ML, predicción, punto de datos, clasificación

**familias exponenciales en red (nExpFam)** Una colección de familias exponenciales, cada una de ellas asignada a un nodo de un red de aprendizaje federado (red FL). Los parametros del modelo están acoplados a través de la estructura de la red al requerir que tengan una pequeña variación total generalizada (GTV) [71].

Vea también: red FL, GTV

**FedGD** Un algoritmo distribuido de aprendizaje federado que puede implementarse como paso de mensajes a través de una red de aprendizaje federado.

Véa también: FL, algoritmo distribuido, red FL, paso de gradiente, métodos de gradiente.

**FedProx** FedProx se refiere a un algoritmo iterativo de aprendizaje federado que alterna entre entrenar modelos locales por separado y combinar los parametros del modelo locales actualizados. A diferencia del promedio federado, que utiliza descenso de gradiente estocástico para entrenar los modelos locales, FedProx usa un operador proximal para el entrenamiento [72].

Vea también: algoritmo, modelo local, parámetros del modelo, FedAvg, SGD, operador proximal

**FedRelax** Un algoritmo distribuido de aprendizaje federado.

Véa también: FL, algoritmo distribuido.

**FedSGD** Un algoritmo distribuido de aprendizaje federado que puede implementarse como paso de mensajes a través de una red de aprendizaje federado.

Véa también: FL, algoritmo distribuido, red FL, paso de gradiente, métodos de gradiente, SGD.

**filtración de privacidad** Consideremos una aplicación de aprendizaje automático que procesa un conjunto de datos  $\mathcal{D}$  y produce una salida, como las predicciones obtenidas para nuevos puntos de datos. Se produce una filtración de privacidad cuando la salida contiene información privada sobre un atributo de un punto de datos (que podría representar a una persona) en  $\mathcal{D}$ . Basado en el modelo de probabilidad para la generación de los datos, podemos medir la filtración de privacidad

usando la información mutua (MI) entre la salida y el atributo sensible. Otra medida cuantitativa de la filtración de privacidad es la privacidad diferencial (DP). Las relaciones entre las diferentes medidas de filtración de privacidad han sido estudiadas en la literatura (véa [73]).

Vea también: ML, conjunto de datos, predicción, punto de datos, atributo, modelo probabilístico, datos, MI, privacidad diferencial (DP)

**frontera de decisión** Consideremos un mapa de hipótesis  $h$  que recibe un vector de atributos  $\mathbf{x} \in \mathbb{R}^d$  y entrega un valor de un conjunto finito  $\mathcal{Y}$ . La frontera de decisión de  $h$  es el conjunto de vectores  $\mathbf{x} \in \mathbb{R}^d$  que se encuentran entre diferentes regiones de decisión. Más precisamente, un vector  $\mathbf{x}$  pertenece a la frontera de decisión si, y solo si, cada entorno  $\{\mathbf{x}' : \|\mathbf{x} - \mathbf{x}'\| \leq \varepsilon\}$ , para cualquier  $\varepsilon > 0$ , contiene al menos dos vectores con diferentes valores de función.

Vea también: hipótesis, atributo, región de decisión, entorno.

**fuertemente convexa** Una función real diferenciable de valor continuo  $f(\mathbf{x})$  es fuertemente convexa con coeficiente  $\sigma$  si cumple:  $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + (\sigma/2) \|\mathbf{y} - \mathbf{x}\|_2^2$  [74], [75, Sec. B.1.1].

Vea también: diferenciable, convexo

**función** Una función entre dos conjuntos  $\mathcal{U}$  y  $\mathcal{V}$  asigna a cada elemento  $u \in \mathcal{U}$  exactamente un elemento  $v \in \mathcal{V}$  [2]. Esto se escribe como  $f : \mathcal{U} \rightarrow \mathcal{V}$ , donde  $\mathcal{U}$  es el dominio y  $\mathcal{V}$  el codominio de  $f$ . Es decir, una función  $f$  define una salida única  $f(u) \in \mathcal{V}$  para cada entrada  $u \in \mathcal{U}$ . Para más detalles,

**función característica** La función característica de una RV real  $x$  es la

función [46, Sección 26]

$$\phi_x(t) := \mathbb{E} \exp(jtx) \text{ con } j = \sqrt{-1}.$$

La función característica determina de forma única la distribución de probabilidad de  $x$ .

Véase también: distribución de probabilidad, RV.

**función cuadrática** Una función  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  de la forma

$$f(\mathbf{w}) = \mathbf{w}^T \mathbf{Q} \mathbf{w} + \mathbf{q}^T \mathbf{w} + a,$$

donde  $\mathbf{Q} \in \mathbb{R}^{d \times d}$  es una matriz,  $\mathbf{q} \in \mathbb{R}^d$  es un vector y  $a \in \mathbb{R}$  es un escalar.

**función de activación** Cada neurona artificial dentro de una red neuronal artificial (RNA) se le asigna una función de activación  $\sigma(\cdot)$  que transforma una combinación ponderada de sus entradas  $x_1, \dots, x_d$  en un único valor de salida:  $a = \sigma(w_1 x_1 + \dots + w_d x_d)$ . Cada neurona está parametrizada por los pesos  $w_1, \dots, w_d$ .

Vea también: ANN, pesos

**función de densidad de probabilidad (pdf)** La probabilidad densidad de probabilidad  $p(x)$  de una variable aleatoria con valores reales  $x \in \mathbb{R}$  es una representación particular de su distribución de probabilidad. Si la función de densidad de probabilidad existe, se puede usar para calcular la probabilidad de que  $x$  tome un valor de un conjunto (medible)  $\mathcal{B} \subseteq \mathbb{R}$  mediante  $\mathbb{P}(x \in \mathcal{B}) = \int_{\mathcal{B}} p(x') dx'$  [5, Ch. 3]. La función de densidad de probabilidad de una variable aleatoria vectorial  $\mathbf{x} \in \mathbb{R}^d$  (si existe)

permite calcular la probabilidad de que  $\mathbf{x}$  pertenezca a una región (medible)  $\mathcal{R}$  mediante  $\mathbb{P}(\mathbf{x} \in \mathcal{R}) = \int_{\mathcal{R}} p(\mathbf{x}') dx'_1 \dots dx'_d$  [5, Ch. 3].

Vea también: RV, distribución de probabilidad, probabilidad.

**función de pérdida** Una función de pérdida es una aplicación

$$L : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \rightarrow \mathbb{R}_+ : ((\mathbf{x}, y), h) \mapsto L((\mathbf{x}, y), h).$$

Asigna un número real no negativo (es decir, la pérdida)  $L((\mathbf{x}, y), h)$  a un par que consiste en un punto de datos, con atributos  $\mathbf{x}$  y una etiqueta  $y$ , y una hipótesis  $h \in \mathcal{H}$ . El valor  $L((\mathbf{x}, y), h)$  cuantifica la discrepancia entre la etiqueta verdadera  $y$  y la predicción  $h(\mathbf{x})$ . Valores bajos (cerca de cero) de  $L((\mathbf{x}, y), h)$  indican una discrepancia menor entre la predicción  $h(\mathbf{x})$  y la etiqueta  $y$ . La Figura 20 muestra una función de pérdida para un punto de datos dado, con atributos  $\mathbf{x}$  y etiqueta  $y$ , como función de la hipótesis  $h \in \mathcal{H}$ .

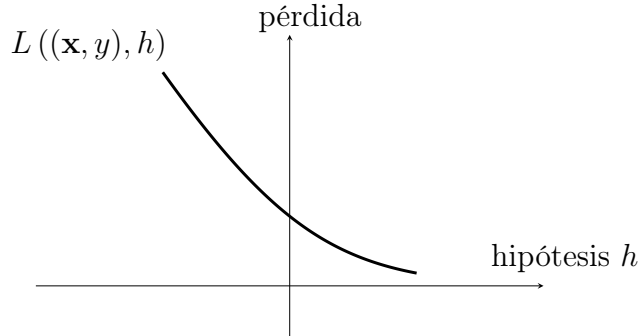


Figure 20: Alguna funcion de perdida  $L((\mathbf{x}, y), h)$  para un punto de datos fijado, con vector de atributos  $\mathbf{x}$  y etiqueta  $y$ , y una hipótesis variable  $h$ . Los métodos de aprendizaje automático intentan encontrar (o aprender) una hipótesis que incurra en una pérdida mínima.

Vea también: pérdida, punto de datos, atributo, etiqueta, hipótesis, predicción, vector de atributos, ML.

**función objetivo** Una función objetivo es un mapa que asigna a cada valor de una variable de optimización, como los parametros del modelo  $\mathbf{w}$  de una hipótesis  $h^{(\mathbf{w})}$ , un valor objetivo  $f(\mathbf{w})$ . El valor objetivo  $f(\mathbf{w})$  podría ser el riesgo o el riesgo empírico de una hipótesis  $h^{(\mathbf{w})}$ .

Vea también: parámetros del modelo, hipótesis, riesgo, riesgo empírico

**generalización** Muchos de los sistemas actuales de aprendizaje automático (y inteligencia artificial) se basan en ERM: En esencia, entrenan un modelo (es decir, aprenden una hipótesis  $\hat{h} \in \mathcal{H}$ ) minimizando la pérdida promedio (o riesgo empírico) en algunos puntos de datos  $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$ , que sirven como un conjunto de entrenamiento  $\mathcal{D}^{(\text{train})}$ . La generalización se refiere a la capacidad de un método de aprendizaje automático para desempeñarse bien fuera del conjunto de entrenamiento. Cualquier teoría matemática de la generalización necesita algún concepto matemático para el "fuera del conjunto de entrenamiento." Por ejemplo, la teoría del aprendizaje estadístico utiliza un modelo de probabilidad como la suposición iid para la generación de datos: los puntos de datos en el conjunto de entrenamiento son realizaciones iid de alguna distribución de probabilidad subyacente  $p(\mathbf{z})$ . Un modelo de probabilidad nos permite explorar el exterior del conjunto de entrenamiento generando adicionales realizaciones iid de  $p(\mathbf{z})$ . Además, el uso de la suposición i.i.d. nos permite definir el riesgo de un modelo entrenado  $\hat{h} \in \mathcal{H}$  como la expectativa de la pérdida  $\bar{L}(\hat{h})$ . Además, podemos utilizar límites de concentración

o resultados de convergencia para secuencias de variables aleatorias iid para limitar la desviación entre el riesgo empírico  $\hat{L}(\hat{h}|\mathcal{D}^{(\text{train})})$  de un modelo entrenado y su riesgo [44]. También es posible estudiar la generalización sin utilizar modelos de probabilidad. Por ejemplo, podríamos utilizar perturbaciones (determinísticas) de los puntos de datos en el conjunto de entrenamiento para estudiar su exterior. En general, deseamos que el modelo entrenado sea robusto, es decir, que sus predicciones no cambien demasiado ante pequeñas perturbaciones de un punto de datos. Considere un modelo entrenado para detectar un objeto en una imagen capturada por un teléfono inteligente. El resultado de la detección no debería cambiar si enmascaramos un pequeño número de píxeles seleccionados aleatoriamente en la imagen [76].

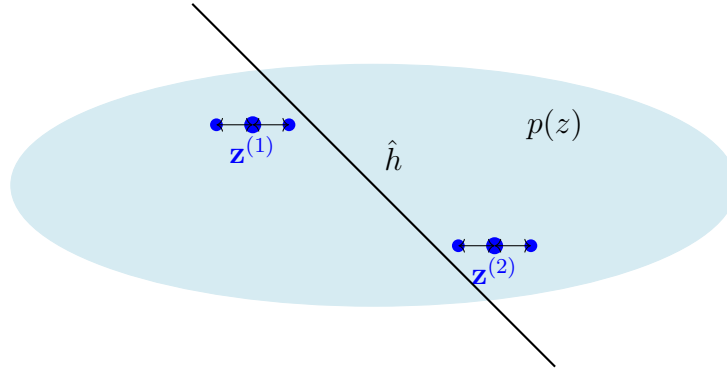


Figure 21: Dos puntos de datos  $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$  que se utilizan como un conjunto de entrenamiento para aprender una hipótesis  $\hat{h}$  mediante ERM. Podemos evaluar  $\hat{h}$  fuera de  $\mathcal{D}^{(\text{train})}$  ya sea mediante una suposición iid con alguna distribución de probabilidad subyacente  $p(\mathbf{z})$  o mediante la perturbación de los puntos de datos.



**gradiente** Para una función de valor real  $f : \mathbb{R}^d \rightarrow \mathbb{R} : \mathbf{w} \mapsto f(\mathbf{w})$ , un vector  $\mathbf{g}$  tal que  $\lim_{\mathbf{w} \rightarrow \mathbf{w}'} \frac{f(\mathbf{w}) - (f(\mathbf{w}') + \mathbf{g}^T(\mathbf{w} - \mathbf{w}'))}{\|\mathbf{w} - \mathbf{w}'\|} = 0$  se denomina gradiente de  $f$  en  $\mathbf{w}'$ . Si existe tal vector, se denota como  $\nabla f(\mathbf{w}')$  o  $\nabla f(\mathbf{w})|_{\mathbf{w}'}$  [2].

**grado de nodo** El grado  $d^{(i)}$  de un nodo  $i \in \mathcal{V}$  en un grafo no dirigido, es el número de sus vecinos, es decir,  $d^{(i)} := |\mathcal{N}^{(i)}|$ .

Vea también: grafo, vecinos

**grado de pertenencia** El grado de pertenencia es un número que indica en qué medida un punto de datos pertenece a un clúster [6, Ch. 8]. Este grado puede interpretarse como una asignación blanda (\*soft\*) al clúster. Los métodos de agrupamiento suave pueden codificar el grado de pertenencia mediante un número real en el intervalo  $[0, 1]$ . El agrupamiento rígido se obtiene como caso extremo, cuando el grado de pertenencia solo toma los valores 0 o 1.

Vea también: punto de datos, cluster, Agrupamiento suave, Agrupamiento rígido

**grafo** Un grafo  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  es un par compuesto por un conjunto de nodos  $\mathcal{V}$  y un conjunto de aristas  $\mathcal{E}$ . En su forma más general, un grafo se especifica por una función que asigna a cada arista  $e \in \mathcal{E}$  un par de nodos [77]. Un grupo importante de grafos son los grafos no dirigidos. Un grafo simple no dirigido es obtenida identificando cada arista  $e \in \mathcal{E}$  con dos nodos diferentes  $\{i, i'\}$ . Los grafos etiquetados asignan un peso numérico específico pesos:  $A_e$  a cada arista  $e \in \mathcal{E}$ .

Vea también: pesos

**grafo conexo** Un grafo no dirigido  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  es conexo si todo subconjunto

no vacío  $\mathcal{V}' \subset \mathcal{V}$  tiene al menos una arista que lo conecta con  $\mathcal{V} \setminus \mathcal{V}'$ .

Vea también: grafo

**grafo de Erdős-Rényi (grafo ER)** Un grafo ER es un modelo de probabilidad para grafos definido sobre un conjunto dado de nodos  $i = 1, \dots, n$ . Una forma de definir el grafo ER es mediante una colección de variables aleatorias iid binarias  $b^{\{i,i'\}} \in \{0, 1\}$ , para cada par de nodos distintos  $i, i'$ . Una realización específica del grafo ER contiene una arista  $\{i, i'\}$  si y solo si  $b^{\{i,i'\}} = 1$ . El grafo ER está parametrizado por el número  $n$  de nodos y la probabilidad  $\mathbb{P}(b^{\{i,i'\}} = 1)$ .

Vea también: grafo, modelo probabilístico, i.i.d., RV, realización, probabilidad.

**grafo de similitud** Algunas aplicaciones de aprendizaje automático generan puntos de datos que están relacionados por una noción de similitud específica del dominio. Estas similitudes pueden ser representadas convenientemente utilizando un grafo de similitud  $\mathcal{G} = (\mathcal{V} := \{1, \dots, m\}, \mathcal{E})$ . El nodo  $r \in \mathcal{V}$  representa el  $r$ -ésimo punto de datos. Dos nodos están conectados por una arista no dirigida si los puntos de datos correspondientes son similares.

Vea también: ML, punto de datos, grafo.

**hipótesis** Una hipótesis se refiere a un mapa (o función)  $h : \mathcal{X} \rightarrow \mathcal{Y}$  que va del espacio de atributos  $\mathcal{X}$  al espacio de etiquetas  $\mathcal{Y}$ . Dado un punto de datos con atributos  $\mathbf{x}$ , utilizamos un mapa de hipótesis  $h$  para estimar (o aproximar) la etiqueta  $y$  mediante la predicción  $\hat{y} = h(\mathbf{x})$ . El aprendizaje automático se centra en aprender (o encontrar) un mapa

de hipótesis  $h$  tal que  $y \approx h(\mathbf{x})$  para cualquier punto de datos (con atributos  $\mathbf{x}$  y etiqueta  $y$ ).

Vea también: espacio de etiquetas, punto de datos, atributo, etiqueta, predicción, ML

**histograma** Un histograma considera un conjunto de datos  $\mathcal{D}$  que consiste en  $m$  puntos de datos  $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$ , cada uno de los cuales pertenece a una celda  $[-U, U] \times \dots \times [-U, U] \subseteq \mathbb{R}^d$  con longitud de lado  $U$ . Dividimos esta celda uniformemente en celdas elementales más pequeñas con longitud de lado  $\Delta$ . El histograma de  $\mathcal{D}$  asigna a cada celda elemental la fracción correspondiente del punto de datos en  $\mathcal{D}$  que caen dentro de esa celda.

Vea también: conjunto de datos, punto de datos

**incertidumbre** En el contexto del aprendizaje automático, la incertidumbre se refiere a la existencia de múltiples resultados o explicaciones plausibles con base en los datos disponibles. Por ejemplo, la predicción  $\hat{h}(\mathbf{x})$  producida por un modelo de aprendizaje automático entrenado  $\hat{h}$  suele reflejar un rango de posibles valores para la verdadera etiqueta de un punto de datos dado. Cuanto más amplio es este rango, mayor es la incertidumbre asociada. La teoría de la probabilidad nos permite representar, cuantificar y razonar sobre la incertidumbre de manera matemáticamente rigurosa.

Véase también: modelo probabilístico, riesgo, entropía, varianza.

**independientes e idénticamente distribuidos (i.i.d.)** Es útil interpretar los puntos de datos  $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$  como realizaciones de variables

aleatorias i.i.d., es decir, independientes e idénticamente distribuidos, con una distribución de probabilidad común. Si estas variables aleatorias son de valor continuo, su función de densidad de probabilidad conjunta se expresa como  $p(\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}) = \prod_{r=1}^m p(\mathbf{z}^{(r)})$ , donde  $p(\mathbf{z})$  es la función de densidad de probabilidad marginal común de las variables aleatorias subyacentes.

Vea también: punto de datos, realización, RV, distribución de probabilidad, pdf.

**información mutua (MI)** La información mutua  $I(\mathbf{x}; y)$  entre dos variables aleatorias  $\mathbf{x}$ ,  $y$  definidas en el mismo espacio de probabilidad está dada por [49]

$$I(\mathbf{x}; y) := \mathbb{E} \left\{ \log \frac{p(\mathbf{x}, y)}{p(\mathbf{x})p(y)} \right\}.$$

Es una medida de qué tan bien podemos estimar  $y$  basándonos únicamente en  $\mathbf{x}$ . Un valor grande de  $I(\mathbf{x}; y)$  indica que  $y$  puede predecirse bien únicamente a partir de  $\mathbf{x}$ . Esta predicción podría obtenerse mediante una hipótesis aprendida por un método de aprendizaje automático basado en ERM.

Vea también: RV, espacio de probabilidad, predicción, hipótesis, ERM, ML.

**Instituto Meteorológico de Finlandia (FMI)** El FMI es una agencia gubernamental responsable de recopilar y reportar datos meteorológicos en Finlandia.

Vea también: datos.

**inteligencia artificial (IA)** La IA se refiere a sistemas que se comportan de

manera racional en el sentido de maximizar una recompensa a largo plazo. El enfoque basado en aprendizaje automático para la IA consiste en entrenar un modelo para predecir acciones óptimas. Estas predicciones se calculan a partir de observaciones sobre el estado del entorno. La elección de la función de pérdida distingue las aplicaciones de IA de las aplicaciones de aprendizaje automático más básicas. Los sistemas de IA rara vez tienen acceso a un conjunto de entrenamiento etiquetado que permita medir la pérdida promedio para cualquier posible elección de parámetros del modelo. En su lugar, los sistemas de IA utilizan señales de recompensa observadas para obtener una estimación puntual de la pérdida incurrida por la elección actual de parámetros del modelo. Vea también: recompensa, ML, modelo, función de pérdida, conjunto de entrenamiento, pérdida, parámetros del modelo.

**inteligencia artificial confiable (IA confiable)** Además de los aspectos computacionales y los aspectos estadísticos, un tercer aspecto principal en el diseño de métodos de aprendizaje automático es su confiabilidad [78]. La Unión Europea ha propuesto siete requisitos clave (KRs) para una inteligencia artificial confiable (que típicamente se basa en métodos de aprendizaje automático) [79]:

- 1) KR1 - Agencia y supervisión humana;
- 2) KR2 - Robustez técnica y seguridad;
- 3) KR3 - Privacidad y gobernanza de los datos;
- 4) KR4 - Transparencia;
- 5) KR5 - Diversidad, no discriminación y equidad;

6) KR6 - Bienestar social y ambiental;

7) KR7 - Responsabilidad.

.

Vea también: aspectos computacionales, aspectos estadísticos, ML, AI

**Interfaz de programación de aplicaciones (API)** Una API es un mecanismo formal para permitir que componentes de software interactúen de manera estructurada [80]. En el contexto de aprendizaje Automático, las APIs se utilizan frecuentemente para hacer accesible un modelo de aprendizaje automático entrenado a diferentes tipos de usuarios. Estos usuarios, que pueden ser otros ordenadores o humanos, pueden solicitar una predicción para la etiqueta de un punto de datos al proporcionar sus atributos. La estructura interna del modelo de aprendizaje automático permanece oculta para el usuario. Por ejemplo, considera un modelo de aprendizaje automático entrenado  $\hat{h}(x) := 2x + 1$ . Una API permite a un usuario enviar el valor de atributo  $x = 3$  y obtener la respuesta  $\hat{h}(3) = 7$  sin conocimiento de la estructura detallada del modelo de aprendizaje automático o su entrenamiento. En la práctica, el modelo de aprendizaje automático suele estar alojado en un ordenador (es decir, un servidor) conectado a internet. Otro ordenador (es decir, un cliente) envía los atributos de un punto de datos al servidor, que luego calcula  $\hat{h}(\mathbf{x})$  y devuelve el resultado al sistema externo. Las APIs ayudan a modularizar el desarrollo de aplicaciones de aprendizaje automático al desacoplar tareas específicas. Por ejemplo, un equipo puede concentrarse en desarrollar y entrenar el modelo, mientras que otro equipo se

encarga de la interacción con el usuario y la integración del modelo en aplicaciones.

Vea también: ML, modelo, vector de atributos, punto de datos, predicción, atributo.

**interpretabilidad** Un método de aprendizaje automático es interpretable para una persona usuaria si esta puede comprender su proceso de decisión. Una forma de desarrollar una definición precisa de interpretabilidad es mediante el concepto de simulabilidad, es decir, la capacidad de una persona de simular mentalmente el comportamiento del modelo [64], [67], [81], [82], [83]. La idea es la siguiente: si una persona entiende un método de aprendizaje automático, entonces debería poder anticipar sus predicciones en un conjunto de prueba. Ilustramos este conjunto de prueba en la Fig. 22, que también muestra dos hipótesis aprendidas:  $\hat{h}$  y  $\hat{h}'$ . El método de aprendizaje automático que produce la hipótesis  $\hat{h}$  es interpretable para una persona familiarizada con el concepto de mapa lineal. Dado que  $\hat{h}$  corresponde a un mapa lineal, la persona puede anticipar sus predicciones sobre el conjunto de prueba. En cambio, el método que produce  $\hat{h}'$  no es interpretable, porque su comportamiento ya no coincide con las expectativas de la persona.

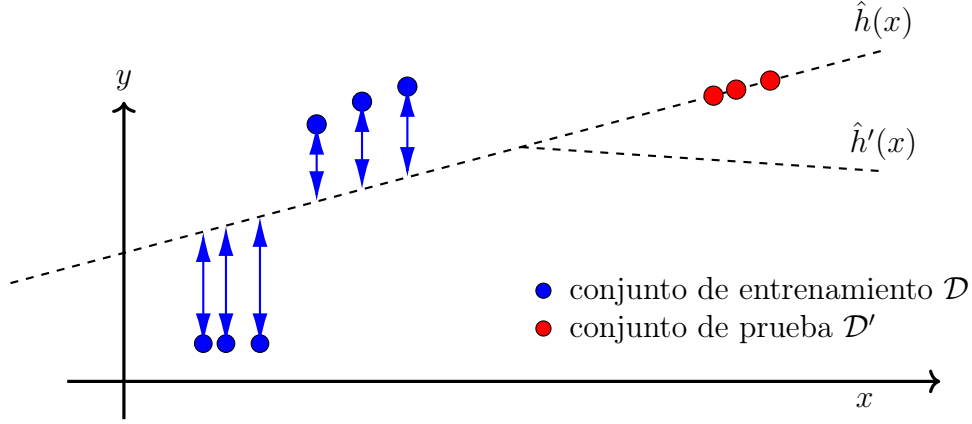


Figure 22: Podemos evaluar la interpretabilidad de los modelos de aprendizaje automático entrenados  $\hat{h}$  y  $\hat{h}'$  comparando sus predicciones con pseudo-etiquetas generadas por una persona para  $\mathcal{D}'$ .

La noción de interpretabilidad está estrechamente relacionada con la de explicabilidad, ya que ambas buscan que los métodos de aprendizaje automático sean más comprensibles para las personas. En el contexto de la Fig. 22, la interpretabilidad de un método  $\hat{h}$  requiere que la persona pueda anticipar sus predicciones sobre cualquier conjunto de prueba. Esto contrasta con la explicabilidad, donde se apoya al usuario mediante explicaciones externas —como mapas de saliencia o ejemplos de referencia del conjunto de entrenamiento— para entender las predicciones de  $\hat{h}$  sobre un conjunto de prueba  $\mathcal{D}'$ .

Vea tambien: explicabilidad, trustworthy AI, regularización, LIME.

**Jacobi method** El método de Jacobi es un algoritmo para resolver sistemas de ecuaciones lineales de la forma  $\mathbf{Ax} = \mathbf{b}$ . Aquí,  $\mathbf{A} \in \mathbb{R}^{d \times d}$  es



una matriz cuadrada con elementos diagonales principales no nulos. El método construye una secuencia  $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots$  actualizando cada componente de  $\mathbf{x}^{(k)}$  de acuerdo con la fórmula

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right).$$

Es importante notar que todos los elementos  $x_1^{(k)}, \dots, x_d^{(k)}$  se actualizan de manera simultánea. La iteración anterior converge a una solución, es decir,  $\lim_{k \rightarrow \infty} \mathbf{x}^{(k)} = \mathbf{x}$ , bajo ciertas condiciones sobre la matriz  $\mathbf{A}$ , por ejemplo, si es estrictamente diagonal dominante o simétrica definida positiva [3], [84], [85]. Los métodos tipo Jacobi son atractivos para sistemas lineales grandes debido a su estructura paralelizable [16]. Podemos interpretar el método de Jacobi como una iteración de punto fijo. En efecto, al descomponer  $\mathbf{A} = \mathbf{D} + \mathbf{R}$ , donde  $\mathbf{D}$  es la parte diagonal de  $\mathbf{A}$ , podemos reescribir la ecuación  $\mathbf{Ax} = \mathbf{b}$  como una ecuación de punto fijo:

$$\mathbf{x} = \underbrace{\mathbf{D}^{-1}(\mathbf{b} - \mathbf{Rx})}_{\mathcal{F}\mathbf{x}},$$

lo que conduce a la iteración  $\mathbf{x}^{(k+1)} = \mathbf{D}^{-1}(\mathbf{b} - \mathbf{Rx}^{(k)})$ .

Por ejemplo, para el sistema lineal

$$\mathbf{Ax} = \mathbf{b}, \quad \text{where} \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix},$$

el método de Jacobi actualiza cada componente de  $\mathbf{x}$  como sigue:

$$\begin{aligned}x_1^{(k+1)} &= \frac{1}{a_{11}} \left( b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} \right), \\x_2^{(k+1)} &= \frac{1}{a_{22}} \left( b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)} \right), \\x_3^{(k+1)} &= \frac{1}{a_{33}} \left( b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)} \right).\end{aligned}$$

Vea tambien: algoritmo, método de optimización.

**kernel** Considere un conjunto de puntos de datos, cada uno representado por un vector de atributos  $\mathbf{x} \in \mathcal{X}$ , donde  $\mathcal{X}$  denota el espacio de atributos. Un kernel (de valores reales) es una función  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  que asigna a cada par de vector de atributos  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$  un número real  $K(\mathbf{x}, \mathbf{x}')$ . Este valor se interpreta típicamente como una medida de similitud entre  $\mathbf{x}$  y  $\mathbf{x}'$ . La propiedad definitoria de un kernel es que es simétrico, es decir,  $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$ , y que para cualquier conjunto finito de vectores de atributos  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$ , la matriz

$$\mathbf{K} = \begin{pmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \dots & K(\mathbf{x}_1, \mathbf{x}_n) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & \dots & K(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{x}_n, \mathbf{x}_1) & K(\mathbf{x}_n, \mathbf{x}_2) & \dots & K(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} \in \mathbb{R}^{n \times n}$$

es semi-positiva definida. Un kernel define naturalmente una transformación de un vector de atributos  $\mathbf{x}$  en una función  $\mathbf{z} = K(\mathbf{x}, \cdot)$ . La función  $\mathbf{z}$  asigna a una entrada  $\mathbf{x}' \in \mathcal{X}$  el valor  $K(\mathbf{x}, \mathbf{x}')$ . Podemos ver la función  $\mathbf{z}$  como un nuevo vector de atributos que pertenece a un espacio de atributos  $\mathcal{X}'$ , que típicamente es diferente de  $\mathcal{X}$ . Este

nuevo espacio de atributos  $\mathcal{X}'$  posee una estructura matemática particular: es un espacio de Hilbert con núcleo reproducible (RKHS, por sus siglas en inglés) [86], [87]. Dado que  $\mathbf{z}$  pertenece a un RKHS, que es un espacio de vectores, podemos interpretarlo como un vector de atributos generalizado. Nótese que un vector de atributos de longitud finita  $\mathbf{x} = (x_1, \dots, x_d)^T \in \mathbb{R}^d$  puede verse como una función  $\mathbf{x} : \{1, \dots, d\} \rightarrow \mathbb{R}$  que asigna un valor real a cada índice  $j \in \{1, \dots, d\}$ . Véa también: vector de atributos, espacio de atributos, espacio de Hilbert, método de kernel.

**ley de los grandes números** La ley de los grandes números se refiere a la convergencia del promedio de un número creciente (grande) de variables aleatorias iid hacia la media de su distribución de probabilidad común. Diferentes instancias de la ley de los grandes números se obtienen utilizando distintas nociones de convergencia [88].  
Vea también: i.i.d., RV, media, distribución de probabilidad.

**lote** En el contexto de descenso de gradiente estocástico, un lote se refiere a un subconjunto elegido al azar del conjunto total de conjunto de entrenamiento. Utilizamos los puntos locales de este subconjunto para estimar el gradiente del error de entrenamiento y, a su vez, actualizar los parámetros del modelo.  
Vea también: SGD, conjunto de entrenamiento, punto de datos, gradiente, error de entrenamiento, parámetros del modelo

**mapa de atributos** Un mapa de atributos se refiere a una función

$$\Phi : \mathcal{X} \rightarrow \mathcal{X}', \quad \mathbf{x} \mapsto \mathbf{x}'$$

que transforma un vector de atributos  $\mathbf{x} \in \mathcal{X}$  de un punto de datos en un nuevo vector de atributos  $\mathbf{x}' \in \mathcal{X}'$ , donde típicamente  $\mathcal{X}'$  es diferente de  $\mathcal{X}$ . La representación transformada  $\mathbf{x}'$  suele ser más útil que la original  $\mathbf{x}$ . Por ejemplo, la geometría de los puntos de datos puede volverse más lineal en  $\mathcal{X}'$ , lo que permite aplicar un modelo lineal a  $\mathbf{x}'$ . Esta idea es central en el diseño de metodos de kernel [86]. Otros beneficios de usar un mapa de atributos incluyen la reducción del sobreajuste y la mejora de la interpretabilidad [68]. Un caso de uso común es la visualización de datos, donde un mapa de atributos con dos dimensiones de salida permite representar punto de datos en un diagrama de dispersión 2D. Algunos métodos de aprendizaje automático emplean mapas de atributos entrenables, cuyos parametros se aprenden a partir de datos. Un ejemplo es el uso de capas ocultas en una red profunda, que actúan como mapa de atributos sucesivos [89]. Una forma fundamentada de entrenar un mapa de atributos es mediante ERM, utilizando una funcion de perdida que mide la calidad de reconstrucción, por ejemplo,  $L = \|\mathbf{x} - r(\mathbf{x}')\|^2$ , donde  $r(\cdot)$  es un mapa entrenable que intenta reconstruir  $\mathbf{x}$  a partir del vector de atributos transformado  $\mathbf{x}'$ . Véa también: atributo, aplicación, método de kernel, aprendizaje de atributos, PCA.

**matrix** A matrix of size  $m \times d$  is a two-dimensional array of numbers, denoted

as

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,d} \\ A_{2,1} & A_{2,2} & \dots & A_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \dots & A_{m,d} \end{bmatrix} \in \mathbb{R}^{m \times d}.$$

Here,  $A_{r,j}$  denotes the matrix entry in the  $r$ -th row and the  $j$ -th column. Matrices are useful representations of various mathematical objects [85]:

- **Systems of Linear Equations.** We can use a matrix to represent a system of linear equations

$$\begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \text{compactly as} \quad \mathbf{A}\mathbf{w} = \mathbf{y}.$$

One important example of systems of linear equations are the optimality condition for the parámetros del modelo within regresión lineal.

- **Aplicación lineals.** Consider a  $d$ -dimensional espacio vectorial  $\mathcal{U}$  and a  $m$ -dimensional espacio vectorial  $\mathcal{V}$ . If we fix a basis  $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(d)}$  for  $\mathcal{U}$  and a basis  $\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(m)}$  for  $\mathcal{V}$ , each matrix  $\mathbf{A} \in \mathbb{R}^{m \times d}$  naturally defines a linear map  $\alpha : \mathcal{U} \rightarrow \mathcal{V}$ :

$$\mathbf{u}^{(j)} \mapsto \sum_{r=1}^m A_{r,j} \mathbf{v}^{(r)}.$$

- **Conjunto de datoss.** We can use a matrix to represent a conjunto de datos: each row corresponds to a single punto de datos, and each column corresponds to a specific atributo or etiqueta of a punto de datos.

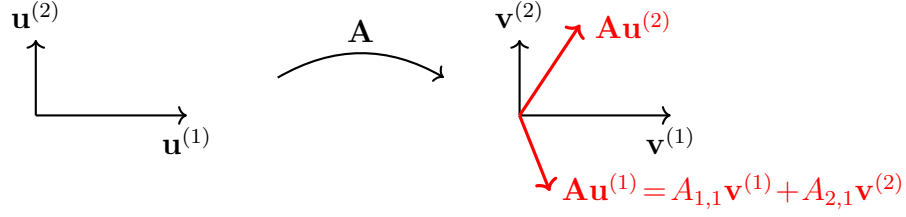


Figure 23: A matrix  $A$  defines a aplicación lineal between two espacio vectorials.

See also: aplicación lineal, conjunto de datos, modelo lineal.

**matriz de atributos** Considere un conjunto de datos  $\mathcal{D}$  con  $m$  puntos de datos caracterizados por vectores de atributos  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)} \in \mathbb{R}^d$ . Es conveniente agrupar los vectores de atributos individuales en una matriz de atributos matrix  $\mathbf{X} := (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)})^T$  de tamaño  $m \times d$ .  
Vea también: conjunto de datos, punto de datos, vector de atributos, atributo.

**matriz de confusión** Considere puntos de datos caracterizados por atributos  $\mathbf{x}$  y sus correspondientes etiquetas  $y$ . Las etiquetas toman valores en un espacio de etiquetas finito  $\mathcal{Y} = \{1, \dots, k\}$ . Para una hipótesis dada  $h$ , la matriz de confusión es una matriz de tamaño  $k \times k$  donde cada fila corresponde a un valor distinto de la etiqueta verdadera  $y \in \mathcal{Y}$  y cada columna a un valor distinto de la predicción  $h(\mathbf{x}) \in \mathcal{Y}$ . La entrada  $(c, c')$ -ésima de la matriz representa la fracción de puntos de datos con etiquetas verdadera  $y = c$  que son clasificadas como  $h(\mathbf{x}) = c'$ . La diagonal principal de la matriz contiene las fracciones de puntos de datos

correctamente clasificadas (es decir, aquellas para las que  $y = h(\mathbf{x})$ ). Las entradas fuera de la diagonal contienen las fracciones de puntos de datos que son mal clasificadas por  $h$ .

Vea tambien: etiqueta, espacio de etiquetas, hipótesis, clasificación.

**matriz de covarianza** La matriz de covarianza de una variable aleatoria  $\mathbf{x} \in \mathbb{R}^d$  se define como  $\mathbb{E}\left\{(\mathbf{x} - \mathbb{E}\{\mathbf{x}\})(\mathbf{x} - \mathbb{E}\{\mathbf{x}\})^T\right\}$ .

Vea también: RV.

**matriz de covarianza muestral** La matriz de matriz de covarianza muestral  $\hat{\Sigma} \in \mathbb{R}^{d \times d}$  para un conjunto dado vectores de atributos  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)} \in \mathbb{R}^d$  se define como

$$\hat{\Sigma} = (1/m) \sum_{r=1}^m (\mathbf{x}^{(r)} - \hat{\mathbf{m}})(\mathbf{x}^{(r)} - \hat{\mathbf{m}})^T.$$

Aqui, usamos la media muestral  $\hat{\mathbf{m}}$ .

Vea también: muestra, matriz de covarianza, vector de atributos, media muestral.

**matriz inversa** Una matriz inversa  $\mathbf{A}^{-1}$  se define para una matriz cuadrada  $\mathbf{A} \in \mathbb{R}^{n \times n}$  que sea de rango completo, es decir, que sus columnas sean linealmente independientes. En este caso, se dice que  $\mathbf{A}$  es invertible, y su inversa satisface

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}.$$

Una matriz cuadrada es invertible si y solo si su determinante es distinto de cero. Las matrices inversas son fundamentales para resolver sistemas de ecuaciones lineales y para la solución en forma cerrada de regresión lineal [42,84]. El concepto de matriz inversa puede extenderse a matrices

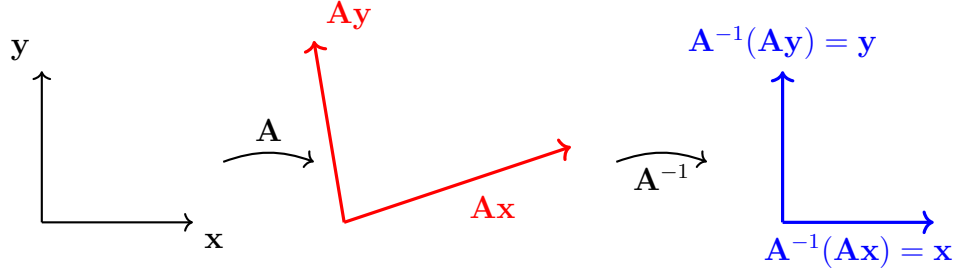


Figure 24: Una matriz  $\mathbf{A}$  representa una transformación lineal de  $\mathbb{R}^2$ . La matriz inversa  $\mathbf{A}^{-1}$  representa la transformación inversa.

que no son cuadradas o que no tienen rango completo. Se puede definir una “inversa por la izquierda”  $\mathbf{B}$  que satisface  $\mathbf{BA} = \mathbf{I}$ , o una “inversa por la derecha”  $\mathbf{C}$  que satisface  $\mathbf{AC} = \mathbf{I}$ . Para matrices rectangulares o singulares en general, la pseudoinversa de Moore–Penrose  $\mathbf{A}^+$  proporciona un concepto unificado de matriz inversa generalizada [3].

Vea tambien: determinante, regresión lineal.

**matriz laplaciana** La estructura de un grafo  $\mathcal{G}$ , con nodos  $i = 1, \dots, n$ , se puede analizar utilizando las propiedades de matrices especiales asociadas con  $\mathcal{G}$ . Una de estas matrices es la matriz laplaciana del grafo  $\mathbf{L}^{(\mathcal{G})} \in \mathbb{R}^{n \times n}$ , la cual se define para un grafo no dirigido y ponderado [10, 90]. Se define de forma elemento por elemento como (Vea la Figura 25)

$$L_{i,i'}^{(\mathcal{G})} := \begin{cases} -A_{i,i'} & \text{para } i \neq i', \{i, i'\} \in \mathcal{E}, \\ \sum_{i'' \neq i} A_{i,i''} & \text{para } i = i', \\ 0 & \text{en otro caso.} \end{cases} \quad (4)$$



Aquí,  $A_{i,i'}$  denota el peso de arista de una arista  $\{i, i'\} \in \mathcal{E}$ .

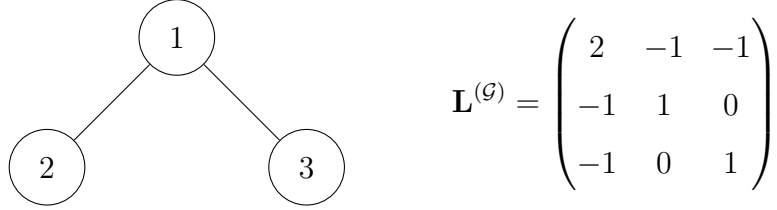


Figure 25: Izquierda: Un grafo no dirigido  $\mathcal{G}$  con tres nodos  $i = 1, 2, 3$ . Derecha: La matriz laplaciana  $\mathbf{L}^{(\mathcal{G})} \in \mathbb{R}^{3 \times 3}$  of  $\mathcal{G}$ .

Vea también grafo, peso de arista.

**media** La esperanza  $\mathbb{E}\{\mathbf{x}\}$  de una variable aleatoria numérica  $\mathbf{x}$ .

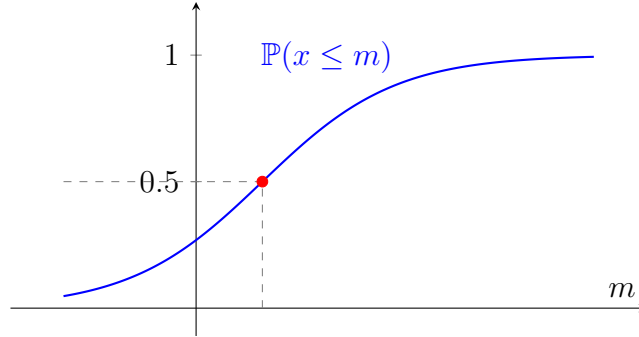
Vea también: expectation, RV

**media muestral** La muestra media  $\mathbf{m} \in \mathbb{R}^d$  para un conjunto de datos dado, con vectores de atributos  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)} \in \mathbb{R}^d$ , se define como

$$\mathbf{m} = (1/m) \sum_{r=1}^m \mathbf{x}^{(r)}.$$

Vea también: muestra, media, conjunto de datos, vector de atributos.

**mediana** Una mediana  $\text{med}(x)$  de una variable aleatoria real  $x$  es cualquier número  $m \in \mathbb{R}$  tal que  $\mathbb{P}(x \leq m) \geq 1/2$  y  $\mathbb{P}(x \geq m) \geq 1/2$  [54]. Podemos definir la mediana  $\text{med}(\mathcal{D})$  de un conjunto de datos  $\mathcal{D} = \{x^{(1)}, \dots, x^{(m)} \in \mathbb{R}\}$  mediante una variable aleatoria específica  $\tilde{x}$  asociada naturalmente con  $\mathcal{D}$ . En particular, esta variable aleatoria se construye como  $\tilde{x} = x^{(I)}$ , donde el índice  $I$  se elige uniformemente



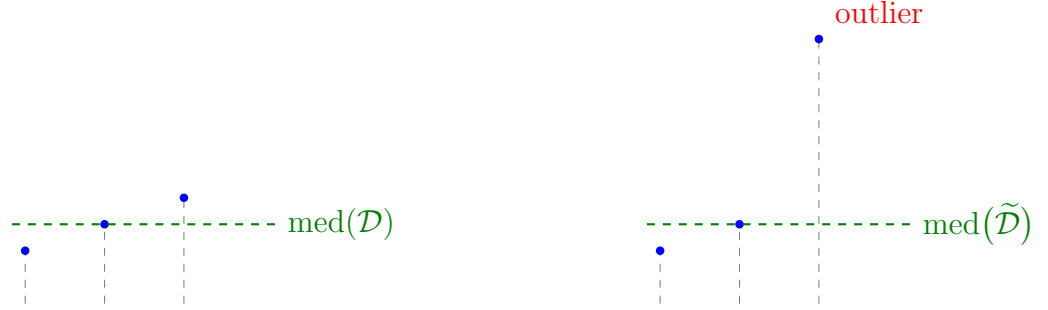
al azar del conjunto  $\{1, \dots, m\}$ , es decir,  $\mathbb{P}(I = r) = 1/m$  para todo  $r = 1, \dots, m$ . Si la variable aleatoria  $x$  es integrable, una mediana de  $x$  es la solución del siguiente problema de optimización:

$$\min_{x' \in \mathbb{R}} \mathbb{E}|x - x'|.$$

Al igual que la media, la mediana de un conjunto de datos  $\mathcal{D}$  también puede usarse para estimar los parámetros del modelo de un modelo de probabilidad subyacente. Comparada con la media, la mediana es más robusta frente a un valor atípico. Por ejemplo, la mediana de un conjunto de datos  $\mathcal{D}$  con más de un punto de datos no cambia aunque aumentemos arbitrariamente el valor del mayor elemento de  $\mathcal{D}$ . En contraste, la media sí lo haría. Véa también: media, robustez, outlier.

**mediana geométrica (GM)** La GM de un conjunto de vectores de entrada  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}$  en  $\mathbb{R}^d$  es un punto  $\mathbf{z} \in \mathbb{R}^d$  que minimiza la suma de distancias a dichos vectores [36], es decir:

$$\mathbf{z} \in \operatorname{argmin}_{\mathbf{y} \in \mathbb{R}^d} \sum_{r=1}^m \|\mathbf{y} - \mathbf{x}^{(r)}\|_2. \quad (5)$$



(a) conjunto de datos original  $\mathcal{D}$  (b) conjunto de datos ruidoso  $\tilde{\mathcal{D}}$  con un valor atípico (outlier).

Figure 26: La mediana es robusta frente a la presencia de punto atípico.

La Figura 27 ilustra una propiedad fundamental de la GM: si  $\mathbf{z}$  no coincide con ninguno de los vectores de entrada, entonces los vectores unitarios que apuntan desde  $\mathbf{z}$  hacia cada  $\mathbf{x}^{(r)}$  deben sumar cero; esta es la condición de subgradiente nulo (óptima) de (??). Además, se ha demostrado que la solución de (5) no puede alejarse arbitrariamente de los vectores de entrada confiables mientras estos constituyan la mayoría [91, Th. 2.2].

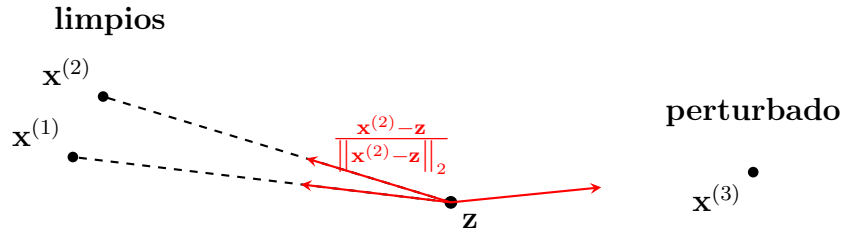


Figure 27: Sea  $\mathbf{z}$  una solución de (??) que no coincide con ninguno de los vectores de entrada. La condición de optimalidad requiere que los vectores unitarios desde  $\mathbf{z}$  hacia los vectores de entrada sumen cero.

Véa también: subgradiente.

**minimización de la pérdida regularizada (RLM)** Consulte minimización del riesgo empírico regularizado (RERM).

**minimización de variación total generalizada (GTVMin)** La minimización de variación total generalizada (GTVMin) es una instancia de minimización del riesgo empírico regularizado que utiliza la variación total generalizada de los parametros del modelo locales como un regularizador [92].

Vea también: RERM, GTV, parámetros del modelo, regularizador

**minimización del riesgo empírico explicable (EERM)** La minimización del riesgo empírico explicable (EERM) es una instancia de minimización del riesgo estructural que agrega un término de regulación a la pérdida promedio en la funcion objetivo de ERM. El término de regulación se elige para favorecer mapas de hipótesis que sean intrínsecamente explicables para un usuario específico. Este usuario se caracteriza por

sus predicciones proporcionadas para los puntos de datos en un conjunto de entrenamiento [65].

Vea también: SRM, regularización, pérdida, función objetivo, ERM, hipótesis, predicción, punto de datos, conjunto de entrenamiento.

### **minimización del riesgo empírico regularizado (RERM)** El ERM básico

aprende una hipótesis (o entrena un modelo)  $h \in \mathcal{H}$  basado únicamente en el riesgo empírico  $\widehat{L}(h|\mathcal{D})$  incurrido en un conjunto de entrenamiento  $\mathcal{D}$ . Para hacer que ERM sea menos propenso al sobreajuste, podemos implementar la regulación incluyendo un regularizador (escalado)  $\mathcal{R}\{h\}$  en el objetivo de aprendizaje. Esto da lugar a la minimización del riesgo empírico regularizado (RERM),

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{L}(h|\mathcal{D}) + \alpha \mathcal{R}\{h\}. \quad (6)$$

El parámetro  $\alpha \geq 0$  controla la intensidad de la regulación. Para  $\alpha = 0$ , recuperamos el ERM estándar sin regulación. A medida que  $\alpha$  aumenta, la hipótesis aprendida se inclina cada vez más hacia valores pequeños de  $\mathcal{R}\{h\}$ . El componente  $\alpha \mathcal{R}\{h\}$  en la función objetivo de (6) se puede entender intuitivamente como un sustituto para el aumento promedio en la pérdida que puede ocurrir al predecir etiquetas para puntos de datos fuera del conjunto de entrenamiento. Esta intuición se puede precisar en varias maneras. Por ejemplo, considere un modelo lineal entrenado usando pérdida de error cuadrático y el regularizador  $\mathcal{R}\{h\} = \|\mathbf{w}\|_2^2$ . En este caso,  $\alpha \mathcal{R}\{h\}$  corresponde al aumento esperado en la pérdida causado por la adición de variables aleatorias gaussianas a los vectores de atributos en el conjunto de entrenamiento [6, Ch.

3]. Una construcción basada en principios para el regularizador  $\mathcal{R}\{h\}$  surge de límites superiores aproximados en el error de generalización. La instancia resultante de RERM se conoce como minimización del riesgo estructural (SRM) [93, Sec. 7.2].

Vea también: ERM, hipótesis, modelo, riesgo empírico, conjunto de entrenamiento, sobreajuste, regularización, regularizador, punto de datos, pérdida, etiqueta, modelo lineal, vector de atributos,

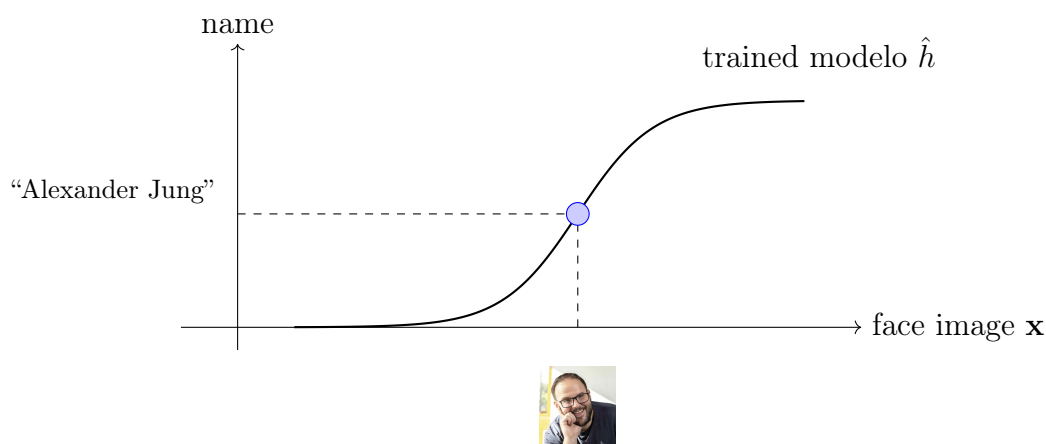
**minimización del riesgo estructural (SRM)** La minimización del riesgo estructural (SRM) es una forma de minimización del riesgo empírico regularizado (RERM), en la que el modelo  $\mathcal{H}$  se puede expresar como una unión contable de submodelos:  $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}^{(n)}$ . Cada submodelo  $\mathcal{H}^{(n)}$  permite la derivación de un límite superior aproximado para el error de generalización que se incurre al aplicar ERM para entrenar  $\mathcal{H}^{(n)}$ . Estos límites superiores individuales para cada submodelo—se combinan para formar un regularizador utilizado en el objetivo de RERM. Estos límites superiores aproximados (uno para cada  $\mathcal{H}^{(n)}$ ) se combinan para construir un regularizador para RERM [44, Sec. 7.2].

Vea también: RERM, modelo, ERM, regularizador

**minimización empírica del riesgo (ERM)** La minimización del riesgo empírico es el problema de optimización que consiste en encontrar una hipótesis (dentro de un modelo) con la mínima pérdida promedio (o riesgo empírico) en un conjunto de datos dado  $\mathcal{D}$  (es decir, el conjunto de entrenamiento). Muchos métodos de aprendizaje automático se obtienen a partir de el riesgo empírico mediante elecciones específicas

de diseño para el conjunto de datos, el modelo, y la perdida [6, Ch. 3].  
 Vea también: Riesgo empírico, hipótesis, modelo, mínimo, pérdida,  
 glsemprisk, conjunto de datos, conjunto de entrenamiento, ML

**model inversion** Una inversión de modelo es una forma de ataque privado dirigida a un sistema de aprendizaje automático. Un adversario intenta inferir un atributo sensible de puntos de datos individuales aprovechando el acceso parcial a un modelo entrenado  $\hat{h} \in \mathcal{H}$ . Este acceso suele consistir en consultar el modelo con predicciones  $\hat{h}(\mathbf{x})$  sobre entradas cuidadosamente seleccionadas. Se han demostrado técnicas básicas de inversión de modelo en el contexto de clasificación de imágenes faciales, donde se reconstruyen imágenes utilizando la (gradiente de la) salida del modelo junto con información auxiliar como el nombre de una persona [94]



Vea tambien: modelo, ataque a la privacidad, ML, atributo sensible,  
 punto de datos, predicción, clasificación, gradiente, trustworthy AI,

protección de la privacidad.

**model selection** En aprendizaje automático, la selección de modelo se refiere al proceso de elegir entre diferentes modelos candidatos. En su forma más básica, la selección de modelo consiste en: 1) entrenar cada modelo candidato; 2) calcular el error de validación para cada modelo entrenado; y 3) elegir el modelo con el menor error de validación [6, Ch. 6].

Vea también: ML, modelo, error de validación

**modelo** En el contexto de métodos de aprendizaje automático, el término modelo típicamente se refiere a el espacio de hipótesis empleado por un método de aprendizaje automático [6, 44]. Sin embargo, el término también se utiliza en otros campos pero con un significado diferente. Por ejemplo, un modelo de probabilidad se refiere a un conjunto parametrizado de distribución de probabilidad.

Vea también: ML, espacio de hipótesis, modelo probabilístico, distribución de probabilidad.

**modelo de lenguaje de gran escala (LLM)** Los modelos de lenguaje de gran escala son un término genérico para métodos de aprendizaje automático que procesan y generan texto similar al humano. Estos métodos suelen usar redes profundas con miles de millones (o incluso billones) de parámetros. Una elección ampliamente utilizada para la arquitectura de red se conoce como Transformers [95]. El entrenamiento de modelos de lenguaje de gran escala a menudo se basa en la tarea de predecir algunas palabras que se eliminan intencionadamente de un corpus de



texto extenso. Así, podemos construir puntos de datos etiquetados simplemente seleccionando algunas palabras de un texto como etiquetas y las palabras restantes como atributos de puntos de datos. Esta construcción requiere muy poca supervisión humana y permite generar conjuntos de entrenamiento suficientemente grandes para modelos de lenguaje de gran escala.

Vea también: ML, red profunda, parámetros, punto de dato etiquetado, etiqueta, atributo, punto de datos, conjunto de entrenamiento, modelo.

**modelo de mezcla gaussiana (GMM)** Un GMM es un tipo particular de modelo de probabilidad para un vector numérico  $\mathbf{x}$  (por ejemplo, los atributos de un punto de datos). Dentro de un GMM, el vector  $\mathbf{x}$  se extrae de una distribución normal multivariante  $p^{(c)} = \mathcal{N}(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)})$  seleccionada aleatoriamente, con  $c = I$ . El índice  $I \in \{1, \dots, k\}$  es una variable aleatoria con probabilidades  $\mathbb{P}(I = c) = p_c$ . Ten en cuenta que un GMM se parametriza por la probabilidad  $p_c$ , el vector medio  $\boldsymbol{\mu}^{(c)}$ , y la matriz de covarianza  $\boldsymbol{\Sigma}^{(c)}$  para cada  $c = 1, \dots, k$ . Los GMM se utilizan ampliamente para agrupamiento, estimación de densidad y como un modelo generativo.

Vea también: modelo probabilístico, atributo, punto de datos, distribución normal multivariante, RV, media, matriz de covarianza, agrupamiento, modelo.

**modelo en red** Un modelo en red sobre un red de aprendizaje federado  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  asigna un modelo local (es decir, un espacio de hipótesis) a cada nodo  $i \in \mathcal{V}$  de la red de aprendizaje federado  $\mathcal{G}$ .

Vea también: red FL, modelo local, espacio de hipótesis

**modelo estocástico de bloques (SBM)** El modelo estocástico de bloques (SBM) es un modelo generativo probabilístico para un grafo no dirigido  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  con conjunto de nodos  $\mathcal{V}$  [96]. En su forma básica, asigna aleatoriamente cada nodo  $i \in \mathcal{V}$  a un clúster  $c_i \in \{1, \dots, k\}$ . Cada par de nodos distintos se conecta con probabilidad  $p_{i,i'}$  que depende únicamente de sus etiquetas  $c_i$  y  $c_{i'}$ . La presencia de aristas entre pares de nodos es estadísticamente independiente.

Vea también: modelo, grafo, cluster, probabilidad, etiqueta

**modelo lineal** Consideremos puntos de datos, cada uno caracterizado por un vector de atributos numérico  $\mathbf{x} \in \mathbb{R}^d$ . Un modelo lineal es un espacio de hipótesis que consiste en todos los mapeos lineales,

$$\mathcal{H}^{(d)} := \{h(\mathbf{x}) = \mathbf{w}^T \mathbf{x} : \mathbf{w} \in \mathbb{R}^d\}. \quad (7)$$

Nótese que (7) define toda una familia de espacios de hipótesis, parametrizados por el número  $d$  de atributos que se combinan linealmente para formar la predicción  $h(\mathbf{x})$ . La elección de diseño de  $d$  se guía por aspectos computacionales (por ejemplo, reducir  $d$  implica menor computación), por aspectos estadísticos (por ejemplo, aumentar  $d$  podría reducir el error de predicción), y por la interpretabilidad. Un modelo lineal que utiliza pocos atributos cuidadosamente elegidas suele considerarse más interpretable [20, 68].

Vea también: punto de datos, vector de atributos, modelo, espacio de hipótesis, atributo, predicción, aspectos computacionales, aspectos estadísticos, interpretabilidad, modelo

**modelo local** Considera una colección de conjuntos de datos locales que están asignados a los nodos de una red de aprendizaje federado. Un modelo local  $\mathcal{H}^{(i)}$  es un espacio de hipótesis asignado a un nodo  $i \in \mathcal{V}$ . Diferentes nodos podrían tener asignados diferentes espacios de hipótesis, es decir, en general  $\mathcal{H}^{(i)} \neq \mathcal{H}^{(i')}$  para diferentes nodos  $i, i' \in \mathcal{V}$ .  
Vea también: dispositivo, red FL, modelo, espacio de hipótesis.

**modelo probabilístico** Un modelo probabilístico interpreta los puntos de datos como realizaciones de variables aleatorias con una distribución de probabilidad conjunta. Esta distribución de probabilidad conjunta típicamente incluye parámetros que deben seleccionarse manualmente o aprenderse usando métodos de inferencia estadística como la estimación por máxima verosimilitud [54].  
Vea también: modelo, punto de datos, realización, RV, distribución de probabilidad, parámetros, máxima verosimilitud

**muestra** Una secuencia finita (o lista) de puntos de datos  $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$  que se obtiene o interpreta como la realización de  $m$  variables aleatorias iid con una distribución de probabilidad común  $p(\mathbf{z})$ . La longitud  $m$  de la secuencia se denomina tamaño de muestra.  
Vea también: punto de datos, realización, i.i.d., RV, distribución de probabilidad, tamaño de la muestra.

**máquina de vectores de soporte (SVM)** La SVM es un método de clasificación binaria que aprende un mapa de hipótesis lineal. Por lo tanto, al igual que la regresión lineal y la regresión logística, también es una instancia de ERM para el modelo lineal. Sin embargo, la SVM utiliza

una función de pérdida diferente a la empleada en esos métodos. Como se ilustra en la Figura 28, su objetivo es separar al máximo los puntos de datos de las dos clases diferentes en el espacio de atributos (es decir, el principio de margen máximo). Maximizar esta separación es equivalente a minimizar una variante regularizada de la pérdida de hinge (11) [62, 87, 97].

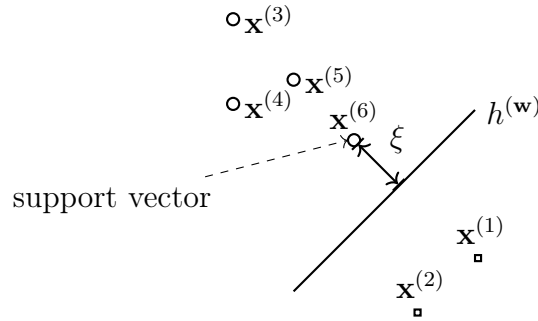


Figure 28: La SVM aprende una hipótesis (o clasificador)  $h^{(w)}$  con pérdida de margen suave promedio mínima pérdida de hinge. Minimizar esta pérdida es equivalente a maximizar el margen  $\xi$  entre la frontera de decisión de  $h^{(w)}$  y cada clase del conjunto de entrenamiento.

La variante básica de la SVM solo es útil si los puntos de datos de diferentes categorías pueden ser (aproximadamente) separables linealmente. Para una aplicación de aprendizaje automático donde las categorías no son separables linealmente basadas en los atributos originales (crudas), es posible aplicar la SVM a atributos transformados. Estos atributos transformados se pueden obtener aplicando un mapa de atributos derivado de un kernel.

Vea también: clasificación, hipótesis, regresión lineal, regresión logística, ERM, modelo lineal, función de pérdida, punto de datos, espacio de atributos, máximo, pérdida de hinge, SVM, clasificador, pérdida, frontera de decisión, conjunto de entrenamiento, ML, kernel.

**máxima verosimilitud** Considera puntos de datos  $\mathcal{D} = \{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}\}$  que se interpretan como las realizaciones de variables aleatorias iid con una distribución de probabilidad común  $\mathbb{P}(\mathbf{z}; \mathbf{w})$  que depende de los parametros del modelo  $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^n$ . Los métodos de máxima verosimilitud aprenden los parametros del modelo  $\mathbf{w}$  al maximizar la probabilidad (densidad)  $\mathbb{P}(\mathcal{D}; \mathbf{w}) = \prod_{r=1}^m \mathbb{P}(\mathbf{z}^{(r)}; \mathbf{w})$  del conjunto de datos observado. Por lo tanto, el estimador de máxima verosimilitud es una solución al problema de optimización  $\max_{\mathbf{w} \in \mathcal{W}} \mathbb{P}(\mathcal{D}; \mathbf{w})$ .

Vea también: punto de datos, realización, i.i.d., RV, distribución de probabilidad, parámetros del modelo, máximo, conjunto de datos.

**máximo** El máximo de un conjunto  $\mathcal{A} \subseteq \mathbb{R}$  de números reales es el elemento más grande en ese conjunto, si tal elemento existe. Un conjunto  $\mathcal{A}$  tiene un máximo si está acotado superiormente y alcanza su supremo [2, Sec. 1.4]

Vea también: supremo (o mínimo de las cotas superiores).

**método de kernel** Un método de kernel es un método de aprendizaje Automático que utiliza un kernel  $K$  para mapear el vector de atributos original (crudo)  $\mathbf{x}$  de un punto de datos a un nuevo (transformado) vector de atributos  $\mathbf{z} = K(\mathbf{x}, \cdot)$  [86, 87]. La motivación para transformar los vectores de atributos es que, al utilizar un kernel adecuado, los pun-

tos de datos presentan una geometría "más conveniente" en el espacio de atributos. Por ejemplo, en un problema de clasificación binaria, el uso de vectores de atributos transformados  $\mathbf{z}$  podría permitirnos utilizar modelos lineales, incluso si los puntos de datos no son linealmente separables en el espacio de atributos original (Vea la Figura 29).

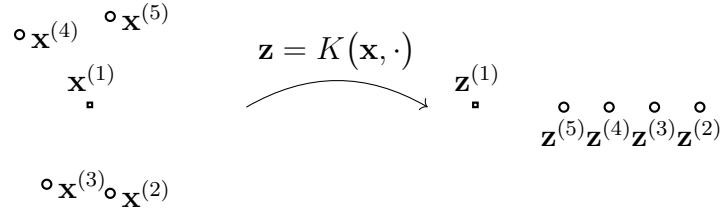


Figure 29: Cinco puntos de datos caracterizados por vectores de atributos  $\mathbf{x}^{(r)}$  y etiquetas  $y^{(r)} \in \{\circ, \square\}$ , para  $r = 1, \dots, 5$ . Con estos vectores de atributos, no hay forma de separar las dos clases mediante una línea recta (rque representa la frontera de decisión de un clasificador lineal). En contraste, los vectores de atributos transformados  $\mathbf{z}^{(r)} = K(\mathbf{x}^{(r)}, \cdot)$  permiten separar los puntos de datos usando un clasificador lineal

Vea tambien: kernel, ML, vector de atributos, punto de datos, espacio de atributos, clasificación, modelo lineal, etiqueta, frontera de decisión, clasificador lineal.

**método de optimización** Un método de optimización es un algoritmo que recibe como entrada una representación de un problema de optimización y entrega como salida una solución (aproximada) [36], [74], [98].

Véa también: algoritmo, problema de optimización.

**métodos de gradiente** Los métodos de gradiente son técnicas iterativas para encontrar el mínimo(o máximo) de una función objetivo diferenciable respecto a los parámetros del modelo. Estos métodos construyen una secuencia de aproximaciones hacia una elección óptima de parámetros del modelo que resulta en un mínimo (o máximo) valor de la función objetivo. Como su nombre indica, los métodos basados en gradiente utilizan los gradientes de la función objetivo evaluados en iteraciones previas para construir nuevos parámetros del modelo (esperablemente) mejorados. Un ejemplo importante de un método basado en gradiente es el descenso por gradiente.

Vea también: gradiente, mínimo, máximo, diferenciable, función objetivo, parámetros del modelo, GD.

**métrica** En su forma más general, una métrica es una medida cuantitativa utilizada para comparar o evaluar objetos. En matemáticas, una métrica mide la distancia entre dos puntos y debe cumplir ciertas reglas: la distancia es siempre no negativa, igual a cero solo si los puntos son idénticos, simétrica y satisface la desigualdad triangular [2]. En el aprendizaje automático, una métrica es una medida cuantitativa del rendimiento de un modelo. Algunos ejemplos son la precisión (accuracy), la precisión y el promedio de la pérdida 0-1 sobre un conjunto de entrenamiento [99], [62]. Una función de pérdida se utiliza para entrenar modelos, mientras que una métrica se utiliza para comparar modelos ya entrenados.

Véa también: ML, modelo, precisión, 0/1 loss, conjunto de prueba, función de pérdida, pérdida, selección de modelo.

**mínimo** Dado un conjunto de números reales, el mínimo es el menor de esos números.

**no suave** Nos referimos a una función como no suave si no es suave [74].

**norma** Una norma es una función que asigna a cada elemento (vectorial) de un espacio vectorial lineal un número real no negativo. Esta función debe ser homogénea, definida positiva y debe cumplir la desigualdad triangular [43].

**norma dual** Toda norma  $\|\cdot\|$  definida en un espacio euclidiano  $\mathbb{R}^d$  tiene una norma dual asociada, denotada por  $\|\cdot\|_*$  y definida como  $\|\mathbf{y}\|_* := \sup_{\|\mathbf{x}\| \leq 1} \mathbf{y}^T \mathbf{x}$ . La norma dual mide el mayor producto interno posible entre  $\mathbf{y}$  y cualquier vector en la bola unitaria de la norma original. Para más detalles, véase [36, Sec. A.1.6].

Véase también: norma, espacio euclidiano.

**normalización de datos** La normalización de datos se refiere a transformaciones aplicadas a los vectores de atributos de puntos de datos para mejorar los aspectos estadísticos o aspectos computacionales del método de aprendizaje automático. Por ejemplo, en regresión lineal con métodos de gradiente que utilizan una tasa de aprendizaje fija, la convergencia depende de controlar la norma de los vectores de atributos en el conjunto de entrenamiento. Un enfoque común es normalizar los vectores de atributos de modo que su norma no exceda de uno [6, Ch. 5].

Vea también: datos, vector de atributos, punto de datos, ML, aspectos estadísticos, aspectos computacionales, regresión lineal, métodos de



gradiente, tasa de aprendizaje, norma, conjunto de entrenamiento.

**número de condición** El número de condición  $\kappa(\mathbf{Q}) \geq 1$  de una matriz definida positiva  $\mathbf{Q} \in \mathbb{R}^{d \times d}$  es el cociente  $\alpha/\beta$  entre el mayor  $\alpha$  y el menor  $\beta$  de  $\mathbf{Q}$ . El número de condición es útil para el análisis de métodos de aprendizaje automático. La complejidad computacional de los métodos de gradiente para regresión lineal depende críticamente del número de condición de la matriz  $\mathbf{Q} = \mathbf{X}\mathbf{X}^T$ , donde  $\mathbf{X}$  es la matriz de atributos del conjunto de entrenamiento. Es por eso que desde una perspectiva computacional, preferimos atributos de los puntos de datos que hagan que  $\mathbf{Q}$  tenga un número de condición cercano a 1. Vea también: ML, métodos de gradiente, regresión lineal, matriz de atributos, conjunto de entrenamiento, atributo, punto de datos

**operador de reducción y selección absoluta mínima (Lasso)** El Lasso es una implementación de minimización del riesgo estructural (SRM). Aprende los pesos  $\mathbf{w}$  de un mapa lineal  $h(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$  basado en un conjunto de entrenamiento. El Lasso se obtiene a partir de regresión lineal al agregar la norma  $\ell_1$  escalado  $\alpha \|\mathbf{w}\|_1$  al promedio de la pérdida de error cuadrático en el conjunto de entrenamiento. Vea también: SRM, pesos, conjunto de entrenamiento, regresión lineal, norma, pérdida de error cuadrático.

**operador proximal** Dada una función convexa  $f(\mathbf{w}')$ , definimos su operador proximal como [100, 101]

$$\mathbf{prox}_{f(\cdot), \rho}(\mathbf{w}) := \underset{\mathbf{w}' \in \mathbb{R}^d}{\operatorname{argmin}} \left[ f(\mathbf{w}') + (\rho/2) \|\mathbf{w} - \mathbf{w}'\|_2^2 \right] \text{ with } \rho > 0.$$

Como se ilustra en la Figura 30, evaluar el operador proximal equivale a minimizar una variante penalizada de  $f(\mathbf{w}')$ . El término de penalización es la distancia euclidiana cuadrada escalada hacia un vector dado  $\mathbf{w}$  (que es la entrada del operador proximal). El operador proximal puede interpretarse como una generalización del paso de gradiente, definido para una función suave y convexo  $f(\mathbf{w}')$ . De hecho, realizar un paso de gradiente con tamaño de paso  $\eta$  en el vector actual  $\mathbf{w}$  es lo mismo que aplicar el operador proximal de la función  $\tilde{f}(\mathbf{w}') = (\nabla f(\mathbf{w}))^T (\mathbf{w}' - \mathbf{w})$  y usar  $\rho = 1/\eta$ .

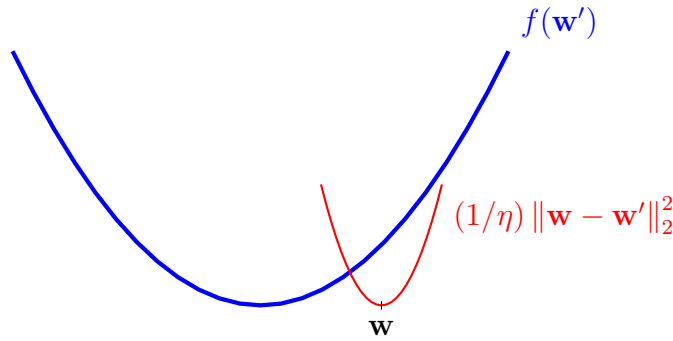


Figure 30: Un paso de gradiente generalizado actualiza un vector  $\mathbf{w}$  minimizando una versión penalizada de la función  $f(\cdot)$ . El término de penalización es la distancia euclidiana cuadrada escalada entre la variable de optimización  $\mathbf{w}'$  y el vector dado  $\mathbf{w}$ .

**optimismo ante la incertidumbre** Los métodos de aprendizaje automático aprenden parámetros del modelo  $\mathbf{w}$  de acuerdo con algún criterio de desempeño  $\bar{f}(\mathbf{w})$ . Sin embargo, normalmente no pueden acceder directamente a  $\bar{f}(\mathbf{w})$  pero dependen de una estimación (o aproximación)

de  $f(\mathbf{w})$ . Por ejemplo, los métodos basados en ERM usan la pérdida promedio en un conjunto de datos (por ejemplo, el conjunto de entrenamiento) como estimación del riesgo de una hipótesis. Usando un modelo de probabilidad, se puede construir un intervalo de confianza  $[l^{(\mathbf{w})}, u^{(\mathbf{w})}]$  para cada elección  $\mathbf{w}$  de los parámetros del modelo. Una construcción simple es  $l^{(\mathbf{w})} := f(\mathbf{w}) - \sigma/2$ ,  $u^{(\mathbf{w})} := f(\mathbf{w}) + \sigma/2$ , donde  $\sigma$  representa una medida de la desviación entre  $f(\mathbf{w})$  y  $\bar{f}(\mathbf{w})$ . También se pueden usar otras construcciones del intervalo, mientras se aseguren que  $\bar{f}(\mathbf{w}) \in [l^{(\mathbf{w})}, u^{(\mathbf{w})}]$  con una probabilidad suficientemente alta. Siendo optimistas, elegimos los parámetros del modelo según el valor más favorable - pero realista - del criterio de desempeño  $\tilde{f}(\mathbf{w}) := l^{(\mathbf{w})}$ . Dos ejemplos de métodos de aprendizaje automático que usan una construcción optimista de una función objetivo son métodos de minimización del riesgo estructural (SRM) [44, Ch. 11] y cota superior de confianza (UCB) para decisiones secuenciales [27, Sec. 2.2].

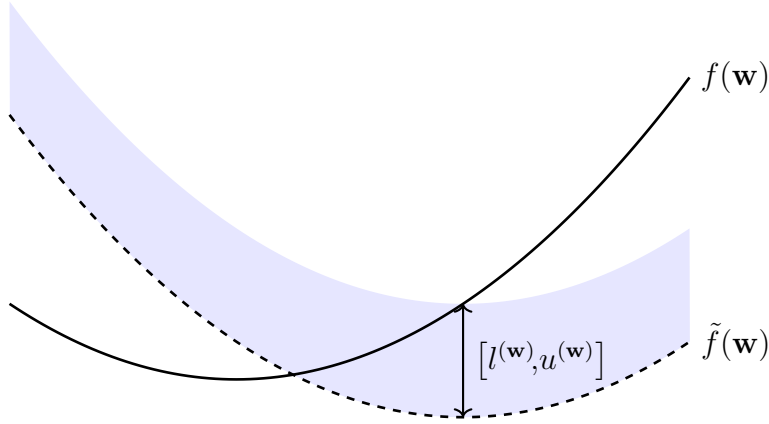


Figure 31: Los métodos de aprendizaje automático aprenden parámetros del modelo  $\mathbf{w}$  usando una estimación de  $f(\mathbf{w})$  como aproximación del criterio de desempeño  $\bar{f}(\mathbf{w})$ . Usando un modelo de probabilidad, se pueden construir intervalos de confianza  $[l^{(\mathbf{w})}, u^{(\mathbf{w})}]$  que contienen  $\bar{f}(\mathbf{w})$  con alta probabilidad. La mejor medida plausible del desempeño para una elección específica  $\mathbf{w}$  es  $\tilde{f}(\mathbf{w}) := l^{(\mathbf{w})}$ .

Vea también: ML, parámetros del modelo, ERM, pérdida, conjunto de datos, conjunto de entrenamiento, riesgo, hipótesis, modelo probabilístico, cota superior de confianza (UCB),

**parámetro** Un parámetro de un modelo de aprendizaje automático es una cantidad ajustable (es decir, que puede aprenderse o modificarse) que nos permite elegir entre diferentes funciones de hipótesis. Por ejemplo, el modelo lineal  $\mathcal{H} := \{h^{(\mathbf{w})} : h^{(\mathbf{w})}(x) = w_1x + w_2\}$  consiste en todas las hipótesis aplicaciones de la forma  $h^{(\mathbf{w})}(x) = w_1x + w_2$  con una elección

particular de los parámetros  $\mathbf{w} = (w_1, w_2)^T \in \mathbb{R}^2$ . Otro ejemplo de un parámetro de modelo es el conjunto de pesos asignados a una conexión entre dos neuronas en una RNA.

Véa también: ML, modelo, hipótesis, aplicación, modelo lineal, pesos, ANN.

**parámetros** Los parámetros de un modelo de aprendizaje automático son cantidades ajustables (es decir, entrenables o modificables) que nos permiten elegir entre diferentes funciones de hipótesis. Por ejemplo, el modelo lineal  $\mathcal{H} := \{h^{(\mathbf{w})} : h^{(\mathbf{w})}(x) = w_1x + w_2\}$  consiste en todas las funciones de hipótesis  $h^{(\mathbf{w})}(x) = w_1x + w_2$  con una elección particular de los parámetros  $\mathbf{w} = (w_1, w_2)^T \in \mathbb{R}^2$ . Otro ejemplo de parámetros son los pesos asignados a las conexiones entre neuronas de una red neuronal artificial.

Vea también: ML, modelo, hipótesis, modelo lineal, pesos, ANN.

**parámetros del modelo** Los parámetros de un modelo son cantidades que se utilizan para seleccionar un mapa de hipótesis específico dentro de un modelo. Podemos pensar en una lista de parámetros del modelo como un identificador único para un mapa de hipótesis, similar a cómo un número de seguridad social identifica a una persona en Finlandia.

Vea también: modelo, parámetros, hipótesis.

**paso de gradiente** Dada una función diferenciable de valores reales  $f(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$  y un vector  $\mathbf{w} \in \mathbb{R}^d$ , el paso de gradiente actualiza  $\mathbf{w}$  sumándole el negativo escalado del gradiente  $\nabla f(\mathbf{w})$  para obtener el nuevo vector

(véase la Figura 32)

$$\hat{\mathbf{w}} := \mathbf{w} - \eta \nabla f(\mathbf{w}). \quad (8)$$

Matemáticamente, el paso de gradiente es un operador (típicamente no lineal)  $\mathcal{T}^{(f,\eta)}$  que está parametrizado por la función  $f$  y el tamaño de paso  $\eta$ .

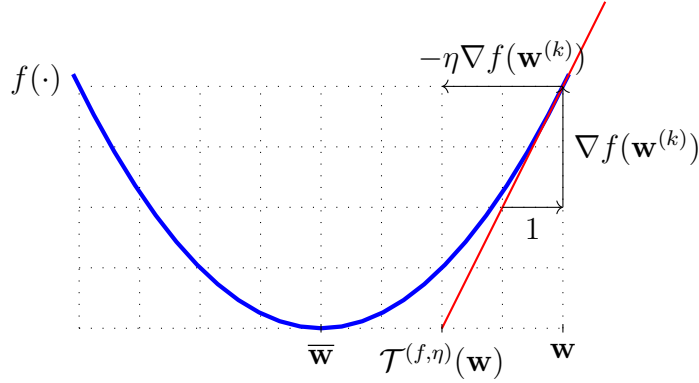


Figure 32: El paso básico de gradiente (8) mapea un vector  $\mathbf{w}$  al vector actualizado  $\mathbf{w}'$ . Define un operador  $\mathcal{T}^{(f,\eta)}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d : \mathbf{w} \mapsto \hat{\mathbf{w}}$ .

Nótese que el paso de gradiente (8) optimiza localmente - en un entorno cuyo tamaño está determinado por el tamaño de paso  $\eta$  - una aproximación lineal de la función  $f(\cdot)$ . Una generalización natural de (8) es optimizar localmente la función misma - en lugar de su aproximación lineal - de tal manera que:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}' \in \mathbb{R}^d}{\operatorname{argmin}} f(\mathbf{w}') + (1/\eta) \|\mathbf{w} - \mathbf{w}'\|_2^2. \quad (9)$$

Intencionalmente usamos el mismo símbolo  $\eta$  para el parámetro en (9) que en el tamaño de paso de (8). Mientras mayor sea el valor de  $\eta$

en (9), más progreso hará la actualización en la reducción del valor de la función  $f(\hat{\mathbf{w}})$ . Nótese que, al igual que el paso de gradiente (8), la actualización (9) también define un operador (típicamente no lineal) parametrizado por la función  $f(\cdot)$  y el parámetro  $\eta$ . Para una función convexa  $f(\cdot)$ , este operador es conocido como el operador proximal de  $f(\cdot)$  [100].

Vea también: diferenciable, gradiente, tamaño de paso, entorno, generalización, convexo, operador proximal

**peso de arista** Cada arista  $\{i, i'\}$  de una red de aprendizaje federado tiene asignado un peso de arista no negativo  $A_{i,i'} \geq 0$ . Un peso de arista cero  $A_{i,i'} = 0$  indica la ausencia de una arista entre los nodos  $i, i' \in \mathcal{V}$ .

Vea también: red FL.

**pesos** Considera un espacio de hipótesis parametrizado  $\mathcal{H}$ . Usamos el término pesos para los parámetros del modelo numéricos que se utilizan para escalar las atributos o sus transformaciones con el fin de calcular  $h^{(\mathbf{w})} \in \mathcal{H}$ . Un modelo lineal utiliza los pesos  $\mathbf{w} = (w_1, \dots, w_d)^T$  para calcular la combinación lineal  $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$ . Los pesos también se utilizan en las red neuronales artificiales para formar combinaciones lineales de los atributos o de las salidas de las neuronas en capas ocultas.

Vea también: espacio de hipótesis, parámetros del modelo, atributo, modelo lineal, ANN.

**precisión (accuracy)** Consideremos puntos de datos caracterizados por atributos  $\mathbf{x} \in \mathcal{X}$  y una etiqueta categórica  $y$  que toma valores de un conjunto finito espacio de etiquetas  $\mathcal{Y}$ . La precisión de una hipóte-

sis  $h : \mathcal{X} \rightarrow \mathcal{Y}$ , cuando se aplica a los puntos de datos en un conjunto de datos  $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(m)}, y^{(m)})\}$ , se define como  $1 - (1/m) \sum_{r=1}^m L^{(0/1)}((\mathbf{x}^{(r)}, y^{(r)}), h)$  usando la **0/1 loss**  $L^{(0/1)}(\cdot, \cdot)$ .

Vea también: punto de datos, atributo, hipótesis, conjunto de datos, 0/1 loss

**predicción** Una predicción es una estimación o aproximación de una cantidad de interés. El Aprendizaje Automático se centra en aprender o encontrar un mapa de hipótesis  $h$  que recibe los atributos  $\mathbf{x}$  de un punto de datos y produce una predicción  $\hat{y} := h(\mathbf{x})$  para su etiqueta  $y$ .

Vea también: ML, hipótesis, atributo, punto de datos

**predictor** Un predictor es un mapa de hipótesis con valores reales. Dado un punto de datos con atributos  $\mathbf{x}$ , el valor  $h(\mathbf{x}) \in \mathbb{R}$  se utiliza como una predicción para la verdadera etiqueta numérica  $y \in \mathbb{R}$  del punto de datos.

Vea también: hipótesis, punto de datos, atributo, predicción, etiqueta

**principio de minimización de datos** La regulación europea de protección de datos incluye un principio de minimización de datos. Este principio requiere que un controlador de datos limite la recolección de información personal a lo que es directamente relevante y necesario para cumplir un propósito específico. Los datos deben retenerse solo durante el tiempo necesario para cumplir ese propósito [102, Article 5(1)(c)], [103].

Vea también: datos.

**privacidad diferencial (DP)** Consideremos un método de aprendizaje automático  $\mathcal{A}$  que recibe como entrada un conjunto de datos (por ejemplo,



el conjunto de entrenamiento usado para ERM) y entrega una salida  $\mathcal{A}(\mathcal{D})$ . La salida puede ser los parámetros del modelo aprendidos o las predicciones para ciertos puntos de datos. DP es una medida precisa de la filtración de privacidad ocasionada al revelar dicha salida. Aproximadamente, un método de aprendizaje automático es diferencialmente privado si la distribución de probabilidad de la salida  $\mathcal{A}(\mathcal{D})$  no cambia significativamente cuando se modifica el atributo sensible de un solo punto de datos del conjunto de entrenamiento. Nótese que la DP se basa en un modelo de probabilidad para un método de aprendizaje automático, es decir, interpretamos su salida  $\mathcal{A}(\mathcal{D})$  como la realización de una variable aleatoria. La aleatoriedad en la salida puede asegurarse añadiendo intencionalmente la realización de una variable aleatoria auxiliar (ruido) a la salida del método de aprendizaje automático.

Vea también: ML, conjunto de datos, conjunto de entrenamiento, ERM, parámetros del modelo, predicción, punto de datos, filtración de privacidad, distribución de probabilidad, atributo sensible, modelo probabilístico, realización, RV

**probabilidad** Asignamos un valor de probabilidad, típicamente elegido en el intervalo  $[0, 1]$ , a cada evento que pueda ocurrir en un experimento aleatorio [5, 46, 61, 104].

**problema de optimización** Un problema de optimización es una estructura matemática que consiste en una función objetivo  $f : \mathcal{U} \rightarrow \mathcal{V}$  definida sobre una variable de optimización  $\mathbf{w} \in \mathcal{U}$ , junto con un conjunto factible  $\mathcal{W} \subseteq \mathcal{U}$ . Se asume que el codominio  $\mathcal{V}$  está ordenado, lo

que significa que para cualesquiera dos elementos  $\mathbf{a}, \mathbf{b} \in \mathcal{V}$ , podemos determinar si  $\mathbf{a} < \mathbf{b}$ ,  $\mathbf{a} = \mathbf{b}$  o  $\mathbf{a} > \mathbf{b}$ . El objetivo de la optimización es encontrar aquellos valores  $\mathbf{w} \in \mathcal{W}$  para los cuales el valor de la función objetivo  $f(\mathbf{w})$  es extremo —es decir, mínimo o máximo [36, 74, 98].

**proceso estocástico** Un proceso estocástico es una colección de variables aleatorias definidas sobre un mismo espacio de probabilidad. Estas variables aleatorias están indexadas por el tiempo o el espacio, y se utilizan para modelar fenómenos aleatorios que evolucionan en el tiempo (por ejemplo, ruido en sensores o series temporales financieras). Grafos aleatorios, como el modelo estocástico de bloques (SBM) o el grafo de Erd, son otro ejemplo de procesos estocásticos. Estos procesos utilizan pares de nodos en un grafo como conjunto de índices. También podemos usar procesos estocásticos para representar algoritmo estocástico como descenso por gradiente estocástico.

Véase también: SGD, incertidumbre, modelo probabilístico, RV, SBM.

**proceso gaussiano (GP)** Un GP es una colección de variables aleatorias  $\{f(\mathbf{x})\}_{\mathbf{x} \in \mathcal{X}}$  indexadas por valores de entrada  $\mathbf{x}$  de un cierto espacio de entrada  $\mathcal{X}$ , tal que, para cualquier subconjunto finito  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)} \in \mathcal{X}$ , las variables aleatorias correspondientes  $f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(m)})$  tienen una distribución conjunta gaussiana multivariante:

$$(f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(m)})) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}).$$

Para un espacio de entrada  $\mathcal{X}$  fijo, un GP queda completamente especificado (o parametrizado) por:

- una función media  $\mu(\mathbf{x}) = \mathbb{E}\{f(\mathbf{x})\}$
- y una función de covarianza  $K(\mathbf{x}, \mathbf{x}') = \mathbb{E}\{(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))\}$ .

Ejemplo: Podemos interpretar la distribución de temperatura en Finlandia (en un instante específico) como la realización de un GP  $f(\mathbf{x})$ , donde cada entrada  $\mathbf{x} = (\text{lat}, \text{lon})$  representa una ubicación geográfica. Las observaciones de temperatura de las estaciones meteorológicas de FMI constituyen muestras de  $f(\mathbf{x})$  en ubicaciones específicas (véase la Figura 33). Un GP nos permite predecir la temperatura en las cercanías de las estaciones meteorológicas y cuantificar la incertidumbre de dichas predicciones.

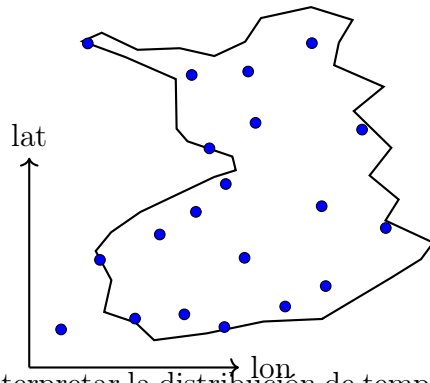


Figure 33: Podemos interpretar la distribución de temperatura sobre Finlandia como una realización de un GP indexado por coordenadas geográficas y muestreado en estaciones meteorológicas de FMI (indicadas por puntos azules).

Véa también: RV, media, función, realización, FMI, muestra, incertidumbre.

**producto de Kronecker** El producto de Kronecker de dos matrices  $\mathbf{A} \in \mathbb{R}^{m \times n}$  y  $\mathbf{B} \in \mathbb{R}^{p \times q}$  es una matriz por bloques denotada como  $\mathbf{A} \otimes \mathbf{B}$  y definida como [3], [43]

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix} \in \mathbb{R}^{mp \times nq}.$$

El producto de Kronecker es un caso particular del producto tensorial para matrices y se utiliza ampliamente en estadística multivariada, álgebra lineal y modelos estructurados de aprendizaje automático. Satisface la identidad  $(\mathbf{A} \otimes \mathbf{B})(\mathbf{x} \otimes \mathbf{y}) = (\mathbf{A}\mathbf{x}) \otimes (\mathbf{B}\mathbf{y})$  para vectores  $\mathbf{x}$  y  $\mathbf{y}$  de dimensiones compatibles.

Vea también: ML, modelo.

**promedio federado (FedAvg)** El promedio federado (FedAvg) se refiere a un algoritmo iterativo de aprendizaje federado que alterna entre entrenar modelos locales por separado y combinar los parámetros del modelo locales actualizados. El entrenamiento de los modelos locales se implementa a través de varios pasos de descenso de gradiente estocástico [105].

Vea también: algoritmo, FL, modelo local, parámetros del modelo, SGD

**protección de la privacidad** Consideremos un método de aprendizaje automático  $\mathcal{A}$  que recibe como entrada un conjunto de datos  $\mathcal{D}$  y entrega una salida  $\mathcal{A}(\mathcal{D})$ . La salida puede ser los parámetros del modelo aprendidos  $\hat{\mathbf{w}}$  o una predicción  $\hat{h}(\mathbf{x})$  obtenida para un punto de datos específico con atributos  $\mathbf{x}$ . Muchas aplicaciones importantes de aprendizaje au-

tomático involucran puntos de datos que representan a personas. Cada punto de datos se caracteriza por atributos  $\mathbf{x}$ , posiblemente una etiqueta  $y$ , y un atributo sensible  $s$  (por ejemplo, un diagnóstico médico). Mas o menos, la protección de la privacidad significa que debería ser imposible inferir, de la salida  $\mathcal{A}(\mathcal{D})$ , cualquier atributo sensible de los puntos de datos en  $\mathcal{D}$ . Matemáticamente, la protección de privacidad requiere que el mapeo  $\mathcal{A}(\mathcal{D})$  no sea invertible. En general, el solo hacer que  $\mathcal{A}(\mathcal{D})$  no sea invertible no es suficiente. Necesitamos que sea suficientemente no invertible.

Vea también: ML, conjunto de datos, parámetros del modelo, predicción, punto de datos, atributo, etiqueta, atributo sensible

**proximable** Una función convexa para la cual el operador proximal puede calcularse de manera eficiente se denomina a veces proximable o simple [106].

Vea también: convexo, operador proximal

**proyección** Consideremos un subconjunto  $\mathcal{W} \subseteq \mathbb{R}^d$  del espacio euclidiano de dimensión  $d$ . Definimos la proyección  $P_{\mathcal{W}}(\mathbf{w})$  de un vector  $\mathbf{w} \in \mathbb{R}^d$  sobre  $\mathcal{W}$  como

$$P_{\mathcal{W}}(\mathbf{w}) = \underset{\mathbf{w}' \in \mathcal{W}}{\operatorname{argmin}} \|\mathbf{w} - \mathbf{w}'\|_2. \quad (10)$$

En otras palabras,  $P_{\mathcal{W}}(\mathbf{w})$  es el vector en  $\mathcal{W}$  más cercano a  $\mathbf{w}$ . La proyección está bien definida solo para aquellos subconjuntos  $\mathcal{W}$  para los cuales existe el mínimo anterior [36].

Vea también: mínimo

**pseudoinversa** La pseudoinversa de Moore–Penrose  $\mathbf{A}^+$  de una matriz  $\mathbf{A} \in$

$\mathbb{R}^{m \times d}$  generaliza la noción de una inversa [3]. La pseudoinversa surge de forma natural en el contexto de la regresión ridge cuando se aplica a un conjunto de datos con etiqueta  $\mathbf{y}$  arbitrarios y una matriz de atributos  $\mathbf{X} = \mathbf{A}$  [28, Cap. 3]. Los parámetros del modelo aprendidos mediante regresión ridge están dados por

$$\hat{\mathbf{w}}^{(\alpha)} = (\mathbf{A}^T \mathbf{A} + \alpha \mathbf{I})^{-1} \mathbf{A}^T \mathbf{y}, \quad \alpha > 0.$$

Podemos entonces definir la pseudoinversa  $\mathbf{A}^+ \in \mathbb{R}^{d \times m}$  vía el límite [107, Cap. 3]

$$\lim_{\alpha \rightarrow 0^+} \hat{\mathbf{w}}^{(\alpha)} = \mathbf{A}^+ \mathbf{y}.$$

Vea también: inverse matrix, regresión ridge, conjunto de datos, etiquetas, matriz de atributos, parámetros del modelo

**puerta trasera (backdoor)** Un ataque de puerta trasera (backdoor) se refiere a la manipulación intencional del proceso de entrenamiento de un método de aprendizaje automático. Esta manipulación se puede implementar perturbando el conjunto de entrenamiento (envenenamiento de datos) o el algoritmo de optimización utilizado por un método basado en ERM. El objetivo de un ataque de puerta trasera es inclinar la hipótesis aprendida  $\hat{h}$  hacia predicciones específicas para un rango determinado de valores de atributos. Este rango de atributos actúa como una clave (o desencadenante) que desbloquea una puerta trasera en el sentido de generar predicciones anómalas. La clave  $\mathbf{x}$  y la predicción anómala correspondiente  $\hat{h}(\mathbf{x})$  son conocidas únicamente por el atacante.

Vea también: ML, conjunto de entrenamiento, datos, algoritmo, ERM, hipótesis, predicción, atributo,

**punto de dato etiquetado** Un punto de datos cuya etiqueta es conocida o ha sido determinada por algún medio, lo que podría requerir trabajo humano.

Vea también: punto de datos, etiqueta

**punto de datos** Un punto de datos es cualquier objeto que transmite información [49]. Ejemplos incluyen estudiantes, señales de radio, árboles, imágenes, variables aleatorias, números reales o proteínas. Los puntos de datos se describen típicamente mediante dos tipos de propiedades (o atributos):

- Atributos son propiedades medibles o computables de un punto de datos. Estos atributos pueden extraerse o calcularse automáticamente mediante sensores, computadoras u otros sistemas de recolección de datos. Para un punto de datos que representa a un paciente, una posible atributo podría ser el peso corporal.
- Etiquetas son hechos de nivel superior (o cantidades de interés) asociadas al punto de datos. Determinar las etiquetas de un punto de datos usualmente requiere experiencia humana o conocimiento del dominio. Para un punto de datos que representa a un paciente, un diagnóstico de cáncer proporcionado por un médico serviría como etiqueta.

La distinción entre atributos y etiquetas no siempre es clara. Una propiedad considerada como etiqueta en un contexto (por ejemplo,

un diagnóstico de cáncer) puede tratarse como una atributo en otro, particularmente si una automatización confiable (por ejemplo, mediante análisis de imágenes) permite calcularla sin intervención humana. El aprendizaje automático busca en términos generales predecir la etiqueta de un punto de datos a partir de sus atributos.

Vea tambien: datos, RV, atributo, etiqueta, ML.

**pérdida** Los métodos de aprendizaje automático usan una funcion de perdida  $L(\mathbf{z}, h)$  para medir el error incurrido al aplicar una hipótesis específica a un punto de datos específico. Con un pequeño abuso de notación, usamos el término pérdida tanto para la función de funcion de perdida  $L$  en sí como para el valor específico  $L(\mathbf{z}, h)$ , para un punto de datos  $\mathbf{z}$  y una hipótesis  $h$ .

Vea también: ML, función de pérdida, hipótesis, punto de datos.

**pérdida de error cuadrático** La perdida de error cuadrático mide el error de predicción de una hipótesis  $h$  al predecir una etiqueta numérica  $y \in \mathbb{R}$  a partir de los atributos  $\mathbf{x}$  de un punto de datos. Se define como

$$L((\mathbf{x}, y), h) := \left(y - \underbrace{h(\mathbf{x})}_{=\hat{y}}\right)^2.$$

.

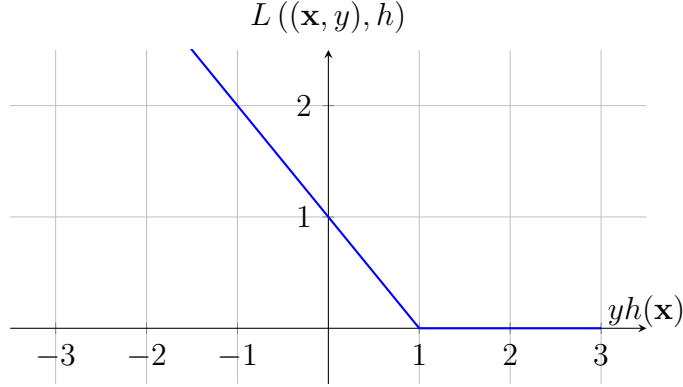
Vea también: pérdida, predicción, hipótesis, etiqueta, atributo, punto de datos

**pérdida de hinge** Consideremos un punto de datos caracterizado por un vector de atributos  $\mathbf{x} \in \mathbb{R}^d$  y una etiqueta binaria  $y \in \{-1, 1\}$ . La perdida de hinge incurrida por un mapa de hipótesis  $h(\mathbf{x})$  se define



como

$$L((\mathbf{x}, y), h) := \max\{0, 1 - yh(\mathbf{x})\}. \quad (11)$$



Una variante regularizada de la perdida de hinge es utilizada por las máquina de vectores de soporte [87]. Vea también: punto de datos, vector de atributos, etiqueta, pérdida, hipótesis, SVM.

**pérdida de Huber** La perdida de Huber unifica la pérdida de error cuadrático y la pérdida por error absoluto.

Vea también: pérdida, pérdida de error cuadrático, pérdida por error absoluto.

**pérdida logística** Consideremos un puntos de datos caracterizado por los atributos  $\mathbf{x}$  y una etiqueta  $y \in \{-1, 1\}$ . Utilizamos una hipótesis con valores reales  $h$  para predecir la etiqueta  $y$  a partir de los atributos  $\mathbf{x}$ . La perdida logística incurrida por esta predicción se define como

$$L((\mathbf{x}, y), h) := \log(1 + \exp(-yh(\mathbf{x}))). \quad (12)$$

Nótese cuidadosamente que la expresión (12) para la perdida logística aplica solo para el espacio de etiquetas  $\mathcal{Y} = \{-1, 1\}$  y cuando se usa la

regla de umbralización (1).

Vea también punto de datos, atributo, etiqueta, hipótesis, pérdida, predicción, espacio de etiquetas.

**pérdida por error absoluto** Considera un punto de datos con atributos  $\mathbf{x} \in \mathcal{X}$  y una etiqueta numérica  $y \in \mathbb{R}$ . La pérdida por error absoluto incurrida por una hipótesis  $h : \mathcal{X} \rightarrow \mathbb{R}$  se define como  $|y - h(\mathbf{x})|$ , es decir, la diferencia absoluta entre la predicción  $h(\mathbf{x})$  y la etiqueta verdadera  $y$ . Vea también: punto de datos, atributo, etiqueta, pérdida, hipótesis, predicción.

**realización** Consideremos una variable aleatoria  $x$  que asigna cada elemento (es decir, resultado o evento elemental)  $\omega \in \mathcal{P}$  de un espacio de probabilidad  $\mathcal{P}$  a un elemento  $a$  de un espacio medible  $\mathcal{N}$  [2, 46, 61]. Una realización de  $x$  es cualquier elemento  $a' \in \mathcal{N}$  tal que existe un elemento  $\omega' \in \mathcal{P}$  con  $x(\omega') = a'$ .

Vea también: RV, espacio de probabilidad

**recompensa** Una recompensa se refiere a alguna cantidad observada (o medida) que nos permite estimar la pérdida incurrida por la predicción (o decisión) de una hipótesis  $h(\mathbf{x})$ . Por ejemplo, en una aplicación de aprendizaje automático para vehículos autónomos,  $h(\mathbf{x})$  podría representar la dirección actual del volante de un vehículo. Podríamos construir una recompensa a partir de las mediciones de un sensor de colisión que indique si el vehículo se dirige hacia un obstáculo. Definimos una recompensa baja para la dirección del volante  $h(\mathbf{x})$  si el vehículo se

mueve peligrosamente hacia un obstáculo.

Vea también: pérdida, predicción, hipótesis, ML.

**red de aprendizaje federado (red FL)** Una red federada es un grafo no dirigido y ponderado, cuyos nodos representan generadores de datos que buscan entrenar un modelo local (o personalizado). Cada nodo de una red federada representa un dispositivo capaz de recopilar un conjunto de datos local y, a su vez, entrenar un modelo local. Los métodos de aprendizaje federado aprenden una hipótesis local  $h^{(i)}$  para cada nodo  $i \in \mathcal{V}$ , de manera que incurra en una pérdida baja sobre su conjunto de datos local.

Vea también: grafo, datos, modelo, dispositivo, conjunto de datos local, modelo local, FL, hipótesis

**red neuronal artificial (RNA)** Una RNA es una representación gráfica (de flujo de señales) de una función que mapea los atributos de un punto de datos en su entrada a una predicción para la correspondiente etiqueta en su salida. La unidad fundamental de una RNA es la neurona artificial, que aplica una función de activación a sus entradas ponderadas. Las salidas de estas neuronas sirven como entradas para otras neuronas, formando capas interconectadas.

Vea también: atributo, punto de datos, predicción, etiqueta, función de activación.

**red profunda** Una red profunda es una red neuronal artificial (RNA) con un número (relativamente) grande de capas ocultas. El aprendizaje profundo (deep learning) es un término general para los métodos de

aprendizaje automático que utilizan una red profunda como modelo [99].

Vea también: ANN, ML, modelo

**reducción de dimensionalidad** La reducción de dimensionalidad se refiere a métodos que aprenden una transformación  $h : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$  de un conjunto (típicamente grande) de atributos originales  $x_1, \dots, x_d$  a un conjunto más pequeño de atributos informativas  $z_1, \dots, z_{d'}$ . Usar un conjunto más pequeño de atributos ofrece múltiples ventajas:

- **Ventaja estadística:** Disminuir la cantidad de atributos reduce usualmente la dimensión efectiva de un modelo, lo cual a su vez disminuye el riesgo de sobreajuste.
- **Ventaja computacional:** Menos atributos implican menos carga computacional durante el entrenamiento de los modelos. Por ejemplo, los métodos de regresión lineal requieren invertir una matriz cuyo tamaño depende del número de atributos.
- **Visualización:** La reducción de dimensionalidad es clave para la visualización de datos. Por ejemplo, se puede aprender una transformación que genere dos atributos  $z_1, z_2$  y emplearlas como coordenadas en un diagrama de dispersión. La Figura 34 ilustra dicho diagrama de dispersión con dígitos manuscritos transformados a partir de imágenes de alta dimensión representadas por valores de escala de grises.

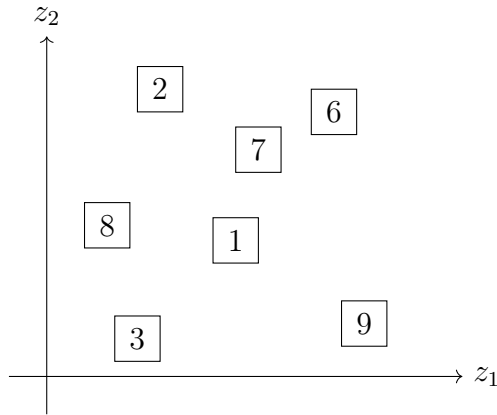


Figure 34: Ejemplo de reducción de dimensionalidad: Datos de imágenes de alta dimensión (p. ej., dígitos manuscritos) incrustados en 2D mediante atributos aprendidas  $(z_1, z_2)$  y visualizados en un diagrama de dispersión.

Véa también: sobreajuste, dimensión efectiva, modelo, diagrama de dispersión.

**referencia (baseline)** Consideremos un método de aprendizaje automático que produce una hipótesis aprendida (o un modelo entrenado)  $\hat{h} \in \mathcal{H}$ . Evaluamos la calidad del modelo entrenado mediante el cálculo de la pérdida promedio en un conjunto de prueba. Pero, ¿cómo saber si ese rendimiento es lo suficientemente bueno? ¿Cómo saber si el modelo entrenado se acerca al óptimo y si tiene sentido o no invertir más recursos (como recopilación de datos o potencia computacional) para mejorarlo? Para ello, es útil contar con un valor de referencia (o \*baseline\*) con el cual comparar el rendimiento del modelo entrenado. Este valor puede provenir del rendimiento humano, como la tasa de error de dermatólogos que diagnostican cáncer mediante inspección visual de la piel [108]. Otra

fuente de referencia puede ser un método de aprendizaje automático ya existente que, por alguna razón, no sea adecuado para la aplicación (por ejemplo, por ser computacionalmente costoso), pero cuya tasa de error en el conjunto de prueba puede servir como baseline. Un enfoque más fundamentado para construir una baseline es utilizar un modelo de probabilidad. En muchos casos, dado un modelo de probabilidad  $p(\mathbf{x}, y)$ , podemos determinar con precisión el mínimo riesgo alcanzable entre todas las hipótesis (incluso aquellas que no pertenecen al espacio de hipótesis  $\mathcal{H}$ ) [54]. Este mínimo alcanzable se conoce como riesgo de Bayes y corresponde al riesgo para la etiqueta  $y$  de un punto de datos, dados sus atributos  $\mathbf{x}$ . Dado una función de pérdida específica, el estimador de Bayes (si existe) está completamente determinado por la distribución de probabilidad  $p(\mathbf{x}, y)$  [54, Cap. 4]. Calcular el estimador de Bayes y el riesgo de Bayes presenta dos desafíos principales:

- 1) La distribución de probabilidad  $p(\mathbf{x}, y)$  desconocida y debe estimarse.
- 2) Incluso si se conoce  $p(\mathbf{x}, y)$  calcular el riesgo de Bayes puede ser computacionalmente muy costoso [109].

Un modelo de probabilidad ampliamente utilizado es la distribución normal multivariante  $(\mathbf{x}, y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  para puntos de datos caracterizados por atributos y etiquetas numéricas. En este caso, bajo la pérdida de error cuadrático, el estimador de Bayes corresponde a la media posterior  $\mu_{y|\mathbf{x}}$  de la etiqueta  $y$ , dado atributos  $\mathbf{x}$  [45, 54]. El riesgo de Bayes asociado es la varianza posterior  $\sigma_{y|\mathbf{x}}^2$  (see Figure 35).

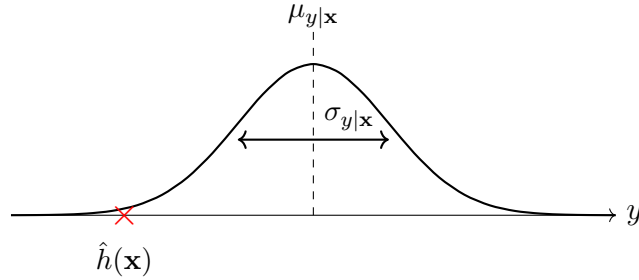


Figure 35: Si los atributos y la etiqueta de un punto de datos siguen una distribución normal multivariante, podemos alcanzar el mínimo riesgo (bajo pérdida de error cuadrático) usando el estimador de bayes  $\mu_{y|\mathbf{x}}$  para predecir la etiqueta  $y$  de un punto de datos con atributos  $\mathbf{x}$ . El mínimo riesgo es dado por la varianza posterior  $\sigma_{y|\mathbf{x}}^2$ . Podemos usar esta cantidad como baseline para evaluar la pérdida promedio del modelo  $\hat{h}$  entrenado.

Vea también: ML, hipótesis, modelo, pérdida, conjunto de prueba, datos, modelo probabilístico, mínimo, riesgo, espacio de hipótesis, riesgo de Bayes, Bayes estimator, distribución de probabilidad, distribución normal multivariante

**región de decisión** Consideremos un mapa de hipótesis  $h$  que entrega valores de un conjunto finito  $\mathcal{Y}$ . Para cada valor de etiqueta (categoría)  $a \in \mathcal{Y}$ , la hipótesis  $h$  determina un subconjunto de valores de atributos  $\mathbf{x} \in \mathcal{X}$  que resultan en el mismo resultado  $h(\mathbf{x}) = a$ . Nos referimos a este subconjunto como una región de decisión de la hipótesis  $h$ .

Vea también: hipótesis, etiqueta, atributo.

**reglamento general de protección de datos (RGPD)** El RGPD fue promulgado por la Union Europea (EU), y entró en efecto el 25 de Mayo de 2018 [102]. Protege la privacidad y los derechos sobre los datos de los individuos dentro de la EU. El RGPD tiene implicaciones significativas sobre cómo se recopilan, almacenan y utilizan los datos en aplicaciones de ML. Las disposiciones clave incluyen:

- Principio de minimización de datos: los sistemas de ML deben utilizar únicamente la cantidad necesaria de datos personal para su propósito.
- Transparencia y explicabilidad: los sistemas de ML deben permitir a sus usuarios comprender cómo se toman las decisiones que los afectan.
- Derechos del titular de los datos: los usuarios deben tener la posibilidad de acceder, rectificar y eliminar sus datos, así como oponerse a decisiones automatizadas y perfiles.
- Responsabilidad: las organizaciones deben garantizar una seguridad robusta de los datos y demostrar cumplimiento mediante documentación y auditorías periódicas.

**regresión** Los problemas de regresión se centran en la predicción de una etiqueta numérica basada únicamente en los atributos de un punto de datos [6, Ch. 2].

Vea también: etiqueta, atributo, punto de datos

**regresión de Huber** La regresión de Huber se refiere a métodos basados en ERM que utilizan la pérdida de Huber como medida del error



de predicción. Dos casos especiales importantes de la regresión de Huber son regresión por desviación absoluta mínima y regresión lineal. Ajustar el parámetro de umbral de la pérdida de Huber permite al usuario balancear la robustez del pérdida por error absoluto frente a los beneficios computacionales de la error cuadrática media suave.

Vea también: regresión, ERM, pérdida de Huber, predicción, regresión, regresión por desviación absoluta mínima, regresión lineal, pérdida por error absoluto, suave, pérdida de error cuadrático.

**regresión lineal** La regresión lineal tiene como objetivo aprender un mapa de hipótesis lineal para predecir una etiqueta numérica basada en los atributos numéricos de un punto de datos. La calidad de un mapa de hipótesis lineal se mide utilizando el promedio de error cuadrado medio incurrido en un conjunto de puntos de datos etiquetados, al que nos referimos como el conjunto de datos.

Vea también: regresión, hipótesis, etiqueta, atributo, punto de datos, pérdida de error cuadrático, punto de dato etiquetado, conjunto de entrenamiento.

**regresión logística** La regresión logística aprende un mapeo hipótesis lineal (o clasificador)  $h(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$  para predecir una etiqueta binaria  $y$  basada en el vector de atributos numérico  $\mathbf{x}$  de algún punto de datos. La calidad de un mapeo hipótesis lineal se mide por el promedio de la pérdida logística sobre algunos puntos de etiquetados (es decir, el conjunto de entrenamiento).

Vea también: regresión, hipótesis, clasificador, etiqueta, vector de

atributos, punto de datos, pérdida logística, punto de dato etiquetado, conjunto de entrenamiento.

**regresión polinómica** La regresión polinómica es una instancia de ERM que aprende una hipótesis aplicación polinómica para predecir una etiqueta numérica a partir de los atributos numéricos de un punto de datos. Para puntos de datos caracterizados por una única atributo numérica, la regresión polinómica utiliza el espacio de hipótesis

$$\mathcal{H}_d^{(\text{poly})} := \left\{ h(x) = \sum_{j=0}^{d-1} x^j w_j \right\}.$$

La calidad de una hipótesis polinómica aplicación se mide mediante la pérdida de error cuadrático promedio incurrida en un conjunto de punto de dato etiquetados (al que nos referimos como el conjunto de entrenamiento).

Vea también: regresión, hipótesis, etiqueta, atributo, punto de datos, espacio de hipótesis, pérdida de error cuadrático, punto de dato etiquetado, conjunto de entrenamiento.

**regresión por desviación absoluta mínima** La regresión por desviación absoluta mínima es una instancia de ERM que utiliza el error absoluto. Es un caso especial de regresión de Huber.

Vea también: ERM, pérdida por error absoluto, Huber regression.

**regresión ridge** La regresión ridge aprende los pesos  $\mathbf{w}$  de un mapa de hipótesis lineal  $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$ . La calidad de una elección particular para los parametros del modelo  $\mathbf{w}$  se mide por la suma de dos componentes. El primer componente es el promedio de error cuadrado medio

incurrido por  $h^{(\mathbf{w})}$  en un conjunto de puntos de datos etiquetados (es decir, el conjunto de entrenamiento). El segundo componente es el cuadrado de la norma Euclidiana escalada  $\alpha\|\mathbf{w}\|_2^2$  con un parámetro de regulación  $\alpha > 0$ . Agregar  $\alpha\|\mathbf{w}\|_2^2$  al promedio de medio cuadrado medio es equivalente a reemplazar cada punto de datos original por la realización de (infinitos) variables aleatorias iid centrados alrededor de estos puntos de datos (vea regulación).

Vea también: regresión, pesos, hipótesis, parámetros del modelo, pérdida de error cuadrático, punto de dato etiquetado, conjunto de entrenamiento, norma, regularización, punto de datos, realización, i.i.d., RV.

**regularización** Un desafío clave de las aplicaciones modernas de aprendizaje automático es que a menudo usan modelos grandes, que tienen un dimension efectiva del orden de miles de millones. Entrenar un modelo de alta dimensión utilizando métodos basados en ERM es propenso al sobreajuste, es decir, la hipótesis aprendida tiene un buen rendimiento en el conjunto de entrenamiento pero un rendimiento pobre fuera del conjunto de entrenamiento. La regularización se refiere a modificaciones de una instancia dada de ERM para evitar el sobreajuste, es decir, para asegurar que la hipótesis aprendida no tenga un rendimiento mucho peor fuera del conjunto de entrenamiento. Existen tres rutas para implementar la regularización:

- 1) Modelo podado: Podemos el modelo original  $\mathcal{H}$  para obtener un modelo más pequeño  $\mathcal{H}'$ . Para un modelo paramétrico, el podado

puede implementarse mediante restricciones en los parametros del modelo (como  $w_1 \in [0.4, 0.6]$  para el peso del atributo  $x_1$  en regresin lineal).

- 2) Perdida penalización: Modificamos la funcion objetivo de ERM añadiendo un término de penalización al error de entrenamiento. El término de penalización estima cuánto mayor es la perdida esperada (o riesgo) en comparación con la perdida promedio en el conjunto de entrenamiento.
- 3) Aumentación de datos: Podemos ampliar el conjunto de entrenamiento  $\mathcal{D}$  añadiendo copias perturbadas de los puntos de datos originales en  $\mathcal{D}$ . Un ejemplo de tal perturbación es añadir la realización de una variable aleatoria al vector de atributos de un punto de datos.

La Fig. 36 ilustra las tres rutas anteriores para la regularización. Estas rutas están estrechamente relacionadas y, a veces, son completamente equivalentes. Aumentación de datos usando VA gaussianas para perturbar los vector de atributos en el conjunto de entrenamiento de regresión lineal tiene el mismo efecto que añadir la penalización  $\lambda \|\mathbf{w}\|_2^2$  al error de entrenamiento (que no es más que regresión ridge). La decisión sobre qué ruta usar para la regularización puede basarse en la infraestructura computacional disponible. Por ejemplo, podría ser mucho más fácil implementar aumentación de datos que el podado de modelo.

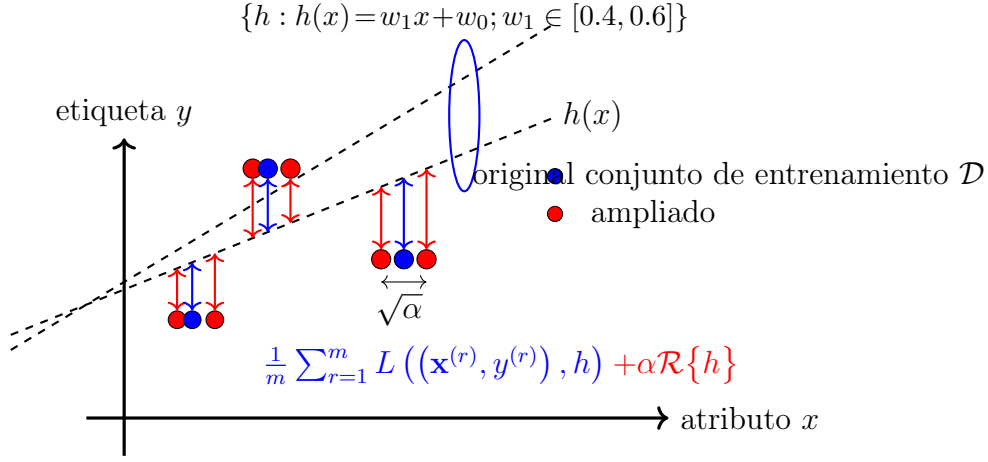


Figure 36: Tres enfoques para la regularización: 1) aumentación de datos; 2) pérdida penalización; y 3) modelo podado (mediante restricciones en parámetros del modelo).

Véa también: ML, modelo, dimensión efectiva, ERM, sobreajuste, hipótesis, conjunto de entrenamiento, parámetros del modelo, atributo, regresión lineal, pérdida, función objetivo, error de entrenamiento, riesgo, aumentación de datos, punto de datos, realización, RV, vector de atributos, VA gaussiana, regresión ridge, etiqueta.

**regularizador** Un regularizador asigna a cada hipótesis  $h$  de un espacio de hipótesis  $\mathcal{H}$  una medida cuantitativa  $\mathcal{R}\{h\}$  que indica cuánto podría diferir su error de predicción en un conjunto de entrenamiento de sus errores de predicción en puntos de datos fuera del conjunto de entrenamiento. La regresión ridge utiliza el regularizador  $\mathcal{R}\{h\} := \|\mathbf{w}\|_2^2$  para mapas de hipótesis lineales  $h^{(\mathbf{w})}(\mathbf{x}) := \mathbf{w}^T \mathbf{x}$  [6, Ch. 3]. El operador

de reducción y selección absoluta mínima (Lasso) utiliza el regularizador  $\mathcal{R}\{h\} := \|\mathbf{w}\|_1$  para mapas de hipótesis lineales  $h^{(\mathbf{w})}(\mathbf{x}) := \mathbf{w}^T \mathbf{x}$  [6, Ch. 3].

Vea también: hipótesis, espacio de hipótesis, predicción, conjunto de entrenamiento, punto de datos, conjunto de entrenamiento, Regresión ridge, Lasso

**divergencia de Rényi** La divergencia de Rényi mide la (dis)similitud entre dos distribuciones de probabilidad [110].

**riesgo** Consideremos una hipótesis  $h$  que se utiliza para predecir la etiqueta  $y$  de un punto de datos a partir de sus atributos  $\mathbf{x}$ . Evaluamos la calidad de una predicción específica usando una función de pérdida  $L((\mathbf{x}, y), h)$ . Si interpretamos los puntos de datos como realizaciones de variables aleatorias iid, entonces  $L((\mathbf{x}, y), h)$  también se convierte en una realización de una variable aleatoria. La suposición de independencia e idéntica distribución nos permite definir el riesgo de una hipótesis como la expectativa de la pérdida  $\mathbb{E}\{L((\mathbf{x}, y), h)\}$ . El riesgo de  $h$  depende tanto de la función de pérdida elegida como de la distribución de probabilidad de los puntos de datos.

Vea también: hipótesis, etiqueta, punto de datos, atributo, predicción, función de pérdida, realización, RV, i.i.d., realization, RV, suposición i.i.d., expectation, pérdida, distribución de probabilidad

**riesgo de Bayes** Considera un modelo de probabilidad con una distribución de probabilidad conjunta  $p(\mathbf{x}, y)$  para los atributos  $\mathbf{x}$  y la etiqueta  $y$  de un punto de datos. El riesgo de Bayes es el riesgo mínimo posible

que puede alcanzarse por cualquier hipótesis  $h : \mathcal{X} \rightarrow \mathcal{Y}$ . Cualquier hipótesis que alcance el riesgo de Bayes se denomina un estimador de Bayes [54].

Vea también: modelo probabilístico, distribución de probabilidad, atributo, etiqueta, punto de datos, riesgo, mínimo, hipótesis, Bayes estimator.

**riesgo empírico** El riesgo empírico  $\hat{L}(h|\mathcal{D})$  de una hipótesis sobre un conjunto de datos  $\mathcal{D}$  es la pérdida promedio incurrida por  $h$  al aplicarse a los puntos de datos en el  $\mathcal{D}$ .

Vea también: riesgo, hipótesis, conjunto de datos, pérdida, punto de datos

**robustez** La robustez es un requisito clave para la IA confiable. Se refiere a la propiedad de un sistema de aprendizaje automático de mantener un rendimiento aceptable incluso cuando se somete a diferentes formas de perturbaciones. Estas perturbaciones pueden afectar los atributos de un punto de datos con el fin de manipular la predicción producida por un modelo de aprendizaje automático ya entrenado. La robustez también abarca la estabilidad de los métodos basados en ERM frente a perturbaciones en el conjunto de entrenamiento. Estas perturbaciones pueden ocurrir como parte de ataque de envenenamiento de datos.

Véa también: trustworthy AI, ML, atributo, punto de datos, predicción, modelo, stability, ERM, conjunto de entrenamiento, envenenamiento de datos, ataque.

**régimen de alta dimensión** El régimen de alta dimensión de minimización empírica del riesgo se caracteriza por que la dimensión efectiva del

modelo es mayor que el tamaño de la muestra, es decir, el número de puntos de datos etiquetados en el conjunto de entrenamiento. Por ejemplo, los métodos de regresión lineal operan en el régimen de alta dimensión cuando el número  $d$  de atributos utilizados para caracterizar los puntos de datos excede el número de puntos de datos en el conjunto de entrenamiento. Otro ejemplo de métodos de aprendizaje automático que operan en el régimen de alta dimensión son las red neuronal artificial grandes, que tienen muchos más pesos ajustables (y términos de sesgo) que el número total de puntos de datos en el conjunto de entrenamiento. La estadística de alta dimensión es una línea principal reciente de la teoría de probabilidad que estudia el comportamiento de los métodos de aprendizaje automático en el régimen de alta dimensión [40, 111].

Vea también: ERM, dimensión efectiva, modelo, tamaño de la muestra, punto de datos, conjunto de entrenamiento, regresión lineal, atributo, ML, ANN, pesos, probabilidad.

**semi-definida positiva (psd)** Una matriz simétrica (con valores reales)

$\mathbf{Q} = \mathbf{Q}^T \in \mathbb{R}^{d \times d}$  se denomina semi-definida positiva (psd) si  $\mathbf{x}^T \mathbf{Q} \mathbf{x} \geq 0$  para todo vector  $\mathbf{x} \in \mathbb{R}^d$ . La propiedad de ser semi-definida positiva puede extenderse desde matrices a funciones kernel simétricas (con valores reales)  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  (con  $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$ ) de la siguiente manera: Para cualquier conjunto finito de vectores de atributos  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}$ , la matriz resultante  $\mathbf{Q} \in \mathbb{R}^{m \times m}$  con entradas  $Q_{r,r'} = K(\mathbf{x}^{(r)}, \mathbf{x}^{(r')})$  es semi-definida positiva [86].

Vea también: kernel, vector de atributos



**sesgo** Consideremos un método de aprendizaje automático que utiliza un espacio de hipótesis  $\mathcal{H}$  parametrizado. Este aprende los parámetros del modelo  $\mathbf{w} \in \mathbb{R}^d$  utilizando el conjunto de datos

$$\mathcal{D} = \left\{ \left( \mathbf{x}^{(r)}, y^{(r)} \right) \right\}_{r=1}^m.$$

Para analizar las propiedades del método de aprendizaje automático, típicamente interpretamos los puntos de datos como realizaciones de iid. variables aleatorias,

$$y^{(r)} = h(\bar{\mathbf{w}})(\mathbf{x}^{(r)}) + \varepsilon^{(r)}, r = 1, \dots, m.$$

Entonces podemos interpretar el método de aprendizaje automático como un estimador  $\hat{\mathbf{w}}$  calculado a partir de  $\mathcal{D}$  (por ejemplo, resolviendo ERM). El sesgo (cuadrado) del estimador  $\hat{\mathbf{w}}$  se define como  $B^2 := \|\mathbb{E}\{\hat{\mathbf{w}}\} - \bar{\mathbf{w}}\|_2^2$ .  
Vea también: ML, espacio de hipótesis, parámetros del modelo, punto de datos, realización, i.i.d., RV, ERM

**sobreajuste** Consideremos un método de aprendizaje automático que utiliza ERM para aprender una hipótesis con el mínimo riesgo empírico en un conjunto de entrenamiento dado. Dicho método presenta sobreajuste del conjunto de entrenamiento si aprende una hipótesis con un riesgo empírico pequeño sobre el conjunto de entrenamiento pero una pérdida significativamente mayor fuera de él conjunto de entrenamiento.  
Vea también: ML, ERM, hipótesis, mínimo, riesgo empírico, conjunto de entrenamiento, pérdida

**stability** Stability is a desirable property of a ML method  $\mathcal{A}$  that maps a conjunto de datos  $\mathcal{D}$  (e.g., a conjunto de entrenamiento) to an output

$\mathcal{A}(\mathcal{D})$ , such as learned parámetros del modelo or the predicción for a specific punto de datos. Intuitively,  $\mathcal{A}$  is stable if small changes in the input conjunto de datos  $\mathcal{D}$  lead to small changes in the output  $\mathcal{A}(\mathcal{D})$ . Several formal notions of stability exist that enable bounds on the generalización error or riesgo of the method; see [44, Ch. 13]. To build intuition, consider the three datasets depicted in Fig. 37, each of which is equally likely under the same datos-generating distribución de probabilidad. Since the optimal parámetros del modelo are determined by this underlying distribución de probabilidad, an accurate ML method  $\mathcal{A}$  should return the same (or very similar) output  $\mathcal{A}(\mathcal{D})$  for all three conjunto de datos. In other words, any useful  $\mathcal{A}$  must be robust to variability in sample realizaciones from the same distribución de probabilidad, i.e., it must be stable. .

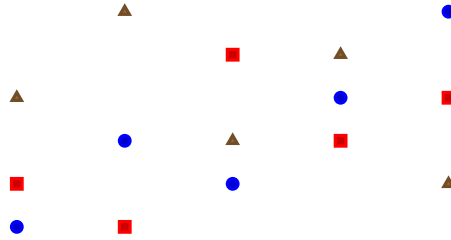


Figure 37: Three conjunto de datoss  $\mathcal{D}^{(*)}$ ,  $\mathcal{D}^{(\square)}$ , and  $\mathcal{D}^{(\Delta)}$ , each sampled independently from the same datos-generating distribución de probabilidad. A stable ML method should return similar outputs when trained on any of these conjunto de datoss.

Vea también: ML, conjunto de datos, conjunto de entrenamiento, parámetros del modelo, predicción, punto de datos, generalización, riesgo, datos, distribución de probabilidad,

**suave** Una función con valores reales  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  es suave si es diferenciable y su gradiente  $\nabla f(\mathbf{w})$  es continuo en todos  $\mathbf{w} \in \mathbb{R}^d$  [74], [112]. Una función suave  $f$  se denomina  $\beta$ -suave si el gradiente  $\nabla f(\mathbf{w})$  es Lipschitz continuo con constante de Lipschitz  $\beta$ , es decir,

$$\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{w}')\| \leq \beta \|\mathbf{w} - \mathbf{w}'\|, \text{ for any } \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d.$$

La constante  $\beta$  cuantifica el grado de suavidad de la función  $f$ : cuanto menor es  $\beta$ , más suave es  $f$ . Los problemas de optimización con una función objetivo suave pueden resolverse eficazmente mediante metodos de descenso de gradiente. De hecho, los metodos de descenso de gradiente aproximan la funcion objetivo localmente alrededor de una elección actual  $\mathbf{w}$  utilizando su gradiente. Esta aproximación funciona bien si el gradiente no cambia demasiado rápido. Podemos precisar esta afirmación informal estudiando el efecto de un solo paso de gradiente con tamaño de paso  $\eta = 1/\beta$  (vea Fig. 38).

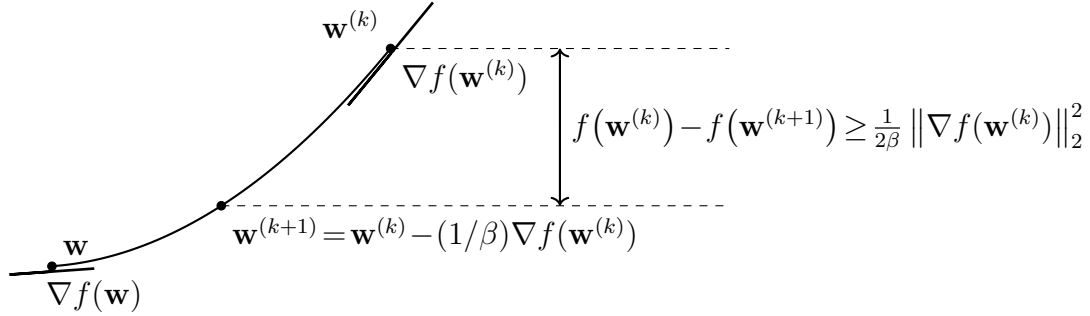


Figure 38: Considere un funcion de objetivo  $f(\mathbf{w})$  que es  $\beta$ -suave. Tomar un paso de gradiente, con tamaño de paso  $\eta = 1/\beta$ , disminuye el objetivo por al menos  $\frac{1}{2\beta} \|\nabla f(\mathbf{w}^{(k)})\|_2^2$  [74], [112], [75]. Nótese que el tamaño de paso  $\eta = 1/\beta$  se hace más grande para un  $\beta$  más pequeño. Por lo tanto, para funciones de objetivo más suaves (es decir, aquellas con un  $\beta$  más pequeño), podemos tomar pasos más grandes.

Vea tambien: diferenciable, gradiente, función objetivo, métodos de gradiente, paso de gradiente, tamaño de paso.

**subajuste** Consideremos un método de aprendizaje automatico que utiliza ERM para aprender una hipótesis con el minimo riesgo empírico en un conjunto de entrenamiento dado. Dicho método presenta subajuste del conjunto de entrenamiento si no es capaz de aprender una hipótesis con un rieso empírico suficientemente pequeño sobre el conjunto de entrenamiento. Si un método sufre de subajuste, típicamente tampoco podrá aprender una hipótesis con un riesgo pequeño.

Vea también: ML, ERM, hipótesis, mínimo, riesgo empírico, conjunto de entrenamiento, riesgo

**subgradiente** Para una función de valor real  $f : \mathbb{R}^d \rightarrow \mathbb{R} : \mathbf{w} \mapsto f(\mathbf{w})$ , un vector  $\mathbf{a}$  tal que  $f(\mathbf{w}) \geq f(\mathbf{w}') + (\mathbf{w} - \mathbf{w}')^T \mathbf{a}$  se denomina subgradiente de  $f$  en  $\mathbf{w}'$  [53, 98].

**suposición de agrupamiento** La suposición de agrupamiento postula que los puntos de datos en un conjunto de datos forman un (pequeño) número de grupos o clústers. Los puntos de datos en el mismo clúster son más similares entre sí que aquellos que están fuera del clúster [23]. Obtenemos diferentes métodos de agrupamiento utilizando diferentes nociones de similitud entre puntos de datos.

Vea también: punto de datos, conjunto de datos, cluster, agrupamiento

**suposición de independencia e idéntica distribución (suposición i.i.d.)**

La suposición iid interpreta los puntos de datos de un conjunto de datos como las realizaciones de variables aleatorias iid.

Vea también: i.i.d., punto de datos, conjunto de datos, realización, RV.

**supremo (o mínimo de las cotas superiores)** El supremo de un conjunto de números reales es el número más pequeño que es mayor o igual que todos los elementos del conjunto. Formalmente, un número real  $a$  es el supremo de un conjunto  $\mathcal{A} \subseteq \mathbb{R}$  si: 1)  $a$  es una cota superior de  $\mathcal{A}$ ; y 2) ningún número menor que  $a$  es una cota superior de  $\mathcal{A}$ . Todo conjunto no vacío de números reales que esté acotado superiormente tiene un supremo, aun si no contiene su supremo como un elemento [2, Sec. 1.4].

**tamaño de la muestra** El número de puntos de datos individuales contenidos en un conjunto de datos.

Vea también: punto de datos, conjunto de datos.

**tamaño de paso** Consulte tasa de aprendizaje.

**tarea de aprendizaje** Considere un conjunto de datos  $\mathcal{D}$  que consiste en múltiples puntos de datos  $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$ . Por ejemplo,  $\mathcal{D}$  podría representar una colección de imágenes en una base de datos. Una tarea de aprendizaje se define especificando aquellas propiedades (o atributos) de un punto de datos que se usan como sus atributos y etiquetas. Dada una elección de modelo  $\mathcal{H}$  y función de pérdida, una tarea de aprendizaje conduce a una instancia de ERM y puede representarse mediante la función de objetivo asociada  $\hat{L}(h|\mathcal{D})$  para  $h \in \mathcal{H}$ . Es importante destacar que se pueden construir múltiples tareas de aprendizaje distintas a partir del mismo conjunto de datos seleccionando diferentes conjuntos de atributos y etiquetas.



Figure 39: \*

Una imagen que muestra vacas pastando en el campo austriaco.

Tarea 1 (Regresión):

- Atributos: Los valores RGB de todos los píxeles de la imagen.
- Etiqueta: El número de vacas representadas.

Tarea 2 (Clasificación):

- Atributos: La intensidad promedio de verde en la imagen.
- Etiqueta: ¿Deben moverse las vacas a otra ubicación (sí/no)?

Figure 40: Dos tareas de aprendizaje construidas a partir de un solo conjunto de datos de imágenes. Estas tareas difieren en la selección de atributo y en la elección de etiqueta (es decir, el objetivo), pero ambas se derivan del mismo conjunto de datos.

. Tales tareas están inherentemente relacionadas, y resolverlas conjun-

tamente, por ejemplo usando métodos de aprendizaje multitarea, suele ser más eficaz que tratarlas de forma independiente [113], [114], [115]. Véase también: conjunto de datos, modelo, función de pérdida, función objetivo, aprendizaje multitarea, espacio de etiquetas.

**tasa de aprendizaje** Considere un método iterativo de aprendizaje automático para encontrar o aprender una hipótesis útil  $h \in \mathcal{H}$ . Dicho método iterativo repite pasos computacionales similares (de actualización) que ajustan o modifican la hipótesis actual para obtener una hipótesis mejorada. Un ejemplo bien conocido de este tipo de método de aprendizaje iterativo es el descenso por gradiente y sus variantes, descenso por gradiente estocástico y descenso por gradiente proyectado. Un parámetro clave de un método iterativo es la tasa de aprendizaje. La tasa de aprendizaje controla la magnitud en que la hipótesis actual puede modificarse durante una sola iteración. Un ejemplo conocido de este parámetro es el tamaño de paso utilizado en descenso por gradiente [6, Ch. 5].

Vea también: ML, hipótesis, GD, SGD, descenso por gradiente proyectado (GD proyectado), tamaño de paso.

**teorema central del límite (CLT)** Consideremos una secuencia de variable aleatoria iid  $x^{(r)}$ , para  $r = 1, 2, \dots$ , cada una con media cero y varianza finita  $\sigma^2 > 0$ . El teorema central del límite (CLT) establece que la suma normalizada

$$s^{(m)} := \frac{1}{\sqrt{m}} \sum_{r=1}^m x^{(r)}$$

converge en distribución a una variable aleatoria gaussiana con media



cero y varianza  $\sigma^2$  cuando  $m \rightarrow \infty$  [116, Proposition 2.17]. Una forma elegante de derivar el CLT es mediante la funcion característica de la suma normalizada  $s^{(m)}$ . Sea  $\phi(t) = \mathbb{E}\{\exp(jtx)\}$  (donde  $j = \sqrt{-1}$  es la unidad imaginaria) la funcion característica común de cada término  $x^{(r)}$ , y sea  $\phi^{(m)}(t)$  la funcion característica de  $s^{(m)}$ . Definimos un operador  $\mathcal{T}$  que actúa sobre funciones de característica tal que

$$\phi^{(m)}(t) = \mathcal{T}(\phi^{(m-1)})(t) := \phi\left(\frac{t}{\sqrt{m}}\right) \cdot \phi^{(m-1)}\left(\frac{\sqrt{m-1}}{\sqrt{m}}t\right).$$

Esta iteracion de punto fijo captura el efecto de añadir recursivamente una variable aleatoria iid  $\mathbf{x}^{(m)}$  y reescalar. Al aplicar  $\mathcal{T}$  iterativamente, se obtiene la convergencia de  $\phi^{(m)}(t)$  hacia el punto fijo

$$\phi^*(t) = \exp(-t^2\sigma^2/2)$$

que es la funcion característica de una variable aleatoria gaussiana con media cero y varianza  $\sigma^2$ . Las generalizaciones del CLT permiten variables aleatorias dependientes o no idénticamente distribuidas [116, Sec. 2.8].

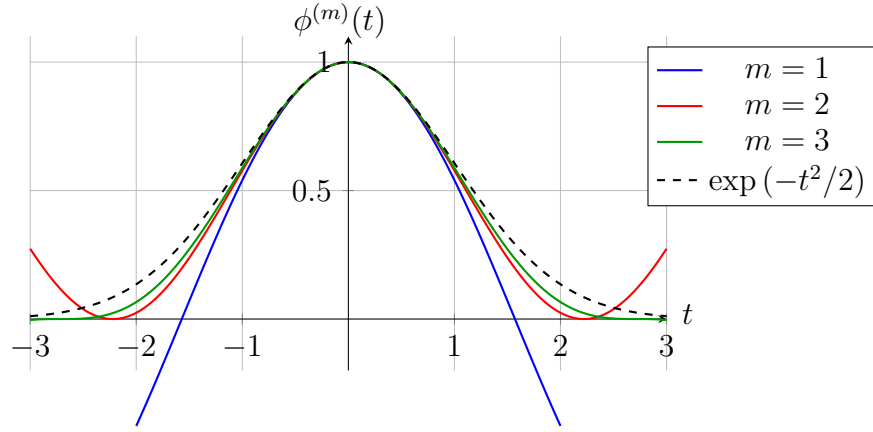


Figure 41: Funciones característica de sumas normalizadas de variables aleatorias iid  $x^{(r)} \in \{-1, 1\}$  para  $r = 1, \dots, m$ , comparadas con el límite gaussiano.

Vea tambien: RV, VA gaussiana.

**transparencia** La transparencia es un requisito fundamental para una inteligencia artificial confiable [117]. En aprendizaje automático, suele utilizarse como sinónimo de explicabilidad [66, 118] pero en el contexto más amplio de sistemas de inteligencia artificial, incluye también información sobre limitaciones, confiabilidad y uso previsto. En sistemas de diagnóstico médico, se requiere informar el nivel de confianza de una predicción. En aplicaciones financieras como la puntuación crediticia, las decisiones automatizadas basadas en inteligencia artificial deben ir acompañadas de explicaciones sobre los factores que influyeron en ellas, como el nivel de ingresos o el historial crediticio. Estas explicaciones permiten que las personas (por ejemplo, un solicitante de crédito) comprendan y, si es necesario, impugnen decisiones

automatizadas.. Algunos métodos de aprendizaje automático ofrecen transparencia de manera intrínseca. Por ejemplo, la regresión logística permite interpretar la fiabilidad de una clasificación mediante el valor absoluto  $|h(\mathbf{x})|$ . Los árboles de decisión, también son consideradas transparentes porque generan reglas comprensibles para los humanos. [20]. La transparencia también requiere que se informe explícitamente cuando una persona está interactuando con un sistema de inteligencia artificial. Por ejemplo, los chatbots impulsados por inteligencia artificial deben indicar claramente que no son humanos. Además, la transparencia incluye documentación exhaustiva que detalle el propósito, las decisiones de diseño y los casos de uso previstos del sistema. Ejemplos de esto son las hojas de datos de modelos [35] y las tarjetas de sistemas de inteligencia artificial [119], que ayudan a los desarrolladores y usuarios a entender las limitaciones y aplicaciones adecuadas de un sistema de inteligencia artificial [120].

Vea también: trustworthy AI, ML, AI, predicción, regresión logística, clasificación, árbol de decisión, modelo

**unidad lineal rectificada (ReLU)** La unidad lineal rectificada (ReLU) es una elección popular para la función de activación de una neurona dentro de una red neuronal artificial. Se define como  $\sigma(z) = \max\{0, z\}$ , donde  $z$  es la entrada ponderada de la neurona artificial.

Vea también: función de activación, ANN

**validación** La validación se refiere a la práctica de evaluar la pérdida incurrida por una hipótesis  $\hat{h}$  que ha sido aprendida mediante algún método

de aprendizaje automático, por ejemplo, resolviendo ERM en un conjunto de entrenamiento  $\mathcal{D}$ . La validación implica evaluar el desempeño de la hipótesis en un conjunto de puntos de datos que no están contenidos en el conjunto de entrenamiento  $\mathcal{D}$ .

Vea también: pérdida, hipótesis, ML, ERM, conjunto de entrenamiento, punto de datos

**validación cruzada de  $k$  particiones ( $k$ -fold CV)** La validación cruzada de  $k$  particiones es un método para aprender y validar una hipótesis utilizando un conjunto de datos dado. Este método divide el conjunto de datos equitativamente en  $k$  subconjuntos o particiones y luego ejecuta  $k$  repeticiones de entrenamiento de modelo (por ejemplo, mediante ERM) y validación. Cada repetición utiliza una partición diferente como conjunto de validación y las  $k - 1$  particiones restantes como conjunto de entrenamiento. El resultado final es el promedio de los errores de validación obtenidos desde las  $k$  repeticiones.

Vea también: hipótesis, conjunto de datos, modelo, ERM, validación, conjunto de validación, conjunto de entrenamiento, error de validación.

**valor atípico (outlier)** Muchos métodos de aprendizaje Automático están motivados por la suposición de independencia e idéntica distribución, que interpreta los puntos de datos como realizaciones de variables aleatorias iid con una distribución de probabilidad común. La suposición de independencia e idéntica distribución es útil en aplicaciones donde las propiedades estadísticas del proceso de generación de datos son estacionarias (o invariantes en el tiempo) [121]. Sin embargo, en algunas

aplicaciones, los datos están compuestos por una mayoría de puntos de datos regulares que cumplen con la suposición de independencia e idéntica distribución y un pequeño número de puntos de datos que presentan propiedades estadísticas fundamentalmente diferentes en comparación con los puntos de datos regulares. Nos referimos a un punto de datos que se desvía significativamente de las propiedades estadísticas de la mayoría como un valor atípico (outlier). Los diferentes métodos de detección de valores atípicos utilizan distintas medidas para evaluar esta desviación. La teoría del aprendizaje estadístico estudia los límites fundamentales sobre la capacidad de mitigar valores atípicos de manera confiable [122, 123].

Vea también: ML, suposición i.i.d., punto de datos, realización, i.i.d., RV, distribución de probabilidad, datos.

**valor propio (eigenvalue)** Nos referimos a un número  $\lambda \in \mathbb{R}$  como un valor propio de una matriz cuadrada  $\mathbf{A} \in \mathbb{R}^{d \times d}$  si existe un vector no nulo  $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$  tal que  $\mathbf{Ax} = \lambda\mathbf{x}$ .

**variable aleatoria (RV)** Una RV es una función que mapea desde un espacio de probabilidad  $\mathcal{P}$  a un espacio de valores [45, 46]. El espacio de probabilidad consiste en eventos elementales y está equipado con una medida de probabilidad que asigna probabilidades a subconjuntos de  $\mathcal{P}$ . Existen diferentes tipos de variables aleatorias (RV), que incluyen:

- RVs binarias, que asignan eventos elementales a un conjunto de dos valores distintos, como  $\{-1, 1\}$  o  $\{\text{cat}, \text{no cat}\}$ ;
- RVs de valor real, que toman valores en los números reales  $\mathbb{R}$ ;

- RVs de valor vectorial, que mapean eventos elementales al espacio euclidiano  $\mathbb{R}^d$ .

La teoría de probabilidad utiliza el concepto de espacios medibles para definir rigurosamente y estudiar las propiedades de (grandes) colecciones de RVs [46].

Vea también: espacio de probabilidad, probabilidad, espacio euclidiano,

**variable aleatoria gaussiana (VA gaussiana)** Una variable aleatoria gaussiana estándar es una variable aleatoria real  $x$  con función de densidad de probabilidad (pdf) [5, 45, 88]

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp^{-x^2/2}.$$

Dada una variable aleatoria gaussiana estándar  $x$ , podemos construir una variable aleatoria gaussiana general  $x'$  con media  $\mu$  y varianza  $\sigma^2$  mediante  $x' := \sigma(x + \mu)$ . La distribución de probabilidad de una variable aleatoria gaussiana se conoce como distribución normal, denotada  $\mathcal{N}(\mu, \sigma)$ .

Un vector aleatorio gaussiano  $\mathbf{x} \in \mathbb{R}^d$  con matriz de covarianza  $\mathbf{C}$  y media  $\boldsymbol{\mu}$  puede construirse como  $\mathbf{x} := \mathbf{A}(\mathbf{z} + \boldsymbol{\mu})$ . Aquí,  $\mathbf{A}$  es cualquier matriz que satisface  $\mathbf{A}\mathbf{A}^T = \mathbf{C}$  y  $\mathbf{z} := (z_1, \dots, z_d)^T$  es un vector cuyos elementos son iid gaussianas estándar variables aleatorias  $z_1, \dots, z_d$ . Los procesos aleatorios gaussianos generalizan los vectores aleatorios gaussianos aplicando transformaciones lineales a a secuencias infinitas de variables aleatorias gaussianas estándar [124].

Las variables aleatorias gaussianas se utilizan ampliamente como modelos de probabilidad en el análisis estadístico de métodos de aprendizaje

automático. Su importancia se debe, en parte, al teorema del límite central, que establece que el promedio de un número creciente de variables aleatorias independientes (aunque no sean gaussianas) converge a una variable aleatoria gaussiana [56].

Vea también: RV, pdf, media, varianza, distribución de probabilidad, matriz de covarianza, i.i.d., ML

**variación total** Vea GTV.

**variación total generalizada (GTV)** GTV es una medida de la variación de los modelos locales entrenados  $h^{(i)}$  (o de sus parametros del modelo  $\mathbf{w}^{(i)}$ ) asignados a los nodos  $i = 1, \dots, n$  de un grafo no dirigido ponderado  $\mathcal{G}$  con aristas  $\mathcal{E}$ . Dada una medida  $d^{(h, h')}$  para la discrepancia entre mapas de hipótesis  $h, h'$ , la GTV se define como:

$$\sum_{\{i, i'\} \in \mathcal{E}} A_{i, i'} d^{(h^{(i)}, h^{(i')})}.$$

Here,  $A_{i, i'} > 0$  denota el peso de la arista no dirigida  $\{i, i'\} \in \mathcal{E}$ .

Vea también: modelo local, parámetros del modelo, grafo, discrepancia, hipótesis

**varianza** La varianza de una variable aleatoria real  $x$  se define como la esperanza  $\mathbb{E}\{(x - \mathbb{E}\{x\})^2\}$  de la diferencia cuadrada entre  $x$  y su esperanza  $\mathbb{E}\{x\}$ . Extendemos esta definición a variables aleatorias vectoriales  $\mathbf{x}$  como  $\mathbb{E}\{\|\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\|_2^2\}$ .

Vea también: RV, expectation

**vecino más cercano (NN)** Los métodos de vecino más cercano (NN) aprenden una hipótesis  $h : \mathcal{X} \rightarrow \mathcal{Y}$  cuyo valor  $h(\mathbf{x})$  se determina únicamente

por los vecinos más cercanos dentro de un conjunto de datos. Distintos métodos usan diferentes medidas para determinar los vecinos más cercanos. Si los puntos de datos se caracterizan por vector de atributos numéricos, podemos usar la distancia euclidiana como medida. the metric.

Vea también: hipótesis, vecinos, conjunto de datos, punto de datos, vector de atributos

**vecinos** Los vecinos de un nodo  $i \in \mathcal{V}$  dentro de una red de aprendizaje federado (red FL) son los nodos  $i' \in \mathcal{V} \setminus \{i\}$  conectados con  $i$  por una arista.

Vea también: red FL

**vector** Un vector es un elemento de un espacio vectorial. En el contexto del aprendizaje automático, un ejemplo particularmente importante de espacios vectoriales es el espacio euclidiano  $\mathbb{R}^d$ , donde  $d \in \mathbb{N}$  representa la dimensión (finita) del espacio. Un vector  $\mathbf{x} \in \mathbb{R}^d$  puede representarse como una lista o arreglo unidimensional de números reales, es decir,  $x_1, \dots, x_d$  con  $x_j \in \mathbb{R}$  para  $j = 1, \dots, d$ . El valor  $x_j$  es la  $j$ -ésima entrada del vector  $\mathbf{x}$ . También puede ser útil considerar un vector  $\mathbf{x} \in \mathbb{R}^d$  como una función que asigna a cada índice  $j \in \{1, \dots, d\}$  un valor  $x_j \in \mathbb{R}$ , es decir,  $\mathbf{x} : j \mapsto x_j$ . Esta perspectiva resulta particularmente útil en el estudio de los metodos de kernel.

Vea tambien: espacio euclidiano, espacio vectorial, aplicación lineal, función, método de kernels

**vector de atributos** El vector de atributos se refiere a un vector  $\mathbf{x} =$



2,-1,3,0,-2,1

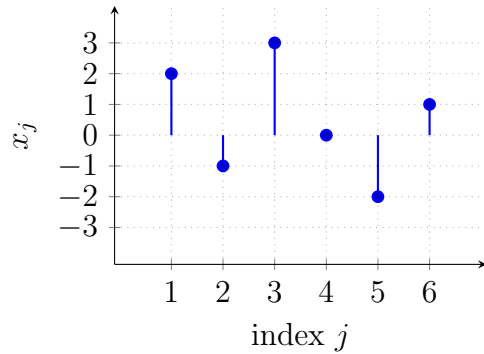


Figure 42: Dos representaciones equivalentes de un vector  $\mathbf{x} = (2, -1, 3, 0, -2, 1)^T \in \mathbb{R}^6$  : como un arreglo numérico (izquierda) y como una función  $j \mapsto x_j$  (derecha).

$(x_1, \dots, x_d)^T$  cuyos elementos son atributos individuales  $x_1, \dots, x_d$ . Muchos métodos de aprendizaje automático utilizan vectores de atributos que pertenecen a algún espacio euclidiano de dimensión finita  $\mathbb{R}^d$ . Sin embargo, para algunos métodos de aprendizaje automático, puede ser más conveniente trabajar con vectores de atributos que pertenezcan a un espacio vectorial de dimensión infinita (por ejemplo, ver método de kernel).

Vea también: ML, espacio euclidiano, ML, método de kernel

**vector normal estándar** Un vector normal estándar es un vector aleatorio  $\mathbf{x} = (x_1, \dots, x_d)^T$  cuyas componentes son variables aleatorias gaussianas iid  $x_j \sim \mathcal{N}(0, 1)$ . Es un caso particular de una distribución normal multivariante, dado que  $\mathbf{x} \sim (\mathbf{0}, \mathbf{I})$ .

Véa también: i.i.d., VA gaussiana, distribución normal multivariante, RV.

**vector propio** Un vector propio de una matriz  $\mathbf{A} \in \mathbb{R}^{d \times d}$  es un vector no nulo  $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$  tal que  $\mathbf{Ax} = \lambda \mathbf{x}$  para algún valor propio  $\lambda$ .  
 Vea también: valor propio.

**0/1 loss** La pérdida 0/1  $L^{(0/1)}((\mathbf{x}, y), h)$  mide la calidad de un clasificador  $h(\mathbf{x})$  que genera una predicción  $\hat{y}$  (por ejemplo, mediante un umbral como en (1)) para la etiqueta  $y$  de un punto de datos con atributos  $\mathbf{x}$ . Es igual a 0 si la predicción es correcta, es decir,  $L^{(0/1)}((\mathbf{x}, y), h) = 0$  cuando  $\hat{y} = y$ . Es igual a 1 si la predicción es incorrecta, es decir,  $L^{(0/1)}((\mathbf{x}, y), h) = 1$  cuando  $\hat{y} \neq y$ .  
 Vea también: pérdida, clasificador, predicción, etiqueta, punto de datos, atributo

**árbol de decisión** Un árbol de decisión es una representación similar a un diagrama de flujo de un mapa de hipótesis  $h$ . Más formalmente, un árbol de decisión es un grafo dirigido que contiene un nodo raíz que lee el vector de atributos  $\mathbf{x}$  de un punto de datos. El nodo raíz luego transfiere el punto de datos a uno de sus nodos hijos basado en alguna prueba elemental sobre los atributos  $\mathbf{x}$ . Si el nodo hijo receptor no es un nodo hoja, es decir, tiene sus propios nodos hijos, representa otra prueba. Según el resultado de la prueba, el punto de datos se transfiere a uno de sus descendientes. Esta prueba y transferencia del punto de datos continúa hasta que el punto de datos termina en un nodo hoja (que no tiene nodos hijos).

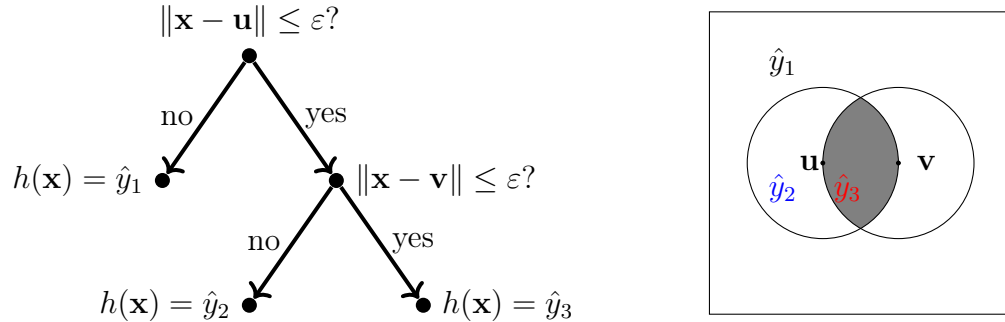


Figure 43: Izquierda: Un árbol de decisión es una representación similar a un diagrama de flujo de una hipótesis  $h : \mathcal{X} \rightarrow \mathbb{R}$  constante por partes. Cada parte es una región de decisión  $\mathcal{R}_{\hat{y}} := \{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) = \hat{y}\}$ . El árbol de decisión mostrado puede aplicarse a vector de atributos numéricos, es decir,  $\mathcal{X} \subseteq \mathbb{R}^d$ . Está parametrizado por el umbral  $\varepsilon > 0$  y los vectores  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ . Derecha: Un árbol de decisión particiona el espacio de atributos  $\mathcal{X}$  en region de decisiones. Cada region de decision  $\mathcal{R}_{\hat{y}} \subseteq \mathcal{X}$  corresponde a un nodo hoja específico en el árbol de decisión.

Vea también: hipótesis, grafo, vector de atributos, punto de datos, atributo, región de decisión, espacio de atributos.

# Index

- 0/1 loss, 162
- $k$ -means, 14
- agrupamiento, 14
- agrupamiento basado en densidad
  - para aplicaciones
  - espaciales con ruido
  - (DBSCAN), 15
- agrupamiento basado en flujo, 15
- agrupamiento convexo, 15
- agrupamiento en grafos, 16
- agrupamiento espectral, 16
- agrupamiento rígido, 18
- agrupamiento suave, 18
- algoritmo, 18
- algoritmo distribuido, 19
- algoritmo en línea, 20
- algoritmo estocástico, 21
- análisis de componentes
  - principales (PCA), 21
- análisis de componentes principales
  - probabilístico (PPCA), 21
- aplicación, 22
- aplicación lineal, 22
- aprendizaje automático (ML), 22
- aprendizaje automático explicable
  - (explainable ML), 23
- aprendizaje de atributos, 23
- aprendizaje en línea, 24
- aprendizaje federado (FL), 27
- aprendizaje federado agrupado
  - (CFL), 24
- aprendizaje federado en red (NFL),  
25
- aprendizaje federado horizontal
  - (horizontal FL), 25
- aprendizaje federado vertical
  - (VFL), 26
- aprendizaje multitarea, 28
- aprendizaje semi-supervisado
  - (SSL), 28
- arrepentimiento (regret), 28
- aspectos computacionales, 29
- aspectos estadísticos, 29
- ataque, 29
- ataque a la privacidad, 30
- ataque de denegación de servicio
  - (DoS), 30

- atributo, 30
- atributo sensible, 31
- aumentación de datos, 31
- autoencoder, 32
- bagging, 32
- bandido multi-brazo (MAB), 33
- boosting, 34
- bootstrap, 35
- bosque aleatorio, 36
- brecha de generalización, 36
- caracterización min-max de
  - Courant–Fischer–Weyl, 36
- clasificador, 37
- clasificador lineal, 38
- classification, 37
- clúster, 38
- condición de gradiente cero, 39
- conectividad algebraica, 40
- conjunto de datos, 40
- conjunto de datos local, 42
- conjunto de entrenamiento, 43
- conjunto de prueba, 43
- conjunto de validación, 43
- contraction operator, 44
- convexo, 44
- cota superior de confianza (UCB),
  - 44
- covarianza, 45
- criterio de parada, 46
- datos, 46
- datos en red, 47
- datos faltantes, 47
- deep net, 131
- descenso de gradiente estocástico
  - (SGD), 47
- descenso por gradiente (GD), 49
- descenso por gradiente en línea
  - (online GD), 50
- descenso por gradiente proyectado
  - (GD proyectado), 52
- descenso por subgradiente, 53
- descomposición en valores propios
  - (EVD), 54
- descomposición en valores
  - singulares (SVD), 54
- desigualdad de concentración, 54
- determinante, 54
- diagrama de dispersión, 56
- diferenciable, 57
- differential privacy (DP), 120

- dimensión de
  - Vapnik–Chervonenkis
  - (dimensión VC), 57
- dimensión efectiva, 57
- discrepancia, 57
- dispositivo, 58
- distribución de probabilidad, 58
- distribución normal multivariante, 58
- divergencia de Kullback-Leibler
  - (divergencia KL), 60
- divergencia de Rényi, 142
- entorno, 60
- entropía diferencial, 60
- envenenamiento de datos, 61
- epígrafo, 61
- error cuadrático medio de
  - estimación (MSEE), 62
- error de entrenamiento, 62
- error de estimación, 63
- error de validación, 63
- espacio de atributos, 64
- espacio de etiquetas, 64
- espacio de Hilbert, 64
- espacio de hipótesis, 65
- espacio de parámetros, 65
- espacio de probabilidad, 66
- espacio euclidiano, 67
- espacio vectorial, 67
- espectrograma, 68
- esperanza-maximización (EM), 69
- estimador de Bayes, 70
- estocástico, 71
- etiqueta, 71
- expectation, 69
- experto, 71
- explicabilidad, 72
- Explicaciones Locales
  - Interpretables e
  - Independientes del Modelo
  - (LIME), 73
- explicación, 73
- familias exponenciales en red
  - (nExpFam), 74
- FedGD, 74
- FedProx, 75
- FedRelax, 75
- FedSGD, 75
- frontera de decisión, 76
- función, 76

función característica, 76  
función cuadrática, 77  
función de activación, 77  
función de densidad de  
    probabilidad (pdf), 77  
función de pérdida, 78  
función objetivo, 79  
  
generalización, 79  
gradient step, 117  
gradiente, 81  
grado de nodo, 81  
grado de pertenencia, 81  
grafo conexo, 81  
grafo de similitud, 82  
graph, 81  
  
hipótesis, 82  
histograma, 83  
  
incertidumbre, 83  
independientes e idénticamente  
    distribuidos (i.i.d.), 83  
información mutua (MI), 84  
Instituto Meteorológico de  
    Finlandia (FMI), 84  
inteligencia artificial (IA), 84  
    inteligencia artificial confiable (IA  
        confiable), 85  
interfaz de programación de  
    aplicaciones (API), 86  
interpretabilidad, 87  
inversión de modelo, 103  
  
kernel, 90  
  
ley de los grandes números, 91  
loperador de reducción y selección  
    absoluta mínima (Lasso),  
        113  
lote, 91  
  
mapa de atributos, 92  
matrix, 92  
matriz de atributos, 94  
matriz de confusión, 94  
matriz de covarianza, 95  
matriz de covarianza muestral, 95  
matriz inversa, 95  
matriz laplaciana, 96  
media, 97  
media muestral, 97  
mediana, 97  
mediana geométrica (GM), 98

minimización de la pérdida	multi-label classification, 37
regularizada (RLM), 100	máquina de vectores de soporte
minimización de variación total	(SVM), 107
generalizada (GTVMin),	máxima verosimilitud, 109
100	máximo, 109
minimización del riesgo empírico	método de Jacobi, 88
explicable (EERM), 100	método de kernel, 109
minimización del riesgo empírico	método de optimización, 110
regularizado (RERM), 101	métodos de gradiente, 111
minimización del riesgo estructural	métrica, 111
(SRM), 102	mínimo, 112
minimización empírica del riesgo	no suave, 112
(ERM), 102	norma, 112
modelo, 104	normalización de datos, 112
modelo de lenguaje de gran escala	número de condición, 113
(LLM), 104	operador proximal, 113
modelo de mezcla gaussiana	optimismo ante la incertidumbre,
(GMM), 105	114
modelo en red, 105	overfitting, 145
modelo estocástico de bloques	parámetro, 116
(SBM), 106	parámetros, 117
modelo lineal, 106	parámetros del modelo, 117
modelo local, 107	peso de arista, 119
modelo probabilístico, 107	precisión (accuracy), 119
muestra, 107	



prediction, 120

predictor, 120

principio de minimización de datos, 120

privacy funnel, 60

privacy leakage, 75

probabilidad, 121

problema de optimización, 121

proceso gaussiano (GP), 122

producto de Kronecker, 124

promedio federado (FedAvg), 124

protección de la privacidad, 124

proximable, 125

proyección, 125

pseudoinversa, 125

puerta trasera (backdoor), 126

punto de dato etiquetado, 127

punto de datos, 127

pérdida, 128

pérdida de error cuadrático, 128

pérdida de hinge, 128

pérdida de Huber, 129

pérdida logística, 129

pérdida por error absoluto, 130

realización, 130

recompensa, 130

red de aprendizaje federado (red FL), 131

red neuronal artificial (RNA), 131

reducción de dimensionalidad, 132

referencia (baseline), 133

región de decisión, 135

reglamento general de protección de datos (RGPD), 136

regresión, 136

regresión de Huber, 136

regresión lineal, 137

regresión logística, 137

regresión polinómica, 138

regresión por desviación absoluta mínima, 138

regresión ridge, 138

regularizador, 141

riesgo, 142

riesgo de Bayes, 142

riesgo empírico, 143

robustez, 143

régimen de alta dimensión, 143

selección de modelo, 104

semi-definida positiva (psd), 144

sesgo, 145

stability, 145

suave, 147

subajuste, 148

subgradiente, 149

suposición de agrupamiento, 149

suposición de independencia e  
    idéntica distribución  
    (suposición i.i.d.), 149

supremo (o mínimo de las cotas  
    superiores), 149

tamaño de la muestra, 149

tamaño de paso, 150

tarea de aprendizaje, 150

tasa de aprendizaje, 152

teorema central del límite (CLT),  
    152

transparency, 154

unidad lineal rectificada (ReLU),  
    155

validación, 155

validación cruzada de  $k$  particiones  
    ( $k$ -fold CV), 156

valor atípico (outlier), 156

valor propio (eigenvalue), 157

variable aleatoria (RV), 157

variable aleatoria gaussiana (VA  
    gaussiana), 158

variación total, 159

variación total generalizada  
    (GTV), 159

varianza, 159

vecino más cercano (NN), 159

vecinos, 160

vector, 160

vector de atributos, 160

vector normal estándar, 161

vector propio, 162

weights, 119

árbol de decisión, 162

## References

- [1] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York, NY, USA: McGraw-Hill, 1987.
- [2] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York, NY, USA: McGraw-Hill, 1976.
- [3] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 2013.
- [4] G. H. Golub and C. F. Van Loan, “An analysis of the total least squares problem,” *SIAM J. Numer. Anal.*, vol. 17, no. 6, pp. 883–893, Dec. 1980, doi: 10.1137/0717073.
- [5] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, 2nd ed. Belmont, MA, USA: Athena Scientific, 2008.
- [6] A. Jung, *Machine Learning: The Basics*. Singapore, Singapore: Springer Nature, 2022.
- [7] Y. SarcheshmehPour, Y. Tian, L. Zhang, and A. Jung, “Flow-based clustering and spectral clustering: A comparison,” in *2021 55th Asilomar Conf. Signals, Syst., Comput.*, M. B. Matthews, Ed. pp. 1292–1296, doi: 10.1109/IEEECONF53345.2021.9723162.
- [8] D. Sun, K.-C. Toh, and Y. Yuan, “Convex clustering: Model, theoretical guarantee and efficient algorithm,” *J. Mach. Learn. Res.*, vol. 22, no. 9, pp. 1–32, Jan. 2021. [Online]. Available: <http://jmlr.org/papers/v22/18-694.html>

- [9] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor, “Convex clustering shrinkage,” presented at the PASCAL Workshop Statist. Optim. Clustering Workshop, 2005.
- [10] U. von Luxburg, “A tutorial on spectral clustering,” *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007, doi: 10.1007/s11222-007-9033-z.
- [11] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press, 2022. [Online]. Available: <http://ebookcentral.proquest.com/lib/aalto-ebooks/detail.action?docID=6925615>
- [12] M. Sipser, *Introduction to the Theory of Computation*, 3rd ed. Andover, U.K.: Cengage Learning, 2013.
- [13] R. Motwani and P. Raghavan, *Randomized Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 1995.
- [14] R. G. Gallager, *Stochastic Processes: Theory for Applications*. New York, NY, USA: Cambridge Univ. Press, 2013.
- [15] G. Tel, *Introduction to Distributed Algorithms*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2000.
- [16] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA, USA: Athena Scientific, 2015.
- [17] E. Hazan, “Introduction to online convex optimization,” *Found. Trends Optim.*, vol. 2, no. 3–4, pp. 157–325, Aug. 2016, doi: 10.1561/24000000013.

- [18] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge Univ. Press, 2006.
- [19] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL, USA: CRC Press, 2015.
- [20] C. Rudin, “Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead,” *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.
- [21] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, “Horizontal federated learning,” in *Federated Learning*. Cham, Switzerland: Springer Nature, 2020, ch. 4, pp. 49–67.
- [22] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, “Vertical federated learning,” in *Federated Learning*. Cham, Switzerland: Springer Nature, 2020, ch. 5, pp. 69–81.
- [23] O. Chapelle, B. Schölkopf, and A. Zien, Eds. *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006.
- [24] D. N. Gujarati and D. C. Porter, *Basic Econometrics*, 5th ed. New York, NY, USA: McGraw-Hill/Irwin, 2009.
- [25] Y. Dodge, Ed. *The Oxford Dictionary of Statistical Terms*. New York, NY, USA: Oxford Univ. Press, 2003.

- [26] B. S. Everitt, *The Cambridge Dictionary of Statistics*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.
- [27] S. Bubeck and N. Cesa-Bianchi, “Regret analysis of stochastic and non-stochastic multi-armed bandit problems,” *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, Dec. 2012, doi: 10.1561/22000000024.
- [28] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer Science+Business Media, 2009.
- [29] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.
- [30] A. Silberschatz, H. F. Korth, and S. Sudarshan, *Database System Concepts*, 7th ed. New York, NY, USA: McGraw-Hill Education, 2019. [Online]. Available: <https://db-book.com/>
- [31] S. Abiteboul, R. Hull, and V. Vianu, *Foundations of Databases*. Reading, MA, USA: Addison-Wesley, 1995.
- [32] S. Hoberman, *Data Modeling Made Simple: A Practical Guide for Business and IT Professionals*, 2nd ed. Basking Ridge, NJ, USA: Technics Publications, 2009.
- [33] R. Ramakrishnan and J. Gehrke, *Database Management Systems*, 3rd ed. New York, NY, USA: McGraw-Hill, 2002.

- [34] E. F. Codd, “A relational model of data for large shared data banks,” *Commun. ACM*, vol. 13, no. 6, pp. 377–387, Jun. 1970, doi: 10.1145/362384.362685.
- [35] T. Gebru et al., “Datasheets for datasets,” *Commun. ACM*, vol. 64, no. 12, pp. 86–92, Nov. 2021, doi: 10.1145/3458723.
- [36] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [37] K. Abayomi, A. Gelman, and M. Levy, “Diagnostics for multivariate imputations,” *J. Roy. Statist. Soc.: Ser. C (Appl. Statist.)*, vol. 57, no. 3, pp. 273–291, Jun. 2008, doi: 10.1111/j.1467-9876.2007.00613.x.
- [38] L. Bottou, “On-line learning and stochastic approximations,” in *On-Line Learning in Neural Networks*, D. Saad, Ed. New York, NY, USA: Cambridge Univ. Press, 1999, ch. 2, pp. 9–42.
- [39] A. Rakhlin, O. Shamir, and K. Sridharan, “Making gradient descent optimal for strongly convex stochastic optimization,” in *Proc. 29th Int. Conf. Mach. Learn.*, J. Langford and J. Pineau, Eds. 2012, pp. 449–456. [Online]. Available: <https://icml.cc/Conferences/2012/papers/261.pdf>
- [40] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
- [41] H. J. Dirschmid, *Tensors and Fields*, (in German). Vienna, Austria: Springer-Verlag, 1996.

- [42] G. Strang, *Computational Science and Engineering*. Wellesley, MA, USA: Wellesley-Cambridge Press, 2007.
- [43] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge Univ. Press, 2013.
- [44] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge Univ. Press, 2014.
- [45] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*, 2nd ed. New York, NY, USA: Springer Science+Business Media, 2009.
- [46] P. Billingsley, *Probability and Measure*, 3rd ed. New York, NY, USA: Wiley, 1995.
- [47] A. Lapidoth, *A Foundation in Digital Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
- [48] A. Lapidoth, *A Foundation in Digital Communication*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2017.
- [49] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
- [50] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, “From the information bottleneck to the privacy funnel,” in *2014 IEEE Inf. Theory Workshop*, pp. 501–505, doi: 10.1109/ITW.2014.6970882.
- [51] X. Liu, H. Li, G. Xu, Z. Chen, X. Huang, and R. Lu, “Privacy-enhanced federated learning against poisoning adversaries,” *IEEE*



- Trans. Inf. Forensics Security*, vol. 16, pp. 4574–4588, 2021, doi: 10.1109/TIFS.2021.3108434.
- [52] J. Zhang, B. Chen, X. Cheng, H. T. T. Binh, and S. Yu, “PoisonGAN: Generative poisoning attacks against federated learning in edge computing systems,” *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3310–3322, Mar. 2021, doi: 10.1109/JIOT.2020.3023126.
  - [53] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA, USA: Athena Scientific, 2003.
  - [54] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York, NY, USA: Springer-Verlag, 1998.
  - [55] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.
  - [56] S. Ross, *A First Course in Probability*, 9th ed. Boston, MA, USA: Pearson Education, 2014.
  - [57] L. Cohen, *Time-Frequency Analysis*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1995.
  - [58] J. Li, L. Han, X. Li, J. Zhu, B. Yuan, and Z. Gou, “An evaluation of deep neural network models for music classification using spectrograms,” *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 4621–4647, Feb. 2022, doi: 10.1007/s11042-020-10465-9.
  - [59] B. Boashash, Ed. *Time Frequency Signal Analysis and Processing: A Comprehensive Reference*. Oxford, U.K.: Elsevier, 2003.

- [60] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Burlington, MA, USA: Academic, 2009.
- [61] P. R. Halmos, *Measure Theory*. New York, NY, USA: Springer-Verlag, 1974.
- [62] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer Science+Business Media, 2006.
- [63] M. J. Wainwright and M. I. Jordan, “Graphical models, exponential families, and variational inference,” *Found. Trends Mach. Learn.*, vol. 1, no. 1–2, pp. 1–305, Nov. 2008, doi: 10.1561/22000000001.
- [64] J. Colin, T. Fel, R. Cadène, and T. Serre, “What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods,” in *Adv. Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. vol. 35, 2022, pp. 2832–2845. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2022/hash/13113e938f2957891c0c5e8df811dd01-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2022/hash/13113e938f2957891c0c5e8df811dd01-Abstract-Conference.html)
- [65] L. Zhang, G. Karakasidis, A. Odnoblyudova, L. Dogruel, Y. Tian, and A. Jung, “Explainable empirical risk minimization,” *Neural Comput. Appl.*, vol. 36, no. 8, pp. 3983–3996, Mar. 2024, doi: 10.1007/s00521-023-09269-3.
- [66] A. Jung and P. H. J. Nardelli, “An information-theoretic approach to personalized explainable machine learning,” *IEEE Signal Process. Lett.*, vol. 27, pp. 825–829, 2020, doi: 10.1109/LSP.2020.2993176.

- [67] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, “Learning to explain: An information-theoretic perspective on model interpretation,” in *Proc. 35th Int. Conf. Mach. Learn.*, J. Dy and A. Krause, Eds. vol. 80, 2018, pp. 883–892. [Online]. Available: <https://proceedings.mlr.press/v80/chen18j.html>
- [68] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you?: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [69] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 3rd ed., 2025. [Online]. Available: <https://christophm.github.io/interpretable-ml-book/>
- [70] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *2017 IEEE Int. Conf. Comput. Vis.*, pp. 618–626, doi: 10.1109/ICCV.2017.74.
- [71] A. Jung, “Networked exponential families for big data over networks,” *IEEE Access*, vol. 8, pp. 202 897–202 909, Nov. 2020, doi: 10.1109/ACCESS.2020.3033817.
- [72] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” in *Proc. Mach. Learn. Syst.*, I. Dhillon, D. Papailiopoulos, and V. Sze, Eds. vol. 2, 2020.

- [Online]. Available: [https://proceedings.mlsys.org/paper\\_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html](https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html)
- [73] A. Ünsal and M. Önen, “Information-theoretic approaches to differential privacy,” *ACM Comput. Surv.*, vol. 56, no. 3, Oct. 2023, Art. no. 76, doi: 10.1145/3604904.
  - [74] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Boston, MA, USA: Kluwer Academic, 2004.
  - [75] D. P. Bertsekas, *Convex Optimization Algorithms*. Belmont, MA, USA: Athena Scientific, 2015.
  - [76] J. Su, D. V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019, doi: 10.1109/TEVC.2019.2890858.
  - [77] R. T. Rockafellar, *Network Flows and Monotropic Optimization*. Belmont, MA, USA: Athena Scientific, 1998.
  - [78] D. Pfau and A. Jung, “Engineering trustworthy AI: A developer guide for empirical risk minimization,” Nov. 2024. [Online]. Available: <https://arxiv.org/abs/2410.19361>
  - [79] High-Level Expert Group on Artificial Intelligence, “The assessment list for trustworthy artificial intelligence (ALTAI): For self assessment,” European Commission, Jul. 17, 2020. [Online]. Available: <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>

- [80] L. Richardson and M. Amundsen, *RESTful Web APIs*. Sebastopol, CA, USA: O'Reilly Media, 2013.
- [81] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," Mar. 2017. [Online]. Available: <https://arxiv.org/abs/1702.08608>
- [82] P. Hase and M. Bansal, "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Jul. 2020, pp. 5540–5552. [Online]. Available: <https://aclanthology.org/2020.acl-main.491>
- [83] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018, doi: 10.1145/3236386.3241340.
- [84] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, UK: Cambridge Univ. Press, 1991.
- [85] G. Strang, *Introduction to Linear Algebra*, 5th ed. Wellesley-Cambridge Press, MA, 2016.
- [86] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.
- [87] C. H. Lampert, "Kernel methods in computer vision," *Found. Trends Comput. Graph. Vis.*, vol. 4, no. 3, pp. 193–285, Sep. 2009, doi: 10.1561/06000000027.

- [88] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. New York, NY, USA: McGraw-Hill Higher Education, 2002.
- [89] S. Mallat, “Understanding deep convolutional networks,” *Philos. Trans. Roy. Soc. A*, vol. 374, no. 2065, Apr. 2016, Art. no. 20150203, doi: 10.1098/rsta.2015.0203.
- [90] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in *Adv. Neural Inf. Process. Syst.*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. vol. 14, 2001, pp. 849–856. [Online]. Available: [https://papers.nips.cc/paper\\_files/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html](https://papers.nips.cc/paper_files/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html)
- [91] H. P. Lopuhaä and P. J. Rousseeuw, “Breakdown points of affine equivariant estimators of multivariate location and covariance matrices,” *Ann. Statist.*, vol. 19, no. 1, pp. 229–248, Mar. 1991, doi: 10.1214/aos/1176347978.
- [92] Y. SarcheshmehPour, Y. Tian, L. Zhang, and A. Jung, “Clustered federated learning via generalized total variation minimization,” *IEEE Trans. Signal Process.*, vol. 71, pp. 4240–4256, 2023, doi: 10.1109/TSP.2023.3322848.
- [93] S. Shalev-Shwartz and A. Tewari, “Stochastic methods for  $\ell_1$  regularized loss minimization,” in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, L. Bottou and M. Littman, Eds. Jun. 2009, pp. 929–936.

- [94] M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333, doi: 10.1145/2810103.2813677.
- [95] A. Vaswani et al., “Attention is all you need,” in *Adv. Neural Inf. Process. Syst.*, I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. vol. 30, 2017, pp. 5998–6008. [Online]. Available: [https://papers.nips.cc/paper\\_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html](https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html)
- [96] E. Abbe, “Community detection and stochastic block models: Recent developments,” *J. Mach. Learn. Res.*, vol. 18, no. 177, pp. 1–86, Apr. 2018. [Online]. Available: <http://jmlr.org/papers/v18/16-480.html>
- [97] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. New York, NY, USA: Cambridge Univ. Press, 2000.
- [98] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.
- [99] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [100] N. Parikh and S. Boyd, “Proximal algorithms,” *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, Jan. 2014, doi: 10.1561/24000000003.
- [101] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone*

*Operator Theory in Hilbert Spaces*, 2nd ed. New York, NY, USA: Springer Science+Business Media, 2017.

- [102] European Union, “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance),” L 119/1, May 4, 2016. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- [103] European Union, “Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC (Text with EEA relevance),” L 295/39, Nov. 21, 2018. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2018/1725/oj>
- [104] O. Kallenberg, *Foundations of Modern Probability*. New York, NY, USA: Springer-Verlag, 1997.
- [105] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, A. Singh and J. Zhu, Eds. vol. 54, 2017, pp. 1273–1282. [Online]. Available: <https://proceedings.mlr.press/v54/mcmahan17a.html>
- [106] L. Condat, “A primal–dual splitting method for convex optimization



- involving lipschitzian, proximable and linear composite terms,” *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 460–479, Aug. 2013, doi: 10.1007/s10957-012-0245-9.
- [107] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, 2nd ed. New York, NY, USA: Springer-Verlag, 2003.
  - [108] M. P. Salinas et al., “A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis,” *npj Digit. Med.*, vol. 7, no. 1, May 2024, Art. no. 125, doi: 10.1038/s41746-024-01103-x.
  - [109] G. F. Cooper, “The computational complexity of probabilistic inference using bayesian belief networks,” *Artif. Intell.*, vol. 42, no. 2–3, pp. 393–405, Mar. 1990, doi: 10.1016/0004-3702(90)90060-D.
  - [110] I. Csiszar, “Generalized cutoff rates and Renyi’s information measures,” *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 26–34, Jan. 1995, doi: 10.1109/18.370121.
  - [111] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin, Germany: Springer-Verlag, 2011.
  - [112] S. Bubeck, “Convex optimization: Algorithms and complexity,” *Found. Trends Mach. Learn.*, vol. 8, no. 3–4, pp. 231–357, Nov. 2015, 10.1561/22000000050.
  - [113] R. Caruana, “Multitask learning,” *Mach. Learn.*, vol. 28, pp. 41–75, Jul. 1997, doi: 10.1023/A:1007379606734.

- [114] A. Jung, G. Hannak, and N. Goertz, “Graphical lasso based model selection for time series,” *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1781–1785, Oct. 2015, doi: 10.1109/LSP.2015.2425434.
- [115] A. Jung, “Learning the conditional independence structure of stationary time series: A multitask learning approach,” *IEEE Trans. Signal Process.*, vol. 63, no. 21, Nov. 2015, doi: 10.1109/TSP.2015.2460219.
- [116] A. van der Vaart, *Asymptotic Statistics*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
- [117] High-Level Expert Group on Artificial Intelligence, “Ethics guidelines for trustworthy AI,” European Commission, Apr. 8, 2019. [Online]. Available: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- [118] C. Gallese, “The AI act proposal: A new right to technical interpretability?,” *SSRN Electron. J.*, Feb. 2023. [Online]. Available: <https://ssrn.com/abstract=4398206>
- [119] M. Mitchell et al., “Model cards for model reporting,” in *Proc. Conf. Fairness, Accountability, Transparency*, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.
- [120] K. Shahriari and M. Shahriari, “IEEE standard review — Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems,” in *2017 IEEE Canada Int. Humanitarian Technol. Conf.*, pp. 197–201, doi: 10.1109/IHTC.2017.8058187.

- [121] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed. New York, NY, USA: Springer-Verlag, 1991.
- [122] M. Kearns and M. Li, “Learning in the presence of malicious errors,” *SIAM J. Comput.*, vol. 22, no. 4, pp. 807–837, Aug. 1993, doi: 10.1137/0222052.
- [123] G. Lugosi and S. Mendelson, “Robust multivariate mean estimation: The optimality of trimmed mean,” *Ann. Statist.*, vol. 49, no. 1, pp. 393–410, Feb. 2021, doi: 10.1214/20-AOS1961.
- [124] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.