

Aprendizaje Supervisado y Modelos Clásicos

ML-6561

October 20, 2023

Contenido

1 Introducción a Aprendizaje Supervisado

- Definición
- Objetivo
- Conceptos básicos

2 Ejemplo sobre el dominio de un problema

- Clasificación de Correo Electrónico

3 Ejemplos Clásicos

- Predicción de Precios de Bienes Raíces
- Reconocimiento de Dígitos Escritos a Mano
- Diagnóstico Médico

4 Definición del problema a resolver

5 Conclusión

Índice

1 Introducción a Aprendizaje Supervisado

- Definición
- Objetivo
- Conceptos básicos

•

2 Ejemplo sobre el dominio de un problema

- Clasificación de Correo Electrónico

3 Ejemplos Clásicos

- Predicción de Precios de Bienes Raíces
- Reconocimiento de Dígitos Escritos a Mano
- Diagnóstico Médico

4 Definición del problema a resolver

5 Conclusión

¿Qué es el Aprendizaje Supervisado?

- El aprendizaje supervisado es un tipo de aprendizaje automático donde el algoritmo aprende a partir de **datos etiquetados** un modelo matemático que llamaremos f .
- **Cada instancia** de datos tiene una etiqueta, valor o clase conocida.

Objetivo del Aprendizaje Supervisado

Encontrar una función, denotada como f , que mapee los datos de entrada X a las etiquetas de salida Y .

$$Y = f(x) \text{ para } x \in X \quad (1)$$

o de igual forma

$$f : X \rightarrow Y \quad \begin{matrix} \nearrow \text{Reales} \\ . \end{matrix} \quad (2)$$

Donde:

- X representa los datos de entrada.
- Y son las etiquetas de salida o las respuestas deseadas.
- $f(\cdot)$ es la función que el modelo de aprendizaje supervisado busca aprender.

Definiciones sobre X y f

- Denominaremos a X como conjunto de datos de entrenamiento o datos de entrada

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

- $y^{(i)}$ son las etiquetas de salida para cada $x^{(i)}$.
- Dada f , la **predicción** de $x^{(i)}$ la representaremos como $\hat{y}^{(i)}$.

Definiciones sobre X y f

- Denominaremos a X como conjunto de datos de entrenamiento o datos de entrada

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

- $y^{(i)}$ son las etiquetas de salida para cada $x^{(i)}$.
- Dada f , la **predicción** de $x^{(i)}$, $f(x^{(i)})$ la representaremos como $\hat{y}^{(i)}$.

Discusión

Cómo definirían el **error** de una predicción i ?

$$|y^i - \hat{y}^i| \quad y^i - \hat{y}^i \quad (\hat{y}^i - y^i)^2$$

Definiciones sobre X y f

- Denominaremos a X como conjunto de datos de entrenamiento o datos de entrada

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

- $y^{(i)}$ son las etiquetas de salida para cada $x^{(i)}$.
- Dada f , la **predicción** de $x^{(i)}$, $f(x^{(i)})$ la representaremos como $\hat{y}^{(i)}$.
- $Error_i(x^{(i)}) = y^{(i)} - \hat{y}^{(i)} = y^{(i)} - f(x^{(i)})$



Objetivo principal equivalente

El **objetivo** es encontrar una función f que minimice los errores entre las predicciones y las etiquetas reales.

Índice

1 Introducción a Aprendizaje Supervisado

- Definición
- Objetivo
- Conceptos básicos

2 Ejemplo sobre el dominio de un problema

- Clasificación de Correo Electrónico

3 Ejemplos Clásicos

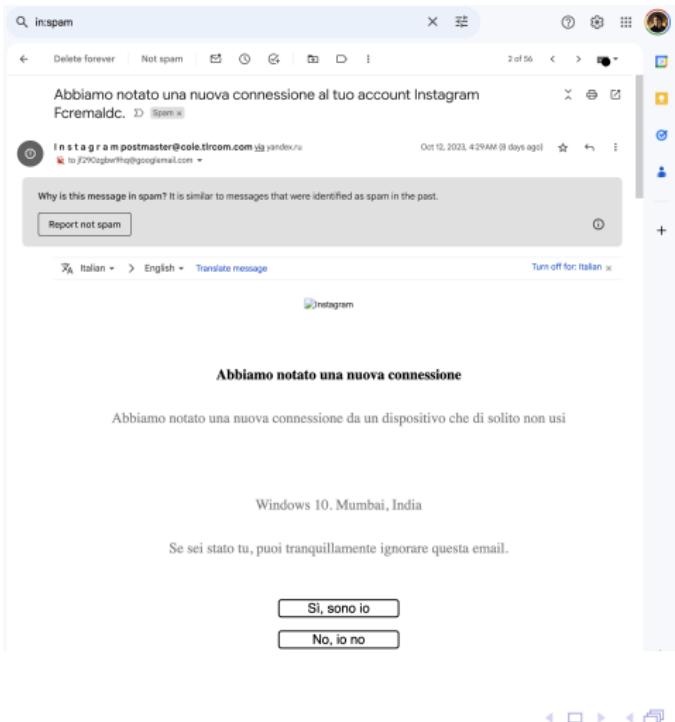
- Predicción de Precios de Bienes Raíces
- Reconocimiento de Dígitos Escritos a Mano
- Diagnóstico Médico

4 Definición del problema a resolver

5 Conclusión

Ejemplo dominio: Clasificador de Spam/No Spam

- **Problema:** Clasificar correos electrónicos como "spam" o "no spam".



in:spam

Delete forever

Not spam



2 of 56



Abbiamo notato una nuova connessione al tuo account Instagram
Fcremaldc. Spam



Instagram postmaster@cole.tlrc.com.com via yandex.ru
to jf290zgbw9hq@googlemail.com

Oct 12, 2023, 4:29AM (8 days ago)



Why is this message in spam? It is similar to messages that were identified as spam in the past.

Report not spam

Italian > English Translate message

Turn off for: Italian



Abbiamo notato una nuova connessione

Abbiamo notato una nuova connessione da un dispositivo che di solito non usi



Windows 10. Mumbai, India

Se sei stato tu, puoi tranquillamente ignorare questa email.

Sì, sono io

No, io no

Ejemplo dominio: Clasificador de Spam/No Spam

Problema: Clasificar correos electrónicos como "spam" o "no spam".

- Recordemos nuestro objetivo:

$$Y = f(X)$$

- Definimos X como?
- Definimos Y como?

Ejemplo dominio: Clasificador de Spam/No Spam

Problema: Clasificar correos electrónicos como "spam" o "no spam".

- Recordemos nuestro objetivo:

$$Y = f \left(\begin{array}{l} \text{spam?} \\ \text{es} \\ \text{email} \end{array} \right)$$

The diagram illustrates a machine learning model. On the left, a variable Y is shown above a red downward arrow pointing to the word "Spam". To the right of an equals sign is a function f . To the right of f is a vertical vector enclosed in parentheses. This vector has three components: a red downward arrow pointing to the question "spam?", another red downward arrow pointing to the word "es", and a third red downward arrow pointing to the word "email".

- Definimos X como?
- Definimos Y como?

Ejemplo dominio: Clasificador de Spam/No Spam

Defino X como el conjunto de secuencias que representan documentos de texto.

$$X = \{x_1, x_2, x_3, \dots, x_N\}$$

Donde:

- x_i representa una secuencia de caracteres arbitraria.
- N es el número de secuencias disponibles.

Discusión

- Cada x_i puede contener palabras, frases o caracteres que conforman el contenido del documento.
- El objetivo del clasificador es etiquetar cada x_i como "spam" o "no spam" según su contenido.

Ejemplo dominio: Clasificador de Spam/No Spam

Discusión

- Cada x_i puede contener palabras, frases o caracteres que conforman el contenido del documento.
- El objetivo del clasificador es etiquetar cada x_i como "spam" o "no spam" según su contenido.
- Qué problema tenemos?

$$f(x_i) \in \{\text{spam}, \text{no spam}\}$$

- ① $f(\textbf{Abbiamo notato una nuova connessione}) = \text{spam}$
- ② $f(\textbf{Abbiamo notato una nuova}) = \text{spam}$
- ③ ...
- ④ $f(\textbf{Abbiamo}) = ?$

Ejemplo dominio: Clasificador de Spam/No Spam

Otro punto de vista

- **Defino un alfabeto:** Sea S un conjunto finito de caracteres individuales utilizados para representar el lenguaje. Por ejemplo, letras mayúsculas y minúsculas, dígitos numéricos, signos de puntuación y espacios en blanco.
- **Defino un documento:** Para crear un documento x_i , tomamos el producto cartesiano de M caracteres dentro del alfabeto S .

$$x : S \times S \times S \times S$$

S

Ejemplo: Supongamos que nuestro alfabeto contiene las letras mayúsculas A, B, C y los números 1, 2, 3. Si $M = 4$, podemos crear una palabra de 4 caracteres:

$$s_i = "AB32"$$

Esta palabra es el resultado de tomar el producto cartesiano de 4 caracteres del alfabeto.

Ejemplo dominio: Clasificador de Spam/No Spam

Discusión

- $X = S \times S \times \cdots \times S$
- Cada $x_i \in X$
- $f(x_i) \in \{\text{spam, no spam}\}$
- Qué problemas tenemos?

- ① $f(\textbf{Abbiamo notato una nuova connessione}) = \text{spam}$
- ② $f(\textbf{Abbiamo notato una nuova}) = \text{spam}$
- ③ ...
- ④ $f(\textbf{Abbiamo}) = ?$

Ejemplo dominio: Clasificador de Spam/No Spam

- **Diccionario de Palabras:** Se tiene un diccionario que asigna cada palabra a un índice único.

Diccionario: $w_i \rightarrow i$

- **Matriz de Embedding:** Se utiliza una matriz de embedding para representar cada palabra como un vector de características.

Embedding de $w_i : \vec{v}_i \in \mathbb{R}^d$

- **Ejemplo:** Si $d = 3$ (vectores tridimensionales), el embedding de la palabra "manzana" puede ser representado como
 $\vec{v}_{\text{manzana}} = [0.2, 0.8, 0.5]$.

Embeddings

El objetivo es que las palabras con significados similares tengan embeddings cercanos en el **espacio vectorial**.

Índice

1 Introducción a Aprendizaje Supervisado

- Definición
- Objetivo
- Conceptos básicos

2 Ejemplo sobre el dominio de un problema

- Clasificación de Correo Electrónico

3 Ejemplos Clásicos

- Predicción de Precios de Bienes Raíces
- Reconocimiento de Dígitos Escritos a Mano
- Diagnóstico Médico

4 Definición del problema a resolver

5 Conclusión

Ejemplo 2: Predicción de Precios de Bienes Raíces

- **Problema:** Predecir el precio de bienes raíces en función de características como área, ubicación, número de habitaciones, etc.
- **Modelos Clásicos:** Regresión Lineal, Árboles de Decisión.

Ejemplo 2: Predicción de Precios de Bienes Raíces

Immobiliare.it Roma + Comune MODIFICA PUBBLICA ANNUNCI Accedi

Tutti i filtri (2) Vendita Case - Appartamenti Prezzo Superficie SALVA RICERCA

Inmobiliare.it > Provincia di Roma > Roma > Scogi la zona

33.530 risultati per: case in vendita Roma

Automatico ESPANDI

€ 699.000
Appartamento via Archimede, Euclide, Roma
NUOVO

Paroli, appartamento in vendita in Via Archimede adiacente Via San Valentino. In palazzo signorile con servizio di portineria, libero tranquillo luminoso ottimo stato, ampio ingresso, salone, sala da pranzo, tre camere, cucina, tripli servizi, due balconi, cantina e posto auto condominiale. Rif. 22999PA Tel: 06.89561424 Paroli, apartment...

4 locali | 150 m² | 3 bagni | 2 piani

€ 379.000
Quadrilocale via Angiolo Cabrini 9, Nuovo Salario, Roma
NUOVO

Nuovo Salario Piazza Pomo Lambertienghi, in zona residenziale immersa nel verde, con servizio di portierato, vicino ai servizi primari (Istituto comprensivo di scuola...

4 locali | 123 m² | 2 bagni | 1 piano

€ 629.000
Quadrilocale viale Glorioso 29, Trastevere, Roma
NUOVO

Nel cuore della Capitale, nel ricerchissimo quartiere di Trastevere, a pochi passi da Via Dandolo e con precisione in Viale Glorioso, HOMELY GROUP cura la vendita di u...

4 locali | 110 m² | 1 bagno | 1 piano

MESSAGGIO VISITA

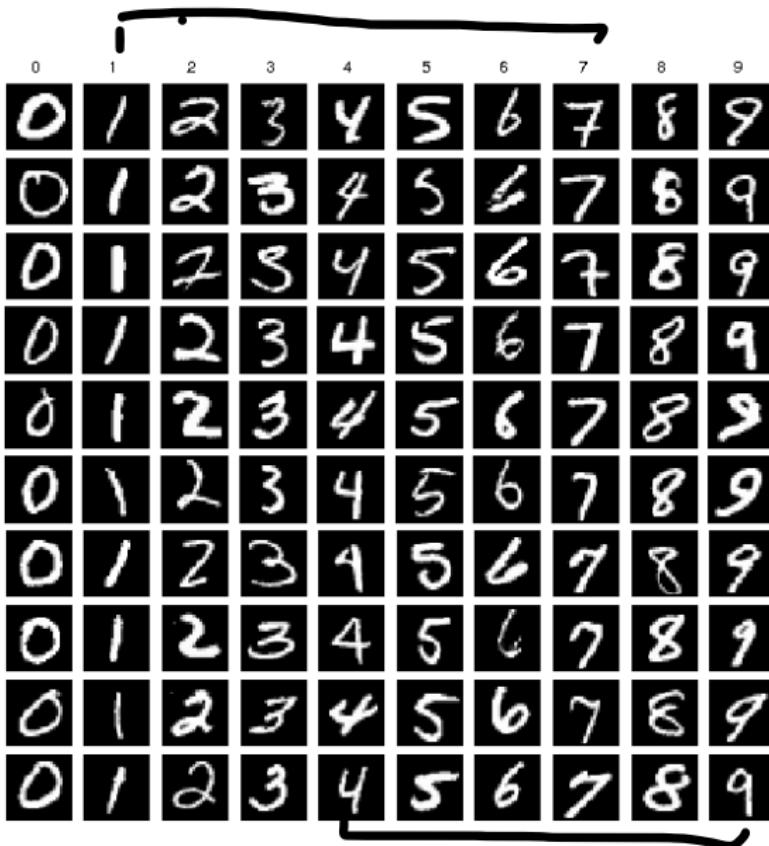
Mappa di Roma con marcatori blu indicanti i risultati della ricerca. Un poligono nero circonda un'area specifica nel centro storico.

Ejemplo 3: Reconocimiento de Dígitos Escritos a Mano

mnis+

- **Problema:** Reconocer dígitos escritos a mano (de 0 a 9).
- **Modelos clásicos:** Redes Neuronales Artificiales (ANN), k-Vectores más Cercanos (k-NN).

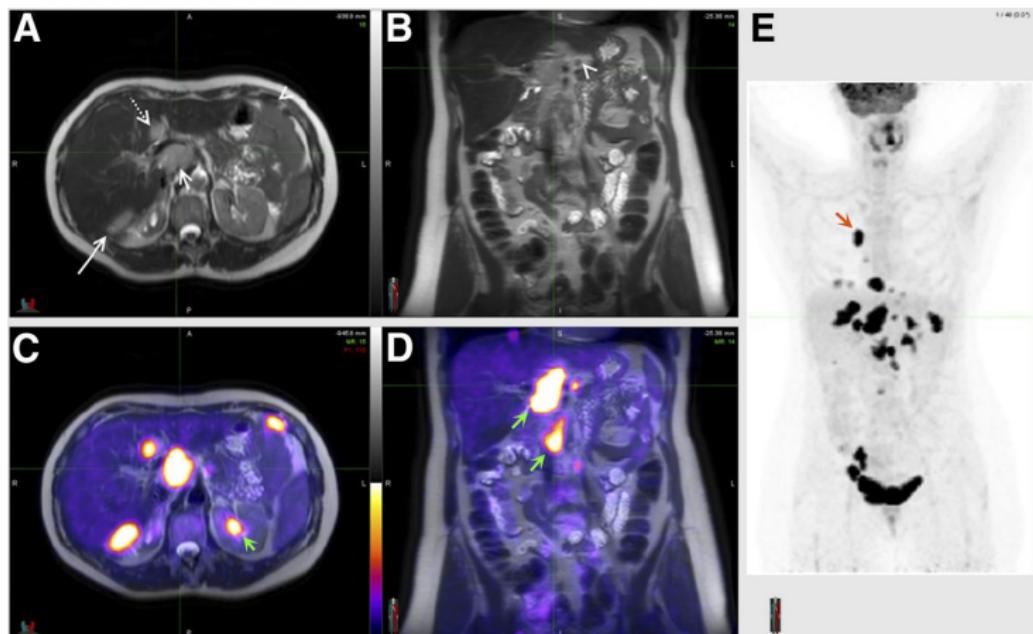
Ejemplo 3: Reconocimiento de Dígitos Escritos a Mano



Ejemplo 4: Diagnóstico Médico

- **Problema:** Diagnosticar enfermedades (por ejemplo, cáncer) a partir de datos médicos.
- **Modelos clásicos:** Redes Neuronales Convolucionales (CNN), Máquinas de Soporte Vectorial (SVM).

Ejemplo 4: Diagnóstico Médico



Índice

1 Introducción a Aprendizaje Supervisado

- Definición
- Objetivo
- Conceptos básicos

2 Ejemplo sobre el dominio de un problema

- Clasificación de Correo Electrónico

3 Ejemplos Clásicos

- Predicción de Precios de Bienes Raíces
- Reconocimiento de Dígitos Escritos a Mano
- Diagnóstico Médico

4 Definición del problema a resolver

5 Conclusión

Función de Coste en Aprendizaje Supervisado

Generalmente definida como $J(\theta)$ o $J_\theta(y^{(i)}, \hat{y}^{(i)})$ evalúa qué tan cerca están las predicciones del modelo $\hat{y}^{(i)}$ de las etiquetas reales $y^{(i)}$ en función de los parámetros del modelo θ .

Algunas funciones de coste comunes incluyen:

- **Clasificación:** Entropía Cruzada, Error Cuadrático Medio.
- **regresión:** Error Cuadrático Medio, Error Absoluto Medio.

Función de Coste, $J_\theta(y^{(i)}, \hat{y}^{(i)})$

El objetivo del Aprendizaje Supervisado es encontrar los parámetros (θ) que minimizan la función de coste.

Problema de Optimización

El objetivo es encontrar los parámetros θ que minimicen una función de costo $J(\theta)$:

$$\theta^* = \arg \min_{\theta} J(\theta) \quad (3)$$

Donde:

- $J(\theta)$ es la función de costo que mide la discrepancia entre las predicciones del modelo y las etiquetas reales.
- θ son los parámetros del modelo a aprender.
- θ^* representa los parámetros óptimos que minimizan $J(\theta)$.

La tarea es encontrar los parámetros que generen un modelo capaz de hacer predicciones precisas.

Error Cuadrático Medio (MSE)

El **Error Cuadrático Medio** (MSE) es una función de coste comúnmente utilizada en problemas de regresión y, en algunos casos, en problemas de clasificación. Se define como:

$$MSE = \frac{1}{m} \sum_{i=1}^m (y^{(i)} - \hat{y}^{(i)})^2 \quad (4)$$

Donde:

- m es el número de ejemplos en el conjunto de entrenamiento.
- $y^{(i)}$ es la etiqueta real del ejemplo i .
- $\hat{y}^{(i)}$ es la predicción del modelo para el ejemplo i .

El objetivo es minimizar el MSE ajustando los parámetros del modelo.

Tarea

Table: Datos de prueba

x	y
1	.7
2	3.1
3	2.6
4	3.1
5	5.1
6	7.3
7	8.3
8	9.2
9	5.4
10	10.9

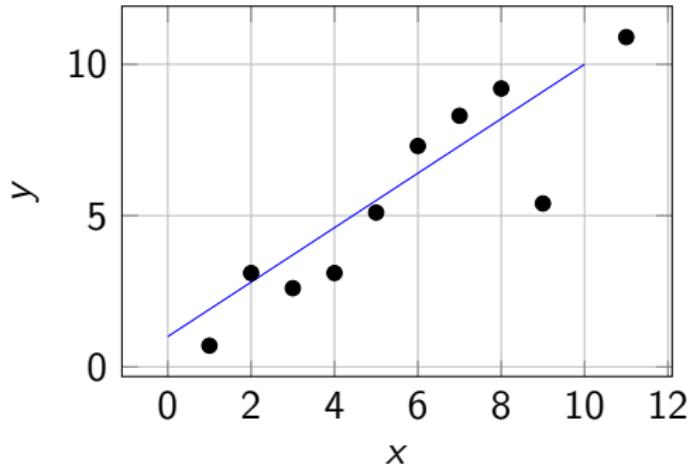


Figure: Calculen el MSE de los siguientes datos asumiendo $f(x) = 0.9 * x + 1$, identifiquen m , cómo saben si el modelo es "bueno"?

Índice

1 Introducción a Aprendizaje Supervisado

- Definición
- Objetivo
- Conceptos básicos

2 Ejemplo sobre el dominio de un problema

- Clasificación de Correo Electrónico

3 Ejemplos Clásicos

- Predicción de Precios de Bienes Raíces
- Reconocimiento de Dígitos Escritos a Mano
- Diagnóstico Médico

4 Definición del problema a resolver

5 Conclusión

Conclusión

- El aprendizaje supervisado es fundamental en numerosos campos.
- Existen diversos modelos para abordar problemas de clasificación y regresión.
- La elección del modelo depende de la naturaleza del problema y los datos disponibles.