Introducción a la teoría de aprendizaje estadístico

Víctor Mijangos

Facultad de Ingeniería

UNAM

Conceptos preliminares

Antecedentes

Algunos puntos históricos relevantes en la historia del aprendizaje son:

- En el Siglo XVII, Gottfried Leibnez propone una marco teórico para llevar a cabo operaciones lógicas (*calculus ratiocinator*).
- En 1887, Charles S. Pierce habla de máquinas lógicas.
- En los 40's, McCulloch y Pitts proponen un modelo lógico para emular la actividad nerviosa [1].
- En los 50's, Rosenblatt plantea los fundamentos del análisis de reconocimiento de patrones [2].
- En los 60's, comienza el desarrollo de modelos dentro de la teoría de aprendizaje.
- En los 80's, se establecen métodos estadísticos para abordar los problemas de este tipo.

Herramientas teóricas del aprendizaje estadístico

El aprendizaje estadístico intersecta con otras áreas del conocimiento, como son:

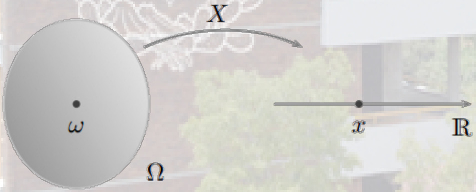
- Ciencias de la computación
- Análisis, cálculo, algebra lineal, topología
- Probabilidad y estadística
- Teoría de la información

La **teoría de la información** aporta herramientas téoricas relevantes para la teoría de aprendizaje estadístico.

Variables aleatorias

Un aspecto importante del aprendizaje son los **datos**.
Los datos se conformarán por varibales que definan una característica de esa instancia.
Cada instancia en los datos será descrita por una serie de variables registradas en un vector:

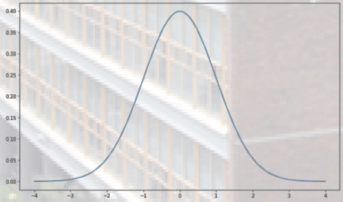
$$x^T = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \end{pmatrix}$$



Funciones de probabilidad y de distribución

La **función de probabilidad** determina la prbabilidad de que una variable aleatoria tome un valor (o conjunto de valores) particular.
Se cumple que:

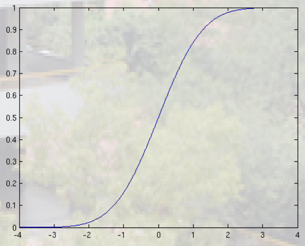
- $p(X = x) \geq 0$
- $\int_{-\infty}^{\infty} p(x)dx = 1$



La función de distribución $F(x)$ puede definirse como:

$$F(x) = \int_{-\infty}^{x} p(X = x)dx$$

Y establece la acumulación de la probabilidadad hasta el valor x.



Valor esperado

Un concepto central en la teoría de probabilidad es la esperanza o valor esperado de una variable aleatoria:

$$\mathbb{E}[X] = \sum_{i=1}^{N} p(X = x_i) \cdot x_i \tag{1}$$

A este valor también se le suele llamar μ (media).

A partir de este valor se puede calcular la varianza:

$$Var(X) = \mathbb{E}[(X - \mu)^2] = \sum_{i=1}^{N} p(X = x_i)(x_i - \mu)^2 \tag{2}$$

También se suele denotar la varianza como σ^2.

Ley de los grandes números

Un resultado importante es la **ley de los grandes números**:

Sea $X_1, X_2, \ldots, X_N$ variables aleatorias convergentes a X y sea $\mathbb{E}[X]$ el valor esperado de X, entonces:

$$\frac{1}{N}\sum_{i=1}^{N} X_i \xrightarrow{P} \mathbb{E}[X]$$

cuando $N \to \infty$

Puede verse que cuando $X \sim Ber(p)$, se justifica la probabilidad frecuentista.

Teorema central del límite

El **teorema central del límite** es otro resultado importante que establece que:

Sea $X_1, X_2, \ldots, X_N$ una secuencia de variables tal que para cada $i = 1, 2, \ldots, N$, la media de X_i es μ_i y su varianza es σ_i^2, entonces:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{X_i - \mu_i}{\sigma_i} \xrightarrow{P} N(0,1)$$

cuando $N \to \infty$. Aquí $N(0,1)$ es la distribución normal estándar.

Estadística descriptiva e inferencial

- **Estadística descriptiva:** La estadística descriptiva busca **describir** un conjunto de datos (muestra) con el fin de organizarlos y presentarlos.
- **Estadística inferencial:** Busca determinar, por medio de la inducción, propiedades de un conjunto de datos (muestra), que describan la familia de datos de forma **general**.

Aprendizaje estadístico

La teoría del aprendizaje estadístico (y sus aplicaciones) tiene sus raíces en el **análisis estadístico**.

El problema de la inferencia a partir de ejemplos puede plantearse como [4]:

> *Dado una colección de datos (empíricos) originados de una dependencia funcional, inferir esta dependencia.*

Se proponen dos acercamiento a la solución de este problema:

1. **Inferencia particular (paramétrica)**: Busca estimar un número finito de parámetros que describan los datos de un problema particular. Asume una distribución (generalmente normal).
2. **Inferencia general**: Busca encontrar un método para aproximar una función a partir de los ejemplos, sin asumir una familia específica de distribuciones.

Métodos paramétricos y no paramétricos

Algunos métodos paramétricos de aprendizaje son:

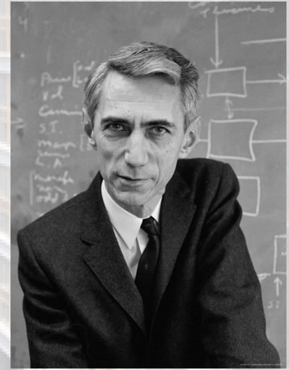
- Métodos de regresión
- Bayes ingenuo
- Perceptrón

Algunos métodos no-paramétricos de aprendizaje son:

- Árboles de decisión
- k-NN
- Redes neuronales profundas

Introducción a la Teoría de la Información

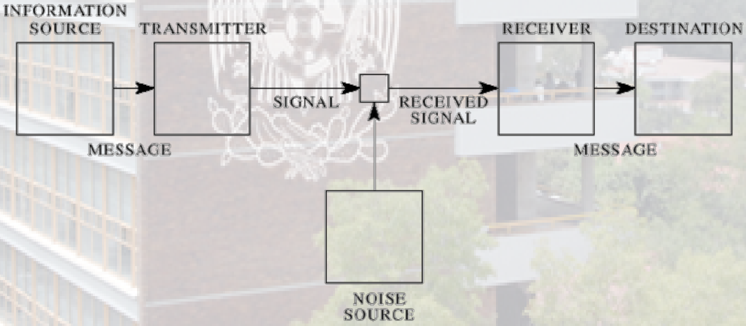
Teoría de la información



- La teoría de información se inaugura con el trabajo pionero de **Claude E. Shannon**: *A mathematical theory of communication* [3].
- Surge en el ámbito de las **telecomunicaciones**. Se plantea el problema de reproducir un mensaje a través de un canal de comunicación.
- Su aplicación se extiende a diferentes áreas del conocimiento (**estadística**, sistemas complejos, lingüística, aprendizaje de máquina, ...).

Modelo matemático de la comunicación

El modelo propuesto por Shannon, llamado **Modelo del Canal Ruidoso**, busca la forma más eficiente de transmitir **información** desde una fuente hacia un destino.



Conceptos básicos de la teoría de la información

Definición (Información)

Si X es una v.a. con función de probabilidad p, definimos la información como:

$$I(X = x) = -\log p(X = x)$$

Definición (Entropia)

Si X es una v.a., definimos la entropia de X como la función

$$H(X) = \mathbb{E}_p[I(X)]$$

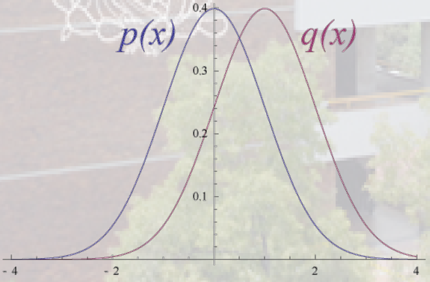
Divergencia de Kullback-Liebler

Un concepto central de la teoría de la información es la divergencia de Kullback-Liebler (KL).

Definición (Divergencia KL)

Dada una v.a. X y dos distribuciones p, q sobre X, la divergencia KL es la función

$$D[p||q] = \mathbb{E}_p[\log \frac{p(X)}{q(X)}]$$



¿Por qué nos interesa la teoría de información?

- En el aprendizaje automático, los datos X muestran distribución p.
- Los algoritmos de ML buscan estimar una distribución empírica $q_\theta = q(\cdot;\theta)$.
- Podemos ver que el objetivo de un algoritmo ML es encontrar la función $\hat{q}_\theta$ que mejor aproxime p. Esta puede estimarse como:

$$\hat{q}_\theta = \arg\min_{q_\theta} D[p||q_\theta]$$

¿Por qué nos interesa la teoría de información?

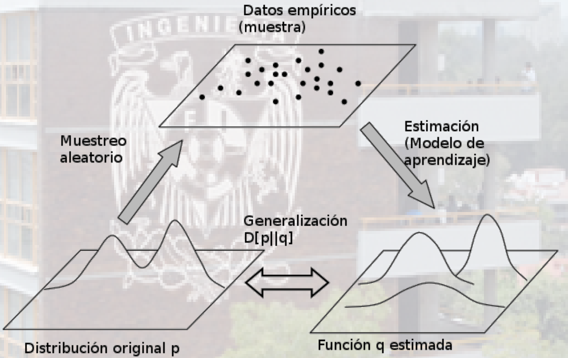


Figura: El objetivo del aprenidzaje es construir un modelo de aprenidzaje que estime una función q más cercana a la distribución real, a partir de datos empíricos $x_1, \ldots, x_N$ [5]

Entropía cruzada

En la práctica, es común utilizar la entropía cruzada, en lugar de la divergencia KL. Ésta última está dada por:

$$H(X; q_\theta) = H(X) + D[p||q_\theta]$$

Se puede observar que:

1. $\arg\min_{q_\theta} D[p||q_\theta] = \arg\min_{q_\theta} H(X; q_\theta)$
2. $H(X; q_\theta) = -\mathbb{E}_p[\log q_\theta(X)]$

Introducción a la teoría de aprendizaje

Teoría de aprendizaje estadístico

La **teoría de aprendizaje estadístico** puede entenderse como la parte teórica del aprendizaje de máquina. Se enfoca en determinar las condiciones en que se puede encontrar una buena aproximación de una **función de predicción** a partir de **datos empíricos**.

Los elementos esenciales para trabajar con esta teoría son:

- X, un conjunto de datos empíricos en $\mathbb{R}^d$.
- Y, un conjunto de clases, generalmente subconjunto de $\mathbb{R}$.
- La función $q : X \to Y$, tal que $q(x) \sim y, x \in X, y \in Y$.

Aprendizaje supervisado y no supervisado

Se pueden distinguir dos grandes familias de modelos de aprendizaje, que dependen del tipo de datos con el que se trabaje:

1. **Aprendizaje supervisado:** Se define por el conjunto de ejemplos:

$$\mathcal{S} = \{(x, y) : x \in \mathbb{R}^d, y \in Y\}$$

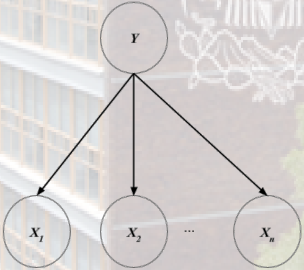
2. **Aprendizaje no-supervisado:** Se define por el conjunto de ejemplos:

$$\mathcal{U} = \{x : x \in \mathbb{R}^d\}$$

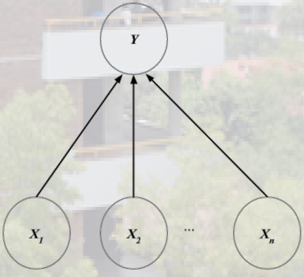
Modelos generativos y discriminativos

Generalmente la función de predicción q puede verse como una función de probabilidad entre $x \in X$ y $y \in Y$. A partir de la forma de estimar esta probabilidad, se tienen dos familias de modelos:

1. **Modelos generativos:** Estiman: $q(y, x) = q(x|y)q(y)$
2. **Modelos discriminativos:** Estiman: $q(y|x)$



Bayes naïve (ingenuo)



Perceptrón

Función de riesgo

En el aprendizaje, es común hablar de la función de riesgo [4]:

$$R(q) = \mathbb{E}_p[L(X, q(X))] \tag{3}$$

En general, la función L está determinada por $-log(\cdot)$; es decir, $R(q)$ es la entropía cruzada.

En la teoría de aprendizaje estadístico se reconocen dos clases de **problemas**:

1. **Problema de regresión:** Está determinado por

$$L = ||y - q(x)||^2$$

2. **Problema de clasificación:** Está determinado por

$$L = \begin{cases} 1 & \text{si } q(x) \neq y \\ 0 & \text{si } q(x) = y \end{cases}$$

Entrenamiento y generalización

El proceso de aprendizaje se divide en dos paradigmas:

1. **Generalización**: cuyo objetivo es determinar la capacidad de generalizar de un modelo. La función de riesgo puede verse como:

$$R(q) = \int_{\Omega} L(x, q(x))dF(x)$$

2. **Entrenamiento**: busca estimar la función q. Se caracteriza por una función de riesgo empírica:

$$R_E(q) = \frac{1}{N}\sum_{i=1}^{N} L(x_i, q(x_i))$$

Parte de la teoría de aprendizaje

La teoría de aprendizaje se puede dividir en las siguientes partes [4]:

1. **Teoría de consistencia:** Busca encontrar las condiciones para que el riesgo de entrenamiento sea igual (o lo más cercano) al riesgo de generalización.
2. **Teoría de cotas:** Busca obtener las cotas en la habilidad de generalización de las máquinas de aprendizaje.
3. **Teoría de control de generalización:** Busca encontrar los mejores métodos que permitan determinar la capacidad de generalizar de una ML a partir de datos empíricos.
4. **Teoría de algoritmos:** Se ocupa de desarrollar algoritmos de máquinas de aprendizaje.

Consistencia y teoría de cotas

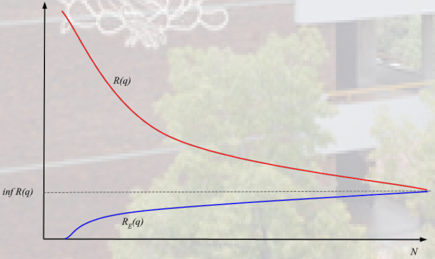
Consistencia

Definición (Consistencia)

Se dice que un proceso de aprendizaje es consistente si el riesgo de entrenamiento y generalización convergen al ambos a un valor mínimos. Esto es:

$$\inf_q R_E(q) \xrightarrow{P} \inf_q R(q) \tag{4}$$

Cuando N $\rightarrow \infty$, siendo N el número de ejemplos.



Convergencia por datos empíricos

Un problema al que nos enfrentamos es determinar qué tantos datos son necesarios para qué se dé la convergencia que pedimos.

Si bien no es factible determinar esto, tenemos resultados que nos dan información sobre el poder de generalización:

$$P(\sup |R(q) - R_E(q)| > \epsilon) \leq 2e^{-2\epsilon^2 N} \tag{5}$$

Esta desigualdad nos hace ver que buscar minimizar el error de evaluación (en base al de entrenamiento) requiere que el número de ejemplos crezca.

Overfiiting

Un problema común en el aprendizaje es el del overfitting.

Definición (Overfiting)

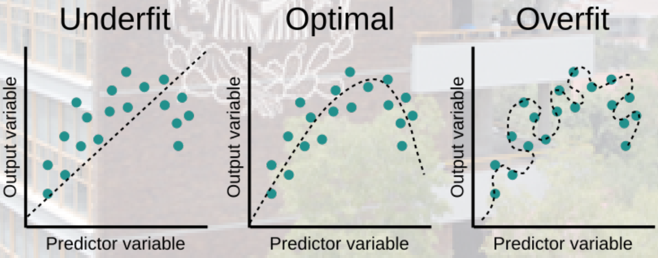
Se da overfitting cuando, en el entrenamiento, se estima una función q que se sobreajusta a los datos ($R_E(q)$ bajo), de tal forma que no es capaz de generalizar ($R(q)$ alto).

Cuando la estimación de la ML es pobre, se habla de **underfitting**.

Overfiiting

El overfitting (y underfitting) dependen en gran medida de la **capacidad** de una ML para tener una buena generalización.

- Una ML con 'poca potencia' no podra representar los datos adecuadamente (undefitting).
- Una ML con 'mucha potencia' requerirá de muchos datos para no sobre representar la muestra de entrenamiento (overfitting).



Por tanto, requerimos de un concepto de **capacidad** o potencia.

Dimensión de Vapnik-Chervonenski

La dimensión de Vapnik-Chervonenski o dimensión VC nos dice la “capacidad” que tiene una ML.

La dimensión VC es el máximo número de datos $x_1, \ldots, x_h$ que la máquina de aprendizaje es capaz de clasificar correctamente sin importar su organización.

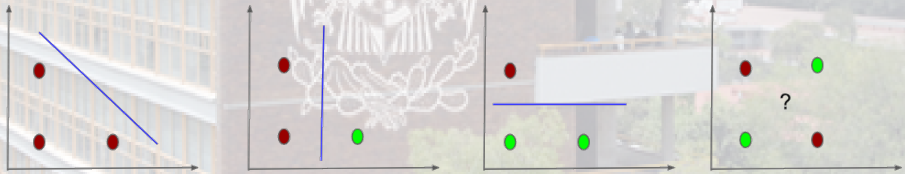


Figura: Una clasificación lineal (binaria) tiene dimensión VC de 3.

Desigualdad de Vapnik-Chervonenski

Un resultado importante sobre la teoría de aprendizaje es la siguiente desigualdad:

$$R(q) \leq R_E(q) + \sqrt{\frac{h[\log(2N/h) + 1]}{N}} \tag{6}$$

donde h es la dimensión de Vapnik-Chervonenski y N el número de ejemplos.

Un algoritmo con mayor capacidad (dimensión VC grande) requiere de un mayor número de datos. De otra forma, puede presentarse overfitting.

Lidiar con el overfitting

Una forma común de minimizar el impacto del overfitting es a partir de la **regularización**. Algunos métodos comúnes son:

- Regularización de Tychonoff:

$$R_\gamma(q) = R(q) + \gamma W(q)$$

- Dropout (para redes neuronales).
- Early stopping.

Regularización de Tychonoff

En la regularización de Tychonof, se deben tomar en cuenta las siguientes condiciones:

- $\gamma > 0$.
- La solución $\hat{q}$ debe ser parte del dominio de W.
- Para todo c, el conjunto $\{q : W(q) \leq c\}$ es compacto.

Algunas formas que puede tomar en la práctica son:

1. Regularización L_1:
$$R_\gamma(q) = R(q) + \gamma||q||_1$$
2. Regularización L_2:
$$R_\gamma(q) = R(q) + \gamma||q||_2^2$$
3. Regularización Elastic net:
$$R_\gamma(q) = R(q) + \gamma_1||q||_1 + \gamma_2||q||_2^2$$

Evaluación

Evaluación de la generalización

Generalmente para evaluar la capacidad de generalizar a partir de datos empíricos se utiliza una **matriz de confusión**:

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Evaluación de la generalización

A partir de la matriz de confusión se pueden proponer diferentes **métricas de evaluación**:

- **Accuracy**:

$$Acc = \frac{VP + VN}{VP + FN + FP + VN}$$

- **Precisión**:

$$Prec = \frac{VP}{VP + FP}$$

- **Recall**:

$$Rec = \frac{VP}{VP + FN}$$

- **Medida F1**:

$$F_1 = 2 \cdot \frac{Prec \cdot Rec}{Prec + Rec}$$

Ejemplo de evaluación

Supongamos que tenemos un conjunto de evaluación supervisado y una predicción de clases dada por la máquina de la siguiente forma:

Número de ejemplo	Observación	Predicción
1	1	1
2	1	1
3	1	0
4	1	1
5	1	0
6	0	0
7	0	1
8	0	1
9	0	0
10	0	1

Ejemplo de evaluación

A partir de las observaciones del cuadro anterior, obtenemos la siguiente matriz de confusión:

	Positivos	**Negativos**
Positivos	3	2
Negativos	3	2

A partir de esta matriz de confusión podemos obtener diferentes métricas de evaluación. El **Accuracy** es:

$$
\begin{aligned}
Acc &= \frac{VP + VN}{VP + FN + FP + VN} \\
&= \frac{3 + 2}{3 + 2 + 3 + 2} \\
&= \frac{5}{10} = 0{,}5
\end{aligned}
$$

Ejemplo de evaluación

Para las métricas de precisión y recall tenemos:

- **Precisión:**

$$Prec = \frac{VP}{VP + FP} = \frac{3}{6} = 0{,}5$$

- **Recall:**

$$Rec = \frac{VP}{VP + FN} = \frac{3}{5} = 0{,}6$$

Ejemplo de evaluación

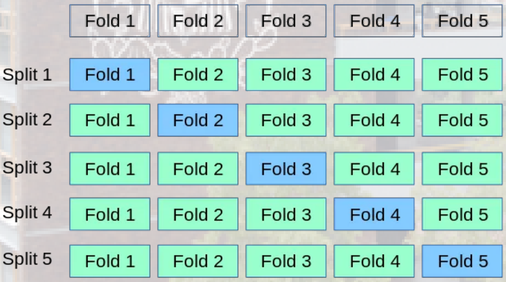
Finalmente, para la **medida $\mathbf{F_1}$** tenemos:

$$
\begin{aligned}
F_1 &= 2 \cdot \frac{Prec \cdot Rec}{Prec + Rec} \\
&= 2 \cdot \frac{0{,}5 \cdot 0{,}6}{0{,}5 + 0{,}6} \\
&= 2 \cdot \frac{0{,}3}{1{,}1} \qquad \approx 0{,}545
\end{aligned}
$$

Holdout y Validación cruzada

Al método que hemos descrito para realizar la evaluación se le conoce como **Holdout**: Sólo se evalúa el conjunto de datos una vez.

Un método que busca representar mejor la generalización es el de **validación cruzada** por **k-folds**: Se define un número *k* de folds. El subconjunto de evaluación se **alterna** en *k* iteraciones.



Evaluación no supervisada

Cuando se realiza evaluación sobre conjuntos no supervisados, generalmente se cuenta con un conjunto **gold standard** $C = \{c_1, .., c_k\}$ y clústers determinados por la máquina $\hat{C} = \{\hat{c}_1, \ldots, \hat{c}_k\}$.

- **Pureza:**

$$Purity(C, \hat{C}) = \frac{1}{N} \sum_i \max_j |\hat{c}_i \cap c_j|$$

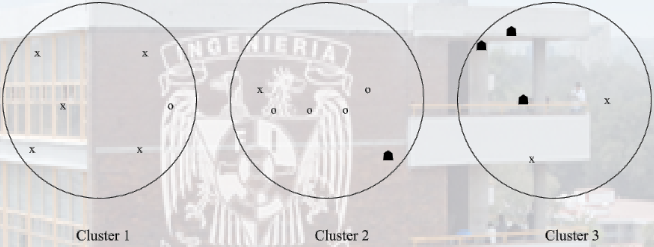
- **Mutual information:**

$$MI(C, \hat{C}) = \sum_i \sum_j p(\hat{c}_i \cap c_j) \log \frac{p(\hat{c}_i \cap c_j)}{p(\hat{c}_i) p(c_j)}$$

- **Normalized Mutual Information:**

$$NMI(C, \hat{C}) = 2 \cdot \frac{MI(C, \hat{C})}{H(C) + H(\hat{C})}$$

Evaluación no supervisada



$$
\begin{aligned}
Purity(C, \hat{C}) &= \frac{1}{N} \sum_i \max_j |\hat{c}_i \cap c_j| \\
&= \frac{1}{17}(4 + 4 + 3) \\
&= \frac{11}{17} \approx 0{,}647
\end{aligned}
$$

Los elementos de un modelo de aprendizaje

Los elementos que conformarán nuestro modelo de aprendizaje son los siguientes:

1. **Conjunto de datos**: Dos tipos de conjuntos:
   1.1 Conjunto supervisado: $\mathcal{S} = \{(x, y) : x \in \mathbb{R}^d, y \in Y\}$.
   1.2 Conjunto no-supervisado: $\mathcal{U} = \{x : x \in \mathbb{R}^d\}$.
2. **Algoritmo de aprendizaje**: Determinado por:
   2.1 Una arquitectura (o función) que depende del tipo de datos; por ejemplo, las funciones de la forma $f(x) = wx + b$, con w y b parámetros de la función.
   2.2 Una función de riesgo $R_E(f)$ para estimar los parámetros que mejor se ajusten a los datos.
3. **Un método de evaluación**. Determinar una métrica y un procedimiento de evaluación.

Warren S McCulloch and Walter Pitts.
A logical calculus of the ideas immanent in nervous activity.
The bulletin of mathematical biophysics, 5(4):115–133, 1943.

Frank Rosenblatt.
The perceptron: a probabilistic model for information storage and organization in the brain.
Psychological review, 65(6):386, 1958.

Claude Elwood Shannon.
A mathematical theory of communication.
Bell system technical journal, 27(3):379–423, 1948.

Vladimir Vapnik.
Statistical learning theory. 1998, volume 3.
Wiley, New York, 1998.

Sumio Watanabe.

Algebraic geometry and statistical learning theory, volume 25.
Cambridge University Press, 2009.

The End