

Introducción a la teoría de aprendizaje estadístico

Mtro. Víctor Mijangos

Facultad de Ingeniería

13-17 de enero, 2020



Antecedentes del aprendizaje estadístico

Algunos puntos históricos relevantes en la historia del aprendizaje estadístico son:

- ▶ En el Siglo XVII, Gottfried Leibnez propone una marco teórico para llevar a cabo operaciones lógicas (*calculus tatiocinato*).
- ▶ 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.



Antecedentes del 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 particula (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



Herramientas teóricas del aprendizaje estadístico

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

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

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



Variables aleatorias

Un aspecto importante del aprendizaje son los **datos**.

Los datos se conformarán por variables que definan una característica de esa instancia.

Cada instancia en los datos será descrita por una serie de variables registradas en un vector:

$$\mathbf{x}^T = (X_1 \quad X_2 \quad \cdots \quad X_d)$$

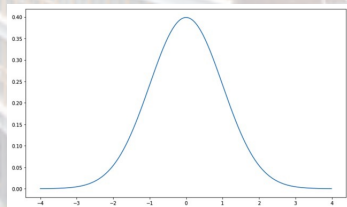


Funciones de probabilidad y de distribución

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

Se cumple que:

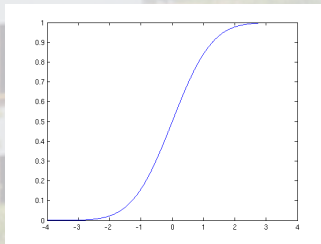
- ▶ $p(X = x) \geq 0$
- ▶ $\int_{-\infty}^{\infty} p(x)dx = 1$



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

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

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



Ley de los grandes números

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

Sea X_1, X_2, \dots, 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 $N \rightarrow \infty$

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



Teorema central del límite

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

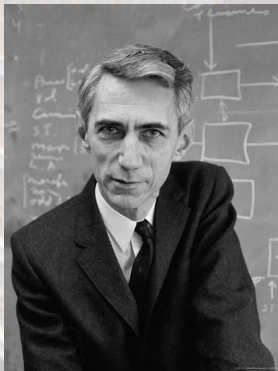
Sea X_1, X_2, \dots, X_N una secuencia de variables tal que para cada $i = 1, 2, \dots, N$, la media de X_i es μ_i y su varianza es σ_i^2 , entonces:

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

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



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.
- ▶ Sin embargo, su aplicación se extiende a diferentes áreas del conocimiento (estadística, sistemas complejos, lingüística, aprendizaje de máquina, ...).



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 (Entropía)

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

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



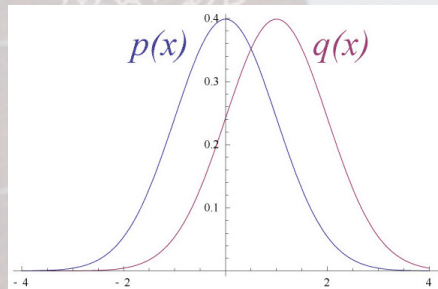
Divergencia de Kullback-Liebler

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

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



¿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}_θ 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?

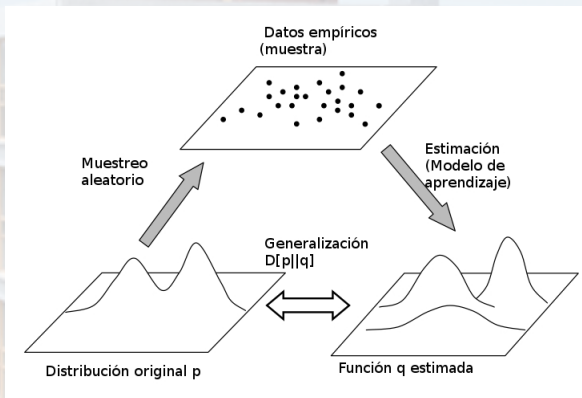


Figura: El objetivo del aprendizaje es construir un modelo de aprendizaje que estime una función q más cercana a la distribución real, a partir de datos empíricos x_1, \dots, x_N [5]



¿Por qué nos interesa la teoría de información?

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. $H(X; q_\theta) = -\mathbb{E}_p[\log q_\theta(X)]$



Función de riesgo

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

$$R(q) = \mathbb{E}_p[L(X, q(X))] \quad (1)$$

En general, la función L está determinada por $-\log(\cdot)$; es decir, $R(q)$ es la entropía cruzada. Otro valor común que puede tomar es $L = \|y - q(x)\|^2$, que es equivalente a la entropía cruzada cuando asumimos una distribución gaussiana.



Entrenamiento y evaluación

El proceso de aprendizaje se divide en dos paradigmas:

1. Evaluación: 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:

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



Datos empíricos

Sin embargo, un problema al que nos enfrentamos es determinar qué tantos datos son necesarios.

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

$$P(\sup |R(q) - R_E(q)| > \epsilon) \leq 2ke^{-2\epsilon^2 N} \quad (2)$$

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



Overfitting

Un problema común en el aprendizaje es el del overfitting.

Definición (Overfitting)

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).

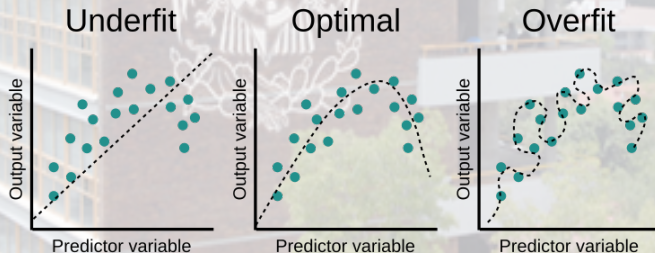
Cuando la estimación de la ML es pobre, se habla de **underfitting**.



Overfitting

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

- ▶ Una ML con 'poca potencia' no podrá representar los datos adecuadamente (underfitting).
- ▶ 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, \dots, x_h que la máquina de aprendizaje es capaz de clasificar correctamente sin importar su clasificación.

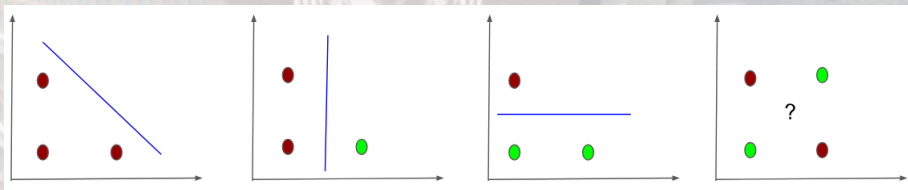


Figura: Una clasificación lineal (binaria) tiene dimensión VC de 3.



Desigualdad de Vapnik-Chervonenski

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

$$R(q) \leq R_E(q) + \sqrt{\frac{h[\log(2N/h) + 1]}{N}} \quad (3)$$

donde h es la dimensión de Vapnik-Chervonenski.

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 comunes 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 Tychonoff, 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:

$$R_\gamma(q) = R(q) + \gamma_1 \|q\|_1 + \gamma_2 \|q\|_2^2$$



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.**





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

