Unidad 2

Rodrigo Barrera

La **Ley de los grandes números** (LLN) es un principio fundamental en la teoría de probabilidad que describe el resultado de realizar el mismo experimento un gran número de veces. Según la LLN, el promedio de los resultados obtenidos de un gran número de pruebas se acercará cada vez más al valor esperado a medida que se aumenta el número de pruebas.

Existen dos versiones de la Ley de los grandes números:

- **Ley débil (LLND):** Supongamos $X_1, X_2, \ldots$ son variables aleatorias i.i.d. con $E[X_i] = \mu$ y $Var(X_i) = \sigma^2 < \infty$. Entonces, para todo $\epsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \to 0 \text{ cuando } n \to \infty.$$

- **Ley fuerte (LLNF):** Bajo las mismas condiciones que la LLND,

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1.$$

Consideremos el experimento de lanzar una moneda justa. Sea X_i la variable aleatoria que representa el resultado del i-ésimo lanzamiento, donde $X_i = 1$ si es cara y $X_i = 0$ si es cruz. La esperanza matemática $E[X_i] = 0{,}5$. Según la Ley de los Grandes Números, si lanzamos la moneda un número suficientemente grande de veces, la media muestral $\frac{1}{n}\sum_{i=1}^{n} X_i$ se aproximará a 0.5.

El **Teorema central del límite** (TCL) es uno de los conceptos más importantes en estadística. Afirma que, bajo ciertas condiciones, la suma de una gran cantidad de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) tiende a seguir una distribución normal, independientemente de la forma de la distribución original de las variables.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n de una población con media μ y varianza σ^2 finita. El Teorema Central del Límite establece que cuando n es suficientemente grande, la distribución de la suma normalizada

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

se aproxima a una distribución normal estándar $N(0, 1)$.

- El TCL permite hacer inferencias sobre poblaciones a partir de muestras. Por ejemplo, el promedio de una muestra de datos puede ser utilizado para estimar el promedio de la población total, y gracias al TCL, sabemos que la distribución de este estimador tiende a ser normal, lo cual facilita la realización de pruebas de hipótesis y la construcción de intervalos de confianza.
- Técnicas como el bootstrap, que implican muestrear repetidamente con reemplazo de un conjunto de datos para estimar la distribución de una estadística, se fundamentan en principios relacionados con el TCL. Esto permite estimar la variabilidad y construir intervalos de confianza para cualquier estadística de interés, incluso cuando no se conoce la distribución subyacente.

- El teorema de aproximación universal establece que una red neuronal artificial con al menos una capa oculta puede aproximar cualquier función continua en un subconjunto compacto de $\mathbb{R}^n$, bajo ciertas condiciones.
- Este teorema subyace en la base teórica del éxito de las redes neuronales en diversas tareas de aprendizaje automático.

Una red neuronal es un modelo computacional inspirado en la forma en que las redes neuronales biológicas del cerebro humano procesan la información. Se compone de unidades básicas llamadas 'neuronas' o 'nodos', organizadas en capas. Estas neuronas se conectan entre sí con 'pesos' que se ajustan durante el proceso de entrenamiento. Las redes neuronales son capaces de aprender y modelar relaciones complejas entre los datos de entrada y salida a través de este proceso de entrenamiento, lo que las hace muy efectivas para tareas de clasificación, regresión, reconocimiento de patrones, y más.

Una capa oculta en una red neuronal es una de las capas que se encuentran entre la capa de entrada (donde los datos se introducen en la red) y la capa de salida (donde se obtienen los resultados finales). Las capas ocultas son el çorazón"de una red neuronal, donde se realiza la mayor parte del procesamiento a través de una combinación lineal de las entradas recibidas, seguida de una aplicación no lineal (función de activación) a estas combinaciones lineales.

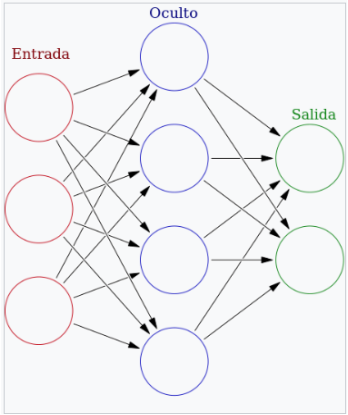

Figura: Red neuronal artificial

Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función continua en un compacto $K \subset \mathbb{R}^n$. Para cualquier $\epsilon > 0$, existe una red neuronal con una sola capa oculta y una función de activación no constante, acotada y continua, tal que para todo $x \in K$,

$$|f(x) - RN(x)| < \epsilon.$$

- Intuitivamente, el teorema nos dice que las redes neuronales tienen la "flexibilidad" para modelar cualquier función continua, ajustando adecuadamente los pesos y sesgos.
- Ejemplo: aproximación de la función $sin(x)$ usando una red neuronal con una capa oculta.

- **Implicaciones:** este teorema fundamenta la capacidad de las redes neuronales para aprender una amplia variedad de tareas complejas.
- **Limitaciones:** el teorema no proporciona una guía sobre la arquitectura de la red óptima ni sobre el tiempo de entrenamiento necesario para alcanzar la aproximación deseada.

Una **función de activación** en una red neuronal es una transformación matemática aplicada a la entrada ponderada de una neurona o a la salida de una capa de neuronas. Su objetivo es introducir no linealidades en el modelo, permitiendo que la red aprenda y modele relaciones complejas entre los datos de entrada y salida.

La **función sigmoide** transforma los valores de entrada a un rango entre 0 y 1, definida matemáticamente como:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

donde x es la entrada y e es la base del logaritmo natural. A pesar de su utilidad para interpretar la salida como una probabilidad, su uso en capas ocultas es limitado debido al problema del **desvanecimiento del gradiente**.

La **función ReLU** es ampliamente utilizada en capas ocultas por su simplicidad y eficacia, definida como:

$$R(x) = \text{máx}(0, x)$$

Esta función devuelve la entrada si es positiva, y cero en caso contrario.

La **función tanh** mapea las entradas a un rango entre -1 y 1, y se define como:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Aunque similar a la sigmoide, su rango simétrico respecto al origen la hace más preferible en ciertas aplicaciones.

El **desvanecimiento del gradiente** es un problema que puede ocurrir durante el entrenamiento de redes neuronales profundas, especialmente en aquellas que utilizan algoritmos de retropropagación para ajustar los pesos de la red. Este fenómeno se caracteriza por una disminución exponencial de los gradientes a medida que se retropropaga a través de las capas de la red, lo que resulta en gradientes muy pequeños para las capas más cercanas a la entrada. Como consecuencia, los pesos en estas capas iniciales se actualizan muy lentamente, lo que ralentiza significativamente el entrenamiento de la red o incluso lo detiene por completo, impidiendo que la red aprenda adecuadamente.

Por ejemplo, en el caso de la función sigmoide, su derivada máxima es 0.25, lo que significa que en cada capa a través de la cual se retropropaga el gradiente, este se reduce al menos en un 75 %, asumiendo que la derivada máxima se alcanza en cada paso. En la práctica, la reducción es a menudo mucho mayor, llevando a gradientes cercanos a cero en las capas iniciales.

Para mitigar el desvanecimiento del gradiente, se han propuesto varias soluciones, como el uso de funciones de activación alternativas como la unidad lineal rectificada (ReLU), que tiene una derivada constante de 1 para todos los valores positivos de entrada, evitando así la disminución rápida de los gradientes. Además, técnicas avanzadas de inicialización de pesos, métodos de normalización como la normalización por lotes, y arquitecturas de red especializadas como las redes neuronales recurrentes de tipo LSTM (Long Short-Term Memory) o GRU (Gated Recurrent Unit) también ayudan a prevenir este problema.

La convergencia uniforme se refiere a una propiedad de las secuencias de funciones. Una secuencia de funciones $\{f_n\}$ converge uniformemente a una función f en un conjunto D si, para todo $\epsilon > 0$, existe un número natural N tal que para todos los $n \geq N$ y para todo x en D, la diferencia absoluta entre $f_n(x)$ y $f(x)$ es menor que ϵ. Matemáticamente, esto se expresa como:

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \forall x \in D, |f_n(x) - f(x)| < \epsilon$$

La convergencia uniforme es más fuerte que la convergencia puntual (donde la condición debe cumplirse para cada punto x de manera individual, posiblemente con diferentes valores de N para diferentes puntos) porque el número N es válido para todos los puntos en D simultáneamente. En el contexto del aprendizaje automático, la convergencia uniforme puede ser relevante al estudiar cómo las estimaciones basadas en muestras finitas (por ejemplo, una función de predicción entrenada) se acercan a la verdadera función subyacente a medida que aumenta el tamaño de la muestra.

La consistencia es un concepto en la teoría de la estimación que describe una propiedad deseable de los estimadores. Un estimador es consistente si, a medida que el tamaño de la muestra n tiende a infinito, converge en probabilidad al parámetro verdadero θ que está estimando. Esto significa que la probabilidad de que la diferencia absoluta entre el estimador $\hat{\theta}_n$ y el verdadero valor θ sea mayor que cualquier $\epsilon > 0$ tiende a cero a medida que n se aproxima a infinito:

$$\forall \epsilon > 0, \lim_{n \to \infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

La Cota de Vapnik-Chervonenkis (VC) ofrece una perspectiva teórica sobre la capacidad de generalización de los modelos de aprendizaje automático, abordando el compromiso entre la complejidad del modelo y el riesgo de sobreajuste. Desde un punto de vista teórico, la cota VC es relevante para entender cómo la estructura de un modelo y la cantidad de datos de entrenamiento disponibles influyen en el rendimiento del modelo en datos no vistos.

Es una medida de la capacidad o complejidad de una clase de funciones, reflejando el número máximo de puntos que pueden ser separados en todas las configuraciones posibles por los modelos dentro de esta clase. Por ejemplo, en el contexto de la clasificación binaria, la dimensión VC sería el número máximo de puntos que se pueden etiquetar de manera completamente correcta en todas las formas posibles usando los modelos de esa clase.

La cota VC proporciona una base teórica para el principio de la complejidad mínima, que sugiere que entre dos modelos que explican igualmente bien los datos de entrenamiento, se debería preferir el más simple porque es más probable que generalice bien. Este principio está en el corazón de muchas técnicas de regularización y selección de modelos en aprendizaje automático, donde se agregan términos de penalización a la función de pérdida para controlar la complejidad del modelo y prevenir el sobreajuste.

Figura: Occam

Pluralitas non est ponenda sine necessitate (La pluralidad no se debe postular sin necesidad).

El compromiso entre sesgo y varianza se manifiesta en que al reducir el sesgo (haciendo el modelo más complejo para ajustarse mejor a los datos) aumentamos la varianza (sensibilidad a los datos de entrenamiento específicos), y viceversa. En términos prácticos, un modelo muy simple no capturará la verdadera estructura de los datos (alto sesgo), mientras que un modelo demasiado complejo capturará patrones aleatorios y ruido, en lugar de la verdadera relación subyacente (alta varianza).

El **sesgo** se refiere al error sistemático introducido por aproximar un problema real y complejo, que puede ser no lineal y de alta dimensión, usando un modelo más simple. Por ejemplo, usar un modelo lineal para datos que en realidad tienen una relación no lineal introduce un alto sesgo. Los modelos con alto sesgo son generalmente simples (baja complejidad de modelo) y tienden a subestimar la complejidad real de los datos, lo que se conoce como 'underfitting' o subajuste.

La **varianza** se refiere a la sensibilidad del modelo a las fluctuaciones en el conjunto de datos de entrenamiento. Un modelo con alta varianza cambia significativamente con pequeñas variaciones en el conjunto de entrenamiento. Estos modelos son complejos y tienden a ajustarse a los ruidos o fluctuaciones aleatorias de los datos de entrenamiento, lo que se conoce como 'verfitting' o sobreajuste.