Chapter 0

Expedition for Alchemists

It's the possibility of having a dream come true that makes life interesting.

– Paulo Coelho, **The Alchemist**



EEP learning has been one of the most popular topics in science and engineering over the past decade and is undergoing rapid advances. This chapter aims to provide beginners, especially "alchemists" who are less concerned with mathematical principles, with a superficial yet mathematically rigorous overview of the essential elements of deep learning.

0.1 Brief Introduction

In brief, *deep learning* is a paradigm based on composed functions of simple functions, called *neural networks*. The composition of multiple (layers) simple functions provides sufficient capacity for representation and flexibility for nonlinear approximation in high dimensions.

The *neural networks* were previously studied by computer scientists, but the real impact of deep learning was revealed only in the last decade. With the advent of GPU architectures, deep learning approaches have begun to outperform classical algorithms in computer vision, data mining, natural language processing, robotics, and related areas. However, this note does not emphasize the practical elements of computation. Instead, we will focus more on the mathematical principles, using basic tools from linear algebra, differential equations, and probability.

0.2 Preparations

For simplicity, we view neural networks as maps from the data space $\mathcal{X}$ to the label space $\mathcal{Y}$. The dependent parameters are denoted by $\theta \in \Theta$, where Θ is an admissible set embedded in some other space. The abstract structure of the neural networks is $f : \mathcal{X} \times \Theta \to \mathcal{Y}$, and the objective is to find a parameter $\theta^* \in \Theta$ such that $f(x, \theta^*)$ achieves a specific goal.

In the context of supervised learning, the ground-truth datapoints $(x_k, y_k) \in \mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$ are provided, and the objective is to approximate a function (neural network) $f(x, \theta^*)$ that can be used to predict the labels for unseen samples in $\mathcal{X}$. The parameter θ^* is found by minimizing the (average) loss function.

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \ell(y, f(x, \theta)),$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ such that $\ell(y, y') = 0$ if and only if $y = y'$, but ℓ is not necessarily a metric or quasi-metric. Typically, the choice of the loss function ℓ and the model f can significantly affect performance and may induce biases.

Remark 0.2.1. *In unsupervised learning, labels are unavailable (partly available in weak supervised learning), and only the input data $x_k \in \mathcal{X}$ are given. The objective is to learn the hidden patterns from the data. It is commonly used for clustering tasks, such as recommendation systems. However, it is not the focus of this note.*

0.2.1 Basic Probability

We will use some basic probability theory later.

Definition 0.2.2. *A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ means*

1. *Ω is the sample space.*
2. *$\mathcal{F}$ is a σ-algebera or event space (set of events).*
3. *$\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure.*

Definition 0.2.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and E be a measurable space. A random variable $X : \Omega \to E$ is a measurable function.*

$$\mathbb{P}[X \in S \subset E] = \mathbb{P}[\{\omega \in \Omega \mid X(\omega) \in S\}]$$

Definition 0.2.4. *The expectation of $X : \Omega \to E$ is*

$$\mathbb{E}[X] = \int_\Omega X(\omega) d\mathbb{P}(\omega).$$

If $f : E \to \mathbb{R}$ is a measurable function, then

$$\mathbb{E}[f(X)] = \int_\Omega f(X(\omega)) d\mathbb{P}(\omega).$$

Remark 0.2.5. *Suppose $\mathbb{E}[f(X)] = m$. It means there exists $\omega \in \Omega$ such that $f(\omega) \geq m$ (or $\leq m$). If the variance $\mathrm{Var}(f(X)) \neq 0$, then the inequality is strict.*

Theorem 0.2.6 (Jensen's Inequality). *Suppose h is a convex function, then*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Definition 0.2.7. *The **probability density function** of a random variable X with*

Continue on next page ⇓

Continue from last page ⇑

respect to a reference measure μ is the Radon-Nikodym derivative

$$dX_*\mathbb{P} = f d\mu$$

and

$$\mathbb{P}[X \in S] = \int_S f d\mu$$

Definition 0.2.8. *Let $p \geq 1$. The random variable $X : \Omega \to E$ satisfies $X \in L^p(\mathbb{P})$ if*

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p} < \infty.$$

Note that if $p < 1$, this does not define a norm.

Theorem 0.2.9 (Interpolation). *Let $p_1, p_2 \geq 1$. If the random variable $X \in L^{p_1}(\mathbb{P}) \cap L^{p_2}(\mathbb{P})$, then for any $p \in (p_1, p_2)$,*

$$\|X\|_p \leq \|X\|_{p_1}^{\alpha} \|X\|_{p_2}^{1-\alpha},$$

where $\frac{1}{p} = \frac{\alpha}{p_1} + \frac{1-\alpha}{p_2}$.

Definition 0.2.10. *Let $X \in L^2(\mathbb{P})$, the variance of X is*

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}|X|^2 - \mathbb{E}[X]^2.$$

If X is multi-dimensional, then it defines the covariance matrix

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T].$$

Theorem 0.2.11 (Chebyshev). *Let $X \in L^2(\mathbb{P})$ be a random variable, then for any $\delta > 0$,*

$$\mathbb{P}(|X| > \delta) \leq \frac{\mathbb{E}|X|^2}{\delta^2}.$$

Definition 0.2.12. *The conditional probability of an event $B \subset \mathcal{F}$ given the event $A \subset \mathcal{F}$ is*

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}.$$

Theorem 0.2.13 (Bayes). *Let $A, B \subset \mathcal{F}$, then*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Definition 0.2.14. *Let X_n be a sequence of random variables.*

1. *Convergence in probability: $X_n \xrightarrow{p} X$ if for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

2. *Convergence almost surely: $X_n \xrightarrow{a.s.} X$ if*

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1.$$

3. *Convergence in L^p: $X_n \xrightarrow{L^p} X$ if*

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0.$$

4. *Convergence in distribution (weak convergence): $X_n \xrightarrow{d} X$ if for any continuous and bounded f,*

$$\lim_{n\to\infty} \mathbb{E}f(X^n) = \mathbb{E}f(X).$$

Theorem 0.2.15. *The relationship between the convergences is well known.*

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X,$$

Continue on next page ⇓

Continue from last page ⇑

and

$$X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X.$$

Definition 0.2.16 (Tightness). *Let $\mathbb{P}_n$ be a sequence of probability measures on Ω. If for any $\varepsilon > 0$ there exists n_0 and a compact set $K \subset \Omega$ that*

$$\mathbb{P}_n(K) > 1 - \varepsilon, \quad \forall n > n_0.$$

Theorem 0.2.17 (Prohorov). *A sequence of probability measures is tight, then it contains a weakly convergent subsequence.*

Theorem 0.2.18 (Law of Large Numbers). *Let X_n be a sequence of independent and identically distributed (i.i.d.) random variables $\sim X$ that $\mathbb{E}[X] = \mu < \infty$. Define*

$$Z_N = \frac{1}{N} \sum_{n=1}^{N} X_n.$$

Then

1. $Z_N \xrightarrow{a.s.} \mu$.
2. *Suppose that $\mathbb{E}|X|^p < \infty$, then $Z_N \xrightarrow{L^p} \mu$.*

Remark 0.2.19. *The weak law of large numbers is directly from Chebyshev's inequality. The strong version needs the Borel-Cantelli theorem. See Durrett (2019) for a proof.*

Question 0.2.20. *Prove the weak version of the LLN.*

Theorem 0.2.21 (Central Limit Theorem). *Let X_n be a sequence of i.i.d. random*

Continue on next page ⇓

Continue from last page ⇑

variables with $\mathbb{E}X = \mu$ and $\operatorname{Var} X = \sigma^2$. Define

$$S_N = \sum_{n=1}^{N} X_n.$$

Then

$$\frac{S_N - N\mu}{N^{1/2}\sigma} \xrightarrow{d} \mathcal{N}(0,1).$$

Remark 0.2.22. *The usual proof uses the characteristic function and the uniqueness of the Hamburger moment problem. See Durrett (2019) for a proof.*

Theorem 0.2.23 (Lindeberg-Feller). *A generalization of the central limit theorem is the following. Let $X_{n,m}$ be a triangular sequence, $1 \le m \le n$, be independent random variables such that $\mathbb{E}X_{n,m} = 0$. If*

1. *$\sum_{m=1}^{n} \mathbb{E}X_{n,m}^2 \to \sigma^2 > 0$;*
2. *For any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{E}[X_{n,m}^2 \mathbf{1}_{|X_{n,m}>\varepsilon|}] = 0$. This is called Lindeberg's condition.*

Then $S_n = \sum_{m=1}^{n} X_{n,m}$ converges weakly to standard normal distribution.

Remark 0.2.24. *This theorem implies that a large number of small, independent events approximately follow a normal distribution. Notice that we do not require identically distributed random variables.*

Corollary 0.2.25 (Lyapunov Central Limit Theorem). *Let X_n be a sequence of independent random variables with $\mathbb{E}X_n = \mu_n$, $\operatorname{Var}(X_n) = \sigma_n^2$, and*

$$\lim_{n\to\infty} \frac{1}{s_N^{2+\delta}} \sum_{n=1}^{N} \mathbb{E}[|X_n - \mu_n|^{2+\delta}] = 0,$$

where $s_N^2 = \sum_{n=1}^{N} \sigma_n^2$. Then

$$\frac{1}{s_N} \sum_{n=1}^{N} (X_n - \mu_n) \xrightarrow{d} \mathcal{N}(0,1).$$

With i.i.d. random variables, the convergence rate of the central limit theorem

can be quantified.

Theorem 0.2.26 (Petrov (2012)). *Let X_n be a sequence of i.i.d. random variables with $\mathbb{E}X = 0, \mathbb{E}X^2 = 1$, and $\mathbb{E}|X|^{2+\delta} = \rho < \infty, \delta \in (0, 1]$. Define*

$$Z_N = \frac{1}{N}(X_1 + \cdots + X_N).$$

Then the cumulative distribution function of $\sqrt{N}Z_N$, denoted by F_N,

$$F_N(x) := \mathbb{P}[x < \sqrt{N}Z_N]$$

satisfies

$$|F_N(x) - \Phi(x)| \leq C\rho N^{-\frac{\delta}{2}},$$

where Φ is the cumulative distribution function of $\mathcal{N}(0, 1)$, and C is an absolute constant.

Remark 0.2.27. *The supremum of the differences between the cumulative distribution functions is known as the Kolmogorov–Smirnov test statistic.*

Question 0.2.28. *Use Bernoulli random variables to show that the rate in Theorem 0.2.26 does not apply for $\delta > 1$. In other words, $\delta = 1$ is sharp.*

When $\delta = 1$, it becomes the Berry-Esseen theorem, which has variants for weakly independent random variables. The convergence rate of $\mathcal{O}(N^{-1/2})$ is optimal.

Theorem 0.2.29 (Berry-Esseen). *Let X_n be a sequence of i.i.d. random variables with $\mathbb{E}X = 0, \mathbb{E}X^2 = 1$, and $\mathbb{E}|X|^3 = \rho < \infty$. Define*

$$Z_N = \frac{1}{N}(X_1 + \cdots + X_N).$$

Then the cumulative distribution function of $\sqrt{N}Z_N$, denoted by F_N satisfies

$$|F_N(x) - \Phi(x)| \leq \frac{C\rho}{\sqrt{N}},$$

where Φ is the cumulative distribution function of $\mathcal{N}(0, 1)$, and $C < 0.4748$ is an absolute constant.

The Berry-Esseen theorem describes the cumulative distribution of the scaled variable $\sqrt{N}Z_N$, whose variance is normalized. The next theorem directly studies the convergence of the empirical cumulative distribution.

Theorem 0.2.30 (Dvoretzky–Kiefer–Wolfowitz Inequality). *Let X_n be a sequence of i.i.d. random variables. The empirical cumulative distribution function F_N is*

$$F_N(x) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}(X_i \leq x),$$

satisfies

$$\mathbb{P}[\sup_{x\in\mathbb{R}} |F_N(x) - F(x)| > \varepsilon] \leq Ce^{-2N^2\varepsilon}, \quad \forall \varepsilon > 0.$$

The constant $C = 2$ is sharp. If $\varepsilon > \sqrt{\frac{\log 2}{2N}}$, the constant $C = 1$ suffices.

Remark 0.2.31. *We consider a simple example where X_n are i.i.d. random variables $\sim X$ with a uniform bound, then*

$$Z_N - \mathbb{E}[X] = \int_{\mathbb{R}} x dF_N(x) - \int_{\mathbb{R}} x dF(x) = -\int_{\mathbb{R}} (F_N(x) - F(x))dx,$$

Therefore, for any $\varepsilon > 0$, $\mathbb{P}[\lim_{N\to\infty} |Z_N - \mathbb{E}[X]| < \varepsilon] = 1$, that is, $Z_N \xrightarrow{a.s.} \mathbb{E}[X]$.

Remark 0.2.32. *The multivariate version of DKW inequality can be found in Naaman (2021), where the constant $C = d(N+1)$ in d dimensions. For sufficiently large N, C can be reduced to $2d$, which is sharp.*

Next, we introduce two useful concentration inequalities. These bounds are sharper than Chebyshev's inequality because of the additional independence requirement.

Theorem 0.2.33 (Chernoff Bound). *Let $X = \sum_{n=1}^{N} X_n$, where X_n are independent Bernoulli random variables with parameters p_n. Let $\mu = \mathbb{E}[X] = \sum_n p_n$. Then,*

1. *Upper Tail.*

$$\mathbb{P}[X \geq (1+\varepsilon)\mu] \leq \exp\left(-\frac{\varepsilon^2}{2+\varepsilon}\mu\right), \quad \forall \varepsilon > 0.$$

Continue on next page ⇓

Continue from last page ⇑

2. *Lower Tail.*

$$\mathbb{P}[X \leq (1-\varepsilon)\mu] \leq \exp\left(-\frac{\varepsilon^2}{2}\mu\right), \quad \forall \varepsilon \in (0,1).$$

A more useful version of the Chernoff bound is the following.

Remark 0.2.34. *If $\varepsilon \in (0,1)$, then*

$$\mathbb{P}[|X-\mu| \leq \varepsilon\mu] \leq 2\exp\left(-\frac{\varepsilon^2}{3}\mu\right).$$

Theorem 0.2.35 (Hoeffding's Inequality)**.** *Let X_n be independent random variables with bounds $a_n \leq X_n \leq b_n$. Let $X = \sum_n X_n$ and $\mu = \sum_n \mathbb{E}[X_n]$. Then,*

$$\mathbb{P}[X \geq (1+\varepsilon)\mu] \leq \exp\left(-\frac{2\varepsilon^2\mu^2}{\sum_n (b_n-a_n)^2}\right)$$

and

$$\mathbb{P}[X \leq (1-\varepsilon)\mu] \leq \exp\left(-\frac{2\varepsilon^2\mu^2}{\sum_n (b_n-a_n)^2}\right).$$

Remark 0.2.36. *Apply the above inequality to i.i.d. Bernoulli random variables of parameter p. It produces a two-sided tail probability*

$$\mathbb{P}_{\textit{Hoeffding}} = 2\exp\left(-2\varepsilon^2 np^2\right), \qquad \mathbb{P}_{\textit{Chernoff}} = 2\exp\left(-\varepsilon^2 np/3\right).$$

When p is small, the Chernoff bound is much sharper. For larger p, the bounds are comparable.

0.2.2 Basic Information Theory

The *entropy* of a random variable is the average level of *uncertainty* or difficulty of encoding about the variable's states. Higher entropy implies greater uncertainty and thus is more difficult to encode.

Definition 0.2.37 (Shannon Entropy)**.** *Let X be a random variable that takes values in a discrete set E. The entropy is*

$$\mathbb{H}(X) = \mathbb{E}[-\log p(X)],$$

Continue on next page ⇓

Continue from last page ⇑

where $p(x) = \mathbb{P}[X = x]$.

Remark 0.2.38. *Note that the* log *function here is taken to be the natural logarithm for convenience; the output is denoted in "nats". Usually, the* log *function should be* $\log_2$*, so the output is the real "bits".*

Remark 0.2.39. *The thermodynamic entropy from statistical mechanics has a similar form (with an extra Boltzmann constant).*

$$S = -k_B \sum p_i \log p_i,$$

where p_i *is the probability of* i*-th microstate in an equilibrium ensemble. The Boltzmann entropy formula takes all* p_i *equal:* $S = -k_B \log W$*, where* W *is the total number of microstates.*

When X is associated with a probability density function $f(x)$, then we can define its differential entropy as a generalization of the Shannon entropy.

Definition 0.2.40 (Differential Entropy)**.** *Let* X *be a random variable with a probability density function* $f(x)$*. The differential entropy is*

$$\mathbb{H}(X) = \mathbb{E}[-\log f(X)] = -\int f(x) \log f(x) dx.$$

Definition 0.2.41 (Conditional Entropy)**.**

$$\mathbb{H}(X \mid Y) = -\sum_{x,y} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_Y(y)},$$

where $p_{X,Y}$ *is the joint probability, and* p_Y *is the marginal probability.*

We can immediately find that $\mathbb{H}(X) \geq \mathbb{H}(X \mid Y)$ and

$$\mathbb{H}(X,Y) = \mathbb{H}(X \mid Y) + \mathbb{H}(Y).$$

It means that the information about X and Y jointly is equivalent to first knowing Y, then knowing X given Y (vice versa).

Definition 0.2.42. *The mutal information*

$$I(X;Y) := \mathbb{H}(X) - \mathbb{H}(X \mid Y) = \mathbb{H}(Y) - \mathbb{H}(Y \mid X).$$

Remark 0.2.43. *The mutual information is symmetric and can be written as*

$$I(X;Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y)$$

Question 0.2.44. *How to use the entropy argument to estimate the upper bound of the balance puzzle?*

Remark 0.2.45. *Any deterministic processing will only reduce the entropy. Since*

$$\mathbb{H}(f(X) \mid X) = 0.$$

Then

$$\mathbb{H}(X) = \mathbb{H}(f(X) \mid X) + \mathbb{H}(X) = \mathbb{H}(f(X)) + \mathbb{H}(X \mid f(X)) \geq \mathbb{H}(f(X)).$$

Question 0.2.46. *A pseudorandom number generator (RNG) is a deterministic algorithm for generating random numbers from a given seed. Therefore, the entropy of the resulting sequence will not be greater than the entropy of the seed. How to understand the randomness there?*

Remark 0.2.47 (Submodularity). *Let X, Y, Z be random variables, then*

$$\mathbb{H}(X,Z) + \mathbb{H}(X,Y) \geq \mathbb{H}(X) + \mathbb{H}(X,Y,Z).$$

Since

$$\mathbb{H}(X,Z) - \mathbb{H}(X) = \mathbb{H}(Z \mid X) \geq \mathbb{H}(Z \mid X,Y) = \mathbb{H}(X,Y,Z) - \mathbb{H}(X,Y).$$

It can be readily generalized to index sets I and J.

$$\mathbb{H}\left(\{X_i\}_{i\in I}\right) + \mathbb{H}\left(\{X_i\}_{i\in J}\right) \geq \mathbb{H}\left(\{X_i\}_{i\in I\cap J}\right) + \mathbb{H}\left(\{X_i\}_{i\in I\cup J}\right).$$

Definition 0.2.48 (Relative Entropy). *The Kullback-Leibler (KL) divergence of two probability distributions P and Q is defined as follows.*

$$D_{KL}(P\|Q) := \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

It defines a quasi-distance (not symmetric) and measures the gain of Q over P.

Theorem 0.2.49 (Gibbs Inequality). $D_{KL}(P\|Q) \geq 0$.

Question 0.2.50. *Therefore, taking Q as uniform distribution leads to $\mathbb{H}(P) \leq \log(|\Omega|)$, $|\Omega|$ is the cardinality of Ω. Prove this.*

Remark 0.2.51. *The mutual information $I(X;Y)$ can be written as*

$$\begin{aligned} I(X;Y) &= D_{KL}(p_{X,Y}\|p_X p_Y) \\ &= \mathbb{E}_Y[D_{KL}(p_{X,Y}\|p_X)] \end{aligned}$$

The cross-entropy measures the mean extra "bits" needed for a distribution Q to encode the true distribution P.

Definition 0.2.52 (Cross-Entropy). *The cross-entropy of two distributions P and Q is*

$$C(P,Q) = -\sum_x P(x) \log Q(x) = \mathbb{H}(P) + D_{KL}(P\|Q).$$

Remark 0.2.53. *For instance, two Bernoulli distributions with parameters p and p' have the cross-entropy as*

$$C(p,p') = -p \log p' - (1-p) \log(1-p').$$

Usually, when comparing a distribution Q to a reference fixed distribution P, the cross-entropy can be regarded as the KL divergence, since $\mathbb{H}(P)$ is fixed. However, the converse (in which Q is fixed) is not equivalent.

Question 0.2.54. *Let P be a uniform distribution on $[-1,1]$ and Q be a uniform distribution on $[-2,2]$, compute $D_{KL}(P\|Q)$ and explain why.*

Empirically, the reference distribution P is not known and must be estimated from a dataset $\mathcal{D} \sim P$, which induces an empirical distribution with density $P(x) = \frac{1}{N}\sum_{x\in\mathcal{D}} \delta(x)$, where $N = |\mathcal{D}|$, that is,

$$\mathcal{C}(\mathcal{D}, Q) = -\frac{1}{N}\sum_{x\in\mathcal{D}} \log Q(x).$$

Remark 0.2.55. *Finally, we note that cross-entropy is related to maximum likelihood estimation. Let $Q_\theta(x)$ be a density function where θ is a parameter to be determined, and the observation data $\mathcal{D}$ is sampled from a distribution P, that is, $\mathcal{D} \sim P$. The likelihood function is*

$$\prod_{i=1}^{N} Q_\theta(x_i) = \exp\left(-N\mathcal{C}(\mathcal{D}, Q_\theta)\right).$$

Therefore, finding the maximum likelihood estimator $\widehat{\theta}_{MLE}$ is equivalent to minimizing the cross-entropy.

0.2.3 Basic Dynamical Systems

Theorem 0.2.56 (Grönwall's Inequality). *Let y be a continuous, differentiable, non-negative function on $[a,b]$ and*

$$y'(t) \le \alpha(t)y(t), \quad a < t < b.$$

Then

$$y(t) \le y(a)\exp\left(\int_a^t \alpha(s)ds\right).$$

Remark 0.2.57. *There are discrete versions and stochastic versions of Grönwall's inequality.*

0.2.4 Basic Functional Analysis

Theorem 0.2.58 (Borsuk-Ulam Theorem). *If $g : \mathbb{S}^d \to \mathbb{R}^d$ is a continuous odd function, then there exists an $x \in \mathbb{S}^d$ that $g(x) = \mathbf{0} \in \mathbb{R}^d$.*

Theorem 0.2.59 (Baire Category Theorem). *Every completely metrizable topological space is a* ***Baire space****, that is, every countable intersection of open dense sets is dense in it.*

Remark 0.2.60. *The Baire Category Theorem uses the* ***Axiom of Choice*** *(AC).*

0.3 Concepts

The following concepts are useful.

0.3.1 Perceptron

A perceptron is an algorithm/function for binary classification. In a nutshell, the given data points x_k, a perceptron is described by a Heaviside (or sign) function that

$$f(x) = \text{sgn}(w \cdot x + b).$$

It identifies the hyperplane $w \cdot x + b = 0$ that separates the data points. In general, if the data points are linearly separable, the usual optimization algorithms (e.g., gradient descent) can identify a separating hyperplane rather quickly.

The multilayer perceptron (MLP) is based on the perceptron concept. MLP is also known as the *fully connected feedforward neural network*. A general MLP with $L+1$ layers means it has L hidden layers (hidden means there are trainable variables).

$$h_j^{(l)} = \sigma(z_j^{(l)}), \quad z_j^{(l)} = \sum_{i=1}^{k_{l-1}} w_{i,j}^{(l)} h_i^{(l-1)} + b_j^{(l)}, \quad l = 1, \cdots, L.$$

In the vector form, we write

$$\mathbf{h}^{(l)} = \sigma\left(\mathbf{w}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right).$$

The weight matrix $\mathbf{w}^{(l)} \in \mathbb{R}^{k_l \times k_{l-1}}$ and $\mathbf{b}^{(l)} \in \mathbb{R}^{k_l}$, k_l is the dimension of the l-th layer. In particular, k_0 is the dimension of the input space and k_L is the dimension of the output space. The function σ is called the activation function. Common choices are ReLU (also known as the hinge function), sigmoidal, etc.

0.3.2 Regression & Classification

Regression and classification are core supervised learning tasks that handle different types of labels (continuous and discrete).

For continuous or numeric labels, the loss functions often measure the mean mismatch, e.g., MAE (absolute), MSE (squared)

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} |y - f(x,\theta)|,$$
$$\mathcal{L}_{\text{MSE}} = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} |y - f(x,\theta)|^2,$$

which are L^1 error or L^2 error squared with respect to the empirical probabilistic measure.

For discrete labels, the model $f(x,\theta)$ often contains a final softmax layer:

$$\text{Softmax}(z) = \frac{1}{M}(e^{z_1/T}, \cdots, e^{z_n/T}), \quad M = \sum_i e^{z_i/T}.$$

to convert the numeric values into a probability distribution over the categories. T denotes the temperature relative to the Boltzmann/Gibbs distribution; the default value is $T = 1$. The loss functions are then defined as the mean/sum of the mismatch between the distributions (e.g., TV and KL divergence). In this case, the data labels are converted into one-hot vectors representing the ground-truth probabilities.

$$\mathcal{L}_{\text{TV}} = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} |y - f(x,\theta)|,$$
$$\mathcal{L}_{\text{KL}} = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} y \cdot \log \frac{y}{f(x,\theta)}.$$

The KL divergence (or relative entropy) is non-symmetric and sensitive to small probabilities. Regarding the implementation, it is often directly using the cross-entropy formulation. Intuitively, the KL divergence quantifies the information gain (difference) between the modeled distribution and the ground-truth distribution, which is more physically meaningful in many tasks.

Problem 0.3.1. *Why is the MSE-type loss NOT used for classification?*

0.3.3 Backpropagation

The training of a neural network $f(x, \theta)$ is basically optimizing the loss function $\mathcal{L}(\theta)$ through gradient descent or its variants, which depend on the gradient information $\partial_\theta \mathcal{L}$.

Taking the MLP structure as an example. We write the neural network as a composition $f(x, \theta) = \mathbf{h}^{(L)}$ and

$$\begin{aligned}\mathbf{h}^{(l)} &= \sigma(\mathbf{w}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}), \quad l = 1, \cdots, L\\ \mathbf{h}^{(0)} &= x.\end{aligned}$$

Each iteration $\mathbf{h}^{(l)}$ is produced by a single-layer perception model with weight variables $\mathbf{w}^{(l)}$ and bias variables $\mathbf{b}^{(l)}$.

For a scalar variable $\theta^{(l)}$ in the l-th layer, the corresponding gradient component is computed as

$$\begin{aligned}\frac{\partial \mathbf{h}^{(k)}}{\partial \theta^{(l)}} &= \sigma'(\mathbf{w}^{(k)}\mathbf{h}^{(k-1)} + \mathbf{b}^{(k)})\mathbf{w}^{(k)}\frac{\partial \mathbf{h}^{(k-1)}}{\partial \theta^{(l)}}, && k > l\\ \frac{\partial \mathbf{h}^{(l)}}{\partial \theta^{(l)}} &= \sigma'(\mathbf{w}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)})\left(\frac{\partial \mathbf{w}^{(l)}}{\partial \theta^{(l)}}\mathbf{h}^{(l-1)} + \frac{\partial \mathbf{b}^{(l)}}{\partial \theta^{(l)}}\right)\\ \frac{\partial \mathbf{h}^{(k)}}{\partial \theta^{(l)}} &= 0, && k < l.\end{aligned}$$

It turns out the gradient $\frac{\partial \mathbf{h}^{(L)}}{\partial \theta^{(l)}}$ can be computed iteratively from the previous layers $k \geq l$.

One potential issue with a deep MLP model is that the gradient with respect to the innermost layers will be multiplied by σ' many times. Activation functions like logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$ have a derivative strictly less than one (why?) will make the gradient $\frac{\partial \mathbf{h}^{(L)}}{\partial \theta^{(l)}}$ vanishing when $L \gg l$. Similar issues arise with ReLU activation, since a portion of the derivative is zero.

0.3.4 Regularization

Regularization is a technique for controlling training. One popular choice is to add a penalty term to the loss function.

$$\widetilde{\mathcal{L}}(\theta) = \mathcal{L}(\theta) + \Omega(\theta).$$

The regularization term $\Omega(\theta)$ can be viewed as a certain prior knowledge about the parameter θ from the Bayesian viewpoint. That is, if we view the data points (x_k, y_k) are independently generated from the following model with independent

noise,

$$Y = f(X, \theta) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \eta^2).$$

Then,

$$\mathbb{P}[Y \mid X, \theta] \propto \exp\left(-\frac{|f(X, \theta) - Y|^2}{\eta^2}\right).$$

The maximum likelihood estimator (MLE) becomes

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta} \prod_{k=1}^{n} \exp\left(-\frac{|f(x_k, \theta) - y_k|^2}{\eta^2}\right),$$

which corresponds to the MSE loss function. We also find out

$$\mathbb{P}[\theta \mid Y, X] = \frac{\mathbb{P}[Y \mid X, \theta] \cdot \mathbb{P}[\theta]}{\mathbb{P}[Y \mid X]} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

Therefore, if we know the prior of θ, say, obeys Gaussian distribution $\mathcal{N}(0, \tau^2)$, then, the maximum *a posteriori* estimate (MAP) becomes

$$\widehat{\theta}_{\text{MAP}} = \arg\max_{\theta} \prod_{k=1}^{n} \exp\left(-\frac{|f(x_k, \theta) - y_k|^2}{\eta^2}\right) \cdot \exp\left(-\frac{|\theta|^2}{\tau^2}\right),$$

which corresponds to the Tikhonov regularization term $\Omega(\theta) = \frac{\eta^2}{\tau^2}|\theta|^2$. There are a few other empirical regularization techniques for neural networks, such as early stopping (to prevent overfitting) and dropout.

Extended Reading

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

Naaman, M. (2021). On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality. *Statistics & Probability Letters*, 173:109088.

Petrov, V. V. (2012). *Sums of independent random variables*, volume 82. Springer Science & Business Media.