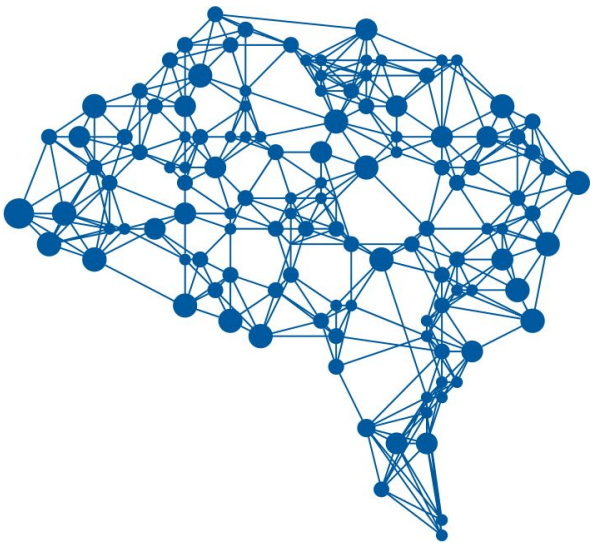


Deep Learning

Mingming Li

First Created: September 29, 2020
Last Modified: March 6, 2023

Contents

List of Figures

List of Tables

Part I

Theory

Chapter 1

Probability

1.1 Markov chain

A **Markov chain** or **Markov process** is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.

1.1.1 Transitions

The changes of state of the system are called **transitions**. The probabilities associated with various state changes are called **transition probabilities**. The process is characterized by a state space, a transition matrix describing the probabilities of particular transitions, and an initial state (or initial distribution) across the state space.

1.1.2 Discrete-time Markov chain

A **discrete-time Markov chain** is a sequence of random variables $X_1, X_2, X_3, \ldots$ with the Markov property, namely that the probability of moving to the next state depends only on the present state and not on the previous states:

$$\mathrm{P}(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathrm{P}(X_{n+1} = x \mid X_n = x_n) \tag{1.1}$$

if both conditional probabilities are well defined, that is, if $\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) > 0$.

The possible values of X_i form a countable set S called the state space of the chain.

1.1.3 Continous-time Markov chain

A **continuous-time Markov chain** X(t) is defined by two components: a jump chain, and a set of holding time parameters λ_i. The jump chain consists of a countable set of states $\mathrm{S} \subset \{0, 1, 2, \cdots\}$ along with transition probabilities $\mathrm{p_{ij}}$. We assume $\mathrm{p_{ii}} = 0$, for all non-absorbing states $\mathrm{i} \in \mathrm{S}$. We assume

1. if $X(t) = i$, the time until the state changes has Exponential (λ_i) distribution;
2. if $X(t) = i$, the next state will be j with probability p_{ij}.

The process satisfies the Markov property. That is, for all $0 \leq t_1 < t_2 < \cdots < t_n < t_{n+1}$, we have

$$\begin{aligned} P\left(X\left(t_{n+1}\right) = j \mid X\left(t_n\right) = i, \mathbf{X}\left(t_{n-1}\right) = i_{n-1}, \cdots, X\left(t_1\right) = i_1\right) \\ = P\left(X\left(t_{n+1}\right) = j \mid \mathbf{X}\left(t_n\right) = i\right) \end{aligned}$$

1.1.4 Markov kernel

Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be **measurable spaces**. A **Markov kernel** with source $(X, \mathcal{A})$ and target $(Y, \mathcal{B})$ is a map $\kappa : \mathcal{B} \times X \to [0, 1]$ with the following properties:

1. For every (fixed) $B \in \mathcal{B}$, the map $x \mapsto \kappa(B, x)$ is $\mathcal{A}$-measurable
2. For every (fixed) $x \in X$, the map $B \mapsto \kappa(B, x)$ is a **probability measure** on $(Y, \mathcal{B})$

In other words it associates to each point $x \in X$ a probability measure $\kappa(dy \mid x) : B \mapsto \kappa(B, x)$ on $(Y, \mathcal{B})$ such that, for every measurable set $B \in \mathcal{B}$, the map $x \mapsto \kappa(B, x)$ is measurable with respect to the σ-algebra $\mathcal{A}$.

1.1.5 Measurable space

A **measurable space** consists of a set and a **σ-algebra**, which defines the subsets that will be measured.

Consider a set X and a σ-algebra $\mathcal{A}$ on X. Then the tuple $(X, \mathcal{A})$ is called a measurable space.

1.1.6 σ-algebra

A **σ-algebra** (also σ-field) on a set X is a nonempty collection Σ of subsets of X closed under complement, countable unions, and countable intersections.

Let X be some set, and let $P(X)$ represent its power set. Then a subset $\Sigma \subseteq P(X)$ is called a σ-algebra if it satisfies the following three properties:

1. X is in Σ, and X is considered to be the universal set in the following context.
2. Σ is closed under complementation: If A is in Σ, then so is its complement, $X \backslash A$.
3. Σ is closed under countable unions: If $A_1, A_2, A_3, \ldots$ are in Σ, then so is $A = A_1 \cup A_2 \cup A_3 \cup \cdots$.

From these properties, it follows that the σ-algebra is also closed under countable intersections (by applying De Morgan's laws).

1.2 Conditional probability

Given two events A and B.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{1.2}$$

$P(A|B)$ stands for the probability of A happening given that B happened. $P(A \cap B)$ stands for the probability of A and B happening at the same time. $P(B)$ stands for the probability of B happening.

1.3 Posterior probability

The **posterior probability** is a type of conditional probability that results from updating the prior probability with information summarized by the likelihood via an application of Bayes' rule.

In variational Bayesian methods, the posterior probability is the probability of the parameters θ given the evidence X, and is denoted $p(\theta \mid X)$. It contrasts with the likelihood function, which is the probability of the evidence given the parameters: $p(X \mid \theta)$.

The two are related as follows: Given a prior belief that a probability distribution function is $p(\theta)$ and that the observations x have a likelihood $p(x \mid \theta)$, then the posterior probability is defined as

$$p(\theta \mid x) = \frac{p(x \mid \theta)}{p(x)} p(\theta) \tag{1.3}$$

where $p(x)$ is the normalizing constant and is calculated as

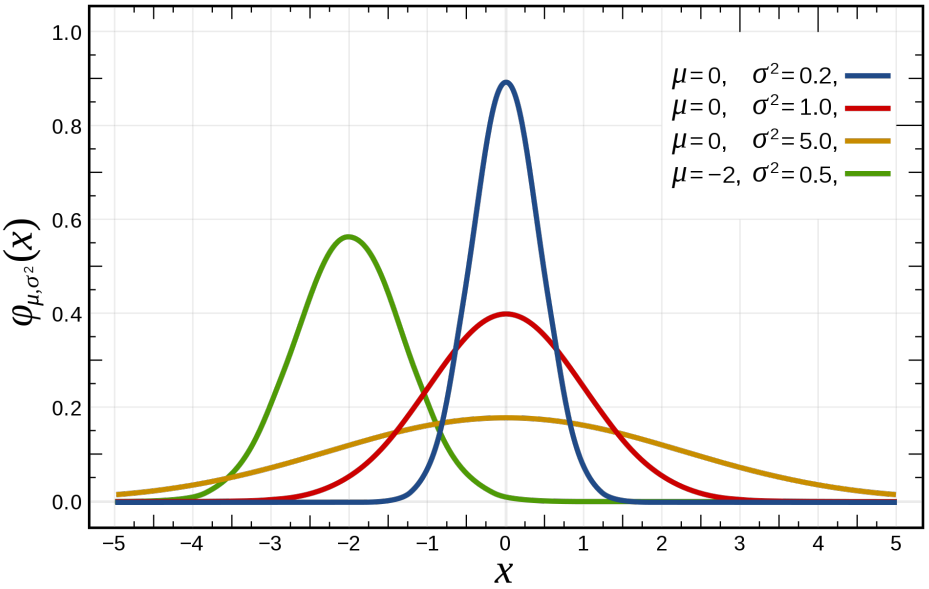
$$p(x) = \int p(x \mid \theta) p(\theta) d\theta \tag{1.4}$$

1.4 Guassian distribution

Gaussian distribution (normal distribution) is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1.5}$$

The parameter μ is the mean or expectation of the distribution (and also its median and mode), while the parameter σ is its standard deviation. The variance of the distribution is σ^2. The Figure 1.1 shows the Gaussian distributions.



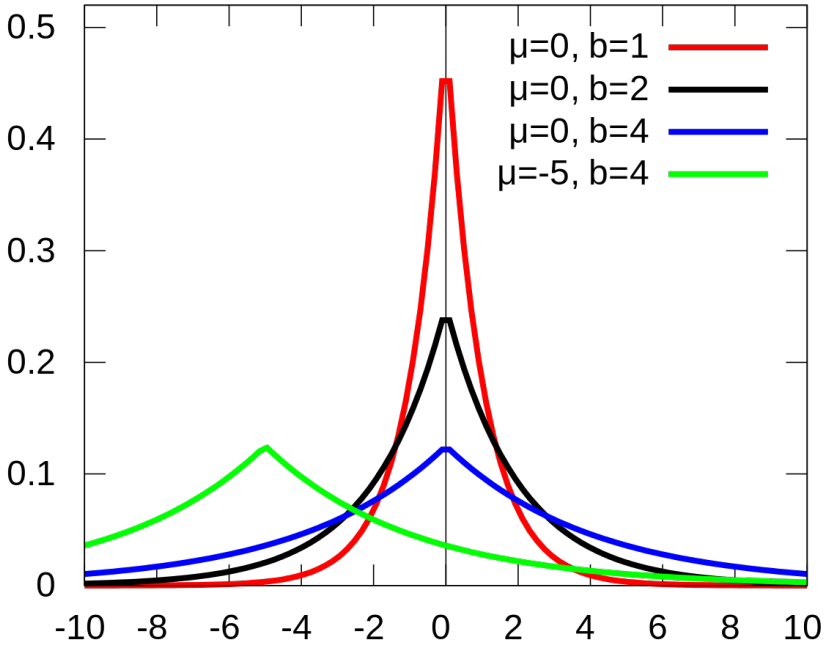
Figures 1.1: Gaussian distribution

1.5 Laplace distribution

A random variable has a Laplace (μ, b) distribution if its probability density function is

$$f(x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) \tag{1.6}$$

Here μ is a location parameter and b is a scale parameter. Figure 1.2 shows the Laplace distributions.



Figures 1.2: Laplace distribution

1.6 Monte Carlo method

Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. The underlying concept is to use randomness to solve problems that might be deterministic in principle.

Monte Carlo methods vary, but tend to follow a particular pattern:

1. Define a domain of possible inputs
2. Generate inputs randomly from a probability distribution over the domain
3. Perform a deterministic computation on the inputs
4. Aggregate the results

1.7 Variational Bayesian methods

It is a simplifying that makes the intractable events into tractable events.

For example, in variational inference, the posterior distribution over a set of unobserved variables $\mathbf{Z}$ =

$\{Z_1 \dots Z_n\}$ given some data $\mathbf{X}$ is approximated by a so-called variational distribution, $Q(\mathbf{Z})$:

$$P(\mathbf{Z} \mid \mathbf{X}) \approx Q(\mathbf{Z}) \tag{1.7}$$

The distribution $Q(\mathbf{Z})$ is restricted to belong to a family of distributions of simpler form than $P(\mathbf{Z} \mid \mathbf{X})$ (e.g. a family of Gaussian distributions), selected with the intention of making $Q(\mathbf{Z})$ similar to the true posterior, $P(\mathbf{Z} \mid \mathbf{X})$.

The similarity (or dissimilarity) is measured in terms of a dissimilarity function $d(Q; P)$ and hence inference is performed by selecting the distribution $Q(\mathbf{Z})$ that minimizes $d(Q; P)$. The most common type of variational Bayes uses the Kullback-Leibler divergence (KL-divergence) of Q from P as the choice of dissimilarity function.

1.8 Bayesian probability

Bayesian probability is an interpretation of the concept of probability, in which, instead of frequency or propensity of some phenomenon, probability is interpreted as reasonable expectation representing a state of knowledge or as quantification of a personal belief.

1.9 Binomial distribution

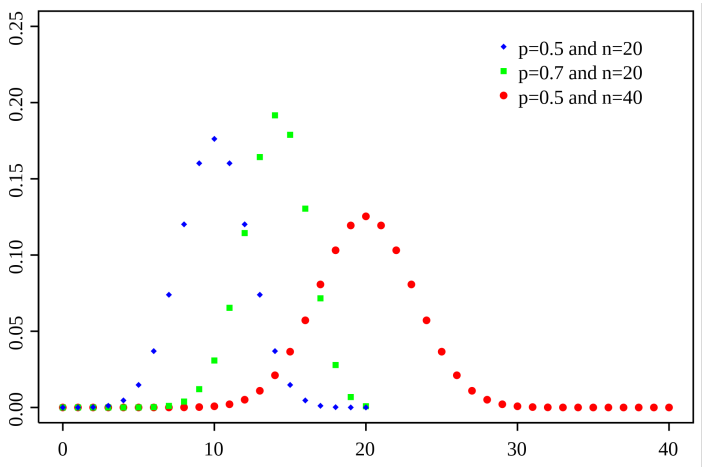
In general, if the random variable X follows the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$, we write $X \sim \mathrm{B}(n, p)$. The probability of getting exactly k successes in n independent Bernoulli trials is given by the probability mass function:

$$f(k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \tag{1.8}$$

for $k = 0, 1, 2, \dots, n$, where

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!} \tag{1.9}$$

The figure 1.3 shows the distribution.



Figures 1.3: Binomial distribution

1.10 Kullback–Leibler divergence

The Kullback-Leibler divergence, denoted $D_{\mathrm{KL}}(P\|Q)$, is a type of statistical distance: a measure of how one probability distribution P is different from a second, reference probability distribution Q.

For discrete probability distributions P and Q defined on the same sample space, $\mathcal{X}$, the relative entropy from Q to P is defined [11] to be

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right). \tag{1.10}$$

which is equivalent to

$$D_{\mathrm{KL}}(P\|Q) = -\sum_{x\in\mathcal{X}} P(x)\log\left(\frac{Q(x)}{P(x)}\right) \tag{1.11}$$

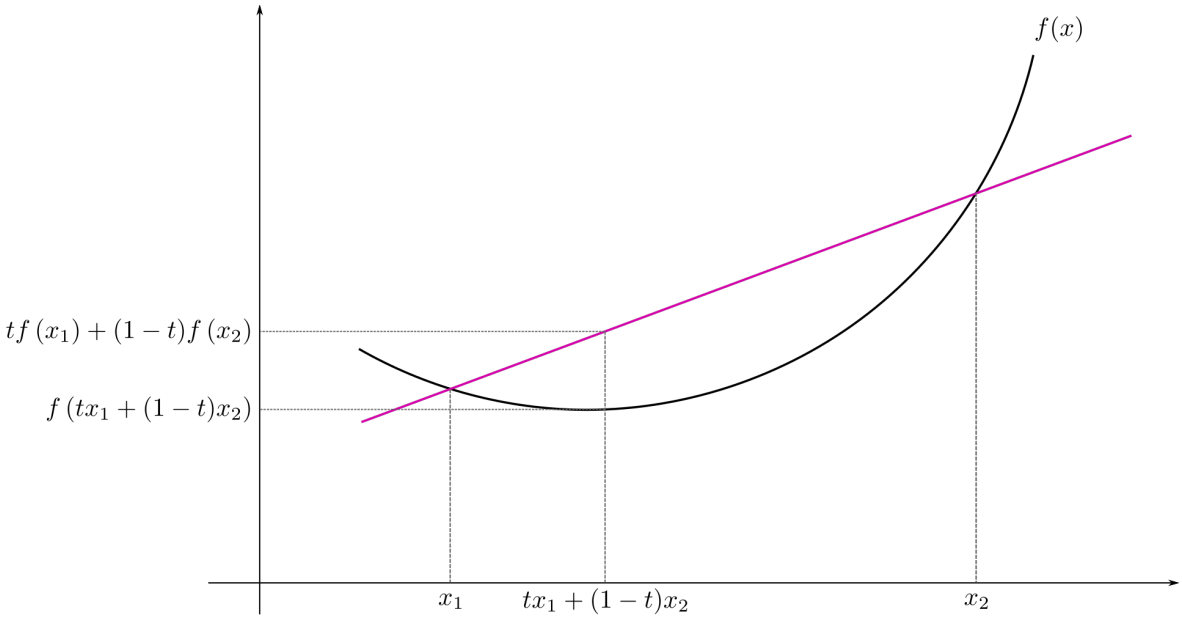
1.11 Jensen's inequality

Jensen's inequality generalizes the statement that a secant line of a convex function lies above its graph.

Thus, Jensen's inequality is

$$f\left(tx_1 + (1-t)x_2\right) \leq tf\left(x_1\right) + (1-t)f\left(x_2\right). \tag{1.12}$$

Shown in Figure 1.4.



Figures 1.4: Jensen's inequality

In the context of probability theory, it is generally stated in the following form: if X is a random variable and φ is a convex function, then

$$\varphi(\mathrm{E}[X]) \leq \mathrm{E}[\varphi(X)] \tag{1.13}$$

The difference between the two sides of the inequality, $\mathrm{E}[\varphi(X)] - \varphi(\mathrm{E}[X])$, is called the Jensen gap.

Chapter 2

Learning

A machine learning algorithm is an algorithm that is able to learn from data. But what do we mean by learning? Mitchell provides a succinct definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"

Now, how does the learn happen? The model or algorithm learns by adjusting the parameters contained in it.

How to adjust the parameters? Refer to Chapter 4.

Chapter 3

Model

What is a model?

You can think a model as a function in math. For example,

$$ax + by = c$$

In this function, a, b and d are the parameters, x and y are input from data. Through adjusting the parameters a, b and d, we can implement the learning process. This is also called training in machine learning.

3.1 Capacity

The capacity is the function space. It defines the limitation that can learn from data.

For example,

$$ax + by = c$$

This function can only learn linear function from data.

$$ax_1^2 + bx_2 = c$$

This function can learn curve function from data.

3.1.1 The no free lunch theorem

For any algorithms (functions) a_1 and a_2, at iteration step m

$$\sum P(d_m^y | f, m, a_1) = \sum P(d_m^y | f, m, a_2) \tag{3.1}$$

where d_m^y denotes the ordered set of size m of the cost values y associated to input values $x \in X$, $f : X \longrightarrow Y$ is the function being optimized and $P(d_m^y | f, m, a)$ is the conditional probability of obtaining a given sequence of cost values from algorithm a run m times on function f.

The no free lunch theorem implies that we must design our machine learning algorithms to perform well on a specific task but not a universal task.

3.1.2 Universal approximation theorem

A feed-forward network with a single hidden layer containing a finite number of neurons can approximate continuous functions on compact subsets of R^n , under mild assumptions on the activation function.

Let $\varphi : \mathrm{R} \to \mathrm{R}$ be a nonconstant, bounded, and continous functions. Let I_m denote the m-dimensional unit hypercube $[0, 1]^m$. The space of real-value continous function on I_m is denoted by $C(I_m)$. Then, given any $\varepsilon > 0$ and function $f \in C(I_m)$, there exist an integer N, real constants $\upsilon_i, b_i \in R$ and real vectors $\omega_i \in R^m$ for $i = 1, \cdots, N$, such that we may define:

$$F(x) = \sum_{i=1}^{N} \upsilon_i \varphi(\omega_i^T x + b_i) \tag{3.2}$$

as an approximate realization of the function f; that is

$$|F(x) - f(x)| < \varepsilon \tag{3.3}$$

for all $x \in I_m$. In other words, functions of the form $F(x)$ are dense in $C(I_m)$.

This still holds when replacing I_m with any compact subset of R^m .

Kurt Hornik showed in 1991 that it is not the specific choice of the activation function, but rather the multi-layer feedforward architecture itself which gives neural networks the potential of being universal approximators. The output units are always assumed to be linear. For notational convenience, only the single output case will be shown. The general case can easily be deduced from the single output case.

In 2017 Lu et al. proved universal approximation theorem for width-bounded deep neural networks.

This is the base of deep learning.

3.2 Overfitting and underfitting

We train model on training data but use test data (not used to train the model) to test out model. The ability to perform on test data is called **generalization**. We can use model on test data because we assume that the train data and the test has the same probability distribution (i.e. they have relationship).

The error on training data is called **training error**. The error on test data is called **test error**. Underfitting occurs when the model is not able to obtain a sufficient low error value on the training set. Overfitting occurs when the gap between the training error and test error is too large.

We can control whether a model is more likely to overfit or underfit by altering its **capacity**. The overfitting and underfitting happen because of the mismatch of capacity and the data.

Chapter 4

Training (optimization)

Training is the process of learning. By inputting the data, we adjust the parameters in the model to achieve better performance. In machine learning this is also called the optimization.

4.1 SGD

Stochastic gradient descent (SGD) is a very basic and important algorithm.

The cost function used by a machining learning algorithm often decomposes as a sum over training examples of some per-example loss function.

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathrm{x},y\sim\hat{p}_{data}} L(\boldsymbol{x}, y, \boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}), \tag{4.1}$$

Where L is the loss function, x and y are the input data and labels, θ is the parameters in the model, m is the number of samples.

The gradient is

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}). \tag{4.2}$$

The computational cost of this operation is $O(m)$. As the training set size grows to billions of examples, the time to take a single gradient step becomes prohibitively long.

The insight of SGD is that the gradient is an expectation. The expectation may be approximately estimated using a small set of samples. Specifically, on each step of the algorithm, we can sample a **minibatch** of examples $\mathbb{B} = \{\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(m)}\}$ drawn uniformly from the training set. The minibatch size m' is typically chosen to be a relatively small number of examples, ranging from one to a few hundred.

The estimator of the gradient is formed as

$$\boldsymbol{g} = \frac{1}{m'}\nabla_{\boldsymbol{\theta}}\sum_{i=1}^{m'} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}) \tag{4.3}$$

using examples from the minibatch $\mathbb{B}$. The stocastic gradient descent algorithm then follows the estimated gradient downhill:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \boldsymbol{g}, \tag{4.4}$$

where ϵ is the learning rate.

4.2 Momentum

The method of momentum is designed to accelerate learning in SGD. The momentum algorithm accumulates an exponentially decaying moving average of past gradients and continues to move in their direction.

$$\boldsymbol{v} \leftarrow \alpha \boldsymbol{v} - \epsilon \boldsymbol{g} \tag{4.5}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{v}. \tag{4.6}$$

Where g is the gradient, $\alpha \in [0, 1]$ determines how quickly the contributions of previous gradients exponentially decay.

4.3 Nesterov momentum

Nesterov Momentum is a variant of the Momentum algorithm. The difference between Nesterov momentum and standard momentum is where the gradient is evaluated. With Nesterov momentum the gradient is evaluated after the current velocity is applied.

$$\boldsymbol{g} = \frac{1}{m'} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m'} L(\boldsymbol{x}, y, \boldsymbol{\theta} + \alpha \boldsymbol{v}) \tag{4.7}$$

4.4 AdaGrad

AdaGrad is designed to converge rapidly when applied to a convex function. Comparing to SGD, AdaGrad algorithm individually adapts the learning rates of all model parameters by scaling them inversely proportional to the square root of the sum of all of their historical squared values.

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \frac{\epsilon}{\sqrt{\delta \boldsymbol{I} + \operatorname{diag}(G)}} \odot \boldsymbol{g} \tag{4.8}$$

where $\boldsymbol{\theta}$ is the parameter to be updated, ϵ is the initial learning rate, δ is some small quantity that used to avoid the division of zero, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{g}$ is the gradient estimate.

$$\boldsymbol{G} = \sum_{\tau=1}^{t} \boldsymbol{g}_\tau \boldsymbol{g}_\tau^T \tag{4.9}$$

AdaGrad shrinks the learning rate according to the entire history of the squared gradient and may have made the learning rate too small before arriving at such a convex structure.

4.5 Adam

The RMSProp algorithm modifies AdaGrad to perform better in the non-convex setting by changing the gradient accumulation into an exponentially weighted moving average. RMSProp uses an exponentially decaying average to discard history from the extreme past so that it can converge rapidly after finding a convex bowl.

$$\boldsymbol{r} \leftarrow \boldsymbol{r} + \boldsymbol{g} \odot \boldsymbol{g} \tag{4.10}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \frac{\epsilon}{\delta + \sqrt{\boldsymbol{r}}} \odot \boldsymbol{g} \tag{4.11}$$

Where $\boldsymbol{g}$ is the gradient, $\boldsymbol{\theta}$ is the parameters in a model, $\boldsymbol{r}$ is initialized to 0..

Chapter 5

Cost function

Loss is the difference between the predicted value and the actual value. The Function used to quantify this loss during the training phase in the form of a single real number is known as **Loss Function**. **Cost function** refers to an average of the loss functions over an entire training dataset. Cost function helps us reach the optimal solution. The cost function is the technique of evaluating "the performance of our algorithm/model".

There are many cost functions in machine learning and each has its use cases depending on whether it is a regression problem or classification problem.

5.1 Regression cost function

Regression models deal with predicting a continuous value. They are calculated on the distance-based error.

$$\text{Error} = y - y' \tag{5.1}$$

5.1.1 Mean squared error

$$\text{MSE} = \frac{\sum_{i=0}^{n}(y - y')^2}{n} \tag{5.2}$$

5.1.2 Mean absolute error

$$\text{MSE} = \frac{\sum_{i=0}^{n}|y - y'|}{n} \tag{5.3}$$

5.2 Classification cost function

In classification, we usually use one hot to encoder the labels. This can eliminate the affect of the distance.

The cross entropy of the distribution q relative to distribution p over a given set is defined as

$$H(p, q) = -E_p[\log q] \tag{5.4}$$

Where the E is the expected value operator respected to the probability q.

For discrete probability distribution p and q with the same support $\mathcal{X}$ this means

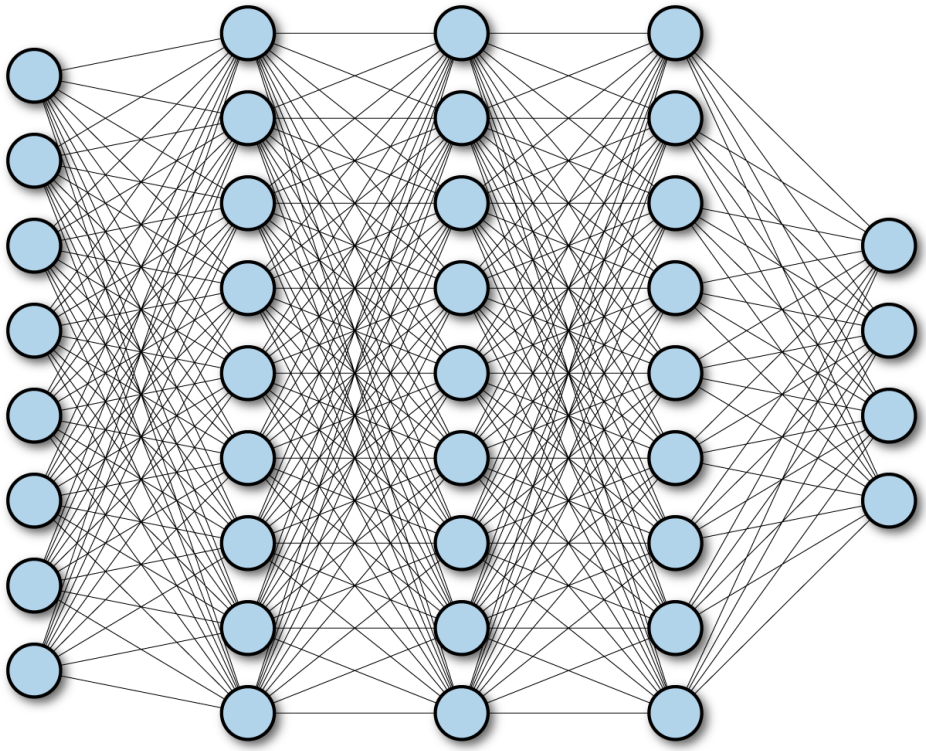
$$H(P, Q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x) \tag{5.5}$$

Chapter 6

Full Connected Networks

A fully connected neural network consists of a series of fully connected layers that connect every neuron in one layer to every neuron in the other layer. Full connected neural network layers use matrix multiplication by a matrix of parameters with a separate parameter describing the interaction between each input unit and each output unit. This means that every output unit interacts with every input unit.

Figure 6.1 show the full connects neural network.



Figures 6.1: Full connected neural network

Chapter 7

CNN

CNN stands for convolutional neural network. Convolutional networks are neural networks that have convolutional layers. A typical convolutional layer consists of three stages:

1. convolution stage: affine transform
2. detector stage: nonlinearty
3. pooling stage

7.1 Convolution

$$s(t) = \int x(a)w(t-a)da. \tag{7.1}$$

This operation is called **convolution**. The convolution operation is typically denoted with an asterisk:

$$s(t) = (x * w)(t). \tag{7.2}$$

In convolutional network terminology, the first argument (in this example, the function x) to the convolution is often referred to as the **input**, and the second argument (int this example, the function w) as the **kernel**. The output is sometimes referred to as the **feature map**.

If we assume that x and w are defined only on integer t, we can define the discrete convolution:

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a). \tag{7.3}$$

We often use convolutions over more than one axis at a time. For example, if we use a two-dimensinal image I as our input, we probably also want to use a two-dimensional kernel K:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n). \tag{7.4}$$

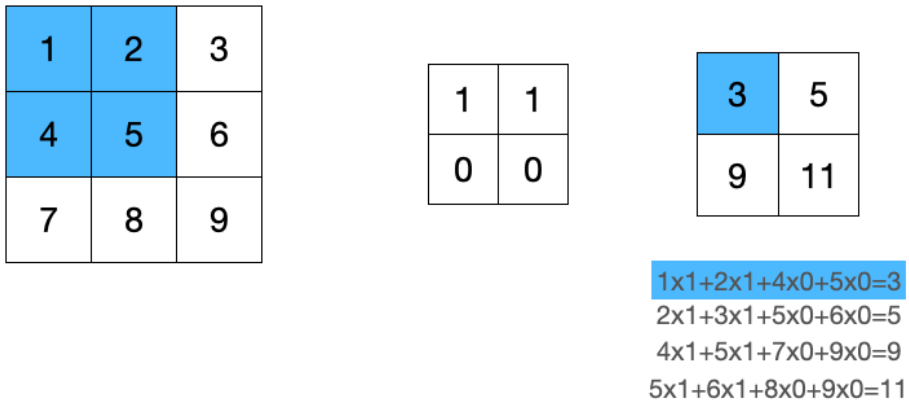
The following formula can be used to calculate the output dimension.

$$h_o = \frac{h_i - h_k}{h_s} + 1 \tag{7.5}$$

$$w_o = \frac{w_i - w_k}{w_s} + 1 \tag{7.6}$$

where h_o is the output height, h_i is the input height, h_k is the kernel height, h_s is the stride height, w_o is the output width, w_i is the input width, w_k is the kernel width, w_s is the stride width.

The convolution operation is shown in Figure 7.1.



Figures 7.1: Convoluation operation

7.2 Properties

CNN leverages three important ideas:

- ♥ sparse interaction.
- ♥ parameter sharing.
- ♥ equivariant representations.

7.2.1 Sparse interaction

This is accomplished by making the kernel smaller than the input.

7.2.2 Parameter sharing

In convolutional layers, the same parameter defined in one kernel are used at every position of the input.

7.2.3 Equivariant representations

In the case of convolution, the particular form of a parameter sharing causes the layer to have a property called **equivariance** to translation. To say a function is equivariant means that if the input changes, the output changes in the same way.

7.3 Pooling

A pooling function replaces the output of the net at a certain location with a summary statistic of the nearby outputs. For example, the max pooling oeration reports the maximum output within a rectangular neighborhood. Pooling helps to make the representation approximately invariant to small translations of the input. Invariant to translation means that if we translate the input by a small amount, the values of most of the pooled outputs do not change.

The following formula can be used to calculate the output dimension.

$$h_o = \frac{h_i - h_k}{h_s} + 1 \tag{7.7}$$

$$w_o = \frac{w_i - w_k}{w_s} + 1 \tag{7.8}$$

where h_o is the output height, h_i is the input height, h_k is the pooling height, h_s is the stride height, w_o is the output width, w_i is the input width, w_k is the pooling width, w_s is the stride width.

Chapter 8

Metric

		Real value	
		positive	negative
Predict value	positive	true positive	false positive
	negative	false negative	true negative

Tables 8.1: Confusion matrix

8.1 Precision

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8.1}$$

Of all the predicted positive values, the ratio of the true positive values (the real value is positive and the predicted value is positive).

8.2 Recall

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8.2}$$

Of all the real positive values, the ratio of the true positive values.

8.3 Accuracy

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8.3}$$

Of all the values, the ratio of correctly predicted values. The disadvantage is that it is not suitable for unbalanced data.

8.4 F-score

$$\mathrm{F} = \frac{(\alpha^2 + 1) \times \text{precision} \times \text{recall}}{\alpha^2 \times \text{precision} + \text{recall}} \tag{8.4}$$

When determining the value of the parameter α, if we pay more attention to recall (compared to precision), we should choose a larger α. The F-1 score is the expression when $\alpha = 1$

Chapter 9

Regularization

We train model on training data but use test data (not used to train the model) to test out model. The ability to perform on test data is called **generalization**. We can use model on test data because we assume that the train data and the test has the same probability distribution (i.e. they have relationship).

Regularization is any modification we make to a learnining algorithm that is intended to reduce its generalization error.

In practice, it is very difficult to find a model with the right number of parameters. Indeed, we often use a larger model that has been regularized appropriately.

9.1 Parameter norm penalties

Parameter norm penalties ($\Omega(\boldsymbol{\theta})$) can be added to the object function J to limit the capacity of the model.

$$\tilde{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = J(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) + \alpha\Omega(\boldsymbol{\theta}) \tag{9.1}$$

9.1.1 L^2 Parameter Regularization

The L^2 parameter norm penalty commonly known as **weight decay**.

$$\Omega(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 \tag{9.2}$$

Where $\boldsymbol{w}$ is the model parameter matrix.

9.1.2 L^1 Regularization

L^1 regularization on the model parameter $\boldsymbol{w}$ is defined as

$$\Omega(\boldsymbol{\theta}) = \|\boldsymbol{w}\|_1 \tag{9.3}$$

9.2 Dataset Augmentation

The best way to make a machine learning model generalize better is to train it on more data. Of course, in practice, the amount of data we have is limited. One way to get around this problem is to create fake data and add it to the training set. Dataset augmentation has been a particular effective technique for a specific classification problem: object recognition.

9.3 Early stopping

Early stopping is used to avoid overfit.

The algorithm terminates when no parameters have improved over the best recorded validation error for some pre-specified number of iterations. This strategy is known as **early stopping**. It is probably the most commonly used form of regularization in deep learning.

9.4 Droupout

Chapter 10

Activation functions

Activation functions are usually added into neural network in order to help the network learn complex patterns. Because many of the patterns we want to learn are non-linear, we usually add non-linear activation functions into neural network to add the ability to learn non-learn patterns. Its function is to determine what information can be passed to the next neuron.

10.1 Desirable features

- ♥ Differentiable. This is needed for optimization.
- ♥ Low computational expense.
- ♥ Zero-centered.

10.2 Sigmoid

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{10.1}$$

It is no longer used in recent deep learning models because it is computationally expensive, causes vanishing gradient problem and not zero-centered.

10.3 Tanh

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{10.2}$$

Comparing to sigmoid, it solve the not zero-centered problem.

10.4 ReLU

$$\text{relu}(x) = \max(0, x) \tag{10.3}$$

This is a widely used activation function.

10.5 Leaky ReLU

$$\text{lrelu}(x) = \max(\alpha x, x) \tag{10.4}$$

10.6 Swish

$$\text{swish}(x) = x * \text{sigmoid}(x) = x * (1 + e^{-x})^{-1} \tag{10.5}$$

Chapter 11

Machine learning process

1. confirm your target
2. observe the data
3. preprocess the data it necessary
4. select the model
5. select the cost function
6. select the optimizer
7. train the model
8. fine-tune the model
9. predict the data using the model
10. deploy the model

Part II

Models

Chapter 12

Diffusion

Diffusion models are generative models, meaning that they are used to generate data similar to the data on which they are trained. Fundamentally, diffusion models work by destroying training data through the successive addition of Gaussian noise, and then learning to recover the data by reversing this noising process. After training, we can use the diffusion model to generate data by simply passing randomly sampled noise through the learned denoising process.

Part III

Tools

Chapter 13

PyTorch

Chapter 14

NumPy

14.1 Load data

```
1 import numpy as np
2
3 data = np.load('foo.npz')
```

Part IV

Practice

Chapter 15

Preprocessing

Preprocessing is not necessary. We preprocess the data because we can not get what we want from the data directly.

15.1 Not valued based data

The computer can only process number-based data. For not number-based data, we need to convert them into number-based data. For example, convert a string into a number list. For null values, we can drop them or compute them with other not null values.

```
from sklearn.preprocessing import LabelBinarizer

data = ['dog', 'cat', 'dog', 'horse']
print('data: ', data)
binarizer = LabelBinarizer()
binarizer.fit(data)
print(binarizer.transform(data))
print(binarizer.transform(['dog']))

"""
data:  ['dog', 'cat', 'dog', 'horse']
[[0 1 0]
 [1 0 0]
 [0 1 0]
 [0 0 1]]
[[0 1 0]]
"""
```

Part V

Projects

Part VI

Papers

Chapter 16

Deep Unsupervised Learning using Nonequilibrium Thermodynamics[1]

The algorithm consists of two trajectories: forward (inference) diffusion process and inverse (generative) diffusion process. The forward diffusion process converts complex data distribution into a simple, tractable distribution and the inverse diffusion process learn a finite-time reversal of the forward diffusion process which define a generative model distribution.

16.1 Algorithm

16.1.1 Forward trajectory

$$\pi(\mathbf{y}) = \int d\mathbf{y}' T_\pi(\mathbf{y} \mid \mathbf{y}'; \beta)\, \pi(\mathbf{y}') \tag{16.1}$$

Where $\pi(\mathbf{y})$ is a well behaved (analyti- cally tractable) distribution. $T_\pi(\mathbf{y} \mid \mathbf{y}'; \beta)$ is a Markov diffusion kernel. β is the diffusion rate.

$$q\left(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)}\right) = T_\pi\left(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)}; \beta_t\right) \tag{16.2}$$

The forward trajectory, corresponding to starting at the data distribution and performing T steps of diffusion is

$$q\left(\mathbf{x}^{(0\cdots T)}\right) = q\left(\mathbf{x}^{(0)}\right) \prod_{t=1}^{T} q\left(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)}\right) \tag{16.3}$$

Where $q\left(\mathbf{x}^{(0)}\right)$ is the initial data distribution.

16.1.2 Reverse Trajectory

The generative distribution will be trained to describe the same trajectory, but in reverse,

$$p\left(\mathbf{x}^{(T)}\right) = \pi\left(\mathbf{x}^{(T)}\right) \tag{16.4}$$

$$p\left(\mathbf{x}^{(0\cdots T)}\right) = p\left(\mathbf{x}^{(T)}\right)\prod_{t=1}^{T} p\left(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}\right) \tag{16.5}$$

For both Gaussian and binomial diffusion, for continuous diffusion (limit of small step size β) the reversal of the diffusion process has the identical functional form as the forward process. Since $q\left(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)}\right)$ is a Gaussian (binomial) distribution, and if β_t is small, then $q\left(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}\right)$ will also be a Gaussian (binomial) distribution. The longer the trajectory the smaller the diffusion rate β can be made.

During learning only the mean and covariance for a Gaussian diffusion kernel, or the bit flip probability for a binomial kernel, need be estimated.

16.2 Model probability

The probability the generative model assigns to the data is

$$p\left(\mathbf{x}^{(0)}\right) = \int d\mathbf{x}^{(1\cdots T)} p\left(\mathbf{x}^{(0\cdots T)}\right) \tag{16.6}$$

Naively this integral is intractable – but taking a cue from annealed importance sampling and the Jarzynski equality, we instead evaluate the relative probability of the forward and reverse trajectories, averaged over forward trajectories

$$\begin{aligned} p\left(\mathbf{x}^{(0)}\right) &= \int d\mathbf{x}^{(1\cdots T)} p\left(\mathbf{x}^{(0\cdots T)}\right) \frac{q\left(\mathbf{x}^{(1\cdots T)} \mid \mathbf{x}^{(0)}\right)}{q\left(\mathbf{x}^{(1\cdots T)} \mid \mathbf{x}^{(0)}\right)} \\ &= \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)} \mid \mathbf{x}^{(0)}\right) \frac{p\left(\mathbf{x}^{(0\cdots T)}\right)}{q\left(\mathbf{x}^{(1\cdots T)} \mid \mathbf{x}^{(0)}\right)} \\ &= \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)} \mid \mathbf{x}^{(0)}\right) \cdot p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)}\right)} \end{aligned} \tag{16.7}$$

16.3 Training

Training amounts to maximizing the model log likelihood,

$$\begin{aligned} L &= \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \\ &= \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \cdot \log \left[\begin{array}{c} \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)} \mid \mathbf{x}^{(0)}\right) \cdot \\ p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)}\right)} \end{array} \right], \end{aligned} \tag{16.8}$$

which has a lower bound provided by Jensen's inequality,

$$L \geq \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \cdot \log \left[p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)}\right)} \right]. \tag{16.9}$$

for our diffusion trajectories this reduces to,

$$L \geq K \tag{16.10}$$

$$\begin{aligned} K = -\sum_{t=2}^{T} \int d\mathbf{x}^{(0)} d\mathbf{x}^{(t)} q\left(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}\right) \\ D_{KL}\left(q\left(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}, \mathbf{x}^{(0)}\right) \| p\left(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}\right)\right) \\ + H_q\left(\mathbf{X}^{(T)} \mid \mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(1)} \mid \mathbf{X}^{(0)}\right) - H_p\left(\mathbf{X}^{(T)}\right) \end{aligned} \tag{16.11}$$

where the entropies and KL divergences can be analytically computed. The derivation of this bound parallels the derivation of the log likelihood bound in variational Bayesian methods.

If the forward and reverse trajectories are identical, corresponding to a quasi-static process, then the inequality in Equation 16.10 becomes an equality.

Training consists of finding the reverse Markov transitions which maximize this lower bound on the log likelihood,

$$\hat{p}\left(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}\right) = \underset{p\left(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}\right)}{\operatorname{argmax}} K. \tag{16.12}$$

Bibliography

[1] Jascha Sohl-Dickstein. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arxiv:arXiv:1503.03585*, 2015.