Inhaltsverzeichnis

Abbildungsverzeichnis

List of Equations

Lecture 1: Introduction

Date: 03/05/2020

Lecturer: David Silver *By: Nithish Moudhgalya*

1.1. Overview

•p.4 expert talks are not relevant for the exam

- **scalar**: x, A, α
- column **vector**: $\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = [x_i] \in \mathbb{R}^N$ with element $x_i = [\underline{x}]_i$
- **matrix**: $\mathbf{A} = [a_{ij}]_{1 \le i \le M, 1 \le j \le N} = [a_{ij}] \in \mathbb{R}^{M \times N}$ with element $a_{ij} = [\mathbf{A}]_{ij}$
- 3D, 4D **tensor**: $\mathbf{A} = [a_{ijk}], \mathbf{A} = [a_{ijkl}]$
- **transpose** of vector and matrix: $\underline{x}^T = [x_1, \ldots, x_N], \mathbf{A}^T = [a_{ji}]_{ij}$
- **determinant** of a square matrix $\mathbf{A}$: $|\mathbf{A}|$
- **inverse** of a square matrix $\mathbf{A}$: $\mathbf{A}^{-1}$
- **trace** of a square matrix $\mathbf{A}$: $\mathrm{tr}(\mathbf{A}) = \sum_i a_{ii}$

Abbildung 1.1: Mathematical notations

•Element notation (element of vector): $[x_i = \underline{a} + \underline{b}]_i$

1.2. Chapter 1.1 - What is machine learning

•Signal processing is not a subset of ML or the other way around, they are identical in the task
•Basic difference is in how to design the processing rule
•Regression means to calculate (SP,ML) a continuous-valued output in the real numbers from a signal, can be one-dimensional
•Classification means to calculate (SP,ML) a discrete-valued output from the natural numbers
•p. 1-5 filter in the middle is the view of a pixel and its 8 neighbours

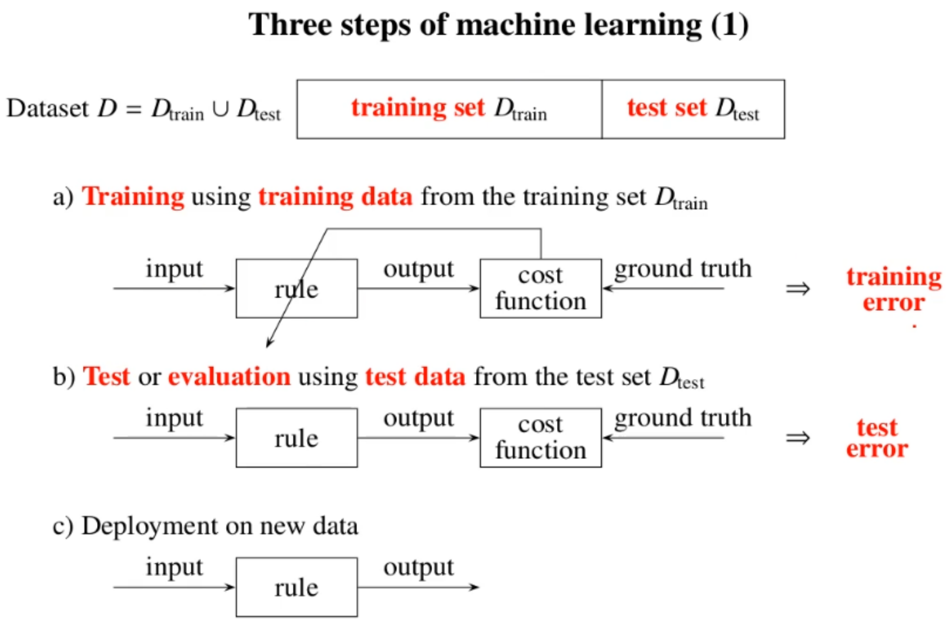


Abbildung 1.2: Three steps of machine learning

•over-fitting just memorizes the training set so test set is needed
•testerrorrate is similar to training error rate $\rightarrow$ no over-fitting
•Test set and training set need to be disjuct concatenated
•TODO: Summary supervised learning and unsupervised learning

1.3. Chapter 1.2 - What is deep learning

•feature extraction: a feature is a clustered subset of information for recognition
•fish example: fish length, color etc.
•Needs human experience, can’t be calculated
•DNN solves the problem without feature extraction

1.3.1. What is a neural network (NN)

•Cascaded pipeline of layers

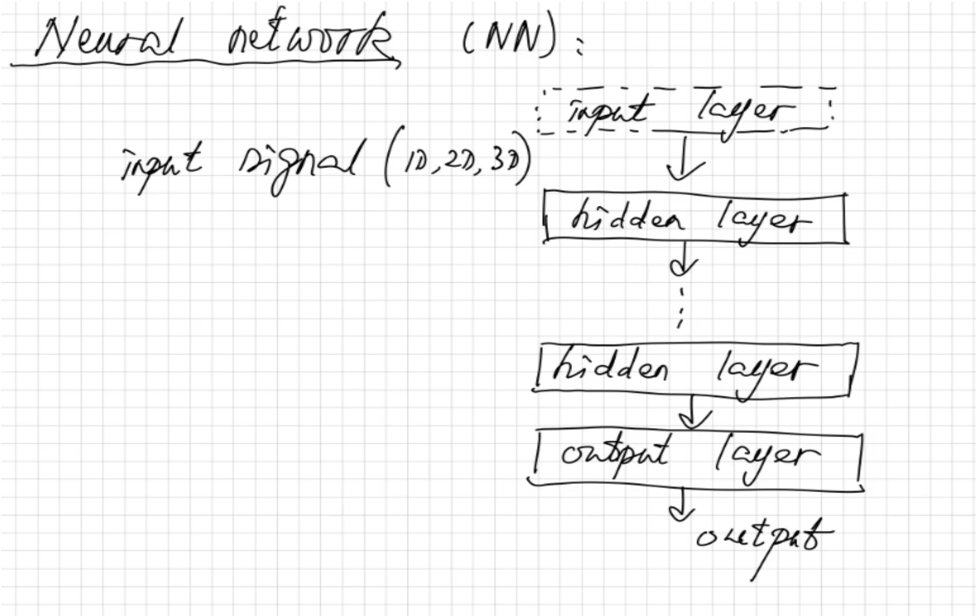


Abbildung 1.3: Neural Network

1.4. Chapter 1.3 - Examples for Deep Learning

•Semantic image segmentation is pixelwise classification

Lecture 2: Tools for Deep Learning

Date: 02/05/2020

Lecturer: David Silver *By: Nithish Moudhgalya*

2.1. Datasets

•Overview for Training Datasets up to ImageNet they are for teaching

Lecture 3: Machine learning basics

Date: 02/05/2020

Lecturer: David Silver *By: Nithish Moudhgalya*

3.1. Linear Algebra

•TODO get Math nicely into the Context book

3.1 Linear algebra 3-4

Vector norms

Given a vector $\underline{x} = [x_i] \in \mathbb{R}^M$. There are different definitions for the vector norm:

- **2-norm** or **l_2-norm** or **Euclidean norm**: $\|\underline{x}\|_2 = \sqrt{\sum_{i=1}^M x_i^2} = \sqrt{\underline{x}^T\underline{x}}$
- **1-norm** or **l_1-norm**: $\|\underline{x}\|_1 = \sum_{i=1}^M |x_i|$
- **0-norm** or **l_0-norm**: $\|\underline{x}\|_0$ = number of non-zero elements in $\underline{x}$

Comments:

- $\|\underline{x}\|_2^2$ represents the energy of $\underline{x}$.
- $\|\underline{x}\|_0$ measures the sparsity of $\underline{x}$.
- Different vector norms have different unit-norm contour lines $\{\underline{x} \mid \|\underline{x}\|_p = 1\}$.

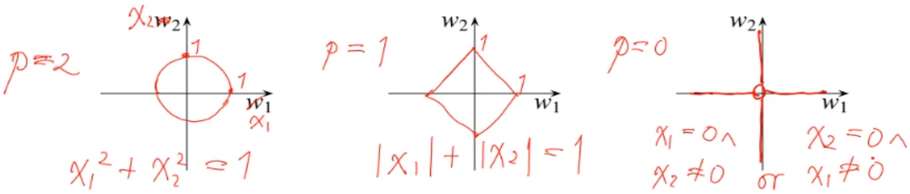


Abbildung 3.1: Vector Normes

3.2. Chapter 3.2 Random variable and probability distribution

3.2.1. 3.2.1 One random vector

- **scalar**: x, A, α
- column **vector**: $\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = [x_i] \in \mathbb{R}^N$ with element $x_i = [\underline{x}]_i$
- **matrix**: $\mathbf{A} = [a_{ij}]_{1 \le i \le M, 1 \le j \le N} = [a_{ij}] \in \mathbb{R}^{M \times N}$ with element $a_{ij} = [\mathbf{A}]_{ij}$
- 3D, 4D **tensor**: $\mathbf{A} = [a_{ijk}], \mathbf{A} = [a_{ijkl}]$
- **transpose** of vector and matrix: $\underline{x}^T = [x_1, \ldots, x_N], \mathbf{A}^T = [a_{ji}]_{ij}$
- **determinant** of a square matrix $\mathbf{A}$: $|\mathbf{A}|$
- **inverse** of a square matrix $\mathbf{A}$: $\mathbf{A}^{-1}$
- **trace** of a square matrix $\mathbf{A}$: $\mathrm{tr}(\mathbf{A}) = \sum_i a_{ii}$

Abbildung 3.2: One random vector

PDF for a discrete-valued RV:
$p(\underline{x}) = \sum_i p_i \delta(\underline{x} - \underline{x}_i,)\ \delta(\underline{x} :)$ Dirac function
cumulative distribution function CCDF
$F(\underline{x}) = \int_{-\infty}^{\infty} p(\underline{z}) d\underline{z}, \quad p(\underline{x}) = \frac{\partial^d F(\underline{x})}{\partial x_i \ldots \partial x_d}$

- **scalar**: x, A, α
- column **vector**: $\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = [x_i] \in \mathbb{R}^N$ with element $x_i = [\underline{x}]_i$
- **matrix**: $\mathbf{A} = [a_{ij}]_{1\leq i\leq M, 1\leq j\leq N} = [a_{ij}] \in \mathbb{R}^{M\times N}$ with element $a_{ij} = [\mathbf{A}]_{ij}$
- 3D, 4D **tensor**: $\mathbf{A} = [a_{ijk}], \mathbf{A} = [a_{ijkl}]$
- **transpose** of vector and matrix: $\underline{x}^T = [x_1, \dots, x_N], \mathbf{A}^T = [a_{ji}]_{ij}$
- **determinant** of a square matrix $\mathbf{A}$: $|\mathbf{A}|$
- **inverse** of a square matrix $\mathbf{A}$: $\mathbf{A}^{-1}$
- **trace** of a square matrix $\mathbf{A}$: $\mathrm{tr}(\mathbf{A}) = \sum_i a_{ii}$

Abbildung 3.3: Moments of a vector

special case d = 1: $\underline{X} \to X \in \mathbb{R}$
$\underline{\mu} \to \mu \in \mathbb{R}$
$\underline{\underline{C}} \to$ variance of $X = Var(X) = \delta^2 = E[(X-\mu)^2] = ... = E(X^2) - \mu^2$
$\delta = \sqrt{Var(X)}$: standard deviation
For any function $g(\underline{X})$ of $\underline{X}$: $E[g(\underline{x})] = \int g(x) \cdot p(\underline{x}) d\underline{x} = (d.v.) \sum_i g(\underline{x}_i \cdot P(\underline{x}_i))$

Multivariate normal (Gaussian) distribution

PDF

$$\underline{X} \in \mathbb{R}^d \sim N(\underline{\mu}, \mathbf{C})$$

$$p(\underline{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \mathbf{C}^{-1}(\underline{x}-\underline{\mu})}$$

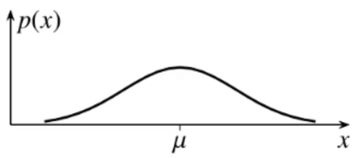
$$\ln(p(\underline{x})) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\mathbf{C}|) - \frac{1}{2}(\underline{x}-\underline{\mu})^T \mathbf{C}^{-1}(\underline{x}-\underline{\mu})$$

$N(\underline{0}, \mathbf{I})$ is called the **standard normal distribution**.

$$\underline{\underline{I}} = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$$

identity matrix

1D-Visualization



Moments

$$E(\underline{X}) = \underline{\mu}, \qquad Cov(\underline{X}) = \mathbf{C}$$

Abbildung 3.4: Multivariate normal (Gaussian) distribution

one-hot coding : Only one bit is 1 e.g. 0100 one cold coding is the inverse
Coding for class label, random vector y of length c so the identity matrix with dimension c is used for class labels
Reforumulation of the PMF (categorical distribution) by one-hot coding of the classes
$\underline{x} = [x_i] \in \{\underline{e}_1, \underline{e}_2, .., \underline{e}_c\}$, i.e. all $x_i = 0$ except for one single element equal to 1

PMF: $P(\underline{X} = \underline{x}) = P(\underline{x}) = \begin{cases} P_i \text{ if } \underline{x} = \underline{e}_1 \text{ or } x_1 = 1 \\ ... \\ P_c \text{ if } \underline{x} = \underline{e}_c \text{ or } x_c = 1 \end{cases} = p_1^{x_1} \cdot p_2^{x_2} \cdot, \dots \cdot p_c^{x_c} = \prod_{i=1}^c p_i^{x_i}$

$$ln(P(\underline{x}) = sum_{i=1}^{c} x_i \cdot ln(p_i) = [x_i, ..., x_c] \cdot \begin{bmatrix} ln(P_i) \\ ... \\ ln(P_c) \end{bmatrix} = \underline{x}^T \cdot ln(\underline{P})$$

ln function applied element wise

3.3. Chapter 3.3 - Multiple random vectors

•3-16 Table for distributions
•product rule for probability $p(\underline{x}, \underline{y}) = p(\underline{x}) \cdot p(\underline{y})$
•Bayes rule $p(\underline{y}|\underline{x}) = p(\underline{y}|\underline{x}) \cdot) \frac{p(\underline{x}}{p(\underline{y}}$
•Independent and identically distributed
$\underline{x}$ and $\underline{y}$ are independent if:
$p(\underline{x}, \underline{y}) = p(\underline{x}) \cdot p(\underline{y}) \leftrightarrow p(\underline{x}|\underline{y})$ or $p(\underline{x}|\underline{y}) = p(\underline{y})$
$\underline{x}_1, ..., \underline{x}_N$ are independent and identically distributed (i.i.d)
$P(\underline{x}_1, ..., \underline{x}_N) = \prod_{i=1}^{N} p_i(\underline{x}_i), \underline{X}_i \sim p_i(\underline{x}_i)$
$p_i(\underline{x}_i) = p(\underline{x}_i) \rightarrow p(\underline{x}_1, ..., \underline{x}_N) = \prod_{i=1}^{N} p(\underline{x}_i)$

3.3.1. Chapter 3.2.3 - Kernel basaed density estimation

PDF: $p(\underline{x})$ of $\underline{X} \in \boldsymbol{R}^d$ unknown , only i.i.d samples $\underline{x}(n), 1 \leq n \leq N$
kernel-based estimate $\hat{p}(\underline{x}) of p(\underline{x})$ from $\underline{x}(n)$
kernel function $k(\underline{x})$, like a PDF
1. $k(\underline{x}) \geq 0 \forall \underline{x}$
2. $\int k(\underline{x}) d\underline{x} = 1$
$\hat{p}(\underline{x}) = \frac{1}{N} \sum_{n=1}^{N} k(\underline{x} - \underline{x}(n))$

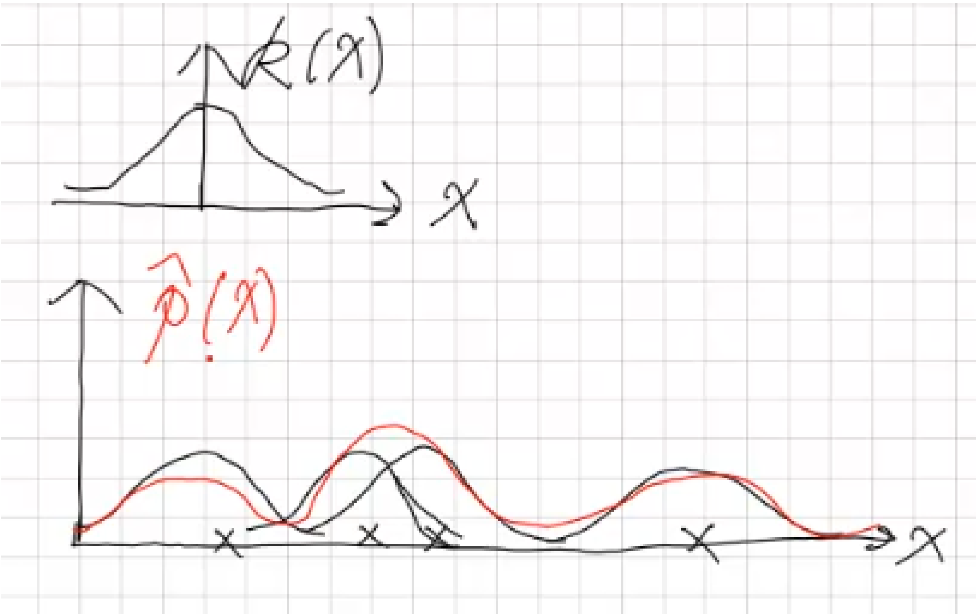


Abbildung 3.5: Kernel function

Smooth Gaussian Kernel:
$N(\underline{0}, \underline{I}) : k(\underline{x}) = \frac{1}{2\pi^{\frac{d}{2}}} \cdot e^{-\frac{1}{2}||\underline{x}||^2}$
Dirac Kernel

$k(\underline{x}) = \delta(\underline{x})$: Dirac function
$\delta(\underline{x}) = \begin{cases} \infty, \underline{x} = \underline{0} \\ 0, \underline{x} \neq \underline{0} \end{cases}$
$\int \delta(\underline{x}) d\underline{x} = 1$
sampling property : $\int \delta(\underline{x} - \underline{x}_0 f(\underline{x}) d\underline{x}) = f(\underline{x}_0)$
empirical distribution
$\hat{p}(\underline{x}) \cdot \frac{1}{N} \sum_{n=1}^{N} \delta(\underline{x} - \underline{x}(n))$

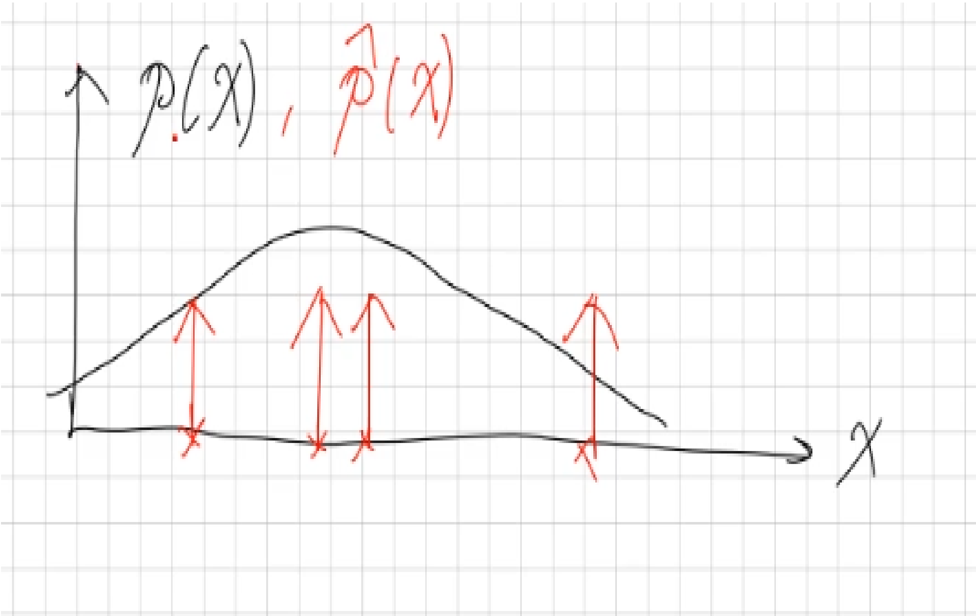
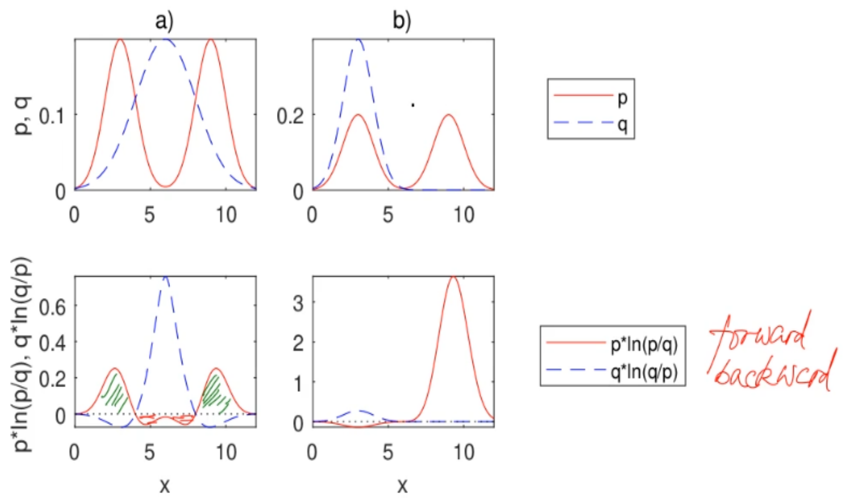


Abbildung 3.6: Estimated PDF function

3.4. Chapter 3.4 Kullback-Leibler divergence and cross entropy

Dissimilarity measure between 2 distributions:
Case A: continuous-valued random variables: PDF
$\underline{X} \sim p(\underline{x})$: true statistical distribution of $\underline{X}$
$q(\underline{x})$: approximation for $p(\underline{x})$, e.g. by DNN
KL divergence (KLD) between p and q:
$D_{KL}(p||q) = \int p(\underline{x}) \cdot ln(\frac{p(\underline{x})}{q(\underline{x})}) d\underline{x}$
$= E_{\underline{X} \sim p}[ln(\frac{p(\underline{X})}{q(\underline{X})})]$
expectation over $p(\underline{x})$
DKL is real valued scalar positive or negative or 0
Case B: discrete-valued random vector: PMF
$\underline{X} \in \{\underline{x}_1, ...\underline{x}_c\} \sim$: true PMF of $\underline{X} \sim Q(\underline{x})$: approximation for $P(\underline{x})$
$D_{KL}(P||Q) = \sum_{i=1}^{c} P(\underline{x}_i) \cdot ln(\frac{P(\underline{x})}{Q(\underline{x})}) = E_{\underline{X} \sim P}[ln(\frac{P(\underline{x})}{Q(\underline{x})})]$
Properties of the KL divergence: P1) Nonnegative $D_{KL}(P||Q) \geq 0 \forall p, q$
P2) Eguality $D_{KL}(P||Q) = 0$ iff(if and only if) $p(\underline{x}) = q(\underline{x})$
proof for ßufficient": $ln(\frac{p(\underline{x})}{q(\underline{x})}) = 0 \forall \underline{x}$
P1 and P2: $D_{KL}(p||1)$ is a suitable metric for approximation p by q
P3 Asymmetry
$D_{KL}(p||q) = E_{\underline{X} \sim p} = ln(\frac{p(\underline{x})}{q(\underline{x})}) \neq D_{KL}(q||p) = E_{\underline{X} \sim q} = ln(\frac{q(\underline{x})}{p(\underline{x})})$
forward KLD backward KLD
D_{KL} is not a true distance measure with $D(\underline{x}, \underline{y}) = D(\underline{y}, \underline{x})$



• When minimizing $D_{\mathrm{KL}}(p\|q)$, a) is better than b) because $q(x)$ is broad.
• When minimizing $D_{\mathrm{KL}}(q\|p)$, b) is better than a) because $q(x)$ is narrow.

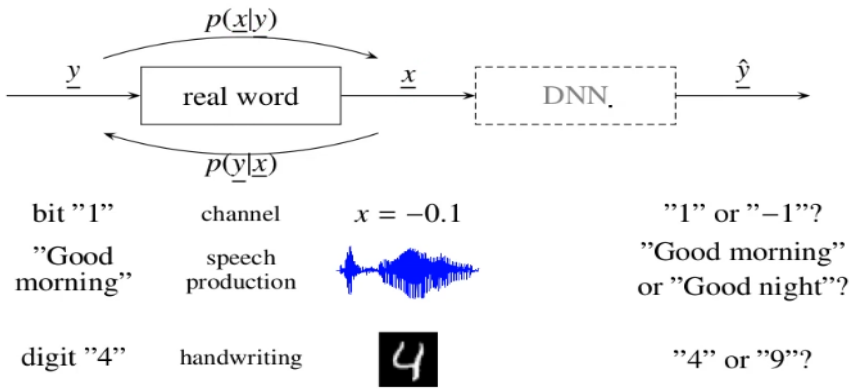
Abbildung 3.7: Forward vs. Backward KL divergence

3.4.1. E3.5 KLD between normal and Laplace distribution

$p(x) =\sim N(0,\sigma^2), p(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{-\frac{x^2}{2\sigma^2}}$
$q(x) \sim$ Laplace: $(0,b)q(x) = \frac{1}{2b}e^{-\frac{|x|}{b}}$
Task: choose b to best approximate p by q.
$\frac{p(x)}{q(x)} = \sqrt{\frac{2}{\pi}} \cdot \frac{b}{\sigma} cdotexp(-\frac{x^2}{2\sigma^2} + \frac{|x|}{b})$
$D_{KL}(p||q) = E_{\underline{X}\ simp} = ln(\frac{p(\underline{x})}{q(\underline{x})}) = ln(\sqrt{\frac{2}{\pi}} \cdot \frac{b}{\sigma}) + E_{X\sim p}((-\frac{x^2}{2\sigma^2} + \frac{|x|}{b}))$
$E_{\underline{X}\ simp} = \sigma^2$
$E_{\underline{X}\ simp}(|x|) = \int_{-\infty}^{\infty} |x| \cdot \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{x^2}{2\sigma^2})dx = 2\int_0^{\infty} | \cdot \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{x^2}{2\sigma^2})dx = \sqrt{\frac{2}{\pi}} \cdot \sigma$
Let $\alpha = \frac{\sigma}{b}, D_{KL}(p||q) = ... = \sqrt{\frac{2}{\pi}} \cdot \alpha - ln(\alpha) + ln(\sqrt{\frac{2}{\pi}}) - \frac{1}{2}\ \frac{dD_{KL}(p||q)}{d\alpha} = \sqrt{\frac{2}{\pi}} - \frac{1}{\alpha} = 0 \rightarrow \alpha = \sqrt{\frac{2}{\pi}}$, i.e. $b \approx 0,86$
$D_{KL,min}(p||q) = D_{KL}(p||q)|\alpha = \sqrt{\frac{2}{\pi}} = .. = \frac{1}{2} - ln(\frac{\pi}{2}) \approx 0,048$
•Probable exam question calculate this for 2 distributions
P4) Additive :
$\underline{X}) = (\underline{x}_1, \underline{x}_2), \underline{x}_1$ and $\underline{x}_2$ are independent, $i.e.$
$p(\underline{x})?p_1(\underline{x}_1) \cdot p_2(\underline{x}_2), q(\underline{x})?q_1(\underline{x}_1) \cdot q_2(\underline{x}_2)$
Then: $D_{KL}(p||q) = D_{KL}(p_1||q_1) + D_{KL}(p_2||q_2)$
P5) Relation to cross entropy :
Definiton of entropy 3-24 probability always greater than 0 but smaller than 1
$D_{KL}(p||q) = \int pln(\frac{p}{q})d\underline{x} = \int pln(p)dx - \int pln(q)dx = -H(p) + H(p,q)$ or
cross entropy: $H(p,q) = D_{KL}(p||q) + H(p) \geq H(p) \geq 0$
For a given (fixed) $p(\underline{x})$: $H(p)$ fixed
Hence: min $D_{KL}(p||q) \leftrightarrow minH(p,q)$
Not the case for backward KLD $D_{KL}(q||p)$!
Minimizing is not the same anymore because then it is $H(q)$ and thats what we are trying to optimize

3.5. Chapter 3.5 Probabilistic framework

valid for both SP and ML
valid for both regression problem and classicication problem



- $\underline{y}$: **latent variable**, hidden, not directly measurable, quantity of interest
- $\underline{x}$: **observed variable**, measurement, input for DNN
- $\hat{\underline{y}}$: output of DNN as estimate for $\underline{y}$
- real world: describes how is $\underline{x}$ generated from $\underline{y}$
- DNN: describes how to estimate $\underline{y}$ from $\underline{x}$

Abbildung 3.8: Probabilistic framework of supervised learning

Both $\underline{y}$ and $\underline{x}$ are modeled as random variables. They are described by the joint **data generating distribution**

$$p(\underline{x}, \underline{y}) = p(\underline{x}|\underline{y})p(\underline{y}) = p(\underline{y}|\underline{x})p(\underline{x}).$$

$p(\underline{y})$ prior PDF of $\underline{y}$ or **prior**, available before any measurement of $\underline{x}$

$p(\underline{x}|\underline{y})$ **likelihood**. It describes the real word, the generation of $\underline{x}$ from $\underline{y}$.
It is a kind of channel-sensor model, e.g.

- bit: communication channel + receiver
- speech: speech production system + microphone
- digit: handwriting + camera

$p(\underline{x})$ prior PDF of $\underline{x}$, also called **evidence**

$p(\underline{y}|\underline{x})$ posterior PDF of $\underline{y}$ after a measurement $\underline{x}$ or **posterior**

Abbildung 3.9: The data generating distribution

Bayes Rule:

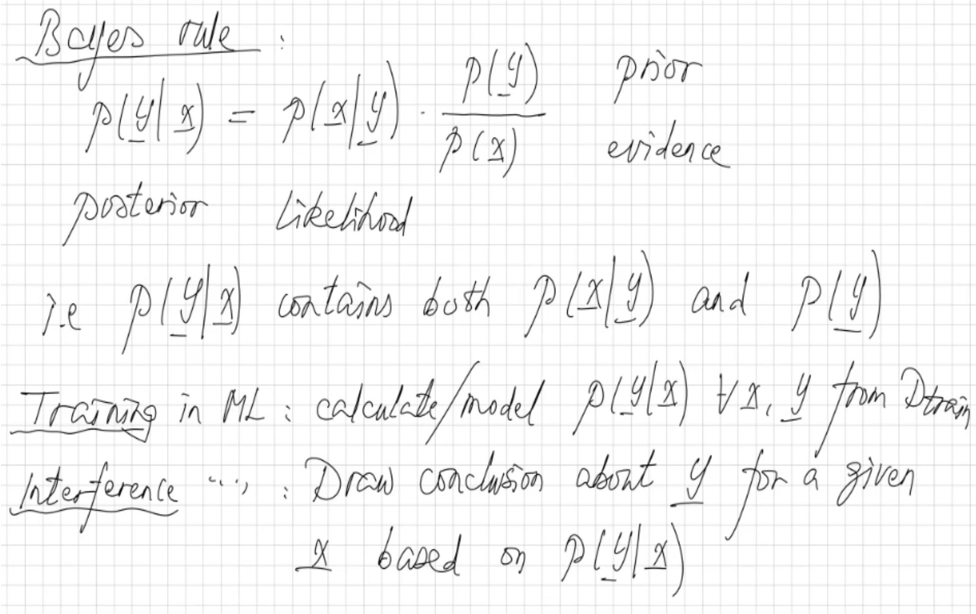


Abbildung 3.10: Bayes Rule in DL

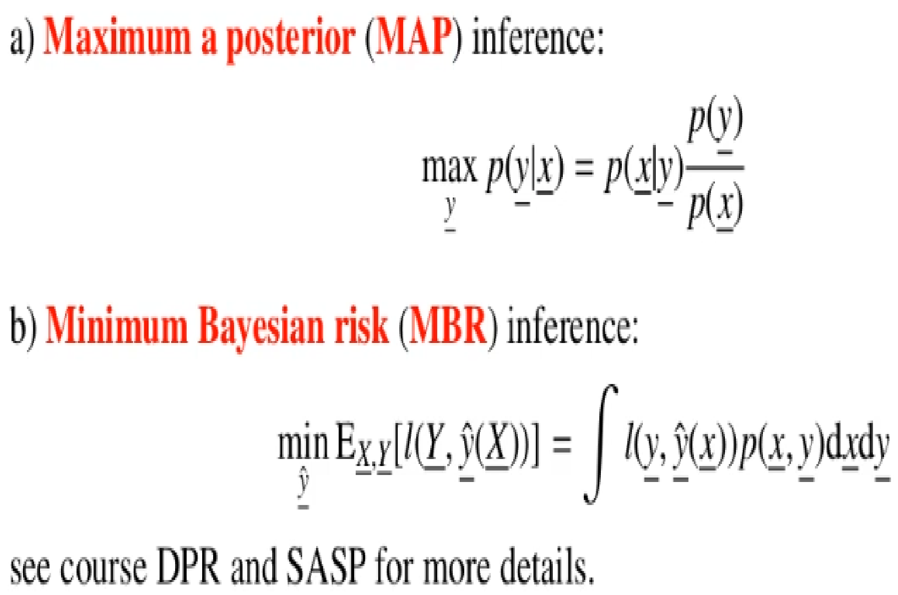


Abbildung 3.11: Bayes decision theorem

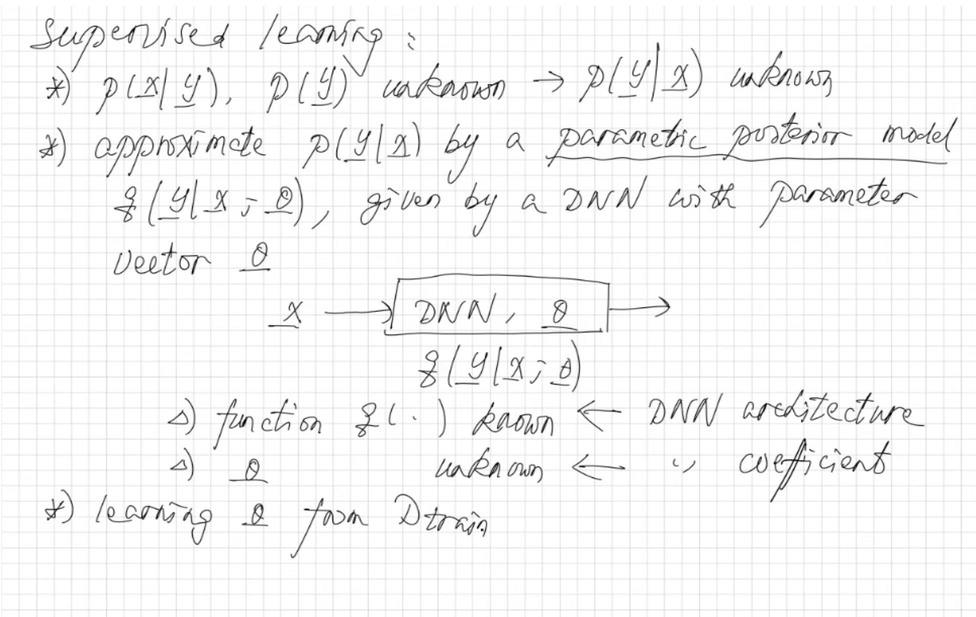


Abbildung 3.12: Supervised learning

Learning Criterion:

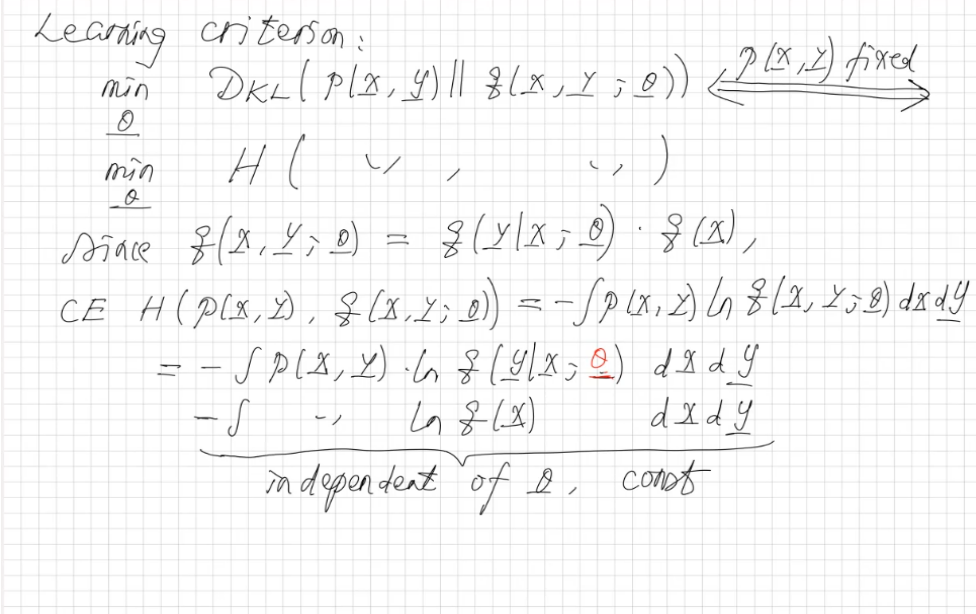


Abbildung 3.13: Calc part1

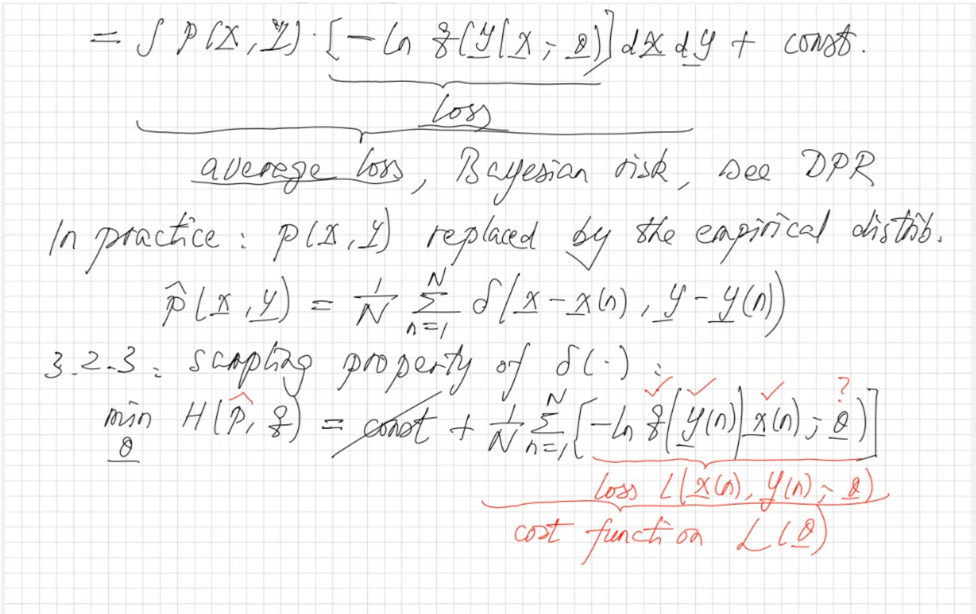


Abbildung 3.14: Calc part2

3.5.1. Role of a NN

1. approximate true posterior $p(\underline{y}|\underline{x})$ by $q(\underline{y}|\underline{x};\Theta)$
2. learn $\underline{\Theta}$ from D_{train}

Lecture 4: Dense Neural Networks

Date: 05/05/2020

Lecturer: David Silver *By: Nithish Moudhgalya*

A general model for $q(\underline{y}|\underline{x};\Theta)$
*) can learn any linear or nonlinear mapping
*) suitable for both regression and classification problems
artificial NN: mimic biological NN (brain)

4.0.1. 4.1 Fully connected neural networks - Neuron

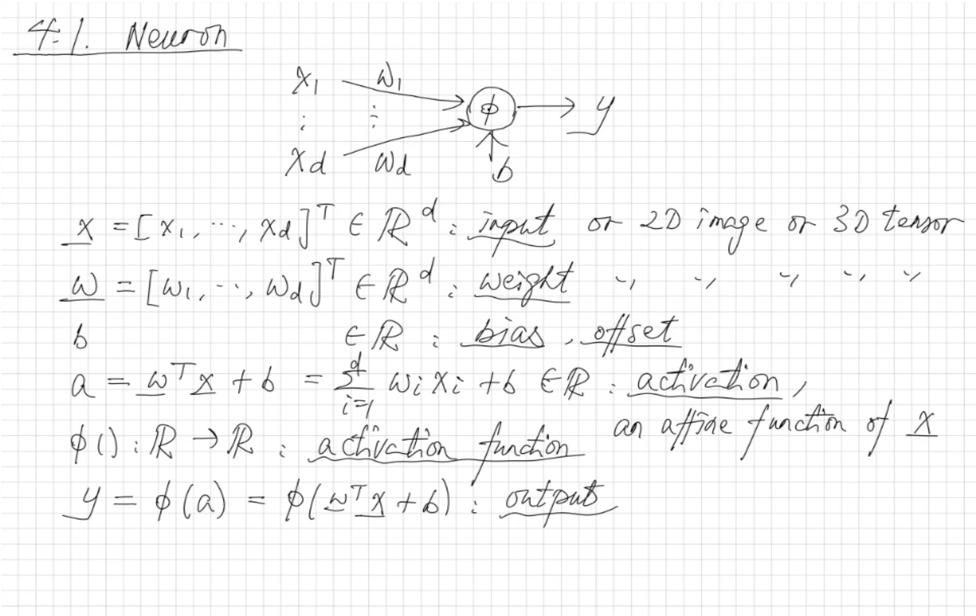


Abbildung 4.1: Neuron

4.0.2. Chapter 4.2 Layer of Nurons

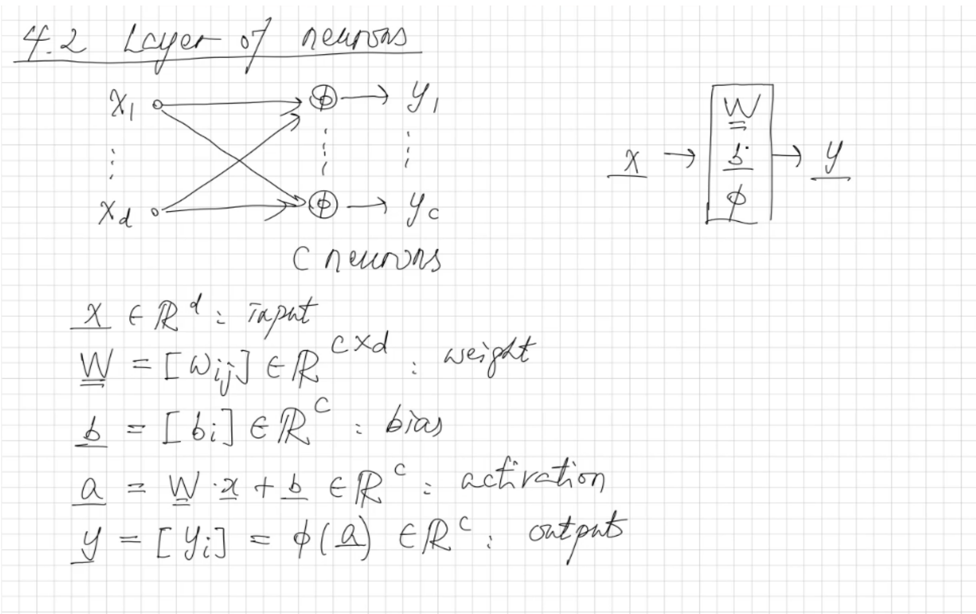


Abbildung 4.2: Layer of Neurons

2 meanings of $\phi()$:

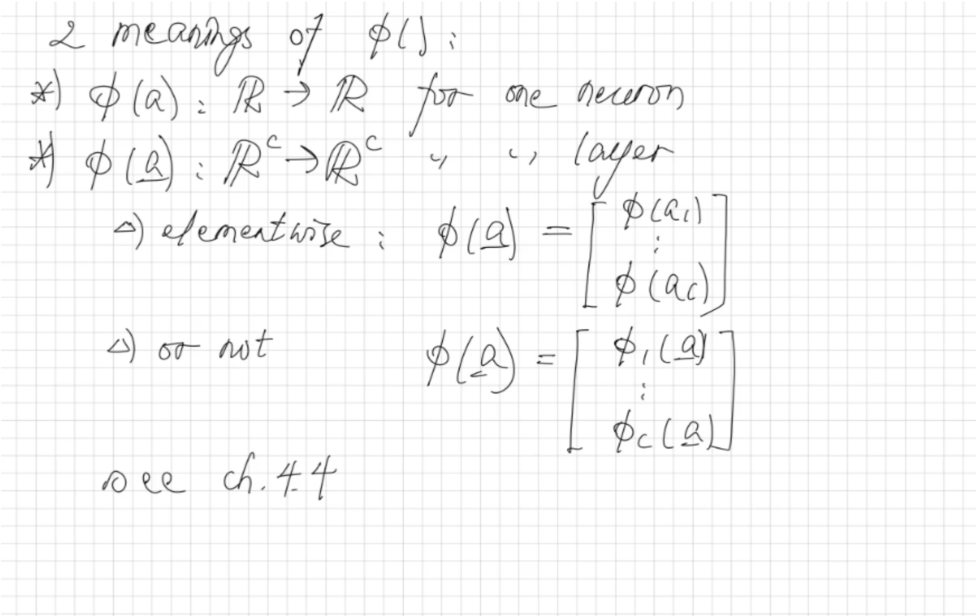


Abbildung 4.3: Meanings of phi

Comments :
•no interconnections between neurons in the same layer
•dense layer, fully connected layer: •each input x_j connected to each neuron i
$\rightarrow c \cdot d$ weights and w_{ij} an c biases b_i , $1 \leq i \leq c, 1 \leq j \leq d$,
$\rightarrow c \cdot (d+1)$ parameters

4.1. 4.3 Feedforward neural network

A cascade of dense layers
Layer $1 \leq l \leq L$:
$M_l - 1$ number of the input neurons

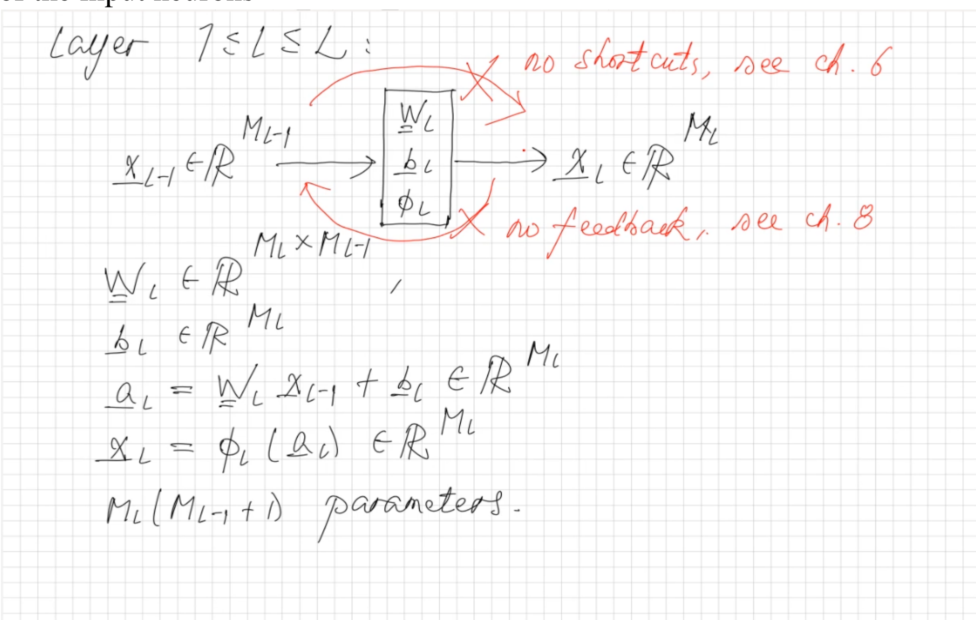


Abbildung 4.4: Feedforward multilayer neural network

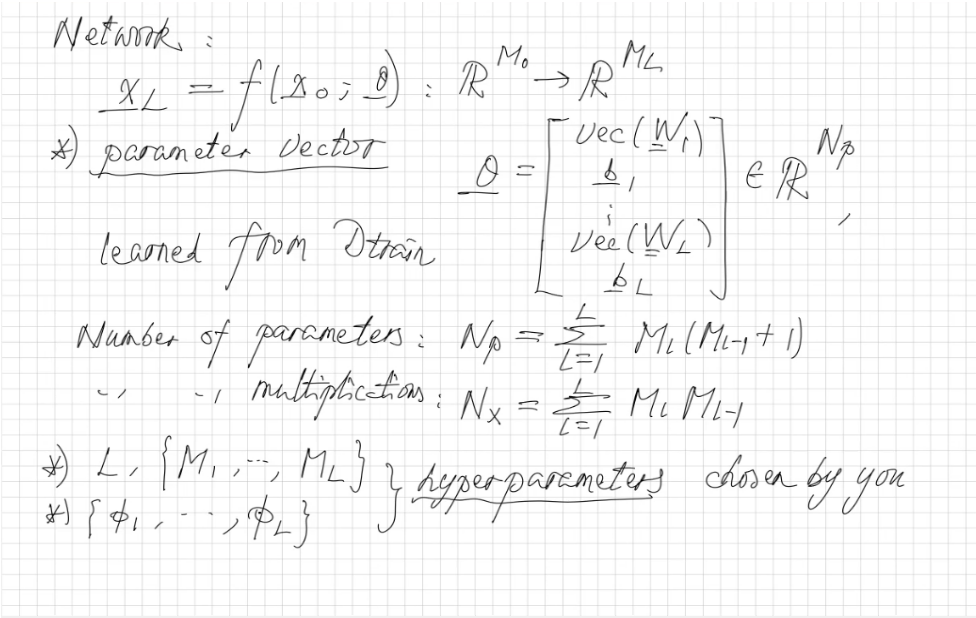


Abbildung 4.5: Network Parameters