

# AI Notes

*Huu Duc Nguyen M.Sc.*

29 February 2022

# Contents

<b>Abbreviations</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
<b>2 Overview of Machine Learning</b>	<b>5</b>
2.1 Tasks . . . . .	5
2.1.1 Classification . . . . .	5
2.1.2 Regression . . . . .	5
2.1.3 Clustering . . . . .	6
2.1.4 Others . . . . .	6
2.2 Performance . . . . .	6
2.3 Experience Types . . . . .	6
2.3.1 Supervised Learning . . . . .	6
2.3.2 Unsupervised Learning . . . . .	7
2.3.3 Semi-supervised Learning . . . . .	7
2.4 Model Parameters and Loss Function . . . . .	7
<b>3 Probabilities</b>	<b>8</b>
3.1 Parameter Estimation . . . . .	8
3.1.1 Maximum Likelihood Estimation . . . . .	9
3.1.2 Maximum A Posteriori . . . . .	9
3.2 Naive Bayes Classifier . . . . .	9
3.3 Views on the Decision Problem . . . . .	10
3.3.1 Generative Methods . . . . .	10
3.3.2 Discriminative Methods . . . . .	10
3.3.3 Classification with Loss Functions . . . . .	10
3.3.4 Minimize the Expected Loss . . . . .	11
3.4 Probability Density Estimation . . . . .	11
3.4.1 Histogram . . . . .	11
3.4.2 Kernel Methods . . . . .	12
3.4.3 K-Nearest Neighbor . . . . .	13
3.4.4 Mixture of Gaussians . . . . .	13
3.4.5 K-Means Clustering . . . . .	14
3.4.6 EM Clustering . . . . .	14

<b>4</b>	<b>Basic ML Problems</b>	<b>15</b>
4.1	Linear Regression . . . . .	15
4.2	Linear Discriminant Functions . . . . .	16
4.2.1	Least-squares classification . . . . .	16
4.2.2	Generalized Linear Discriminant . . . . .	17
4.2.3	Generalized Linear Models . . . . .	17
4.3	Logistics Regression . . . . .	17
4.4	Softmax Regression . . . . .	19
4.5	SVM . . . . .	20
4.5.1	Hard Margin State Vector Machine (SVM) . . . . .	20
4.5.2	Soft Margin SVM . . . . .	22
4.5.3	Kernel Support Vector Machine . . . . .	23
<b>5</b>	<b>Ensembles</b>	<b>24</b>
5.1	Error Reduction . . . . .	24
5.2	Bagging . . . . .	24
5.3	Boosting . . . . .	25
5.3.1	Adaboost . . . . .	25
5.3.2	Gradient Boosting (Regression) . . . . .	27
5.4	Information . . . . .	27
5.5	Decision Trees . . . . .	27
5.5.1	Stopping Conditions . . . . .	28
5.5.2	Pruning . . . . .	29
5.5.3	Computational Complexity . . . . .	29
5.5.4	Summary . . . . .	30
5.6	Random Forest . . . . .	30
5.7	Bayesian Model Averaging . . . . .	30
5.8	Distillation . . . . .	31
<b>6</b>	<b>Feature Engineering</b>	<b>32</b>
6.1	Principle Component Analysis . . . . .	32
6.2	Linear Discriminant Analysis . . . . .	32
6.3	Word Representation . . . . .	32
6.3.1	The trigram (n-gram method) . . . . .	32
6.3.2	Word Embedding . . . . .	33
6.3.3	word2vec . . . . .	33
6.3.4	Hierarchical Softmax . . . . .	33

<b>7</b>	<b>Error Functions</b>	<b>34</b>
7.1	Ideal Miss-classification Error . . . . .	34
7.2	Squared Error - $L_2$ Loss . . . . .	34
7.3	Cross Entropy Error . . . . .	35
7.4	Squared Error on Sigmoid / Tanh . . . . .	35
7.5	Hinge Error . . . . .	35
7.6	$L_1, L_0$ Loss . . . . .	35
7.7	Average Loss . . . . .	35
<b>8</b>	<b>Neural Network</b>	<b>36</b>
8.1	General . . . . .	36
8.2	Gradient Descent . . . . .	37
8.2.1	Vanilla Gradient Descent . . . . .	37
8.2.2	Momentum . . . . .	37
8.2.3	Nesterov Accelerated Gradient . . . . .	37
8.2.4	Root mean squared prop (RMSprop) . . . . .	38
8.2.5	Adaptive moment estimation (Adam) . . . . .	38
8.3	Tips and Tricks . . . . .	38
8.4	Network Structures . . . . .	39
8.4.1	Recurrent Neural Network . . . . .	39
8.4.2	Residual Network . . . . .	39
<b>9</b>	<b>Generative Models</b>	<b>40</b>
9.1	Definitions . . . . .	40
9.2	Variational Inference . . . . .	40
9.2.1	The Variational Approximation . . . . .	41
9.2.2	Algorithm . . . . .	42
9.3	Amortized Variational Inference . . . . .	42
9.3.1	Algorithm . . . . .	43
9.3.2	Reparameterization Trick . . . . .	44
9.4	Conditional Variational Inference Models . . . . .	44
9.5	Generative Models . . . . .	45
9.5.1	Generative Adversarial Network . . . . .	45
9.6	References . . . . .	45
	<b>Bibliography</b>	<b>I</b>

# Abbreviations

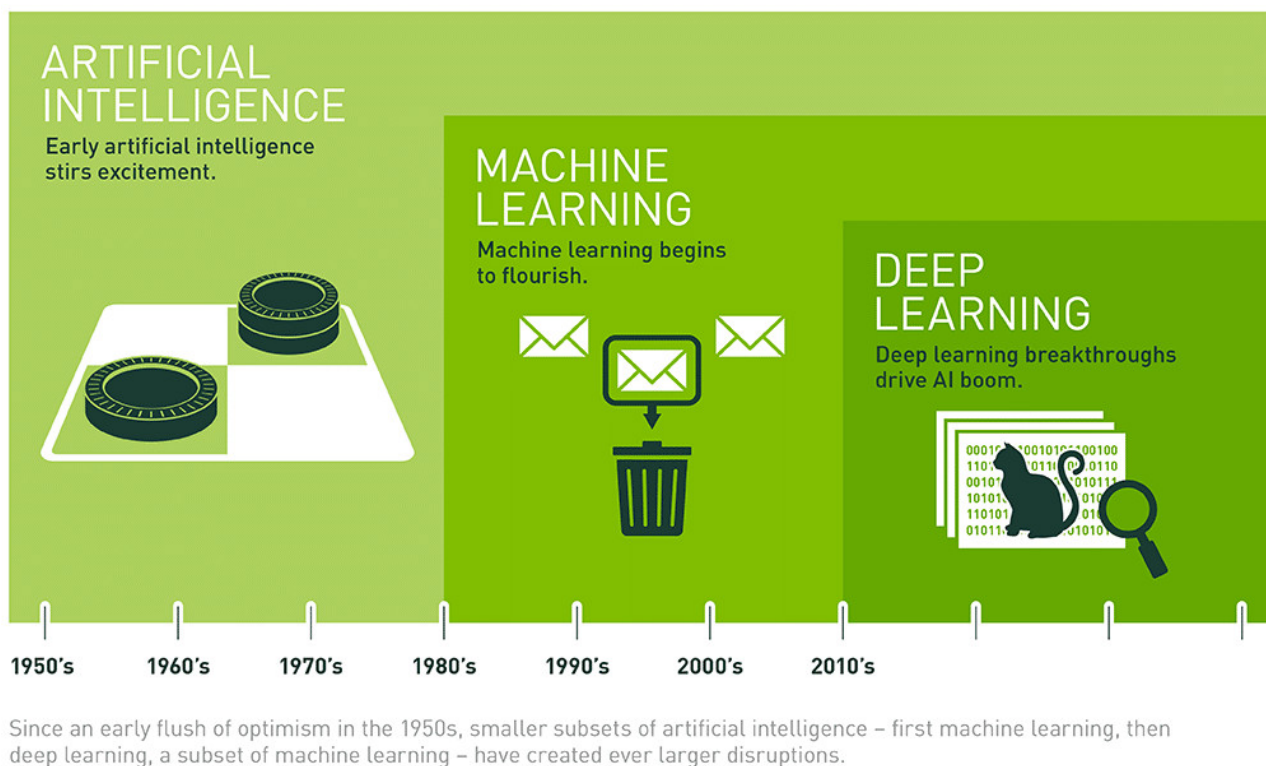
<b>AI</b>	Artificial Intelligence
<b>ML</b>	Machine Learning
<b>DL</b>	Deep Learning
<b>CS</b>	Computer Science
<b>CV</b>	Computer Vision
<b>RL</b>	Reinforcement Learning
<b>NLP</b>	Natural Language Processing
<b>prob.</b>	probability
<b>params.</b>	parameters
<b>algor.</b>	algorithms
<b>a.k.a.</b>	also known as
<b>w.r.t.</b>	with regard to
<b>no.</b>	number of
<b>func.</b>	function
<b>vs.</b>	versus
<b>pdf.</b>	Probability Density Function
<b>MLE</b>	Maximum Likelihood Estimation
<b>MAP</b>	Maximum A Posteriori
<b>MoG</b>	Mixture of Gaussians
<b>SVM</b>	State Vector Machine
<b>SGD</b>	Stochastic Gradient Descent
<b>NAG</b>	Nestorov Accelerated Gradient
<b>RMSprop</b>	Root mean squared prop
<b>Adam</b>	Adaptive moment estimation
<b>PCA</b>	Principal Component Analysis
<b>LDA</b>	Linear Discriminant Analysis
<b>K-L</b>	Kullback–Leibler
<b>KKT</b>	Karush-Kuhn-Tucker
<b>RBF</b>	Radial basic function
<b>ReLU</b>	Rectified Linear Unit
<b>BPTT</b>	Backpropagation through time
<b>VAE</b>	Variational Auto-Encoders
<b>GAN</b>	Generative Adversarial Network
<b>DPM</b>	Deformable Part Model

## *Contents*

**HOG**      Histogram of Oriented Gradients

# 1 Introduction

Artificial Intelligence (AI) is the study field that leverages the ability of machines to mimic the problem-solving skill of human. In other words, AI pushing machines to think, act like people, in a rational way. It lies in the core of countless novel applications in real life, self-driving cars, virtual assistant, face recognition, etc. Machine Learning (ML) is a sub-field of Computer Science (CS) and AI, that “gives computers the ability to learn without being explicitly programmed” (Wikipedia). As a great amount of collected data and powerful computational hardware arise, Deep Learning (DL) is then a subset of ML (Fig. 1.1). Advanced applications which relate to Natural Language Processing (NLP), Computer Vision (CV), robotic learning, etc., are with in this DL subset.



**Figure 1.1:** The relation between AI, ML and DL (src).

It's important to understand that there are more to AI and ML than just neural networks. In the end, to create a meaningful and working neural network, one should have strong background in the basics of ML as well.

These notes are my way of keeping record of what I have learn in the field. The structure of the notes is as follows:

## 1 Introduction

- Chap. 2 introduces common ideas in ML.
- Chap. 3 presents the mathematics background on probabilities, matrix.
- **[TODO: chapter 3]** explain basic concepts, the branching of different classes in ML.  
Later chapters presents each smaller branches.

**[TODO: The structure of the notes]**



## 2 Overview of Machine Learning

A machine learning algorithm is an algorithm that has the ability to *learn* from the data. A computer program is said to **learn**, if its performance at tasks in  $T$ , measured by  $P$ , improves with experience  $E$  (in which the experience is equivalent to the data). [GBC16]

### 2.1 Tasks

A *task* is usually described by how the ML model process a single *data point*. This section presents some common ML tasks. [Vu18]

#### 2.1.1 Classification

The task is to specify a label for the given data point. The labels are usually members of a list. E.g.:

- Hand-written digit classification
  - The data point: images of hand-written numbers with their labels
  - The task: tell which number is in a unseen image
  - Labels: there are 10 possible labels, i.e.,  $\{0, 1, \dots, 9\}$ .
- Hand-written letter classification
  - The data point: images of hand-written letters with their labels
  - The task: tell which letter is in a unseen image
  - Labels: there are 26 (or more) possible labels, i.e.,  $\{a, b, \dots, z\}$ .

#### 2.1.2 Regression

If the desired output is a real value, instead of a label in a list, then it's a regression problem. E.g.:

- Input data: a person image  $\longrightarrow$  Output: predicts the age of the person
- Input data: a feature vector  $\longrightarrow$  Output: generates an image
- Input data: dimension of a block of gold  $\longrightarrow$  Output: the price of it

## 2 Overview of Machine Learning

### 2.1.3 Clustering

This is the task of grouping relevant data points based on some relationship between them.

E.g., find the pattern in customer shopping behaviors.

### 2.1.4 Others

Other worth-mentioning tasks:

- Recommendation System
- Machine Translation
- Completion
- Ranking
- Information Retrieval
- Denoising

## 2.2 Performance

Usually, the dataset is divided into *training set* and *test set*. The model uses the training set to tune/update the model parameters ([params.](#)), and the test set to examine the performance.

*Online training* is the approach when new data will continuously arise and introduce for the model to learn. E.g. in Reinforcement Learning ([RL](#)). *Offline training* is the opposite case where the model learns from the a fixed training set.

## 2.3 Experience Types

### 2.3.1 Supervised Learning

**Data with labels:** Supervised Learning is the approach that predict the outputs of new data points based on pairs of known inputs and outputs. This is the most common type of [ML](#) algorithms ([algor.](#)).

### 2.3.2 Unsupervised Learning

**Data without labels:** On the opposite, with unsupervised Learning, there isn't known output, just inputs. Unsupervised Learning [algor.](#) will carry on some tasks which base on the characteristics of the dataset, e.g. clustering, dimension reduction.

### 2.3.3 Semi-supervised Learning

**Some data with, some without labels**

## 2.4 Model Parameters and Loss Function

Each [ML](#) model is described by a set of model [params.](#). E.g., in the problem of finding a line passing through points in the 2D plane, the model [params.](#) are  $a, b$  in the line equation  $y = ax + b$ . The of training aim is to find the model [params.](#) which leads to the best performance or the minimum loss. For classification problems, it can be having the least number of incorrect classified data points. For regression problems, it can be having the smallest difference with the actual output. It is then equivalent to having an optimization problem, in which we try to minimize that loss/cost function.

# 3 Probabilities

Many ML algorithms and problems concern with the probability ([prob.](#)) or Probability Density Function ([pdf.](#)) of the outputs from complex inputs. We all aware of some basic examples of the output [prob.](#) from simple inputs, e.g., flipping a coin, rolling a dice. With these simple inputs, due to uncertainty, one could only say about the [prob.](#) that the output could be, e.g., flipping a coin has 50% of getting head, 50% of getting tail. Comparing to flipping a coin, inputs for ML algorithms are usually more complex, in higher dimensions, in different modals, etc. The inputs can be datasets, images, audio, etc. However, as with the simple inputs, we concern ourselves with the [prob.](#) of the outputs, e.g.:

- Given a dog image, output the [prob.](#) of the breed of the dog.
- Image segmentation: given an image, output the [prob.](#) of object class for each pixel.
- Cancer detection: given a medical image, output the [prob.](#) of the patient having cancer.

This chapter presents important [prob.](#) problems relating to ML. For mathematics foundation on [prob.](#), check the mathematics notes.

[TODO: Add image: image with [prob.](#) of being cat or dog]

## 3.1 Parameter Estimation

Many of ML problems are boiled down to finding *statistical models*. Those models predict the [prob.](#) of the data belongs to one class (in classification problem), [prob.](#) of events that will happen, etc. It all end up with finding a suitable set of [params.](#) for these *statistical models*.

### 3.1.1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) finds the parameters that maximize the [prob.](#) of the existing data.

$$\text{Model parameter } \theta: \quad \theta = \arg \max_{\theta} p(x_1, x_2, \dots, x_N | \theta) \quad (3.1)$$

$$\text{Assuming independent variables:} \quad \theta = \arg \max_{\theta} \prod_{n=1}^N p(x_n | \theta) \quad (3.2)$$

$$\text{Maximum log-likelihood:} \quad \theta = \arg \max_{\theta} \sum_{n=1}^N [\log p(x_n | \theta)] \quad (3.3)$$

$$\text{Minimum negative log-likelihood:} \quad \theta = \arg \min_{\theta} \sum_{n=1}^N [-\log p(x_n | \theta)] \quad (3.4)$$

### 3.1.2 Maximum A Posteriori

Sometimes, prior knowledge of the [pdf.](#) is presented. E.g., the [prob.](#) of getting head when flipping a coin is around 50%. Maximum A Posteriori (MAP) takes advantage of the prior knowledge  $p(\theta)$  on the parameters  $\theta$  by applying Bayes rule.

$$\theta = \arg \max_{\theta} \prod_{n=1}^N p(x_n | \theta) p(\theta) \quad (3.5)$$

[MLE \(Sec. 3.1.1\)](#) suffers when there is not enough data  $\Rightarrow$  use MAP

## 3.2 Naive Bayes Classifier

The word *naive* implies having the independence assumption on the variables.

$$c = \arg \max_{c \in \mathbb{C}} p(c | x) \quad (3.6)$$

$$= \arg \max_{c \in \mathbb{C}} p(x | c) p(c) \quad (3.7)$$

If  $x$  is:

- continuous variable  $\Rightarrow$  Gaussian Naive Bayes
- feature vector  $\Rightarrow$  Multinomial Naive Bayes
- binary vector  $\Rightarrow$  Bernoulli Naive Bayes

Minimize the expected loss:  $\mathbb{E}[L] = \sum_k \sum_j \int_{R_j} L_{kj} p(x, C_k) dx$  by choosing region  $R_j$  such that

$$\mathbb{E}[L] = \sum_k L_{kj} p(C_k | x)$$

### 3.3 Views on the Decision Problem

#### 3.3.1 Generative Methods

First determine the class-conditional densities and separately infer the prior class [prob.](#)  $\Rightarrow$  Bayes theorem  $\Rightarrow$  class membership

$$p(x|C_k) p(C_k) \Rightarrow y_k(x)$$

E.g., Mixture of Gaussians

#### 3.3.2 Discriminative Methods

First solve the inference problem of determined the posterior class [prob.](#)

#### 3.3.3 Classification with Loss Functions

For classification problem with more than 2 classes, the output's predicted class  $C_k$  is the one with the greatest posterior [prob.](#) among all classes  $C_j$ , which can still be wrong:

$$\begin{aligned} p(C_k|x) &> p(C_j|x) && \forall j \neq k \\ p(x|C_k)p(C_k) &> p(x|C_j)p(C_j) && \forall j \neq k \end{aligned}$$

Loss function differentiate the possible decisions and the possible true classes. E.g.:

- Decisions: *sick* or *healthy*
- Classes: patient is *sick* or *healthy*

The cost may be symmetric. In this case however, it is asymmetric:

$$\text{loss}(\text{decision}=\text{healthy} \mid \text{patient}=\text{sick}) \gg \text{loss}(\text{decision}=\text{sick} \mid \text{patient}=\text{healthy})$$

The classification problem could be formalized by introducing loss matrix  $L_{kj}$

$$L_{kj} = \text{loss for decision } C_j \text{ if truth is } C_k$$

E.g., cancer diagnosis:

		<b>Decision</b>	
		<i>cancer</i>	<i>normal</i>
<b>Truth</b>	<i>cancer</i>	$\begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}$	
	<i>normal</i>		

$$L_{\text{cancer diagnosis}} =$$

Loss functions maybe different for different actors:

$$L_{\text{stocktrader}}(\text{subprime}) \neq L_{\text{bank}}(\text{subprime})$$

### 3.3.4 Minimize the Expected Loss

But regardless, the optimal solution is the one that minimizes the *expected loss*. It is *expected* loss because it depends on the true class, which is unknown.

⇒ Minimize the expected loss

E.g., 2 class  $C_1, C_2$ , 2 decisions  $\alpha_1, \alpha_2$ .

The loss:  $L(\alpha_j|C_k) = L_{kj}$ .

The expected loss is equal to the risk  $R$ .

$$\mathbb{E}_{\alpha_1}[L] = R(\alpha_1|x) = L_{11}p(C_1|x) + L_{21}p(C_2|x)$$

$$\mathbb{E}_{\alpha_2}[L] = R(\alpha_2|x) = L_{12}p(C_1|x) + L_{22}p(C_2|x)$$

Choose  $\alpha_1$  if  $R(\alpha_1|x) < R(\alpha_2|x)$

## 3.4 Probability Density Estimation

The following approaches' aim is to estimate an unknown pdf..

- Non-parametric approaches: represent the pdf. without parameter, usually as a lookup table. Using lookup table is fast, but could also be memory-consuming.  
Among the presented ones, only histogram is non-parametric.
- Parametric approaches: within sufficiently small region  $\mathcal{R}$ , the prob. can be estimated as  $p = \int_{\mathcal{R}} p(y)dy \approx p(x)V = \frac{K}{N.V}$ , where  $K$  is the number of data points in the region,  $V$  is the volume of the region.

### 3.4.1 Histogram

The prob. of a bin:

$$p_i = \frac{n_i}{N \cdot \Delta_i} \tag{3.8}$$

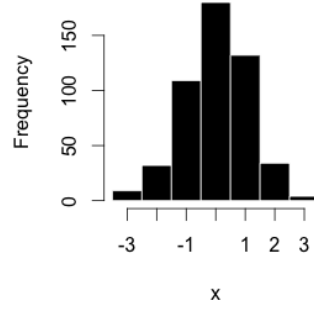
in which  $n_i$  is the number of data points in that bin,  $N$  is the total number of data point,  $\Delta_i$  is the width of the bin, often  $\Delta_i = \Delta$ .

Notes:

- $\Delta$  serves as the **smoothing factor**

### 3 Probabilities

- With  $D$  as the dimensions of the data points. The number of bins grow exponentially with  $\mathcal{O}(k^D)$



**Figure 3.1:** Example of a histogram,  $\Delta = 1$  ([src](#)).

#### 3.4.2 Kernel Methods

The kernel methods fix the volume  $V$  and determine the number of data points  $K$ . The volume  $V$  is the space restricted within a parzen window  $k(u)$  that satisfies  $k(u) \geq 0$ .

A hyper-space cube: 
$$k(u) = \begin{cases} 1 & \text{if } |u_i| \leq \frac{1}{2}h, \quad i = 1, 2, \dots, D \\ 0 & \text{else} \end{cases} \quad (3.9)$$

The number of points inside: 
$$K = \sum_{n=1}^N k(x - x_n) \quad (3.10)$$

The region *volume*: 
$$V = \int k(u) du = h^D \quad (3.11)$$

The probability: 
$$\Rightarrow p(x) \approx \frac{K}{N.V} = \frac{1}{N.h^D} \sum_{n=1}^N k(x - x_n) \quad (3.12)$$

**NOTE:** the above parzen window is asymmetric.

The ***symmetric Gaussian kernel is a better substitution.***

A Gaussian kernel: 
$$k(u) = \frac{1}{\sqrt{2\pi}h^2} \exp\left(\frac{-u^2}{2h^2}\right) \quad (3.13)$$

The region *volume*: 
$$V = \int k(u) du = 1 \quad (3.14)$$

The probability: 
$$\Rightarrow p(x) \approx \frac{1}{N} \sum_{n=1}^N \frac{1}{(2\pi)^{D/2}h} \exp\left(\frac{-||x - x_n||^2}{2h^2}\right) \quad (3.15)$$

For Kernel methods,  $h$  is the ***smoothing factor***.

Generalization:  $k(u) \geq 0, \int k(u) du = 1$ .

**Size of the hypersphere is proportional to  $h^2$ .**



### 3.4.3 K-Nearest Neighbor

When you fix the number of data points  $K$  and determine the volume  $V$ , it leads to K-Nearest Neighbor.

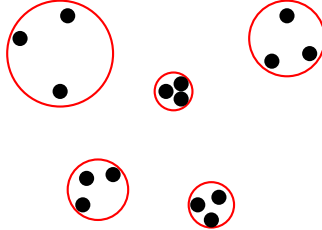


Figure 3.2: K-Nearest Neighbor with  $K = 3$ .

$$p(x) \approx \frac{K}{NV} \quad (3.16)$$

Here,  $K$  is the smoothing factor.

- Too much bias  $\Rightarrow$  too smooth
- Too much variance  $\Rightarrow$  NOT smooth enough

$\Rightarrow$  combine parametric methods to a mixture model

Mixture distribution = multi-parametric model

### 3.4.4 Mixture of Gaussians

Mixture of Gaussians (MoG), as **Generative Model**, is defined from the sum **prob.** of elemental Gaussians:  $p(x|\theta) = \sum_{j=1}^M p(x|\theta_j)p(j)$ , where  $p(x|\theta_j)$  is a **mixture component**,  $p(j) = \pi_j$  is the **weight of the component**

$$p(x|\theta_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(x - \mu_j)^2}{2\sigma_j^2}\right], \quad p(j) = \pi_j, \quad \sum \pi_j = 1 \quad (3.17)$$

$$p(x|\theta_j) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j)\right] \quad (3.18)$$

### 3.4.5 K-Means Clustering

There are 3 steps. Step 2 and 3 are repeated until there is no change.

1. Pick  $K$  centroids
2. Assign sample to the centroid
3. Adjust centroids

**NOTE:** Check [machinelearningcoba.com](http://machinelearningcoba.com).

- This leads to a local optimum, depends on initialization.
- It's sensitive to outliers, detects spherical clusters only.
- Application: e.g., image compression.

### 3.4.6 EM Clustering

Assuming  $N$  data points and  $K$  Gaussians.

- **E-Step:** Fix the Gaussians, find  $\gamma_j(x)$ , which represent the **responsibility of component  $j$  for data point  $x$** .

$$\gamma_j(x_n) = \frac{\pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)}{\sum_{k=1}^K \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)} \quad \forall j = 1, 2, \dots, K, \quad n = 1, 2, \dots, N \quad (3.19)$$

- **M-Step:** Fix  $\gamma_j(x)$ , update the Gaussians.

$$\hat{N}_j = \sum_{n=1}^N \gamma_j(x_n) \quad (3.20)$$

$$\hat{\mu}_j = \frac{1}{\hat{N}_j} \sum_{n=1}^N \gamma_j(x_n) x_n \quad (3.21)$$

$$\hat{\pi}_j = \frac{\hat{N}_j}{N} \quad (3.22)$$

$$\hat{\Sigma}_j = \frac{1}{\hat{N}_j} \sum_{n=1}^N \gamma_j(x_n) (x_n - \hat{\mu}_j) (x_n - \hat{\mu}_j)^T \quad (3.23)$$

**Notes:**

- EM is short for Expectation-Maximization
- Regularization with  $\Sigma + \sigma_{min} I$
- Initialization  $\mu_j$  with K-Means
- Compare with K-Means
  - Hard-assignment  $\Leftrightarrow$  K-Means: each data point to 1 class
  - Soft-assignment  $\Leftrightarrow$  EM Clustering: each data point  $\Rightarrow$  **prob.** to fall into many classes
  - With **more params.**, EM **needs more iteration**.

# 4 Basic ML Problems

## 4.1 Linear Regression

This is the simplest regression problem in [ML](#).

**Problem statement:** Given data points  $\mathbf{x}_i \in \mathbb{R}^D$  and their labels  $y_i \in \mathbb{R}$ , find the *line* that fits these data points. The line is represented via parameters  $\mathbf{w} = [w_0, w_1, \dots, w_n]^T$ . For each data point  $\mathbf{x}$  and its label  $y$ .

**Approach:**

$$\bar{\mathbf{x}} = [1, x_0, \dots, x_n] \quad (\text{x bar - extended variable})$$

$$y \approx \hat{y} = \bar{\mathbf{x}} \cdot \mathbf{w} \quad (\text{y hat - predicted label})$$

$$\Rightarrow \frac{1}{2}e^2 = \frac{1}{2}(y - \bar{\mathbf{x}} \cdot \mathbf{w})^2 \quad (\text{the squared error between true and predicted labels})$$

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (y_i - \bar{\mathbf{x}}_i \cdot \mathbf{w})^2 \quad (\text{the loss function for all points}) \quad (4.1)$$

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2N} \sum_{i=1}^N (y_i - \bar{\mathbf{x}}_i \cdot \mathbf{w})^2 \quad (\text{the average loss}) \quad (4.2)$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \mathcal{L}(\mathbf{w}) \quad (\text{the weights that minimize the loss function}) \quad (4.3)$$

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \|\mathbf{y} - \bar{\mathbf{X}} \cdot \mathbf{w}\|_2^2 \quad (\text{using matrix form}) \quad (4.4)$$

$$\text{with } \bar{\mathbf{X}} = \begin{bmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \\ \vdots \\ \bar{\mathbf{x}}_n \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

**Solution:**

$$\frac{\partial \mathcal{L}(\mathbf{w}^*)}{\partial \mathbf{w}} = \bar{\mathbf{X}}^T (\bar{\mathbf{X}} \mathbf{w}^* - \mathbf{y}) = 0 \quad (4.5)$$

$$\Leftrightarrow \bar{\mathbf{X}}^T \bar{\mathbf{X}} \mathbf{w}^* = \bar{\mathbf{X}}^T \mathbf{y} \quad (4.6)$$

$$\Leftrightarrow \mathbf{w}^* = (\bar{\mathbf{X}}^T \bar{\mathbf{X}})^\dagger \bar{\mathbf{X}}^T \mathbf{y} \quad (4.7)$$

in which,  $A^\dagger$  (A dagger) is the pseudo inverse of a matrix, because  $A$  might not be inverse-able.

$A^\dagger = (A^T A)^{-1} A^T$

**NOTE:**

- Sensitive to outliers  $\Rightarrow$  pre-processing
- Multi-variable:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2N} \|\mathbf{y} - \bar{\mathbf{X}} \cdot \mathbf{w}\|_2^2$$

$$\Rightarrow \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \frac{1}{N} \bar{\mathbf{X}}^T (\bar{\mathbf{X}} \mathbf{w} - \mathbf{y})$$

## **4.2 Linear Discriminant Functions**

**Problem statement:** of general classification problem

- Given: training set  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  with target values (labels)  $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_N\}$ .
- Goal: take a new input  $\mathbf{x}$  and assign it to one of  $K$  classes  $C_k$   
 $\Rightarrow$  Learn a discriminant function  $y(x)$  to perform the classification

**Approach:**

- 2-class problem: Binary target values:  $t_n \in \{0, 1\}$   
E.g. of discriminant function:  $y(x) > 0 \Rightarrow$  class  $C_1$ , else  $C_2$
- K-class problem: 1-of-K coding scheme, e.g.:  $\mathbf{t}_n = [0, 1, 0, 0, 0]^T$
- Extension to multiple classes: one-vs-all, one-vs-one classifiers

### **4.2.1 Least-squares classification**

With above problem statement, consider  $K$  classes described by linear models:

$$y_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0} \quad k = 1, \dots, K$$

$$\Rightarrow \mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^T \widetilde{\mathbf{x}} \quad (\text{using vector notation})$$

$$\mathbf{Y}(\widetilde{\mathbf{X}}) = \widetilde{\mathbf{X}} \widetilde{\mathbf{W}} \quad (\text{for the entire dataset})$$

In which output  $\mathbf{y}$  in 1-of-K notation, and can be compared to the target value

$$\mathbf{t} = [t_1, \dots, t_k]^T$$

**Solution:** minimize the sum-of-squares error  $E(\mathbf{w})$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (\mathbf{y} - \mathbf{t})^2 = \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K (\mathbf{w}_k^T \mathbf{x}_n - t_{kn})^2$$

$$\Rightarrow \widetilde{\mathbf{W}}^* = \left( \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}} \right)^{-1} \widetilde{\mathbf{X}}^T \mathbf{T} = \widetilde{\mathbf{X}}^\dagger \mathbf{T}$$

**NOTE:** Least-squares is very sensitive to outliers!

The error function penalizes predictions that are “too correct”.

### 4.2.2 Generalized Linear Discriminant

$$y_k(x) = \sum_{j=1}^M w_{kj} \phi_j(x) + w_{k0} = \sum_{j=0}^M w_{kj} \phi_j(x), \quad \phi_0(x) = 1 \quad (4.8)$$

### 4.2.3 Generalized Linear Models

- Linear Model:

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 \quad (4.9)$$

- Generalized Linear Model:

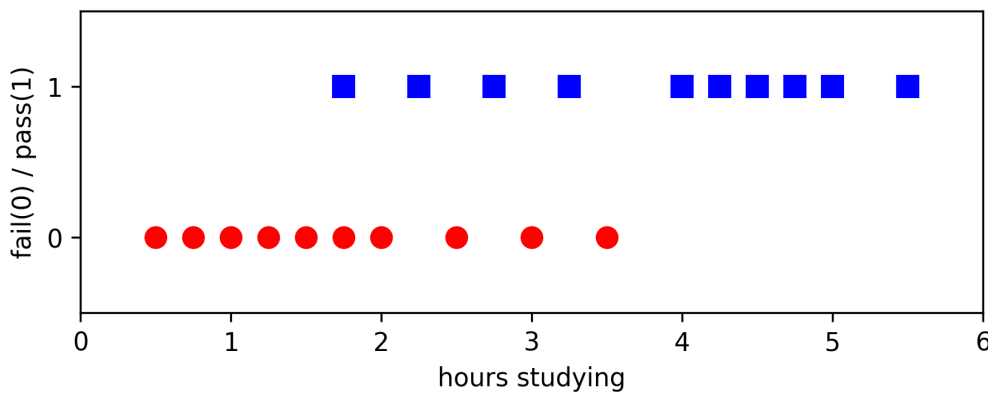
$$y(\mathbf{x}) = g(\mathbf{w}^T \mathbf{x} + w_0) \quad (4.10)$$

In which  $g(\cdot)$  is called an **activation function** and may be nonlinear

## 4.3 Logistics Regression

Though named as "regression", logistics regression is used for classification problem, just like Linear Discriminant (Sec. 4.2).

**Problem:** In many of binary classification problems (classification problem with 2 classes), it's rather hard (or even impossible) to be certain about the class of the output, and the training data is not linear separable. Instead of a hard threshold, we could have a soft one, represent by the [prob.](#) belonging to either classes.



**Figure 4.1:** Example of exam results based on study hours ([src](#)). Given the number of hours, instead of predicting whether fail or pass, the model predicts the [prob.](#) that the student will pass, or fail.

The sigmoid function gives a nice nonlinear transition for the [prob.](#). The further the data point is from the threshold, the higher the [prob.](#) it belongs to one class, and small to the others.

$$f(s) = \frac{1}{1 + e^{-s}} \triangleq \sigma(s) \quad (4.11)$$

$$s = \ln \left( \frac{\sigma}{1 - \sigma} \right) \quad (4.12)$$

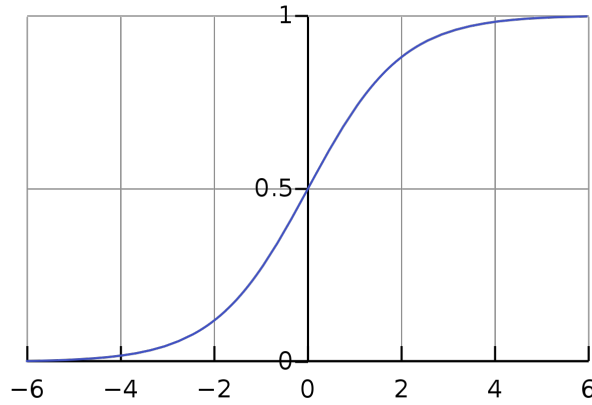
$$\sigma'(s) = \sigma(s) (1 - \sigma(s)) \quad (4.13)$$

Thus, if we use the sigmoid function as the activation function for Eq. 4.10:

$$s = \mathbf{w}^T \mathbf{x} \quad (4.14)$$

$$y = \sigma(s) \quad (4.15)$$

$$\Rightarrow \frac{\partial y}{\partial \mathbf{w}} = \frac{\partial y}{\partial s} \frac{\partial s}{\partial \mathbf{w}} = y(1 - y)\mathbf{x} \quad (4.16)$$



**Figure 4.2:** Sigmoid function ([src](#)).

### Approach:

- Design of the error function.

Assume that the [prob.](#) of data point  $\mathbf{x}$  falls into class 1 is  $f(\mathbf{w}^T \mathbf{x})$  and class 0 is  $1 - f(\mathbf{w}^T \mathbf{x})$ .

With  $z_i = f(\mathbf{w}^T \mathbf{x}_i)$ :

$$\begin{cases} P(y_i = 1 | \mathbf{x}_i; \mathbf{w}) &= f(\mathbf{w}^T \mathbf{x}_i) = z_i \\ P(y_i = 0 | \mathbf{x}_i; \mathbf{w}) &= 1 - f(\mathbf{w}^T \mathbf{x}_i) = 1 - z_i \end{cases} \Leftrightarrow P(y_i | \mathbf{x}_i; \mathbf{w}) = z_i^{y_i} (1 - z_i)^{1 - y_i} \quad (4.17)$$

With the aim to find the model [params.](#) that maximizes the data [prob.](#) (Subsec. 3.1.1):

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} P(\mathbf{y}|\mathbf{X}; \mathbf{w})$$

$$\iff \mathbf{w}^* = - \arg \min_{\mathbf{w}} \log P(\mathbf{y}|\mathbf{X}; \mathbf{w}) \quad (\text{negative log-likelihood})$$

$$P(\mathbf{y}|\mathbf{X}; \mathbf{w}) = \prod_{i=1}^N P(y_i|\mathbf{x}_i; \mathbf{w}) = \prod_{i=1}^N z_i^{y_i} (1 - z_i)^{1-y_i} \quad (\text{independence assumption})$$

$$\Rightarrow -\log P(\mathbf{y}|\mathbf{X}; \mathbf{w}) = -\sum_{i=1}^N [y_i \log z_i + (1 - y_i) \log(1 - z_i)] \quad (\text{the cross entropy error})$$

- Optimization with Stochastic Gradient Descent ([SGD](#)):

$$\mathbf{w} = \mathbf{w} - \eta \cdot \frac{\partial J}{\partial \mathbf{w}}, \quad (\eta \text{ as the learning rate}) \quad (4.18)$$

$$J(\mathbf{w}, \mathbf{x}_i, y) = -(y_i \log z_i + (1 - y_i) \log(1 - z_i)), \quad \text{data point } (\mathbf{x}_i, y_i) \quad (4.19)$$

$$\Rightarrow \frac{\partial J}{\partial \mathbf{w}} = - \left( \frac{y_i}{z_i} - \frac{1 - y_i}{1 - z_i} \right) \frac{\partial z_i}{\partial \mathbf{w}} = \frac{z_i - y_i}{z_i(1 - z_i)} \frac{\partial z_i}{\partial \mathbf{w}} \quad (4.20)$$

$$\frac{\partial z_i}{\partial \mathbf{w}} = z_i(1 - z_i)\mathbf{x}_i \quad (\text{the beauty of sigmoid function, Eq. 4.16}) \quad (4.21)$$

$$\Rightarrow \frac{\partial J}{\partial \mathbf{w}} = (z_i - y_i)\mathbf{x}_i \quad (4.22)$$

$$\Rightarrow \mathbf{w} = \mathbf{w} + \eta(y_i - z_i)\mathbf{x}_i \quad (\text{replace into Eq. 4.18}) \quad (4.23)$$

**NOTE:** Require less [params.](#), only  $D$  with  $D$  as the number of ([no.](#)) dimensions, compared to Gaussians with  $\left\lceil \frac{M(M+5)}{2} + 1 \right\rceil$  [params.](#).

## 4.4 Softmax Regression

Softmax Regression is the generalization of Logistics Regression for multiple-class classification problem. Thus, it is also known as: **Multinomial Logistics Regression, Maximum Entropy Classifier**. Logistics Regression can only be applied for binary classification problem. Given  $C$  classes, we would need multiple one-vs-all or one-vs-one classifiers. Softmax regression offers a better alternative.

$$z_i = \mathbf{w}^T \mathbf{x}_i \quad (4.24)$$

$$a_i = \frac{\exp(z_i)}{\sum_{j=1}^C \exp(z_j)} \quad (4.25)$$

$$\begin{cases} a_i > 0, & a_i \text{ represents the } \text{prob.} \text{ the input belongs to class } C_i \\ \sum a_i = 1 \\ z_m > z_n \iff a_m > a_n \quad (\text{order}) \end{cases} \quad (4.26)$$

There are cases that one output  $z_i$  is significantly greater than the others, which will lead to numerical error while coding. In these cases, the following Softmax version is more stable:

$$a_i = \frac{\exp(z_i)}{\sum_{j=1}^C \exp(z_j)} = \frac{\exp(z_i - c)}{\sum_{j=1}^C \exp(z_j - c)}, \quad c = \max_i z_i \quad (4.27)$$

For the one hot coding representation, for each input  $\mathbf{x}$ , the softmax regression model calculate a output vector  $\mathbf{a} = \mathbf{W}^T \mathbf{x}$ . The loss function will be build to represent the difference between output  $\mathbf{a}$  and the real label  $\mathbf{y}$ . Vector  $\mathbf{a}$  itself is a [prob.](#) distribution of the input belongs to different classes. Instead of choosing the squared error function, cross-entropy is a better alternative to represent the difference between two [prob.](#) distributions (Sec. ??). As mentioned before,  $\mathbf{q} > 0$ , thus, we can't choose  $\mathbf{q} = \mathbf{y}$ .

$$J(\mathbf{W}, \mathbf{x}_i, \mathbf{y}_i) = - \sum_{i=1}^N \sum_{j=1}^C y_{ij} \log(a_{ij}) \quad \text{for each data point } (\mathbf{x}_i, \mathbf{y}_i) \quad (4.28)$$

$$\frac{\partial J_i(\mathbf{W})}{\partial \mathbf{W}} = \mathbf{x}_i \mathbf{e}_i^T = \mathbf{x}_i (\mathbf{a}_i - \mathbf{y}_i)^T \quad (\text{the gradient}) \quad (4.29)$$

$$\mathbf{W} = \mathbf{W} + \eta \mathbf{x}_i (\mathbf{y}_i - \mathbf{a}_i)^T \quad (\text{SGD, } \text{machinelearningcoban.com}) \quad (4.30)$$

## 4.5 SVM

### SVM

#### 4.5.1 Hard Margin SVM

**Problem:** Give  $N$  data points  $\mathbf{x}_i$  and their labels  $y_i \in \{-1, 1\}$ , find the line that separate these points into two classes while maximize the margin.

$$\text{margin} = \min_n \frac{y_n(\mathbf{w}^T \mathbf{x}_n) + b}{\|\mathbf{w}\|_2} \quad (4.31)$$

$$\Rightarrow (\mathbf{w}, b) = \arg \max_{\mathbf{w}, b} \left\{ \min_n \frac{y_n(\mathbf{w}^T \mathbf{x}_n) + b}{\|\mathbf{w}\|_2} \right\} \quad (4.32)$$

$$\Rightarrow (\mathbf{w}, b) = \arg \max_{\mathbf{w}, b} \left\{ \frac{1}{\|\mathbf{w}\|_2} \min_n [y_n(\mathbf{w}^T \mathbf{x}_n + b)] \right\} \quad (4.33)$$

For simplicity, we can assume that:  $\min_n [y_n(\mathbf{w}^T \mathbf{x}_n + b)] = 1$ . Thus, the original problem can be transform to:

**Problem:** Find  $(\mathbf{w}, b) = \arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2$  that subject to:

$$1 - y_n(\mathbf{w}^T \mathbf{x}_n + b) \leq 0, \quad \forall n = 1, 2, \dots, N$$



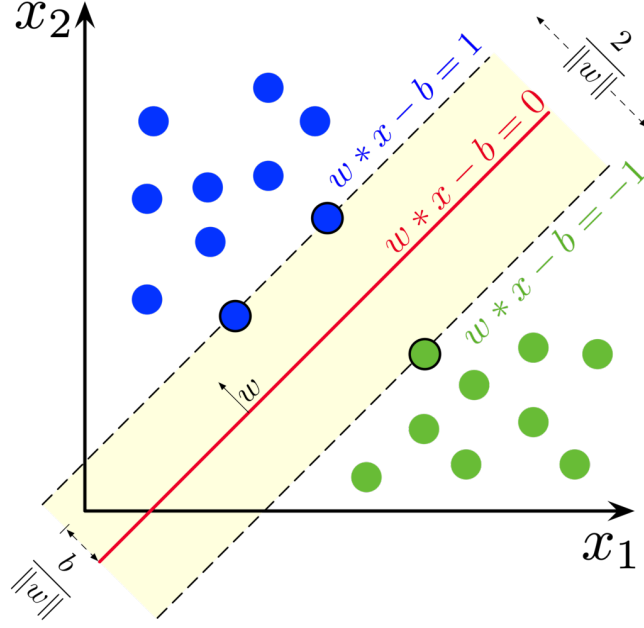


Figure 4.3: Example of SVM (src).

This is a well-studied optimization problem, solved with Lagrange's method.

**Approach:** The Lagrange's method

**The primal formulation of SVM:**

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_{n=1}^N \lambda_n [1 - y_n(\mathbf{w}^T \mathbf{x}_n + b)], \quad \lambda_n \geq 0 \quad \forall n \quad (4.34)$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda})}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^N \lambda_n y_n \mathbf{x}_n \quad (4.35)$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda})}{\partial b} = 0 \Rightarrow \sum_{n=1}^N \lambda_n y_n = 0 \quad (4.36)$$

**The dual formulation of SVM:**

$$\Rightarrow g(\boldsymbol{\lambda}) = \sum_{n=1}^N \lambda_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m \quad (4.37)$$

$$\text{Set } \mathbf{V} = [y_1 \mathbf{x}_1, y_2 \mathbf{x}_2, \dots, y_N \mathbf{x}_N] \quad (4.38)$$

$$\Rightarrow g(\boldsymbol{\lambda}) = -\frac{1}{2} \boldsymbol{\lambda}^T \mathbf{V}^T \mathbf{V} \boldsymbol{\lambda} + \mathbf{1}^T \boldsymbol{\lambda} \quad \text{is } \underline{\text{concave}} \text{ with } \lambda_i \geq 0, \quad \sum_{n=1}^N \lambda_n y_n = 0 \quad (4.39)$$

**$\Rightarrow$  Find  $\lambda$  by solving Quadratic Programming**

Then use Karush-Kuhn-Tucker (KKT) conditions to find  $\mathbf{w}, b$ :

$$\mathbf{S} = \{\mathbf{n} \mid \lambda_n \neq 0\} \Rightarrow \begin{cases} b = \frac{1}{N_S} \sum_{n \in S} (y_n - \sum_{m \in S} \lambda_m y_m \mathbf{x}_m^T \mathbf{x}_n) \\ \mathbf{w} = \sum_{m \in S} \lambda_m y_m \mathbf{x}_m \end{cases}$$

## 4.5.2 Soft Margin SVM

- Hard margin SVM:

$$(\mathbf{w}, b) = \arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2$$

$$\text{subject to } y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \quad \forall n$$

- Soft margin SVM:

$$(\mathbf{w}, b, \xi) = \arg \min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{n=1}^N \xi_n$$

$$\text{subject to } \begin{cases} y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n \\ \xi_n \geq 0 \end{cases} \quad \forall n$$

**$C$  and the margins are in-proportional:**  $\begin{cases} C \uparrow \Rightarrow \text{margin} \downarrow \\ C \downarrow \Rightarrow \text{margin} \uparrow \end{cases}$

The "Soft" constraints:

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n \quad \Longleftrightarrow \quad 1 - \xi_n - y_n(\mathbf{w}^T \mathbf{x}_n + b) \leq 0 \quad \forall n$$

**Problem:** Find  $(\mathbf{w}, b, \xi) = \arg \min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{n=1}^N \xi_n$  that subject to:

$$\begin{cases} 1 - \xi_n - y_n(\mathbf{w}^T \mathbf{x}_n + b) \leq 0 \\ -\xi_n \leq 0 \end{cases} \quad \forall n$$

**Approach:** Lagrange's method with  $\lambda_i \geq 0$  and  $\mu_i \geq 0$ :

$$\mathcal{L}(\mathbf{w}, b, \xi, \lambda, \mu) = \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{n=1}^N \xi_n + \sum_{n=1}^N \lambda_n [1 - \xi_n - y_n(\mathbf{w}^T \mathbf{x}_n + b)] - \sum_{n=1}^N \mu_n \xi_n \quad (4.40)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \Longleftrightarrow \mathbf{w} = \sum_{n=1}^N \lambda_n y_n \mathbf{x}_n \quad (4.41)$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Longleftrightarrow \sum_{n=1}^N \lambda_n y_n = 0 \quad (4.42)$$

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = 0 \Longleftrightarrow \lambda_n = C - \mu_n \quad (4.43)$$

$$g(\lambda, \mu) = \inf_{\mathbf{w}, b, \xi} \mathcal{L}(\mathbf{w}, b, \xi, \lambda, \mu) \quad (4.44)$$

$$\Rightarrow g(\lambda, \mu) = \sum_{n=1}^N \lambda_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m = g(\lambda) \quad (4.45)$$

$$\Rightarrow \lambda = \arg \max_{\lambda} g(\lambda) \quad \text{subject to} \quad \begin{cases} \sum_{n=1}^N \lambda_n y_n = 0 \\ 0 \leq \lambda_n \leq C \quad \forall n \end{cases} \quad (4.46)$$

After finding  $\lambda$ :

$$\mathcal{M} = \{n \mid 0 < \lambda_n < C\} \quad \Rightarrow b = \frac{1}{N_M} \sum_{n \in \mathcal{M}} \left( y_n - \sum_{m \in \mathcal{S}} \lambda_m y_m \mathbf{x}_m^T \mathbf{x}_n \right) \quad (4.47)$$

$$\mathcal{S} = \{m \mid 0 < \lambda_m \leq C\} \quad \Rightarrow \mathbf{w} = \sum_{m \in \mathcal{S}} \lambda_m y_m \mathbf{x}_m \quad (4.48)$$

- $\begin{cases} \lambda_n = 0 \\ \xi_n = 0 \end{cases} \Rightarrow \text{Safe points}$
- $\begin{cases} 0 < \lambda_n < C \\ \xi_n = 0 \end{cases} \Rightarrow \text{Marginal points}$
- $\lambda_n = C \quad \begin{cases} \xi_n \leq 1 & \Rightarrow \text{still correct} \\ \xi_n > 1 & \Rightarrow \text{incorrect} \end{cases}$

### 4.5.3 Kernel Support Vector Machine

**Mercer's Conditions:** kernel function ([func.](#)) must theoretically satisfy.

In practice, however, some acceptable kernels don't satisfy the Mercer's conditions.

$$\sum_{n=1}^N \sum_{m=1}^N k(x_n, x_m) c_n c_m \geq 0 \quad \forall c_i \in \mathbb{R}, i = 1, 2, \dots, N \quad (4.49)$$

E.g. kernel functions:

$k(x, z) = x^T z$	<b>linear</b>
$k(x, z) = (\gamma x^T z + r)^d$	<b>polynomial</b>
$k(x, z) = \exp(-\gamma \ x - z\ _2^2), \quad \gamma > 0$	<b>RBF</b>
$k(x, z) = \tanh(\gamma x^T z + r)$	<b>sigmoid</b>

# 5 Ensembles

Ensemble of Models are also known as additive models.

## 5.1 Error Reduction

[TODO: Explanation]

E.g.: bagging

$$y_{COM}(x) = \frac{1}{M} \sum_{m=1}^M y_m(x) \quad (5.1)$$

$$y(x) = h(x) + \varepsilon(x) \quad (5.2)$$

$$\mathbb{E}_x = [y_m(x) - h(x)]^2 = \mathbb{E}_x [\varepsilon_m(x)^2] \quad (5.3)$$

$$\Rightarrow \mathbb{E}_{AV} = \frac{1}{M} \sum_{m=1}^M \mathbb{E}_x [\varepsilon_m(x)^2] \quad (5.4)$$

$$\mathbb{E}_{COM} = \mathbb{E}_x \left[ \left\{ \frac{1}{M} \sum_{m=1}^M y_m(x) - h(x) \right\}^2 \right] = \mathbb{E}_x \left[ \left\{ \frac{1}{M} \sum_{m=1}^M \varepsilon_m(x) \right\}^2 \right] \quad (5.5)$$

$$\Rightarrow \text{if } \begin{cases} \text{errors have 0 mean:} & \mathbb{E}_x[\varepsilon_m(x)] = 0 \\ \text{errors are uncorrelated:} & \mathbb{E}_x[\varepsilon_m(x)\varepsilon_j(x)] = 0 \text{ (unrealistic??)} \end{cases} \Rightarrow \mathbb{E}_{COM} = \frac{1}{M} \mathbb{E}_{AV}$$

However, in general,  $\mathbb{E}_{COM} < \mathbb{E}_{AV}$

**NOTE:** The weak classifiers have to be unstable algorithms, i.e., decision trees, neural network;  
**NOT** nearest neighbor, SVM, linear regression.

## 5.2 Bagging

- also known as (a.k.a.) Bootstrap aggregating
- Average  $\approx 63\%$

[TODO: Image]

- Split the given data set into train and test set
- Assume train set with  $n$  data points
- Pick  $n'$  data points into each bag  $D_i$ ,  $\frac{n'}{n} < 1 (\approx 60\%)$

- Create  $m$  data bags:  $D_1, D_2, \dots, D_m$
- Learn a model  $M_i$  from each bag  $D_i$
- The final output is the average of models' outputs:

$$y = \frac{1}{m} \sum_{i=1}^m f(M_i)$$

- **Random with replacement:** bag  $i$  can have multiple times data point  $\mathbf{x}_i$
- **Simple, easy to implement, commonly used**

**NOTE:** Practically, resampling with replacement is usually unnecessary, since SGD and random initialization usually makes the models sufficiently independent.

## 5.3 Boosting

- Adaboost
- Gradient boosting
- Xgboost (extreme gradient boosting)
- Gentle Boost (cross entropy error)

### 5.3.1 Adaboost

1. First weak classifier
2. Calculate error  $J \Rightarrow \varepsilon$   
in case of normalization  $w \Rightarrow \sum w = 1 \Rightarrow \varepsilon = J$
3. Calculate amount of say  $\alpha$  from  $\varepsilon$
4. Update  $w$  with  $\alpha$  (normalize  $w$ )
5. 2nd weak classifier with new weight  $w$  or sample on new  $w$

- Ensembles of Classifiers:  $K$  independent classifiers, error [prob.](#)  $< 0.5$
- Suitable for unstable algorithms (i.e., decision trees, neural networks)
- Not good with stable methods (i.e., nearest neighbors, [SVMs](#), linear regression)

[TODO: EXPLANATION?]

$$-g(x_i) = -\frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)} \quad (5.6)$$

$$\Rightarrow \text{Regression model: } h_i \text{ for } (x_i, -g(x_i)) \quad (5.7)$$

$$\Rightarrow F := F + \rho h, \quad \rho = 1 \quad (5.8)$$

[TODO: EXPLANATION?]

**Problem:**

## 5 Ensembles

- Given:  $N$  data points and their labels  $(x_n, t_n)$
- Goal: find  $M$  weak classifiers  $h_m(x)$

$$w_n^{(0)} = \frac{1}{N} \quad \text{initial data weights} \quad (5.9)$$

$$J_m = \sum_{n=1}^N w_n^{(m)} I(h_m(x) \neq t_n) \quad \text{weighted error function} \quad (5.10)$$

$$\varepsilon_m = \frac{\sum_{n=1}^N w_n^{(m)} I(h_m(x) \neq t_n)}{\sum_{n=1}^N w_n^{(m)}} \quad \text{(normalized) weighted error} \quad (5.11)$$

$$\alpha_m = \ln \left( \frac{1 - \varepsilon_m}{\varepsilon_m} \right) \quad \text{classifier weights} \quad (5.12)$$

$$w_n^{(m+1)} = w_n^{(m)} \cdot \exp \{ \alpha_m I(h_m(x) \neq t_n) \} \quad \text{updated data weights} \quad (5.13)$$

$$H(x) = \text{sign} \left( \sum_{m=1}^M \alpha_m h_m(x) \right) \quad \text{final classification} \quad (5.14)$$

[TODO: EXPLANATION?] **Adaboost**

$$F(x) = \text{sign} \left( \sum_{m=1}^M \theta_m f_m(x) \right) \quad (5.15)$$

$$w(x_i, y_i) = \frac{1}{n} \quad \text{initial weights for each data} \quad (5.16)$$

$$\varepsilon_m = \mathbb{E}_{\varepsilon_m} [1_{y \neq f(x)}] \quad (5.17)$$

$$\Rightarrow \theta_m = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_m}{\varepsilon_m} \right) \quad \text{Update rule} \quad (5.18)$$

$$w_{m+1}(x_i, y_i) = \frac{w_m(x_i, y_i) \exp[-\theta_m y_i f_m(x_i)]}{z_m} \quad \text{update weights} \quad (5.19)$$

$$(z_m \text{ is the normalization factor}) \quad (5.20)$$

[TODO: Explanation:] **Exponential error used in Adaboost**

- Fast convergence
- No penalty for too correct
- Less robust to outlier

[TODO: Add function graph]

### 5.3.2 Gradient Boosting (Regression)

$$(x_i, y_i) \text{ and first model } F_1(x) \quad (5.21)$$

$$\Rightarrow \text{fit } F_2(x) \text{ to } (x_i, y_i - F_1(x_i)), \quad y_i - F_1(x_i) = h(x_i) \text{ as the residuals} \quad (5.22)$$

$$J = \sum \mathcal{L}(y_i, F(x_i)) \quad (5.23)$$

$$\Rightarrow F_{i+1} = F_i + h_i \quad (5.24)$$

$$\mathcal{L} = \frac{(y_i - F_i)^2}{2} \Rightarrow -g = y_i - F_i(x_i) \quad (5.25)$$

$$\mathcal{L} = |y - F| \quad (\text{absolute loss}) \quad (5.26)$$

$$\mathcal{L} = \begin{cases} \frac{1}{2}(y - F)^2 & \text{if } |y - F| \geq \delta \\ \delta \left( |y - F| - \frac{\delta}{2} \right) & \text{if } |y - F| < \delta \end{cases} \quad (\text{Huber loss}) \quad (5.27)$$

The two loss functions Eq. 5.26 and Eq. 5.27 are less sensitive to outliers

## 5.4 Information

Mutual Information: [YouTube](#), [YouTube](#).

## 5.5 Decision Trees

Iterative Dichotomiser 3: [MLCoBan](#) Detail Example calculation: [medium.com](#), [blog](#)

CART: Gini Index: [blog](#)

Random forest: [YouTube](#)

[**TODO: Add image**] root node, non-leaf node has 2 (or more) child node, leaf/terminal node.

All non-leaf nodes have 2 child nodes  $\Rightarrow$  binary decision tree

Iterative Dichotomiser 3 (ID3) only for categorical attribute (discrete)

Classification and Regression Tree (CART) for both categorical and continuous.

6 questions:

- Bin / multi valued
- when node  $\rightarrow$  leaf
- Deal impure nodes?
- how to select query
- pruned?

- missing attribute?

Impurity measures:

- Misclassification:  $i(s_j) = 1 - \max_k p(C_k|s_j)$
- Information gain:  $C$  classes with  $N_C$  as the number of members in each class

$$H(S) = - \sum_{c=1}^C \frac{N_C}{N} \log \left( \frac{N_C}{N} \right) \quad \text{entropy at a node} \quad (5.28)$$

Choose attribute  $X \Rightarrow K$  child nodes:  $S_1, S_2, \dots, S_k$  with  $m_k$  elements

$$H(x, S) = \sum_{k=1}^K \frac{m_k}{N} H(S_k) \quad \text{entropy sum with weights} \quad (5.29)$$

$$\Rightarrow G(x, S) = H(S) - H(x, S) \quad \text{information gain} \quad (5.30)$$

$$x^* = \arg \max_x G(x, S) = \arg \min_x H(x, S) \quad (5.31)$$

**Reduction in entropy = gain in information**

- Gini Index:

$$i(s_j) = \sum_{k \neq l} p(C_k|s_j)p(C_l, s_j) \quad \text{variance impurity at node } s_j \quad (5.32)$$

$$= \frac{1}{2} \left[ 1 - \sum_k p^2(C_k, s_j) \right] \quad (5.33)$$

$$H(x, s_j) = \sum_{k=1}^K \frac{m_k}{N} i(s_j) \quad (5.34)$$

$$\Rightarrow x^* = \min H(x, s_j) \quad (5.35)$$

- Chi-square:  $= \sqrt{\frac{(\text{actual} - \text{expected})^2}{\text{expected}}}$ ,  $\arg \max$   
CHAID: Chi-square Automatic Interaction Detector
- Reduction in Variance,  $\arg \min$

$$i(s_j) = \frac{\sum (x - \bar{x})^2}{n} \quad (5.36)$$

[TODO: Add image]

**Stopping and Pruning is more IMPORTANT**

### 5.5.1 Stopping Conditions

- Prevent over-fitting = **Prepruning**
- $H(\mathcal{S}) = 0$ : entropy = 0  $\Rightarrow \forall$  points  $\in$  each class
- **no.** members in each node  $<$  a certain threshold  
The leaf node's class = the dominant class
- Distance from node  $\rightarrow$  root =  $c \Rightarrow$  to limit the tree depth



- The total **no.** of node exceed a threshold
- Further expansion induces insignificant entropy reduction

### 5.5.2 Pruning

#### Post-pruning

1. Validation Set: Prune a node if it increases the **precision** for VS (Validation Set). This is also called Reduction error pruning method
2. Regularized loss function

$$\mathcal{L} = \sum_{k=1}^K \frac{|\mathcal{S}_k|}{\mathcal{S}} H(\mathcal{S}_k) + \lambda K \quad (5.37)$$

- Minimum error: sensitive to **no.** classes, the least accurate in practice
- Pessimistic: the most crude and the quickest, no need for validation set, but extra caution needed
- Error complexity:  $R(t) = r(t)p(t) + \alpha N_t$ , with  $r(t)$  as the error,  $N_t$  as the **no.** leafs
- Critical value: the value we choose to decide the query at each nodes  
 $\Rightarrow$  depends on how the tree is created
- Reduced error: ...

**NOTE:** The last three error functions are more stable and accurate

### 5.5.3 Computational Complexity

Given:

$$\begin{cases} N \text{ data points} \\ D \text{ dimensions} \end{cases} \Rightarrow \begin{cases} \text{Storage: } \mathcal{O}(N) \\ \text{Test time: } \mathcal{O}(\log N) \\ \text{Training time: } \mathcal{O}(DN^2 \log N) \end{cases}$$

#### **Explanation:**

- Test time: in the worst case scenario, each leaf has one data  
 $\Rightarrow$  After  $k$  step, the number of nodes are:  $2^k = N$   
 $\Rightarrow k \in \mathcal{O}(\log N)$
- Training time:  
 maximum  $\mathcal{O}(DN \log N)$  times each node  
 maximum  $N$  nodes

### 5.5.4 Summary

- Simple
- Interpretable results
- Resistance to overfitting
- **Memory consumption  $\Rightarrow$  suitable for problems with little data**
- Noisy weak classifiers not generated well
- Sensitive to outliers ??
- Expensive learning step

## 5.6 Random Forest

- Handle missing values while maintaining accuracy
- Won't overfit
- Not good with regression
- Have little control to modify

Steps:

- Sample from training set
- Choose  $m < M$  (input features). At each node, select  $m$  random data from  $M$  to decide query attribute to split the node
- Grow tree to the largest, no pruning
- Predict data output:
  - classification  $\Rightarrow$  majority vote
  - regression  $\Rightarrow$  average

Choose randomly  $K$  attributes:

Training time:  $\mathcal{O}(KN^2 \log N)$ ,  $K \ll D$  ( $K = \sqrt{N_\delta}$ )

Typically:  $\begin{cases} K = 10 & \text{root node} \\ K = 100d & \text{level-}d \text{ node} \end{cases}$

## 5.7 Bayesian Model Averaging

Given  $H$  different models with prior **prob.**  $p(h)$ , the final output would be the weighted average of these models:

$$p(X) = \sum_{h=1}^H p(X|h)p(h) \quad (5.38)$$

## 5.8 Distillation

Ensemble models is usually more robust than single models. However, the big question is:

*Can we make a single model that is as good as an ensemble?*

⇒ **Distillation:** train on the ensemble's predictions as "soft" targets [HVD+15]

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (5.39)$$

**Intuition:** more knowledge in soft targets than hard labels!

# 6 Feature Engineering

[TODO: Add explanation]

Dimensionality Reduction is an important technique in [ML](#). Actual feature vectors can be in great dimension, in great number. Thus, filtering is crucial for storage, calculation. Dimension reduction is necessary and also useful in data compression.

[TODO: ]

## 6.1 Principle Component Analysis

Principal Component Analysis ([PCA](#)) is a method for Feature Extraction

*Learning Resources:*

- [setosa.io](#): for Visualization

## 6.2 Linear Discriminant Analysis

Linear Discriminant Analysis ([LDA](#))

[blog](#)

## 6.3 Word Representation

The idea of represent a word as a vector, but not as one-hot coding. With this representation

- similar words have similar vector values
- similar word's relationships have similar vector values

### 6.3.1 The trigram (n-gram method)

- Huge amount of n-tuples of words  $\Rightarrow$  predict relative [prob.](#)
- Problems: scalability, observability

### 6.3.2 Word Embedding

$$\mathbf{x}_{V \times 1} \longrightarrow \mathbf{W}_{V \times d} \longleftrightarrow \mathbf{h}_{D \times 1} \quad (6.1)$$

1 of K encoding

### 6.3.3 word2vec

- CBOW: **syntactic (grammar)**  
Only care which word occurs  
Don't care about order of occurrence
- SKIP Gram **semantic (meaning)**  
less weight to more distance word

[TODO: Add image]

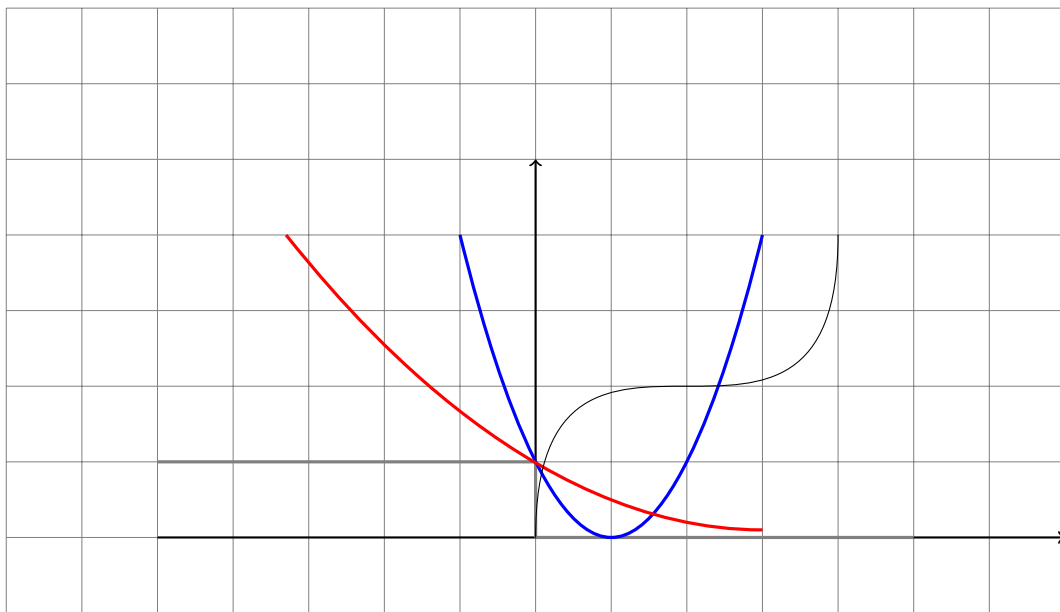
### 6.3.4 Hierarchical Softmax

Organizing words in binary search tree

# 7 Error Functions

As mention in Sec. 2.4, the loss function is the metric to assess the performance of the ML model with a certain task. Different loss functions fit with different tasks. In many cases, the design of the loss function is what makes the differences.

[TODO: Add graph and explanation]



**Figure 7.1:** Different error functions: Ideal Miss-classification Error (gray line)

## 7.1 Ideal Miss-classification Error

Gradient = 0  $\Rightarrow$  can't use gradient descent.

It simply counts incorrectly classified points.

## 7.2 Squared Error - $L_2$ Loss

- Leads to closed form solutions
- Sensitive to outliers
- Penalize "too correct" data points

### 7.3 Cross Entropy Error

- Concave function  $\Rightarrow$  unique minimum exists
- Robust to outliers, error increases only roughly linear
- No closed-form solution, requires iterative method

### 7.4 Squared Error on Sigmoid / Tanh

- No penalty for "too correct" points
- Zero gradient for confidently incorrect classifications

$\Rightarrow$  **Do NOT** use  $L_2$  loss with sigmoid outputs, instead, use cross-entropy.

### 7.5 Hinge Error

- Robust to outliers
- Zero error for points outside margin  $\Rightarrow$  sparsity
- Not differentiable around  $z_n = 1$

**NOTE:** Want the correct class to have a score that is higher than incorrect class by a fixed margin  $\Delta$ .

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta) \quad (7.1)$$

in which,  $s_j$  is other classes score,  $s_{y_i}$  is real class score.

### 7.6 $L_1, L_0$ Loss

Median, no wrong points

$$L_1 = \sum |t - y| \quad (7.2)$$

$$L_2 = \sum (t - y)^2 \quad (7.3)$$

### 7.7 Average Loss

Mathematically, dividing the loss by the amount of data  $N$  (in each batch or epoch) doesn't have any effect on the result. However, it's usually advisable to take the average to have more meaningful judgment and avoiding overflow when there are numerous data points.

# 8 Neural Network

Deep neural network takes care of the complex feature engineering process (Chap. 6). E.g., in classical computer vision, for people detection, the following process is applied:

1. Input: Image
2. Low-level feature extraction: Histogram of Oriented Gradients (HOG)
3. Mid-level features: Deformable Part Model (DPM)
4. Classifier: SVM
5. Output: final label

Using deep neural network, the process is truncated to simply:

1. Input: Image
2. Network training (end-to-end)
3. Output: final label

Different tasks require special expertise to design feature extraction, e.g., designing a program playing gammon would need someone knowing the rules, tips and tricks. On the other hands, we can assured that the human-proposed features are sufficient and helpful. Deep neural network alleviate the human-effort of hand-designing feature extraction process. In addition, it also learns prioritize important features for specific task.

## 8.1 General

**Forward pass:**

$$\mathbf{y}^{(0)} = \mathbf{x} \tag{8.1}$$

$$\mathbf{z}^{(k)} = \mathbf{W}^{(k)}\mathbf{y}^{(k-1)}, \quad k = 1, \dots, l \tag{8.2}$$

$$\mathbf{y}^{(k)} = g_k(\mathbf{z}^{(k)}) \tag{8.3}$$

$$\mathbf{y} = \mathbf{y}^{(l)} \tag{8.4}$$

$$E = L(\mathbf{t}, \mathbf{y}) + \lambda\Omega(\mathbf{W}) \tag{8.5}$$



**Backward pass:**

$$h \leftarrow \frac{\partial E}{\partial \mathbf{y}} = \frac{\partial}{\partial \mathbf{y}} L(\mathbf{t}, \mathbf{y}) + \lambda \frac{\partial}{\partial \mathbf{y}} \Omega \quad (8.6)$$

$$\text{for } k = l \leftarrow 1 : \quad (8.7)$$

$$h \leftarrow \frac{\partial E}{\partial \mathbf{z}^{(k)}} = h \odot g(\mathbf{y}^{(k)}) \quad (8.8)$$

$$\frac{\partial E}{\partial \mathbf{w}^{(k)}} = h \mathbf{y}^{(k-1)T} + \lambda \frac{\partial \Omega}{\partial \mathbf{w}^{(k)}} \quad (8.9)$$

$$h \leftarrow \frac{\partial E}{\partial \mathbf{y}^{(k-1)}} = \mathbf{W}^{(k)T} h \quad (8.10)$$

## 8.2 Gradient Descent

**NOTE:** Just use ADAM??

### 8.2.1 Vanilla Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} f(\theta_t) \quad (8.11)$$

Check derivative:

$$f'(x) \approx \frac{f(x + \varepsilon) - f(x - \varepsilon)}{2\varepsilon} \quad (\text{numerical gradient}) \quad (8.12)$$

### 8.2.2 Momentum

- Init:  $v_{dW_0} = 0, v_{db_0} = 0$
- Calculate  $dW, db$
- Update  $W, b$

$$\Rightarrow \begin{cases} v_{dW} &= \beta v_{dW} + (1 - \beta) dW \\ v_{db} &= \beta v_{db} + (1 - \beta) db \end{cases} \Rightarrow \begin{cases} W &= W - \alpha v_{dW} \\ b &= b - \alpha v_{db} \end{cases} \quad (8.13)$$

The above formulas are to calculate the moving average of  $v_{dW}$  and  $v_{db}$ .

- Tips: Choose  $\beta_1 = \mathbf{0.9}$ , implying taking average of the last 10 steps.
- Reference source: [DeepLearning.AI](#).

### 8.2.3 Nesterov Accelerated Gradient

Nesterov Accelerated Gradient ([NAG](#)):

$$v_t = \gamma v_{t-1} + \eta \nabla_t J(\theta - \gamma v_{t-1}) \quad (8.14)$$

### 8.2.4 RMSprop

- Init  $s_{dW_0} = 0, s_{db_0} = 0$
- Calculate  $dW, db$
- Update  $W, b$

$$\begin{cases} s_{dW} = \beta s_{dW} + (1 - \beta) dW^2 \\ s_{db} = \beta s_{db} + (1 - \beta) db^2 \end{cases} \Rightarrow \begin{cases} W = W - \alpha \frac{dW}{\sqrt{s_{dW} + \varepsilon}} \\ b = b - \alpha \frac{db}{\sqrt{s_{db} + \varepsilon}} \end{cases} \quad (8.15)$$

- Tips: choose  $\beta_2 = 0.999$ ,  $\varepsilon = 10^{-7}$
- Reference source: [DeepLearning.AI](#).

### 8.2.5 Adam

[Adam](#) is basically the combination of Momentum and [RMSprop](#).

- Init  $v_{dW_0}, s_{dW_0}, v_{db_0}, s_{db_0} = 0$
- Calculate  $dW, db$
- Update  $W, b$

$$\begin{cases} v_{dW} = \beta_1 v_{dW} + (1 - \beta_1) dW \\ v_{db} = \beta_1 v_{db} + (1 - \beta_1) db \\ s_{dW} = \beta_2 s_{dW} + (1 - \beta_2) dW^2 \\ s_{db} = \beta_2 s_{db} + (1 - \beta_2) db^2 \end{cases} \Rightarrow \begin{cases} v_{dW}^{cor.} = \frac{v_{dW}}{1 - \beta_1^t} \\ v_{db}^{cor.} = \frac{v_{db}}{1 - \beta_1^t} \\ s_{dW}^{cor.} = \frac{s_{dW}}{1 - \beta_2^t} \\ s_{db}^{cor.} = \frac{s_{db}}{1 - \beta_2^t} \end{cases} \Rightarrow \begin{cases} W = W - \alpha \frac{v_{dW}^{cor.}}{\sqrt{s_{dW}^{cor.} + \varepsilon}} \\ b = b - \alpha \frac{v_{db}^{cor.}}{\sqrt{s_{db}^{cor.} + \varepsilon}} \end{cases} \quad (8.16)$$

- Tips: choose  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\varepsilon = 10^{-7}$
- Reference source: [DeepLearning.AI](#).

## 8.3 Tips and Tricks

- Shuffling
- Data Augmentation: reshape, rescale, crops, zooming, change color (color [PCA](#))
- Normalizing the inputs

Convergence is the fastest if

- The mean of each input variable = 0
- Scale  $\Rightarrow$  same covariance

Mean cancellation  $\Rightarrow$  Kullback–Leibler ([K-L](#)) expansion  $\Rightarrow$  covariance equalization (if possible)

- Leaky Rectified Linear Unit ([ReLU](#)) is better a bit than [ReLU](#), ELU

- Weights initialization: Xavier-Glorot:

$$W \sim U\left(0, \sqrt{\frac{6}{n_{in} + n_{out}}}\right)$$

- Batch Norm(alization): Normalize after each layer  
⇒ learn the moving average
- Drop out

**NOTE:** When in inferencing (after training), must multiply the activation output with the [prob.](#) that the weights are set to 0

## 8.4 Network Structures

This section present some common network structures and their capabilities.

### 8.4.1 Recurrent Neural Network

Backpropagation through time ([BPTT](#)):

[**TODO:** Add image, content]

$$\frac{\partial E_t}{\partial w_{ij}} = \sum_{1 \leq k \leq t} \left( \frac{\partial E_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial w_{ij}} \right) \quad (8.17)$$

$$E = \sum_{1 \leq t \leq T} E_t \quad (8.18)$$

$$\frac{\partial h_t}{\partial h_k} = \prod_{t \geq i \geq k} \frac{\partial h_i}{\partial h_{i-1}} \quad (8.19)$$

### 8.4.2 Residual Network

[**TODO:** Residual:]

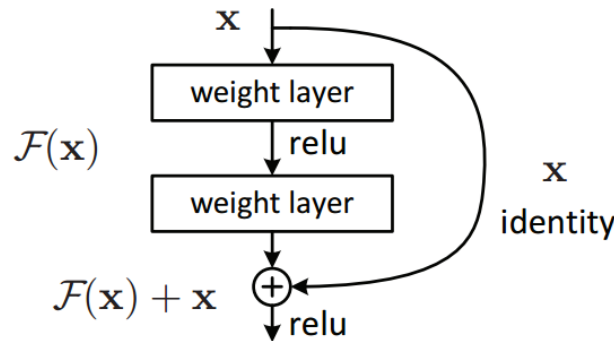


Figure 8.1: Residual block ([\[TODO: \]](#)).

# 9 Generative Models

## 9.1 Definitions

- *Probabilistic model*: is a model that represents a probability distribution. It could be a marginal [prob.](#) distribution  $p(x)$  or a conditional [prob.](#) distribution  $p(y|x)$ .
- *Evidence, query, and latent variables*:

The word ***latent*** implies ***existing, but hidden***. E.g.:

- In  $p(x)$ ,  $x$  is the query variable and there is no evidence variable.
- In  $p(y|x)$ ,  $y$  is the query variable and  $x$  is the evidence variable.
- In [MoG](#)  $p(x) = \sum_z p(x|z)p(z)$ ,  $x$  is the query variable,  $z$  is the latent variable to represent the mixture elements.
- In  $p(y|x) = \sum_z p(y|x, z)p(z)$ ,  $z$  is the latent variable.
- In a neural network, the inputs are evidence variables, the output are the query variables. In Variational Auto-Encoders ([VAE](#)),  $z$  is considered as the latent variable with a simple Gaussian distribution.
- In model-based [RL](#) with latent variable models, state  $x_t$  is the latent variable.
- *Latent variable models*: is the type of probabilistic model that represent a complex [pdf](#). as the product of multiple simpler [pdf](#)s, one of which belongs to a latent variable.

E.g.: in [VAE](#), the complex distribution  $p(x)$  (of an image) is represented as the product of two simple distribution.  $p(z)$  is a Gaussian distribution, and  $p(x|z)$  as a learnable neural network

$$p(x) = \int p(x|z)p(z)dz$$

## 9.2 Variational Inference

***Goal***: to model a complex distribution  $p_\theta(x)$ , given the data  $\mathcal{D} = \{x_1, \dots, x_N\}$

$$\text{Maximum likelihood fit:} \quad \theta \leftarrow \arg \max_{\theta} \frac{1}{N} \sum_i \log p_\theta(x_i) \quad (9.1)$$

$$\text{With latent variable } z: \quad p(x) = \int p(x|z)p(z)dz \quad (9.2)$$

$$\Rightarrow (\text{sadly, completely } \textbf{intractable}) \quad \theta \leftarrow \arg \max_{\theta} \frac{1}{N} \sum_i \log \left( \int p(x|z)p(z)dz \right) \quad (9.3)$$

$$\Rightarrow \text{alternative: } \textit{expected} \text{ log-likelihood} \quad \theta \leftarrow \arg \max_{\theta} \frac{1}{N} \mathbb{E}_{z \sim p(z|x_i)} [\log p_\theta(x_i, z)] \quad (9.4)$$

This section describes how to calculate  $p(z|x_i)$

## 9.2.1 The Variational Approximation

**Idea:** Approximating  $p(z|x_i)$  with  $q_i(z) = \mathcal{N}(\mu_i, \sigma_i)$

$$\log p(x_i) = \log \int_z p(x_i|z)p(z)dz \quad (9.5)$$

$$= \log \int_z p(x_i|z)p(z) \frac{q_i(z)}{q_i(z)} dz \quad (9.6)$$

$$= \log \mathbb{E}_{z \sim q_i(z)} \left[ \frac{p(x_i|z)p(z)}{q_i(z)} \right] \quad (9.7)$$

$$\geq \mathbb{E}_{z \sim q_i(z)} \left[ \log \frac{p(x_i|z)p(z)}{q_i(z)} \right] \quad (\text{Jensen's inequality: } \log \mathbb{E}[y] \geq \mathbb{E}[\log y]) \quad (9.8)$$

$$= \mathbb{E}_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)] - \mathbb{E}_{z \sim q_i(z)} [\log q_i(z)] \quad (9.9)$$

$$= \underbrace{\mathbb{E}_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)]}_{\mathcal{L}_i(p, q_i)} + \mathcal{H}(q_i) \quad (9.10)$$

- $\mathbb{E}_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)] + \mathcal{H}(q_i)$  is the lower bound of  $\log p(x_i)$ .  
Thus, maximize  $\mathbb{E}_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)] + \mathcal{H}(q_i)$  will indirectly maximize  $\log p(x_i)$ .
- Maximize the 1st term  $\mathbb{E}_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)]$  will find the place with high  $p(x_i, z)$   
Maximize the 2nd term  $\mathcal{H}(q_i)$  will increase the randomness of  $z$

In another point of view, as we approximate  $p(z|x_i)$  with  $q_i(z)$ , the **K-L-divergence**  $D_{KL}(q_i(z)||p(z|x_i))$  can measure how good is the approximation:

$$D_{KL}(q_i(z)||p(z|x_i)) = \mathbb{E}_{z \sim q_i(z)} \left[ \log \frac{q_i(z)}{p(z|x_i)} \right] = \mathbb{E}_{z \sim q_i(z)} \left[ \log \frac{q_i(z)p(x_i)}{p(x_i, z)} \right] \quad (9.11)$$

$$= -\mathbb{E}_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)] + \mathbb{E}_{z \sim q_i(z)} [\log q_i(z)] + \mathbb{E}_{z \sim q_i(z)} [\log p(x_i)] \quad (9.12)$$

$$= -\underbrace{\mathbb{E}_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)]}_{\mathcal{L}_i(p, q_i)} - \mathcal{H}(q_i) + \log p(x_i) \quad (9.13)$$

$$\Rightarrow \log p(x_i) = D_{KL}(q_i(z)||p(z|x_i)) + \mathcal{L}_i(p, q_i) \quad (9.14)$$

$$\Rightarrow \log p(x_i) \geq \mathcal{L}_i(p, q_i), \quad \text{since } D_{KL}(q||p) \geq 0 \quad (9.15)$$

In addition, since  $\log p(x_i)$  is independent of  $q_i$ , maximizing  $\mathcal{L}_i(p, q_i)$  with regard to (**w.r.t.**)  $q_i$  will minimize the **K-L-divergence**.

### 9.2.2 Algorithm

Again, the goal is to model the complex distribution  $p(x)$ . Since the common maximum likelihood is intractable, we maximize its lower bound.

$$\begin{aligned}\theta &\leftarrow \arg \max_{\theta} \frac{1}{N} \sum_i \log p_{\theta}(x_i) && - \text{intractable maximum likelihood} \\ \theta &\leftarrow \arg \max_{\theta} \frac{1}{N} \sum_i \mathcal{L}_i(p, q_i) && - \text{tractable lower bound}\end{aligned}$$

**The algorithm:**

1. for each  $x_i$  (or mini-batch):
2.   calculate  $\nabla_{\theta} \mathcal{L}_i(p, q_i)$ :
3.   sample  $z \sim q_i(z)$
4.    $\nabla_{\theta} \mathcal{L}_i(p, q_i) \approx \nabla_{\theta} \log p_{\theta}(x_i|z)$
5.    $\theta \leftarrow \theta + \alpha \nabla_{\theta} \mathcal{L}_i(p, q_i)$
6.   update  $q_i$  to maximize  $\mathcal{L}_i(p, q_i)$

In the above 6<sup>th</sup> step, assuming  $q_i(z) = \mathcal{N}(\mu_i, \sigma_i)$ :

- calculate the gradient  $\nabla_{\mu_i} \mathcal{L}_i(p, q_i)$  and  $\nabla_{\sigma_i} \mathcal{L}_i(p, q_i)$
- apply gradient ascent on  $\mu_i$  and  $\sigma_i$  to maximize  $\mathcal{L}_i(p, q_i)$ .

**Problem:** for each  $x_i$ , there will be one  $\mu_i$  and one  $\sigma_i$ , with the appropriate dimension. Thus, the total number of [params.](#) would be huge:  $|\theta| + (|\mu_i| + |\sigma_i|) \times N$

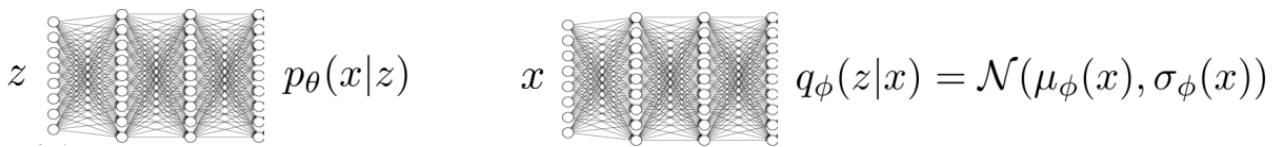
## 9.3 Amortized Variational Inference

The word *amortize* means to reduce or pay off (a debt) with regular payments.

As mentioned above, the problem with using a Gaussian for  $q_i(z)$  is the huge number of [params.](#). This section proposes an alternative, which is to use a neural network (Fig. 9.1)

$$q_i(z) = q(z|x_i) \approx q_{\phi}(z|x) = \mathcal{N}(\mu_{\phi}(x), \sigma_{\phi}(x)) \quad (9.16)$$

$$\Rightarrow \log p(x_i) \geq \underbrace{\mathbb{E}_{z \sim q_{\phi}(z|x_i)} [\log p_{\theta}(x_i|z) + \log p(z)] + \mathcal{H}(q_{\phi}(z|x_i))}_{\mathcal{L}_i(p_{\theta}(x_i|z), q_{\phi}(z|x_i))} \quad (9.17)$$



**Figure 9.1:** Two neural networks:  $p_{\theta}(x|z)$  and  $q_{\phi}(z|x)$ .

## 9.3.1 Algorithm

The algorithm:

1. for each  $x_i$  (or mini-batch):
2.   calculate  $\nabla_{\theta} \mathcal{L}_i(p_{\theta}(x_i|z), q_{\phi}(z|x_i))$ :
3.   sample  $z \sim q_{\phi}(z|x_i)$
4.    $\nabla_{\theta} \mathcal{L} \approx \nabla_{\theta} \log p_{\theta}(x_i|z)$
5.    $\theta \leftarrow \theta + \alpha \nabla_{\theta} \mathcal{L}$
6.    $\phi \leftarrow \phi + \alpha \nabla_{\phi} \mathcal{L}$

Let's consider the gradient  $\nabla_{\phi} \mathcal{L}$  in the above 6<sup>th</sup> step:

$$\mathcal{L}_i = \underbrace{\mathbb{E}_{z \sim q_{\phi}(z|x_i)} [\log p_{\theta}(x_i|z) + \log p(z)]}_{J(\phi) = \mathbb{E}_{z \sim q_{\phi}(z|x_i)} [r(x_i, z)]} + \mathcal{H}(q_{\phi}(z|x_i)) \quad (9.18)$$

$$= \nabla_{\phi} J(\phi) + \nabla_{\phi} \mathcal{H}(q_{\phi}(z|x_i)) \quad (9.19)$$

- There is a formula to find the gradient of the entropy of a Gaussian

$$\nabla_{\phi} \mathcal{H}(q_{\phi}(z|x_i)) = \nabla_{\phi} \mathcal{H}(\mathcal{N}(\mu_{\phi}, \sigma_{\phi})) \quad (9.20)$$

- To find the gradient  $\nabla_{\phi} J(\phi)$ , we could use similar approach as in policy gradient, however, with the problem of high variance. Instead, we could use the reparameterization trick:

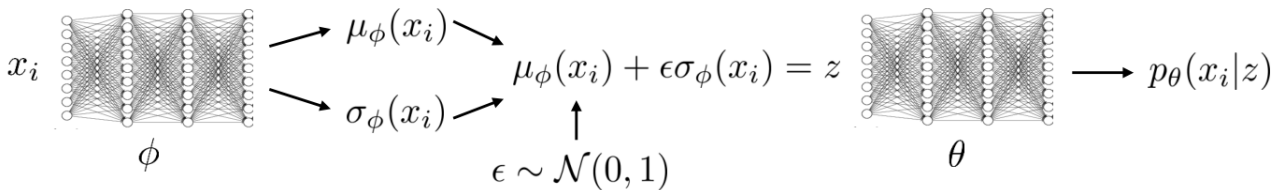
$$J(\phi) = \mathbb{E}_{z \sim q_{\phi}(z|x_i)} [r(x_i, z)] \quad q_{\phi}(z|x) = \mathcal{N}(\mu_{\phi}(x), \sigma_{\phi}(x)) \quad (9.21)$$

$$= \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} [r(x_i, \mu_{\phi}(x) + \epsilon \sigma_{\phi}(x))] \quad z = \mu_{\phi}(x) + \epsilon \sigma_{\phi}(x) \quad (9.22)$$

with  $\epsilon \sim \mathcal{N}(0, 1)$  is independent of  $\phi$ !

$\Rightarrow$  estimate  $\nabla_{\phi} J(\phi)$ :

- sample  $\epsilon_1, \dots, \epsilon_M$  from  $\mathcal{N}(0, 1)$  (a single sample works just well!)
- $\nabla_{\phi} J(\phi) \approx \frac{1}{M} \sum_j \nabla_{\phi} r(x_i, \mu_{\phi}(x) + \epsilon \sigma_{\phi}(x))$



**Figure 9.2:** The complete structure for amortized variational inference.

### 9.3.2 Reparameterization Trick

Another way to look:

$$\mathcal{L}_i = \mathbb{E}_{z \sim q_\phi(z|x_i)} [\log p_\theta(x_i|z) + \log p(z)] + \mathcal{H}(q_\phi(z|x_i)) \quad (9.23)$$

$$= \mathbb{E}_{z \sim q_\phi(z|x_i)} [\log p_\theta(x_i|z)] + \mathbb{E}_{z \sim q_\phi(z|x_i)} [\log p(z)] + \mathcal{H}(q_\phi(z|x_i)) \quad (9.24)$$

$$= \mathbb{E}_{z \sim q_\phi(z|x_i)} [\log p_\theta(x_i|z)] - D_{KL}(q_\phi(z|x_i) || p(z)) \quad (9.25)$$

$$= \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} [\log p_\theta(x_i | \mu_\phi(x) + \epsilon \sigma_\phi(x))] - D_{KL}(q_\phi(z|x_i) || p(z)) \quad (9.26)$$

$$\approx \log p_\theta(x_i | \mu_\phi(x) + \epsilon \sigma_\phi(x)) - D_{KL}(q_\phi(z|x_i) || p(z)) \quad (\text{single sample estimate}) \quad (9.27)$$

**NOTE:** Reparameterization trick versus (vs.) policy gradient

- Policy Gradient

$$\nabla_\phi J(\phi) \approx \frac{1}{M} \sum_j \nabla_\phi \log q_\phi(z_j|x_i) r(x_i, z_j)$$

+can handle both discrete and continuous latent variables

−high variance, requires multiple samples & small learning rates

- Reparameterization trick

$$\nabla_\phi J(\phi) \approx \frac{1}{M} \sum_j \nabla_\phi r(x_i, \mu_\phi(x) + \epsilon \sigma_\phi(x))$$

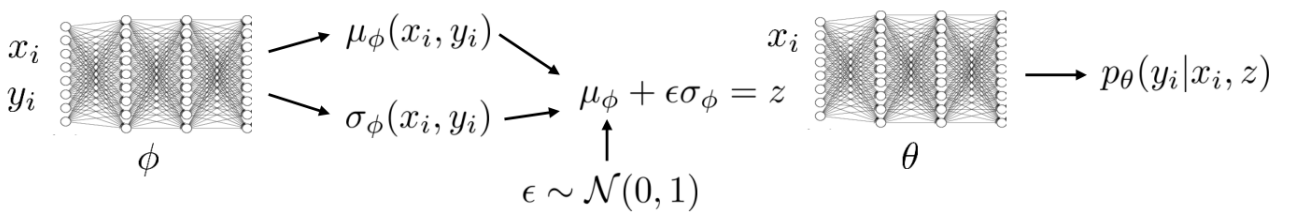
−only continuous latent variables

+very simple to implement

+low variance

## 9.4 Conditional Variational Inference Models

$$\mathcal{L}_i = \mathbb{E}_{z \sim q_\phi(z|x_i, y_i)} [\log p_\theta(y_i|x_i, z) + \log p(z|x_i)] + \mathcal{H}(q_\phi(z|x_i, y_i)) \quad (9.28)$$



**Figure 9.3:** The conditional model.



## 9.5 Generative Models

*Generative models* are models that generate data  $x$ . E.g.:  $p(x)$  is a generative model, because knowing  $p(x)$ , we can sample  $x$ .

**NOTE:** Not all generative models are not necessary latent variable models, and not all latent variable models are generative models. But it's common for a generative model to be a latent variable model, because sometimes, to generate data, we usually want to know the [prob.](#) distribution of it. When that [prob.](#) distribution is complex, we would represent it as a product of multiple simple [prob.](#) distribution, using some latent variables.

### 9.5.1 Generative Adversarial Network

[TODO: Generative Adversarial Network ([GAN](#))]

## 9.6 References

- [Variational Inference | CS 285, Stanford | YouTube](#)
- [Generative models | CS231, Stanford | YouTube](#)
- [Variational Autoencoders | Arxiv Insights | YouTube](#)
- [Variational Autoencoders | CodeEmporium | YouTube](#)
- [Real "deconv" layer | GitHub](#)

# Bibliography

- [GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.
- [HVD+15] G. Hinton, O. Vinyals, J. Dean, et al. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* 2.7 (2015).
- [Vu18] H. T. Vu. *Deep learning*. 2018.