

# Super VIP Cheatsheet: Deep Learning

Afshine AMIDI and Shervine AMIDI

November 25, 2018

## Contents

### 1 Convolutional Neural Networks

**2**

1.1 Overview . . . . .	2
1.2 Types of layer . . . . .	2
1.3 Filter hyperparameters . . . . .	2
1.4 Tuning hyperparameters . . . . .	3
1.5 Commonly used activation functions . . . . .	3
1.6 Object detection . . . . .	4
1.6.1 Face verification and recognition . . . . .	5
1.6.2 Neural style transfer . . . . .	5
1.6.3 Architectures using computational tricks . . . . .	6

### 2 Recurrent Neural Networks

**7**

2.1 Overview . . . . .	7
2.2 Handling long term dependencies . . . . .	8
2.3 Learning word representation . . . . .	9
2.3.1 Motivation and notations . . . . .	9
2.3.2 Word embeddings . . . . .	9
2.4 Comparing words . . . . .	9
2.5 Language model . . . . .	10
2.6 Machine translation . . . . .	10
2.7 Attention . . . . .	10

### 3 Deep Learning Tips and Tricks

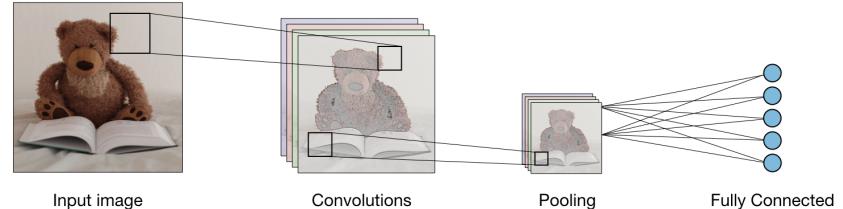
**11**

3.1 Data processing . . . . .	11
3.2 Training a neural network . . . . .	12
3.2.1 Definitions . . . . .	12
3.2.2 Finding optimal weights . . . . .	12
3.3 Parameter tuning . . . . .	12
3.3.1 Weights initialization . . . . .	12
3.3.2 Optimizing convergence . . . . .	12
3.4 Regularization . . . . .	13
3.5 Good practices . . . . .	13

## 1 Convolutional Neural Networks

### 1.1 Overview

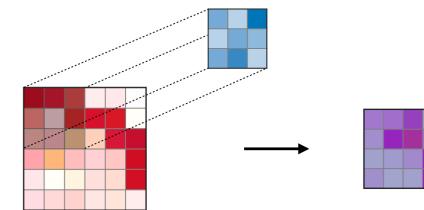
□ **Architecture of a traditional CNN** – Convolutional neural networks, also known as CNNs, are a specific type of neural networks that are generally composed of the following layers:



The convolution layer and the pooling layer can be fine-tuned with respect to hyperparameters that are described in the next sections.

### 1.2 Types of layer

□ **Convolutional layer (CONV)** – The convolution layer (CONV) uses filters that perform convolution operations as it is scanning the input  $I$  with respect to its dimensions. Its hyperparameters include the filter size  $F$  and stride  $S$ . The resulting output  $O$  is called *feature map* or *activation map*.

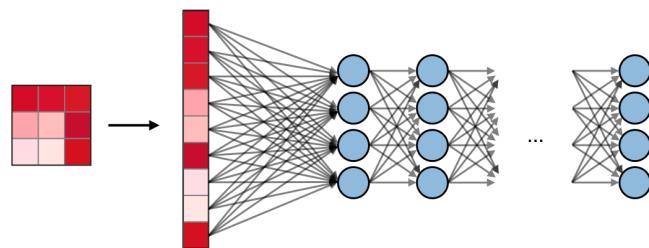


*Remark: the convolution step can be generalized to the 1D and 3D cases as well.*

□ **Pooling (POOL)** – The pooling layer (POOL) is a downsampling operation, typically applied after a convolution layer, which does some spatial invariance. In particular, max and average pooling are special kinds of pooling where the maximum and average value is taken, respectively.

	Max pooling	Average pooling
Purpose	Each pooling operation selects the maximum value of the current view	Each pooling operation averages the values of the current view
Illustration		
Comments	<ul style="list-style-type: none"> <li>- Preserves detected features</li> <li>- Most commonly used</li> </ul>	<ul style="list-style-type: none"> <li>- Downsamples feature map</li> <li>- Used in LeNet</li> </ul>

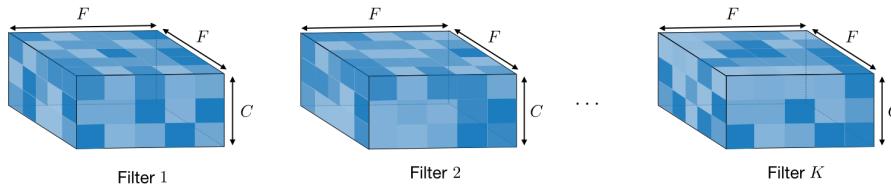
□ **Fully Connected (FC)** – The fully connected layer (FC) operates on a flattened input where each input is connected to all neurons. If present, FC layers are usually found towards the end of CNN architectures and can be used to optimize objectives such as class scores.



### 1.3 Filter hyperparameters

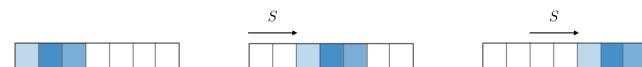
The convolution layer contains filters for which it is important to know the meaning behind its hyperparameters.

□ **Dimensions of a filter** – A filter of size  $F \times F$  applied to an input containing  $C$  channels is a  $F \times F \times C$  volume that performs convolutions on an input of size  $I \times I \times C$  and produces an output feature map (also called activation map) of size  $O \times O \times 1$ .



Remark: the application of  $K$  filters of size  $F \times F$  results in an output feature map of size  $O \times O \times K$ .

□ **Stride** – For a convolutional or a pooling operation, the stride  $S$  denotes the number of pixels by which the window moves after each operation.



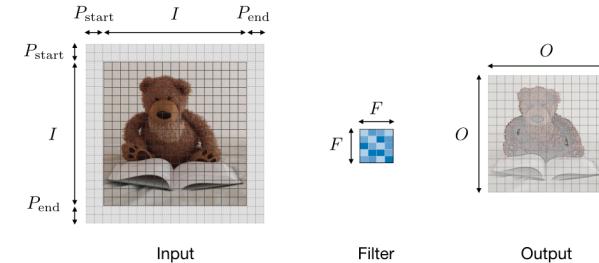
□ **Zero-padding** – Zero-padding denotes the process of adding  $P$  zeroes to each side of the boundaries of the input. This value can either be manually specified or automatically set through one of the three modes detailed below:

	Valid	Same	Full
Value	$P = 0$	$P_{\text{start}} = \left\lceil \frac{S \lceil \frac{I}{S} \rceil - I + F - S}{2} \right\rceil$ $P_{\text{end}} = \left\lceil \frac{S \lceil \frac{I}{S} \rceil - I + F - S}{2} \right\rceil$	$P_{\text{start}} \in [0, F - 1]$ $P_{\text{end}} = F - 1$
Illustration			
Purpose	<ul style="list-style-type: none"> <li>- No padding</li> <li>- Drops last convolution if dimensions do not match</li> </ul>	<ul style="list-style-type: none"> <li>- Padding such that feature map size has size <math>\lceil \frac{I}{S} \rceil</math></li> <li>- Output size is mathematically convenient</li> <li>- Also called 'half' padding</li> </ul>	<ul style="list-style-type: none"> <li>- Maximum padding such that end convolutions are applied on the limits of the input</li> <li>- Filter 'sees' the input end-to-end</li> </ul>

### 1.4 Tuning hyperparameters

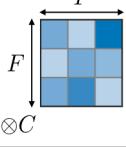
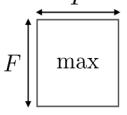
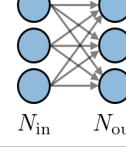
□ **Parameter compatibility in convolution layer** – By noting  $I$  the length of the input volume size,  $F$  the length of the filter,  $P$  the amount of zero padding,  $S$  the stride, then the output size  $O$  of the feature map along that dimension is given by:

$$O = \frac{I - F + P_{\text{start}} + P_{\text{end}}}{S} + 1$$



Remark: often times,  $P_{\text{start}} = P_{\text{end}} \triangleq P$ , in which case we can replace  $P_{\text{start}} + P_{\text{end}}$  by  $2P$  in the formula above.

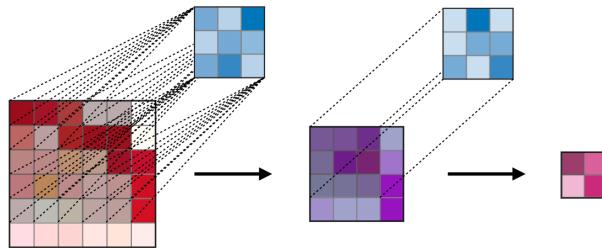
**□ Understanding the complexity of the model** – In order to assess the complexity of a model, it is often useful to determine the number of parameters that its architecture will have. In a given layer of a convolutional neural network, it is done as follows:

	CONV	POOL	FC
Illustration			
Input size	$I \times I \times C$	$I \times I \times C$	$N_{\text{in}}$
Output size	$O \times O \times K$	$O \times O \times C$	$N_{\text{out}}$
Number of parameters	$(F \times F \times C + 1) \cdot K$	0	$(N_{\text{in}} + 1) \times N_{\text{out}}$
Remarks	<ul style="list-style-type: none"> <li>- One bias parameter per filter</li> <li>- In most cases, <math>S &lt; F</math></li> <li>- A common choice for <math>K</math> is <math>2C</math></li> </ul>	<ul style="list-style-type: none"> <li>- Pooling operation done channel-wise</li> <li>- In most cases, <math>S = F</math></li> </ul>	<ul style="list-style-type: none"> <li>- Input is flattened</li> <li>- One bias parameter per neuron</li> <li>- The number of FC neurons is free of structural constraints</li> </ul>

**□ Receptive field** – The receptive field at layer  $k$  is the area denoted  $R_k \times R_k$  of the input that each pixel of the  $k$ -th activation map can ‘see’. By calling  $F_j$  the filter size of layer  $j$  and  $S_i$  the stride value of layer  $i$  and with the convention  $S_0 = 1$ , the receptive field at layer  $k$  can be computed with the formula:

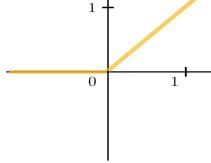
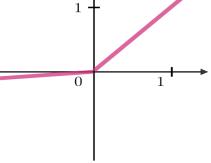
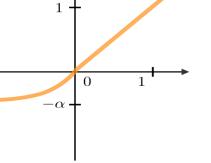
$$R_k = 1 + \sum_{j=1}^k (F_j - 1) \prod_{i=0}^{j-1} S_i$$

In the example below, we have  $F_1 = F_2 = 3$  and  $S_1 = S_2 = 1$ , which gives  $R_2 = 1 + 2 \cdot 1 + 2 \cdot 1 = 5$ .



## 1.5 Commonly used activation functions

**□ Rectified Linear Unit** – The rectified linear unit layer (ReLU) is an activation function  $g$  that is used on all elements of the volume. It aims at introducing non-linearities to the network. Its variants are summarized in the table below:

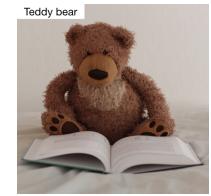
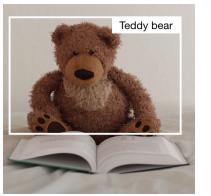
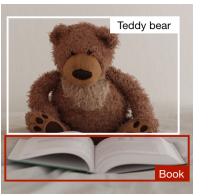
ReLU	Leaky ReLU	ELU
$g(z) = \max(0, z)$	$g(z) = \max(\epsilon z, z)$ with $\epsilon \ll 1$	$g(z) = \max(\alpha(e^z - 1), z)$ with $\alpha \ll 1$
		
Non-linearity complexities biologically interpretable	Addresses dying ReLU issue for negative values	Differentiable everywhere

**□ Softmax** – The softmax step can be seen as a generalized logistic function that takes as input a vector of scores  $x \in \mathbb{R}^n$  and outputs a vector of output probability  $p \in \mathbb{R}^n$  through a softmax function at the end of the architecture. It is defined as follows:

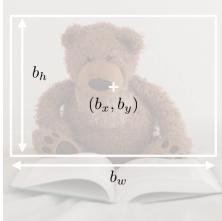
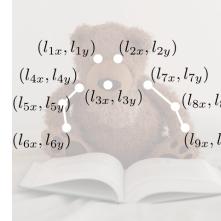
$$p = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} \quad \text{where} \quad p_i = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}$$

## 1.6 Object detection

**□ Types of models** – There are 3 main types of object recognition algorithms, for which the nature of what is predicted is different. They are described in the table below:

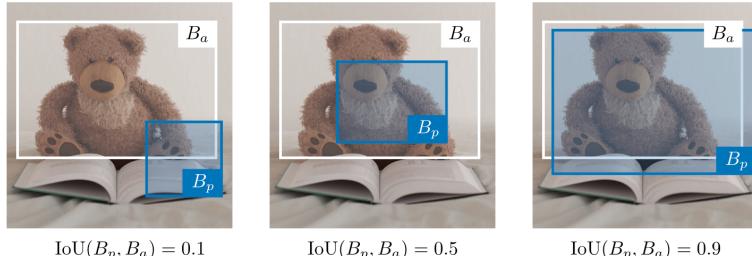
Image classification	Classification w. localization	Detection
		
<ul style="list-style-type: none"> <li>- Classifies a picture</li> <li>- Predicts probability of object</li> </ul>	<ul style="list-style-type: none"> <li>- Detects object in a picture</li> <li>- Predicts probability of object and where it is located</li> </ul>	<ul style="list-style-type: none"> <li>- Detects up to several objects in a picture</li> <li>- Predicts probabilities of objects and where they are located</li> </ul>
Traditional CNN	Simplified YOLO, R-CNN	YOLO, R-CNN

**□ Detection** – In the context of object detection, different methods are used depending on whether we just want to locate the object or detect a more complex shape in the image. The two main ones are summarized in the table below:

Bounding box detection	Landmark detection
Detects the part of the image where the object is located	<ul style="list-style-type: none"> <li>- Detects a shape or characteristics of an object (e.g. eyes)</li> <li>- More granular</li> </ul>
	
Box of center $(b_x, b_y)$ , height $b_h$ and width $b_w$	Reference points $(l_1x, l_1y), \dots, (l_nx, l_ny)$

□ **Intersection over Union** – Intersection over Union, also known as IoU, is a function that quantifies how correctly positioned a predicted bounding box  $B_p$  is over the actual bounding box  $B_a$ . It is defined as:

$$\text{IoU}(B_p, B_a) = \frac{B_p \cap B_a}{B_p \cup B_a}$$

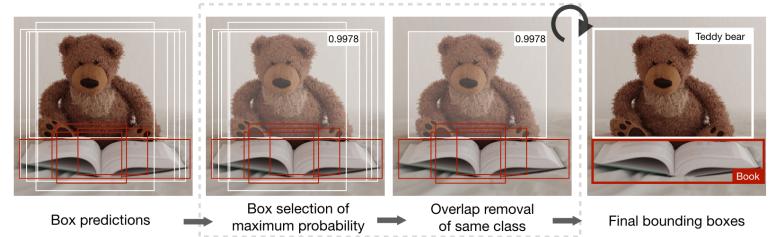


Remark: we always have  $\text{IoU} \in [0, 1]$ . By convention, a predicted bounding box  $B_p$  is considered as being reasonably good if  $\text{IoU}(B_p, B_a) \geq 0.5$ .

□ **Anchor boxes** – Anchor boxing is a technique used to predict overlapping bounding boxes. In practice, the network is allowed to predict more than one box simultaneously, where each box prediction is constrained to have a given set of geometrical properties. For instance, the first prediction can potentially be a rectangular box of a given form, while the second will be another rectangular box of a different geometrical form.

□ **Non-max suppression** – The non-max suppression technique aims at removing duplicate overlapping bounding boxes of a same object by selecting the most representative ones. After having removed all boxes having a probability prediction lower than 0.6, the following steps are repeated while there are boxes remaining:

- Step 1: Pick the box with the largest prediction probability.
- Step 2: Discard any box having an  $\text{IoU} \geq 0.5$  with the previous box.



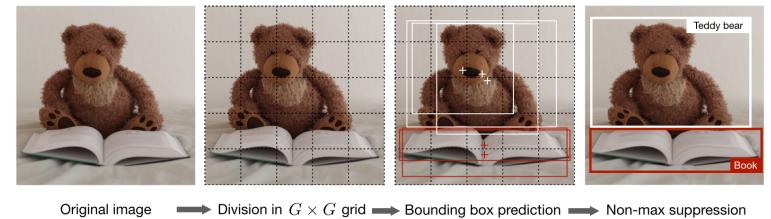
□ **YOLO** – You Only Look Once (YOLO) is an object detection algorithm that performs the following steps:

- Step 1: Divide the input image into a  $G \times G$  grid.
- Step 2: For each grid cell, run a CNN that predicts  $y$  of the following form:

$$y = \underbrace{[p_c, b_x, b_y, b_h, b_w, c_1, c_2, \dots, c_p, \dots]}_{\text{repeated } k \text{ times}}^T \in \mathbb{R}^{G \times G \times k \times (5+p)}$$

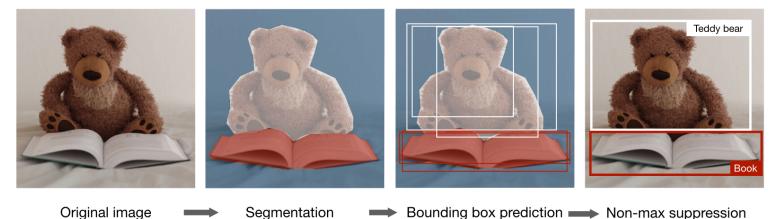
where  $p_c$  is the probability of detecting an object,  $b_x, b_y, b_h, b_w$  are the properties of the detected bounding box,  $c_1, \dots, c_p$  is a one-hot representation of which of the  $p$  classes were detected, and  $k$  is the number of anchor boxes.

- Step 3: Run the non-max suppression algorithm to remove any potential duplicate overlapping bounding boxes.



Remark: when  $p_c = 0$ , then the network does not detect any object. In that case, the corresponding predictions  $b_x, \dots, c_p$  have to be ignored.

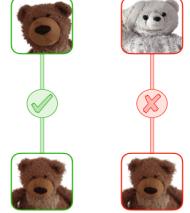
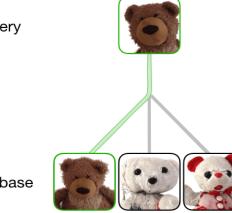
□ **R-CNN** – Region with Convolutional Neural Networks (R-CNN) is an object detection algorithm that first segments the image to find potential relevant bounding boxes and then run the detection algorithm to find most probable objects in those bounding boxes.



Remark: although the original algorithm is computationally expensive and slow, newer architectures enabled the algorithm to run faster, such as Fast R-CNN and Faster R-CNN.

### 1.6.1 Face verification and recognition

□ **Types of models** – Two main types of model are summed up in table below:

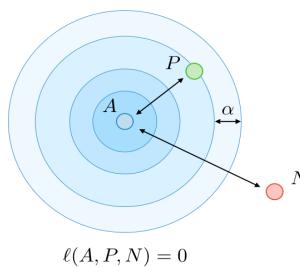
Face verification	Face recognition
- Is this the correct person? - One-to-one lookup	- Is this one of the $K$ persons in the database? - One-to-many lookup
 	

□ **One Shot Learning** – One Shot Learning is a face verification algorithm that uses a limited training set to learn a similarity function that quantifies how different two given images are. The similarity function applied to two images is often noted  $d(\text{image 1}, \text{image 2})$ .

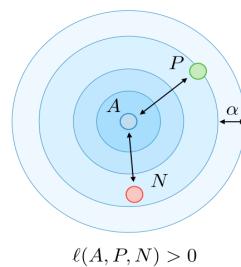
□ **Siamese Network** – Siamese Networks aim at learning how to encode images to then quantify how different two images are. For a given input image  $x^{(i)}$ , the encoded output is often noted as  $f(x^{(i)})$ .

□ **Triplet loss** – The triplet loss  $\ell$  is a loss function computed on the embedding representation of a triplet of images  $A$  (anchor),  $P$  (positive) and  $N$  (negative). The anchor and the positive example belong to a same class, while the negative example to another one. By calling  $\alpha \in \mathbb{R}^+$  the margin parameter, this loss is defined as follows:

$$\ell(A, P, N) = \max(d(A, P) - d(A, N) + \alpha, 0)$$



$$\ell(A, P, N) = 0$$



$$\ell(A, P, N) > 0$$

### 1.6.2 Neural style transfer

□ **Motivation** – The goal of neural style transfer is to generate an image  $G$  based on a given content  $C$  and a given style  $S$ .



□ **Activation** – In a given layer  $l$ , the activation is noted  $a^{[l]}$  and is of dimensions  $n_H \times n_w \times n_c$

□ **Content cost function** – The content cost function  $J_{\text{content}}(C, G)$  is used to determine how the generated image  $G$  differs from the original content image  $C$ . It is defined as follows:

$$J_{\text{content}}(C, G) = \frac{1}{2} \|a^{[l]}(C) - a^{[l]}(G)\|^2$$

□ **Style matrix** – The style matrix  $G^{[l]}$  of a given layer  $l$  is a Gram matrix where each of its elements  $G_{kk'}^{[l]}$  quantifies how correlated the channels  $k$  and  $k'$  are. It is defined with respect to activations  $a^{[l]}$  as follows:

$$G_{kk'}^{[l]} = \sum_{i=1}^{n_H^{[l]}} \sum_{j=1}^{n_w^{[l]}} a_{ijk}^{[l]} a_{ijk'}^{[l]}$$

Remark: the style matrix for the style image and the generated image are noted  $G^{[l](S)}$  and  $G^{[l](G)}$  respectively.

□ **Style cost function** – The style cost function  $J_{\text{style}}(S, G)$  is used to determine how the generated image  $G$  differs from the style  $S$ . It is defined as follows:

$$J_{\text{style}}^{[l]}(S, G) = \frac{1}{(2n_H n_w n_c)^2} \|G^{[l](S)} - G^{[l](G)}\|_F^2 = \frac{1}{(2n_H n_w n_c)^2} \sum_{k,k'=1}^{n_c} \left( G_{kk'}^{[l](S)} - G_{kk'}^{[l](G)} \right)^2$$

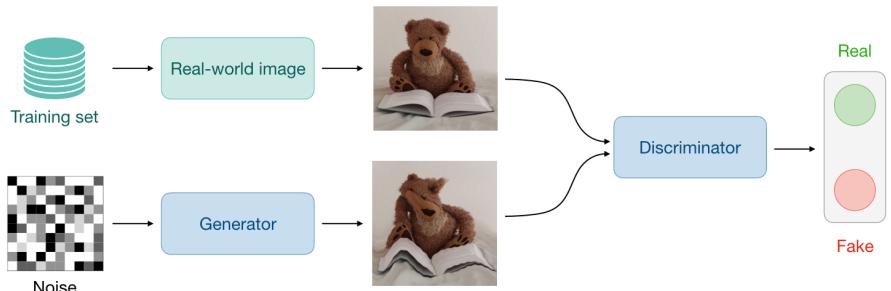
□ **Overall cost function** – The overall cost function is defined as being a combination of the content and style cost functions, weighted by parameters  $\alpha, \beta$ , as follows:

$$J(G) = \alpha J_{\text{content}}(C, G) + \beta J_{\text{style}}(S, G)$$

Remark: a higher value of  $\alpha$  will make the model care more about the content while a higher value of  $\beta$  will make it care more about the style.

### 1.6.3 Architectures using computational tricks

□ **Generative Adversarial Network** – Generative adversarial networks, also known as GANs, are composed of a generative and a discriminative model, where the generative model aims at generating the most truthful output that will be fed into the discriminative which aims at differentiating the generated and true image.



*Remark: use cases using variants of GANs include text to image, music generation and synthesis.*

□ **ResNet** – The Residual Network architecture (also called ResNet) uses residual blocks with a high number of layers meant to decrease the training error. The residual block has the following characterizing equation:

$$a^{[l+2]} = g(a^{[l]} + z^{[l+2]})$$

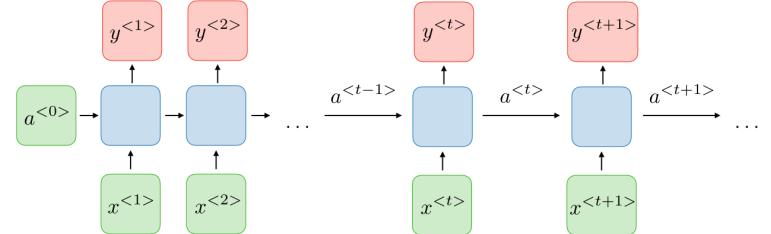
□ **Inception Network** – This architecture uses inception modules and aims at giving a try at different convolutions in order to increase its performance. In particular, it uses the  $1 \times 1$  convolution trick to lower the burden of computation.

\* \* \*

## 2 Recurrent Neural Networks

### 2.1 Overview

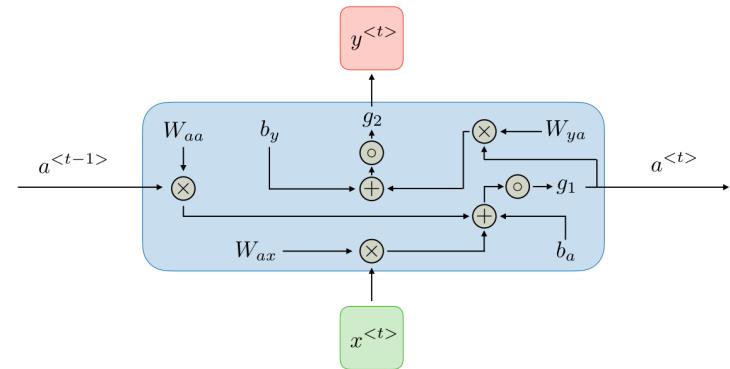
□ **Architecture of a traditional RNN** – Recurrent neural networks, also known as RNNs, are a class of neural networks that allow previous outputs to be used as inputs while having hidden states. They are typically as follows:



For each timestep  $t$ , the activation  $a^{<t>}$  and the output  $y^{<t>}$  are expressed as follows:

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a) \quad \text{and} \quad y^{<t>} = g_2(W_{ya}a^{<t>} + b_y)$$

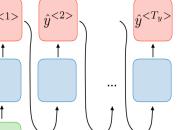
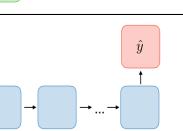
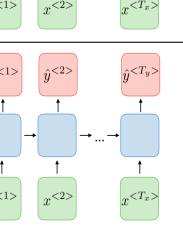
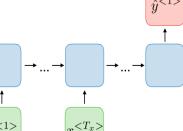
where  $W_{ax}, W_{aa}, W_{ya}, b_a, b_y$  are coefficients that are shared temporally and  $g_1, g_2$  activation functions



The pros and cons of a typical RNN architecture are summed up in the table below:

Advantages	Drawbacks
<ul style="list-style-type: none"> <li>- Possibility of processing input of any length</li> <li>- Model size not increasing with size of input</li> <li>- Computation takes into account historical information</li> <li>- Weights are shared across time</li> </ul>	<ul style="list-style-type: none"> <li>- Computation being slow</li> <li>- Difficulty of accessing information from a long time ago</li> <li>- Cannot consider any future input for the current state</li> </ul>

□ **Applications of RNNs** – RNN models are mostly used in the fields of natural language processing and speech recognition. The different applications are summed up in the table below:

Type of RNN	Illustration	Example
One-to-one $T_x = T_y = 1$		Traditional neural network
One-to-many $T_x = 1, T_y > 1$		Music generation
Many-to-one $T_x > 1, T_y = 1$		Sentiment classification
Many-to-many $T_x = T_y$		Name entity recognition
Many-to-many $T_x \neq T_y$		Machine translation

□ **Loss function** – In the case of a recurrent neural network, the loss function  $\mathcal{L}$  of all time steps is defined based on the loss at every time step as follows:

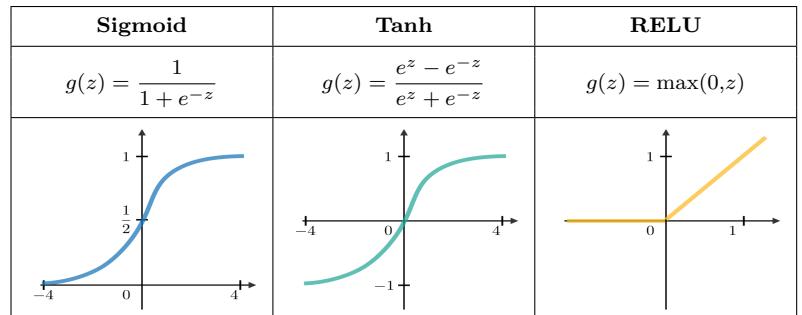
$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}(\hat{y}^{<t>}, y^{<t>})$$

□ **Backpropagation through time** – Backpropagation is done at each point in time. At timestep  $T$ , the derivative of the loss  $\mathcal{L}$  with respect to weight matrix  $W$  is expressed as follows:

$$\frac{\partial \mathcal{L}^{(T)}}{\partial W} = \sum_{t=1}^T \frac{\partial \mathcal{L}^{(T)}}{\partial W} \Big|_{(t)}$$

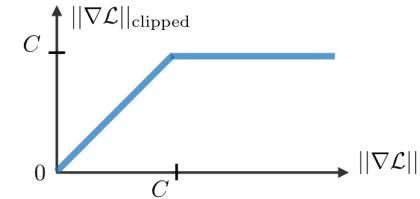
## 2.2 Handling long term dependencies

□ **Commonly used activation functions** – The most common activation functions used in RNN modules are described below:



□ **Vanishing/exploding gradient** – The vanishing and exploding gradient phenomena are often encountered in the context of RNNs. The reason why they happen is that it is difficult to capture long term dependencies because of multiplicative gradient that can be exponentially decreasing/increasing with respect to the number of layers.

□ **Gradient clipping** – It is a technique used to cope with the exploding gradient problem sometimes encountered when performing backpropagation. By capping the maximum value for the gradient, this phenomenon is controlled in practice.



□ **Types of gates** – In order to remedy the vanishing gradient problem, specific gates are used in some types of RNNs and usually have a well-defined purpose. They are usually noted  $\Gamma$  and are equal to:

$$\Gamma = \sigma(Wx^{<t>} + Ua^{<t-1>} + b)$$

where  $W, U, b$  are coefficients specific to the gate and  $\sigma$  is the sigmoid function. The main ones are summed up in the table below:

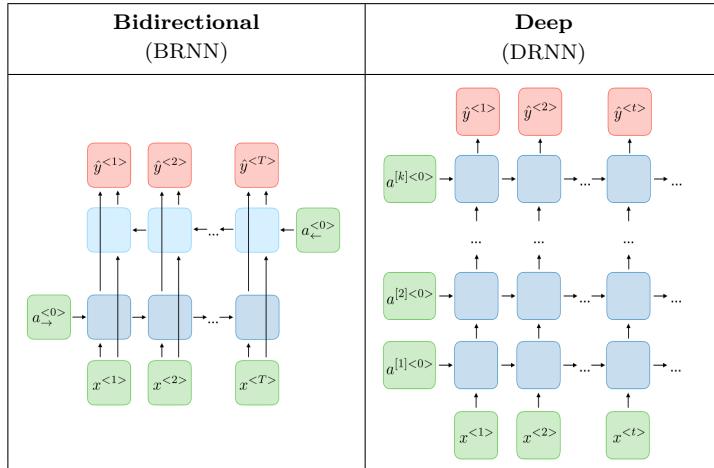
Type of gate	Role	Used in
Update gate $\Gamma_u$	How much past should matter now?	GRU, LSTM
Relevance gate $\Gamma_r$	Drop previous information?	GRU, LSTM
Forget gate $\Gamma_f$	Erase a cell or not?	LSTM
Output gate $\Gamma_o$	How much to reveal of a cell?	LSTM

□ **GRU/LSTM** – Gated Recurrent Unit (GRU) and Long Short-Term Memory units (LSTM) deal with the vanishing gradient problem encountered by traditional RNNs, with LSTM being a generalization of GRU. Below is a table summing up the characterizing equations of each architecture:

	Gated Recurrent Unit (GRU)	Long Short-Term Memory (LSTM)
$\tilde{c}^{<t>}$	$\tanh(W_c[\Gamma_r \star a^{<t-1>}, x^{<t>}] + b_c)$	$\tanh(W_c[\Gamma_r \star a^{<t-1>}, x^{<t>}] + b_c)$
$c^{<t>}$	$\Gamma_u \star \tilde{c}^{<t>} + (1 - \Gamma_u) \star c^{<t-1>}$	$\Gamma_u \star \tilde{c}^{<t>} + \Gamma_f \star c^{<t-1>}$
$a^{<t>}$	$c^{<t>}$	$\Gamma_o \star c^{<t>}$
Dependencies		

Remark: the sign  $\star$  denotes the element-wise multiplication between two vectors.

□ **Variants of RNNs** – The table below sums up the other commonly used RNN architectures:



## 2.3 Learning word representation

In this section, we note  $V$  the vocabulary and  $|V|$  its size.

### 2.3.1 Motivation and notations

□ **Representation techniques** – The two main ways of representing words are summed up in the table below:

1-hot representation	Word embedding
<ul style="list-style-type: none"> <li>- Noted <math>o_w</math></li> <li>- Naive approach, no similarity information</li> </ul>	<ul style="list-style-type: none"> <li>- Noted <math>e_w</math></li> <li>- Takes into account words similarity</li> </ul>

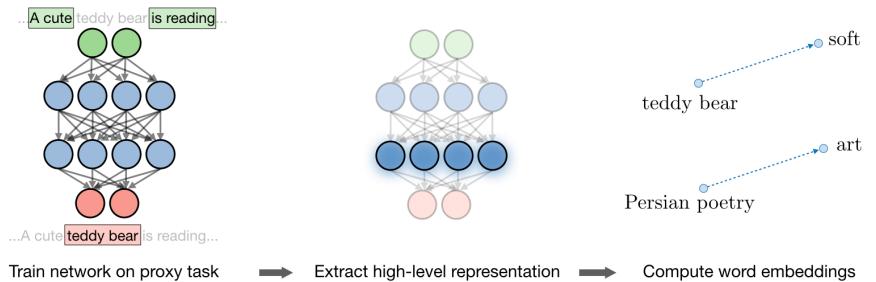
□ **Embedding matrix** – For a given word  $w$ , the embedding matrix  $E$  is a matrix that maps its 1-hot representation  $o_w$  to its embedding  $e_w$  as follows:

$$e_w = E o_w$$

Remark: learning the embedding matrix can be done using target/context likelihood models.

### 2.3.2 Word embeddings

□ **Word2vec** – Word2vec is a framework aimed at learning word embeddings by estimating the likelihood that a given word is surrounded by other words. Popular models include skip-gram, negative sampling and CBOW.



□ **Skip-gram** – The skip-gram word2vec model is a supervised learning task that learns word embeddings by assessing the likelihood of any given target word  $t$  happening with a context word  $c$ . By noting  $\theta_t$  a parameter associated with  $t$ , the probability  $P(t|c)$  is given by:

$$P(t|c) = \frac{\exp(\theta_t^T e_c)}{\sum_{j=1}^{|V|} \exp(\theta_j^T e_c)}$$

*Remark:* summing over the whole vocabulary in the denominator of the softmax part makes this model computationally expensive. CBOW is another word2vec model using the surrounding words to predict a given word.

**Negative sampling** – It is a set of binary classifiers using logistic regressions that aim at assessing how a given context and a given target words are likely to appear simultaneously, with the models being trained on sets of  $k$  negative examples and 1 positive example. Given a context word  $c$  and a target word  $t$ , the prediction is expressed by:

$$P(y = 1|c,t) = \sigma(\theta_t^T e_c)$$

*Remark:* this method is less computationally expensive than the skip-gram model.

**GloVe** – The GloVe model, short for global vectors for word representation, is a word embedding technique that uses a co-occurrence matrix  $X$  where each  $X_{i,j}$  denotes the number of times that a target  $i$  occurred with a context  $j$ . Its cost function  $J$  is as follows:

$$J(\theta) = \frac{1}{2} \sum_{i,j=1}^{|V|} f(X_{ij})(\theta_i^T e_j + b_i + b'_j - \log(X_{ij}))^2$$

here  $f$  is a weighting function such that  $X_{i,j} = 0 \implies f(X_{i,j}) = 0$ .

Given the symmetry that  $e$  and  $\theta$  play in this model, the final word embedding  $e_w^{(\text{final})}$  is given by:

$$e_w^{(\text{final})} = \frac{e_w + \theta_w}{2}$$

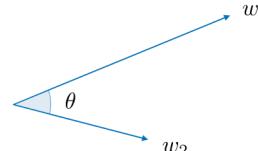
*Remark:* the individual components of the learned word embeddings are not necessarily interpretable.

## 2.4 Comparing words

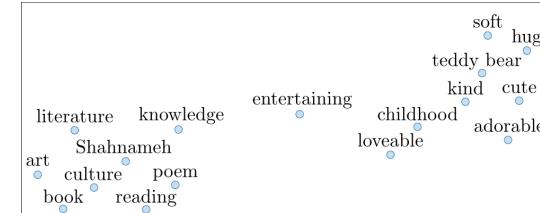
**Cosine similarity** – The cosine similarity between words  $w_1$  and  $w_2$  is expressed as follows:

$$\text{similarity} = \frac{w_1 \cdot w_2}{\|w_1\| \|w_2\|} = \cos(\theta)$$

*Remark:*  $\theta$  is the angle between words  $w_1$  and  $w_2$ .



**t-SNE** – t-SNE (t-distributed Stochastic Neighbor Embedding) is a technique aimed at reducing high-dimensional embeddings into a lower dimensional space. In practice, it is commonly used to visualize word vectors in the 2D space.



## 2.5 Language model

**Overview** – A language model aims at estimating the probability of a sentence  $P(y)$ .

**n-gram model** – This model is a naive approach aiming at quantifying the probability that an expression appears in a corpus by counting its number of appearance in the training data.

**Perplexity** – Language models are commonly assessed using the perplexity metric, also known as PP, which can be interpreted as the inverse probability of the dataset normalized by the number of words  $T$ . The perplexity is such that the lower, the better and is defined as follows:

$$\text{PP} = \prod_{t=1}^T \left( \frac{1}{\sum_{j=1}^{|V|} y_j^{(t)} \cdot \hat{y}_j^{(t)}} \right)^{\frac{1}{T}}$$

*Remark:* PP is commonly used in t-SNE.

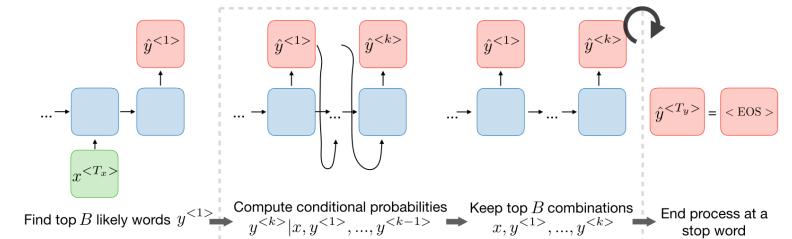
## 2.6 Machine translation

**Overview** – A machine translation model is similar to a language model except it has an encoder network placed before. For this reason, it is sometimes referred as a conditional language model. The goal is to find a sentence  $y$  such that:

$$y = \arg \max_{y^{<1>} \dots, y^{<T_y>}} P(y^{<1>} \dots, y^{<T_y>} | x)$$

**Beam search** – It is a heuristic search algorithm used in machine translation and speech recognition to find the likeliest sentence  $y$  given an input  $x$ .

- Step 1: Find top  $B$  likely words  $y^{<1>}$
- Step 2: Compute conditional probabilities  $y^{<k>} | x, y^{<1>} \dots, y^{<k-1>}$
- Step 3: Keep top  $B$  combinations  $x, y^{<1>} \dots, y^{<k>}$



*Remark: if the beam width is set to 1, then this is equivalent to a naive greedy search.*

**Beam width** – The beam width  $B$  is a parameter for beam search. Large values of  $B$  yield to better result but with slower performance and increased memory. Small values of  $B$  lead to worse results but is less computationally intensive. A standard value for  $B$  is around 10.

**Length normalization** – In order to improve numerical stability, beam search is usually applied on the following normalized objective, often called the normalized log-likelihood objective, defined as:

$$\text{Objective} = \frac{1}{T_y^\alpha} \sum_{t=1}^{T_y} \log \left[ p(y^{<t>} | x, y^{<1>}, \dots, y^{<t-1>}) \right]$$

*Remark: the parameter  $\alpha$  can be seen as a softener, and its value is usually between 0.5 and 1.*

**Error analysis** – When obtaining a predicted translation  $\hat{y}$  that is bad, one can wonder why we did not get a good translation  $y^*$  by performing the following error analysis:

Case	$P(y^* x) > P(\hat{y} x)$	$P(y^* x) \leq P(\hat{y} x)$
<b>Root cause</b>	Beam search faulty	RNN faulty
<b>Remedies</b>	Increase beam width	- Try different architecture - Regularize - Get more data

**Bleu score** – The bilingual evaluation understudy (bleu) score quantifies how good a machine translation is by computing a similarity score based on  $n$ -gram precision. It is defined as follows:

$$\text{bleu score} = \exp \left( \frac{1}{n} \sum_{k=1}^n p_k \right)$$

where  $p_n$  is the bleu score on  $n$ -gram only defined as follows:

$$p_n = \frac{\sum_{\substack{\text{n-gram} \in \hat{y}}} \text{count}_{\text{clip}}(\text{n-gram})}{\sum_{\substack{\text{n-gram} \in \hat{y}}} \text{count}(\text{n-gram})}$$

*Remark: a brevity penalty may be applied to short predicted translations to prevent an artificially inflated bleu score.*

## 2.7 Attention

**Attention model** – This model allows an RNN to pay attention to specific parts of the input that is considered as being important, which improves the performance of the resulting model in practice. By noting  $\alpha^{<t,t'>}$  the amount of attention that the output  $y^{<t>}$  should pay to the activation  $a^{<t'>}$  and  $c^{<t>}$  the context at time  $t$ , we have:

$$c^{<t>} = \sum_{t'} \alpha^{<t,t'>} a^{<t'>} \quad \text{with} \quad \sum_{t'} \alpha^{<t,t'>} = 1$$

*Remark: the attention scores are commonly used in image captioning and machine translation.*



A cute teddy bear is reading Persian literature



A cute teddy bear is reading Persian literature

**Attention weight** – The amount of attention that the output  $y^{<t>}$  should pay to the activation  $a^{<t'>}$  is given by  $\alpha^{<t,t'>}$  computed as follows:

$$\alpha^{<t,t'>} = \frac{\exp(e^{<t,t'>})}{\sum_{t''=1}^{T_x} \exp(e^{<t,t''>})}$$

*Remark: computation complexity is quadratic with respect to  $T_x$ .*

\* \* \*

### 3 Deep Learning Tips and Tricks

#### 3.1 Data processing

**◻ Data augmentation** – Deep learning models usually need a lot of data to be properly trained. It is often useful to get more data from the existing ones using data augmentation techniques. The main ones are summed up in the table below. More precisely, given the following input image, here are the techniques that we can apply:

Original	Flip	Rotation	Random crop
- Image without any modification	- Flipped with respect to an axis for which the meaning of the image is preserved	- Rotation with a slight angle - Simulates incorrect horizon calibration	- Random focus on one part of the image - Several random crops can be done in a row

Color shift	Noise addition	Information loss	Contrast change
- Nuances of RGB is slightly changed - Captures noise that can occur with light exposure	- Addition of noise - More tolerance to quality variation of inputs	- Parts of image ignored - Mimics potential loss of parts of image	- Luminosity changes - Controls difference in exposition due to time of day

**◻ Batch normalization** – It is a step of hyperparameter  $\gamma, \beta$  that normalizes the batch  $\{x_i\}$ . By noting  $\mu_B, \sigma_B^2$  the mean and variance of that we want to correct to the batch, it is done as follows:

$$x_i \leftarrow \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

It is usually done after a fully connected/convolutional layer and before a non-linearity layer and aims at allowing higher learning rates and reducing the strong dependence on initialization.

#### 3.2 Training a neural network

##### 3.2.1 Definitions

**◻ Epoch** – In the context of training a model, epoch is a term used to refer to one iteration where the model sees the whole training set to update its weights.

**◻ Mini-batch gradient descent** – During the training phase, updating weights is usually not based on the whole training set at once due to computation complexities or one data point due to noise issues. Instead, the update step is done on mini-batches, where the number of data points in a batch is a hyperparameter that we can tune.

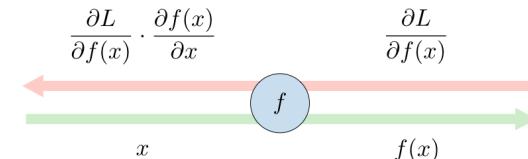
**◻ Loss function** – In order to quantify how a given model performs, the loss function  $L$  is usually used to evaluate to what extent the actual outputs  $y$  are correctly predicted by the model outputs  $z$ .

**◻ Cross-entropy loss** – In the context of binary classification in neural networks, the cross-entropy loss  $L(z,y)$  is commonly used and is defined as follows:

$$L(z,y) = -[y \log(z) + (1-y) \log(1-z)]$$

##### 3.2.2 Finding optimal weights

**◻ Backpropagation** – Backpropagation is a method to update the weights in the neural network by taking into account the actual output and the desired output. The derivative with respect to each weight  $w$  is computed using the chain rule.

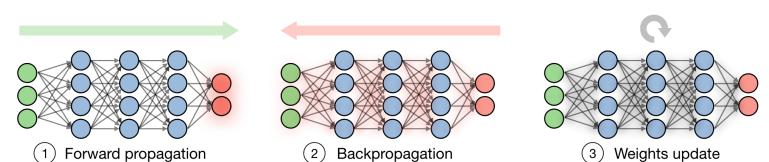


Using this method, each weight is updated with the rule:

$$w \leftarrow w - \alpha \frac{\partial L(z,y)}{\partial w}$$

**◻ Updating weights** – In a neural network, weights are updated as follows:

- Step 1: Take a batch of training data and perform forward propagation to compute the loss.
- Step 2: Backpropagate the loss to get the gradient of the loss with respect to each weight.
- Step 3: Use the gradients to update the weights of the network.



### 3.3 Parameter tuning

#### 3.3.1 Weights initialization

**Xavier initialization** – Instead of initializing the weights in a purely random manner, Xavier initialization enables to have initial weights that take into account characteristics that are unique to the architecture.

**Transfer learning** – Training a deep learning model requires a lot of data and more importantly a lot of time. It is often useful to take advantage of pre-trained weights on huge datasets that took days/weeks to train, and leverage it towards our use case. Depending on how much data we have at hand, here are the different ways to leverage this:

Method	Explanation	Update of $w$	Update of $b$
Momentum	- Dampens oscillations - Improvement to SGD - 2 parameters to tune	$w \leftarrow w - \alpha v_{dw}$	$b \leftarrow b - \alpha v_{db}$
RMSprop	- Root Mean Square propagation - Speeds up learning algorithm by controlling oscillations	$w \leftarrow w - \alpha \frac{dw}{\sqrt{s_{dw}}}$	$b \leftarrow b - \alpha \frac{db}{\sqrt{s_{db}}}$
Adam	- Adaptive Moment estimation - Most popular method - 4 parameters to tune	$w \leftarrow w - \alpha \frac{v_{dw}}{\sqrt{s_{dw}} + \epsilon}$	$b \leftarrow b - \alpha \frac{v_{db}}{\sqrt{s_{db}} + \epsilon}$

*Remark: other methods include Adadelta, Adagrad and SGD.*

Training size	Illustration	Explanation
Small		Freezes all layers, trains weights on softmax
Medium		Freezes most layers, trains weights on last layers and softmax
Large		Trains weights on layers and softmax by initializing weights on pre-trained ones

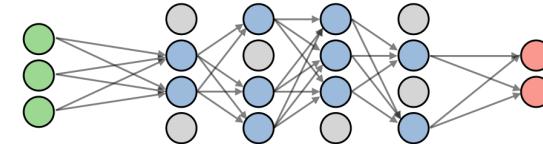
#### 3.3.2 Optimizing convergence

**Learning rate** – The learning rate, often noted  $\alpha$  or sometimes  $\eta$ , indicates at which pace the weights get updated. It can be fixed or adaptively changed. The current most popular method is called Adam, which is a method that adapts the learning rate.

**Adaptive learning rates** – Letting the learning rate vary when training a model can reduce the training time and improve the numerical optimal solution. While Adam optimizer is the most commonly used technique, others can also be useful. They are summed up in the table below:

### 3.4 Regularization

**Dropout** – Dropout is a technique used in neural networks to prevent overfitting the training data by dropping out neurons with probability  $p > 0$ . It forces the model to avoid relying too much on particular sets of features.

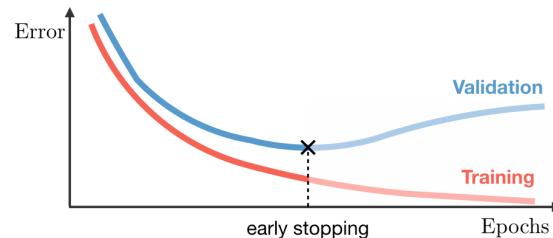


*Remark: most deep learning frameworks parametrize dropout through the 'keep' parameter 1 – p.*

**Weight regularization** – In order to make sure that the weights are not too large and that the model is not overfitting the training set, regularization techniques are usually performed on the model weights. The main ones are summed up in the table below:

LASSO	Ridge	Elastic Net
- Shrinks coefficients to 0 - Good for variable selection	Makes coefficients smaller	Tradeoff between variable selection and small coefficients
$\dots + \lambda   \theta  _1$ $\lambda \in \mathbb{R}$	$\dots + \lambda   \theta  _2^2$ $\lambda \in \mathbb{R}$	$\dots + \lambda [(1-\alpha)  \theta  _1 + \alpha  \theta  _2^2]$ $\lambda \in \mathbb{R}, \alpha \in [0,1]$

- **Early stopping** – This regularization technique stops the training process as soon as the validation loss reaches a plateau or starts to increase.



### 3.5 Good practices

- **Overfitting small batch** – When debugging a model, it is often useful to make quick tests to see if there is any major issue with the architecture of the model itself. In particular, in order to make sure that the model can be properly trained, a mini-batch is passed inside the network to see if it can overfit on it. If it cannot, it means that the model is either too complex or not complex enough to even overfit on a small batch, let alone a normal-sized training set.

- **Gradient checking** – Gradient checking is a method used during the implementation of the backward pass of a neural network. It compares the value of the analytical gradient to the numerical gradient at given points and plays the role of a sanity-check for correctness.

	Numerical gradient	Analytical gradient
Formula	$\frac{df}{dx}(x) \approx \frac{f(x+h) - f(x-h)}{2h}$	$\frac{df}{dx}(x) = f'(x)$
Comments	<ul style="list-style-type: none"> <li>- Expensive; loss has to be computed two times per dimension</li> <li>- Used to verify correctness of analytical implementation</li> <li>- Trade-off in choosing <math>h</math>: not too small (numerical instability), nor too large (poor gradient approx.)</li> </ul>	<ul style="list-style-type: none"> <li>- 'Exact' result</li> <li>- Direct computation</li> <li>- Used in the final implementation</li> </ul>

\* \* \*

# Super VIP Cheatsheet: Machine Learning

Afshine AMIDI and Shervine AMIDI

October 6, 2018

## Contents

### 1 Supervised Learning

**2**

1.1	Introduction to Supervised Learning . . . . .	2
1.2	Notations and general concepts . . . . .	2
1.3	Linear models . . . . .	2
1.3.1	Linear regression . . . . .	2
1.3.2	Classification and logistic regression . . . . .	3
1.3.3	Generalized Linear Models . . . . .	3
1.4	Support Vector Machines . . . . .	3
1.5	Generative Learning . . . . .	4
1.5.1	Gaussian Discriminant Analysis . . . . .	4
1.5.2	Naive Bayes . . . . .	4
1.6	Tree-based and ensemble methods . . . . .	4
1.7	Other non-parametric approaches . . . . .	4
1.8	Learning Theory . . . . .	5

### 2 Unsupervised Learning

**6**

2.1	Introduction to Unsupervised Learning . . . . .	6
2.2	Clustering . . . . .	6
2.2.1	Expectation-Maximization . . . . .	6
2.2.2	$k$ -means clustering . . . . .	6
2.2.3	Hierarchical clustering . . . . .	6
2.2.4	Clustering assessment metrics . . . . .	6
2.3	Dimension reduction . . . . .	7
2.3.1	Principal component analysis . . . . .	7
2.3.2	Independent component analysis . . . . .	7

### 3 Deep Learning

**8**

3.1	Neural Networks . . . . .	8
3.2	Convolutional Neural Networks . . . . .	8
3.3	Recurrent Neural Networks . . . . .	8
3.4	Reinforcement Learning and Control . . . . .	9

<b>4</b>	<b>Machine Learning Tips and Tricks</b>	<b>10</b>
4.1	Metrics . . . . .	10
4.1.1	Classification . . . . .	10
4.1.2	Regression . . . . .	10
4.2	Model selection . . . . .	11
4.3	Diagnostics . . . . .	11
<b>5</b>	<b>Refreshers</b>	<b>12</b>
5.1	Probabilities and Statistics . . . . .	12
5.1.1	Introduction to Probability and Combinatorics . . . . .	12
5.1.2	Conditional Probability . . . . .	12
5.1.3	Random Variables . . . . .	13
5.1.4	Jointly Distributed Random Variables . . . . .	13
5.1.5	Parameter estimation . . . . .	14
5.2	Linear Algebra and Calculus . . . . .	14
5.2.1	General notations . . . . .	14
5.2.2	Matrix operations . . . . .	15
5.2.3	Matrix properties . . . . .	15
5.2.4	Matrix calculus . . . . .	16

## 1 Supervised Learning

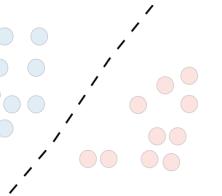
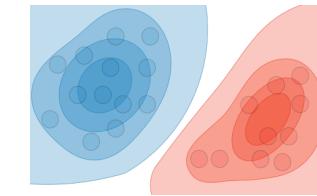
### 1.1 Introduction to Supervised Learning

Given a set of data points  $\{x^{(1)}, \dots, x^{(m)}\}$  associated to a set of outcomes  $\{y^{(1)}, \dots, y^{(m)}\}$ , we want to build a classifier that learns how to predict  $y$  from  $x$ .

□ **Type of prediction** – The different types of predictive models are summed up in the table below:

	Regression	Classifier
Outcome	Continuous	Class
Examples	Linear regression	Logistic regression, SVM, Naive Bayes

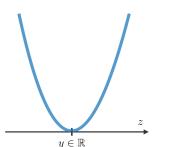
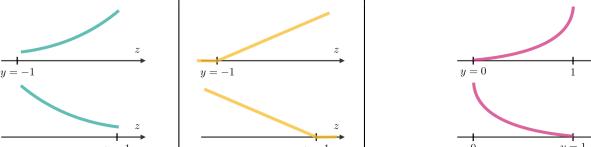
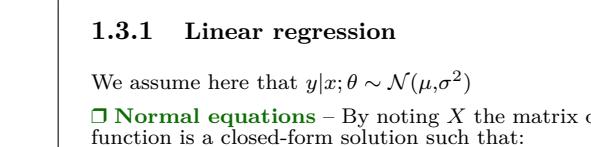
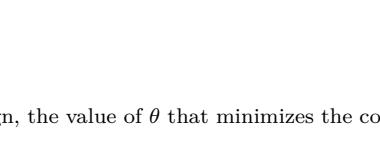
□ **Type of model** – The different models are summed up in the table below:

	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

### 1.2 Notations and general concepts

□ **Hypothesis** – The hypothesis is noted  $h_\theta$  and is the model that we choose. For a given input data  $x^{(i)}$ , the model prediction output is  $h_\theta(x^{(i)})$ .

□ **Loss function** – A loss function is a function  $L : (z,y) \in \mathbb{R} \times Y \mapsto L(z,y) \in \mathbb{R}$  that takes as inputs the predicted value  $z$  corresponding to the real data value  $y$  and outputs how different they are. The common loss functions are summed up in the table below:

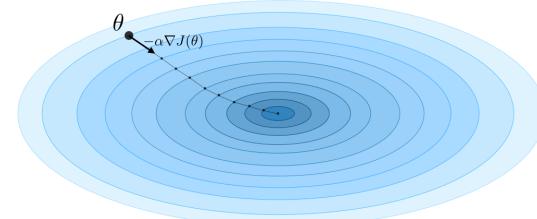
Least squared	Logistic	Hinge	Cross-entropy
$\frac{1}{2}(y - z)^2$	$\log(1 + \exp(-yz))$	$\max(0, 1 - yz)$	$-\left[y \log(z) + (1 - y) \log(1 - z)\right]$
			
Linear regression	Logistic regression	SVM	Neural Network

□ **Cost function** – The cost function  $J$  is commonly used to assess the performance of a model, and is defined with the loss function  $L$  as follows:

$$J(\theta) = \sum_{i=1}^m L(h_\theta(x^{(i)}), y^{(i)})$$

□ **Gradient descent** – By noting  $\alpha \in \mathbb{R}$  the learning rate, the update rule for gradient descent is expressed with the learning rate and the cost function  $J$  as follows:

$$\theta \leftarrow \theta - \alpha \nabla J(\theta)$$



Remark: Stochastic gradient descent (SGD) is updating the parameter based on each training example, and batch gradient descent is on a batch of training examples.

□ **Likelihood** – The likelihood of a model  $L(\theta)$  given parameters  $\theta$  is used to find the optimal parameters  $\theta$  through maximizing the likelihood. In practice, we use the log-likelihood  $\ell(\theta) = \log(L(\theta))$  which is easier to optimize. We have:

$$\theta^{\text{opt}} = \arg \max_{\theta} L(\theta)$$

□ **Newton's algorithm** – The Newton's algorithm is a numerical method that finds  $\theta$  such that  $\ell'(\theta) = 0$ . Its update rule is as follows:

$$\theta \leftarrow \theta - \frac{\ell'(\theta)}{\ell''(\theta)}$$

Remark: the multidimensional generalization, also known as the Newton-Raphson method, has the following update rule:

$$\theta \leftarrow \theta - (\nabla_{\theta}^2 \ell(\theta))^{-1} \nabla_{\theta} \ell(\theta)$$

### 1.3 Linear models

#### 1.3.1 Linear regression

We assume here that  $y|x; \theta \sim \mathcal{N}(\mu, \sigma^2)$

□ **Normal equations** – By noting  $X$  the matrix design, the value of  $\theta$  that minimizes the cost function is a closed-form solution such that:

$$\theta = (X^T X)^{-1} X^T y$$

□ **LMS algorithm** – By noting  $\alpha$  the learning rate, the update rule of the Least Mean Squares (LMS) algorithm for a training set of  $m$  data points, which is also known as the Widrow-Hoff learning rule, is as follows:

$$\forall j, \quad \theta_j \leftarrow \theta_j + \alpha \sum_{i=1}^m [y^{(i)} - h_\theta(x^{(i)})] x_j^{(i)}$$

*Remark: the update rule is a particular case of the gradient ascent.*

□ **LWR** – Locally Weighted Regression, also known as LWR, is a variant of linear regression that weights each training example in its cost function by  $w^{(i)}(x)$ , which is defined with parameter  $\tau \in \mathbb{R}$  as:

$$w^{(i)}(x) = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

### 1.3.2 Classification and logistic regression

□ **Sigmoid function** – The sigmoid function  $g$ , also known as the logistic function, is defined as follows:

$$\forall z \in \mathbb{R}, \quad g(z) = \frac{1}{1 + e^{-z}} \in ]0,1[$$

□ **Logistic regression** – We assume here that  $y|x; \theta \sim \text{Bernoulli}(\phi)$ . We have the following form:

$$\phi = p(y=1|x; \theta) = \frac{1}{1 + \exp(-\theta^T x)} = g(\theta^T x)$$

*Remark: there is no closed form solution for the case of logistic regressions.*

□ **Softmax regression** – A softmax regression, also called a multiclass logistic regression, is used to generalize logistic regression when there are more than 2 outcome classes. By convention, we set  $\theta_K = 0$ , which makes the Bernoulli parameter  $\phi_i$  of each class  $i$  equal to:

$$\phi_i = \frac{\exp(\theta_i^T x)}{\sum_{j=1}^K \exp(\theta_j^T x)}$$

### 1.3.3 Generalized Linear Models

□ **Exponential family** – A class of distributions is said to be in the exponential family if it can be written in terms of a natural parameter, also called the canonical parameter or link function,  $\eta$ , a sufficient statistic  $T(y)$  and a log-partition function  $a(\eta)$  as follows:

$$p(y; \eta) = b(y) \exp(\eta T(y) - a(\eta))$$

*Remark: we will often have  $T(y) = y$ . Also,  $\exp(-a(\eta))$  can be seen as a normalization parameter that will make sure that the probabilities sum to one.*

Here are the most common exponential distributions summed up in the following table:

Distribution	$\eta$	$T(y)$	$a(\eta)$	$b(y)$
Bernoulli	$\log\left(\frac{\phi}{1-\phi}\right)$	$y$	$\log(1 + \exp(\eta))$	1
Gaussian	$\mu$	$y$	$\frac{\eta^2}{2}$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$
Poisson	$\log(\lambda)$	$y$	$e^\eta$	$\frac{1}{y!}$
Geometric	$\log(1 - \phi)$	$y$	$\log\left(\frac{e^\eta}{1-e^\eta}\right)$	1

□ **Assumptions of GLMs** – Generalized Linear Models (GLM) aim at predicting a random variable  $y$  as a function of  $x \in \mathbb{R}^{n+1}$  and rely on the following 3 assumptions:

$$(1) \quad y|x; \theta \sim \text{ExpFamily}(\eta) \quad (2) \quad h_\theta(x) = E[y|x; \theta] \quad (3) \quad \eta = \theta^T x$$

*Remark: ordinary least squares and logistic regression are special cases of generalized linear models.*

### 1.4 Support Vector Machines

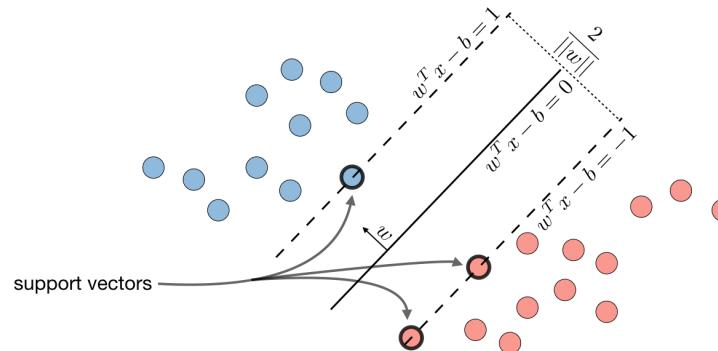
The goal of support vector machines is to find the line that maximizes the minimum distance to the line.

□ **Optimal margin classifier** – The optimal margin classifier  $h$  is such that:

$$h(x) = \text{sign}(w^T x - b)$$

where  $(w, b) \in \mathbb{R}^n \times \mathbb{R}$  is the solution of the following optimization problem:

$$\min \frac{1}{2} \|w\|^2 \quad \text{such that} \quad y^{(i)}(w^T x^{(i)} - b) \geq 1$$



*Remark: the line is defined as  $w^T x - b = 0$ .*

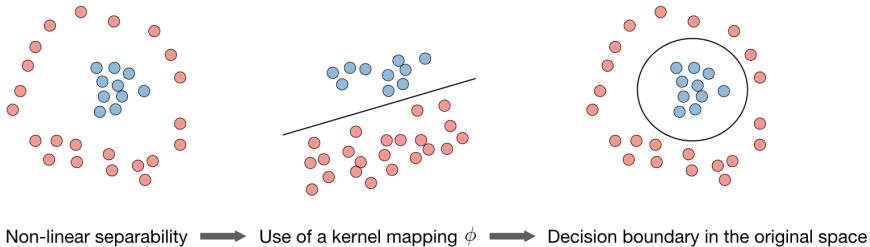
□ **Hinge loss** – The hinge loss is used in the setting of SVMs and is defined as follows:

$$L(z, y) = [1 - yz]_+ = \max(0, 1 - yz)$$

□ **Kernel** – Given a feature mapping  $\phi$ , we define the kernel  $K$  to be defined as:

$$K(x,z) = \phi(x)^T \phi(z)$$

In practice, the kernel  $K$  defined by  $K(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$  is called the Gaussian kernel and is commonly used.



*Remark: we say that we use the "kernel trick" to compute the cost function using the kernel because we actually don't need to know the explicit mapping  $\phi$ , which is often very complicated. Instead, only the values  $K(x,z)$  are needed.*

□ **Lagrangian** – We define the Lagrangian  $\mathcal{L}(w,b)$  as follows:

$$\mathcal{L}(w,b) = f(w) + \sum_{i=1}^l \beta_i h_i(w)$$

*Remark: the coefficients  $\beta_i$  are called the Lagrange multipliers.*

## 1.5 Generative Learning

A generative model first tries to learn how the data is generated by estimating  $P(x|y)$ , which we can then use to estimate  $P(y|x)$  by using Bayes' rule.

### 1.5.1 Gaussian Discriminant Analysis

□ **Setting** – The Gaussian Discriminant Analysis assumes that  $y$  and  $x|y = 0$  and  $x|y = 1$  are such that:

$$y \sim \text{Bernoulli}(\phi)$$

$$x|y = 0 \sim \mathcal{N}(\mu_0, \Sigma) \quad \text{and} \quad x|y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$$

□ **Estimation** – The following table sums up the estimates that we find when maximizing the likelihood:

$\hat{\phi}$	$\hat{\mu}_j \quad (j = 0,1)$	$\hat{\Sigma}$
$\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{y^{(i)}=1\}}$	$\frac{\sum_{i=1}^m \mathbf{1}_{\{y^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^m \mathbf{1}_{\{y^{(i)}=j\}}}$	$\frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$

### 1.5.2 Naive Bayes

□ **Assumption** – The Naive Bayes model supposes that the features of each data point are all independent:

$$P(x|y) = P(x_1, x_2, \dots | y) = P(x_1|y)P(x_2|y)\dots = \prod_{i=1}^n P(x_i|y)$$

□ **Solutions** – Maximizing the log-likelihood gives the following solutions, with  $k \in \{0,1\}$ ,  $l \in [1, L]$

$$P(y = k) = \frac{1}{m} \times \#\{j | y^{(j)} = k\}$$

and

$$P(x_i = l | y = k) = \frac{\#\{j | y^{(j)} = k \text{ and } x_i^{(j)} = l\}}{\#\{j | y^{(j)} = k\}}$$

*Remark: Naive Bayes is widely used for text classification and spam detection.*

## 1.6 Tree-based and ensemble methods

These methods can be used for both regression and classification problems.

□ **CART** – Classification and Regression Trees (CART), commonly known as decision trees, can be represented as binary trees. They have the advantage to be very interpretable.

□ **Random forest** – It is a tree-based technique that uses a high number of decision trees built out of randomly selected sets of features. Contrary to the simple decision tree, it is highly uninterpretable but its generally good performance makes it a popular algorithm.

*Remark: random forests are a type of ensemble methods.*

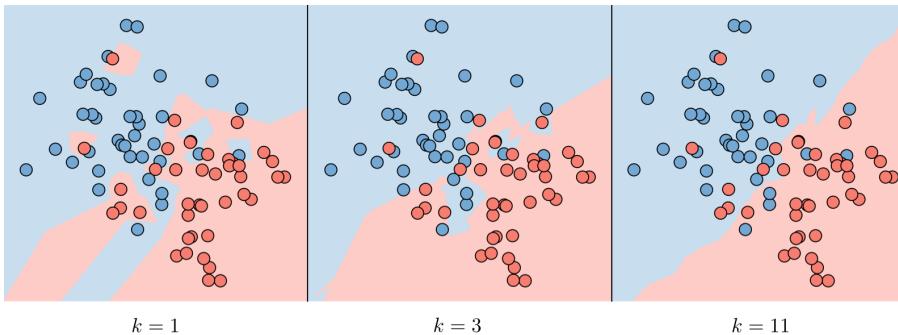
□ **Boosting** – The idea of boosting methods is to combine several weak learners to form a stronger one. The main ones are summed up in the table below:

Adaptive boosting	Gradient boosting
- High weights are put on errors to improve at the next boosting step - Known as Adaboost	- Weak learners trained on remaining errors

## 1.7 Other non-parametric approaches

□  **$k$ -nearest neighbors** – The  $k$ -nearest neighbors algorithm, commonly known as  $k$ -NN, is a non-parametric approach where the response of a data point is determined by the nature of its  $k$  neighbors from the training set. It can be used in both classification and regression settings.

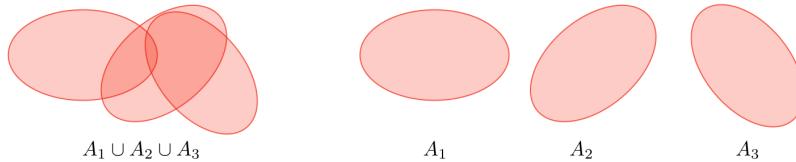
*Remark: The higher the parameter  $k$ , the higher the bias, and the lower the parameter  $k$ , the higher the variance.*



## 1.8 Learning Theory

**□ Union bound** – Let  $A_1, \dots, A_k$  be  $k$  events. We have:

$$P(A_1 \cup \dots \cup A_k) \leq P(A_1) + \dots + P(A_k)$$



**□ Hoeffding inequality** – Let  $Z_1, \dots, Z_m$  be  $m$  iid variables drawn from a Bernoulli distribution of parameter  $\phi$ . Let  $\hat{\phi}$  be their sample mean and  $\gamma > 0$  fixed. We have:

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2 \exp(-2\gamma^2 m)$$

Remark: this inequality is also known as the Chernoff bound.

**□ Training error** – For a given classifier  $h$ , we define the training error  $\hat{\epsilon}(h)$ , also known as the empirical risk or empirical error, to be as follows:

$$\hat{\epsilon}(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{h(x^{(i)}) \neq y^{(i)}\}}$$

**□ Probably Approximately Correct (PAC)** – PAC is a framework under which numerous results on learning theory were proved, and has the following set of assumptions:

- the training and testing sets follow the same distribution
- the training examples are drawn independently

**□ Shattering** – Given a set  $S = \{x^{(1)}, \dots, x^{(d)}\}$ , and a set of classifiers  $\mathcal{H}$ , we say that  $\mathcal{H}$  shatters  $S$  if for any set of labels  $\{y^{(1)}, \dots, y^{(d)}\}$ , we have:

$$\exists h \in \mathcal{H}, \quad \forall i \in \llbracket 1, d \rrbracket, \quad h(x^{(i)}) = y^{(i)}$$

**□ Upper bound theorem** – Let  $\mathcal{H}$  be a finite hypothesis class such that  $|\mathcal{H}| = k$  and let  $\delta$  and the sample size  $m$  be fixed. Then, with probability of at least  $1 - \delta$ , we have:

$$\epsilon(\hat{h}) \leq \left( \min_{h \in \mathcal{H}} \epsilon(h) \right) + 2 \sqrt{\frac{1}{2m} \log \left( \frac{2k}{\delta} \right)}$$

**□ VC dimension** – The Vapnik-Chervonenkis (VC) dimension of a given infinite hypothesis class  $\mathcal{H}$ , noted  $\text{VC}(\mathcal{H})$  is the size of the largest set that is shattered by  $\mathcal{H}$ .

Remark: the VC dimension of  $\mathcal{H} = \{\text{set of linear classifiers in 2 dimensions}\}$  is 3.



**□ Theorem (Vapnik)** – Let  $\mathcal{H}$  be given, with  $\text{VC}(\mathcal{H}) = d$  and  $m$  the number of training examples. With probability at least  $1 - \delta$ , we have:

$$\epsilon(\hat{h}) \leq \left( \min_{h \in \mathcal{H}} \epsilon(h) \right) + O \left( \sqrt{\frac{d}{m} \log \left( \frac{m}{d} \right)} + \frac{1}{m} \log \left( \frac{1}{\delta} \right) \right)$$

## 2 Unsupervised Learning

### 2.1 Introduction to Unsupervised Learning

**Motivation** – The goal of unsupervised learning is to find hidden patterns in unlabeled data  $\{x^{(1)}, \dots, x^{(m)}\}$ .

**Jensen's inequality** – Let  $f$  be a convex function and  $X$  a random variable. We have the following inequality:

$$E[f(X)] \geq f(E[X])$$

### 2.2 Clustering

#### 2.2.1 Expectation-Maximization

**Latent variables** – Latent variables are hidden/unobserved variables that make estimation problems difficult, and are often denoted  $z$ . Here are the most common settings where there are latent variables:

Setting	Latent variable $z$	$x z$	Comments
Mixture of $k$ Gaussians	Multinomial( $\phi$ )	$\mathcal{N}(\mu_j, \Sigma_j)$	$\mu_j \in \mathbb{R}^n, \phi \in \mathbb{R}^k$
Factor analysis	$\mathcal{N}(0, I)$	$\mathcal{N}(\mu + \Lambda z, \psi)$	$\mu_j \in \mathbb{R}^n$

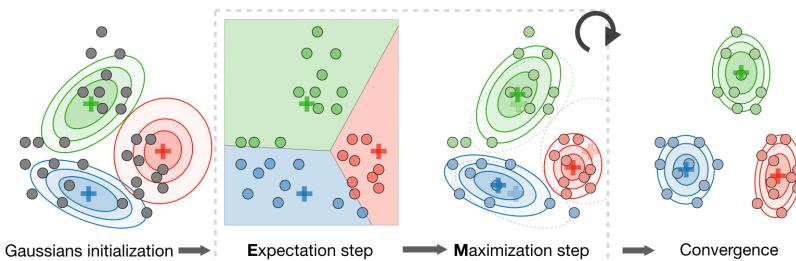
**Algorithm** – The Expectation-Maximization (EM) algorithm gives an efficient method at estimating the parameter  $\theta$  through maximum likelihood estimation by repeatedly constructing a lower-bound on the likelihood (E-step) and optimizing that lower bound (M-step) as follows:

- E-step: Evaluate the posterior probability  $Q_i(z^{(i)})$  that each data point  $x^{(i)}$  came from a particular cluster  $z^{(i)}$  as follows:

$$Q_i(z^{(i)}) = P(z^{(i)}|x^{(i)}; \theta)$$

- M-step: Use the posterior probabilities  $Q_i(z^{(i)})$  as cluster specific weights on data points  $x^{(i)}$  to separately re-estimate each cluster model as follows:

$$\theta_i = \underset{\theta}{\operatorname{argmax}} \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log \left( \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right) dz^{(i)}$$



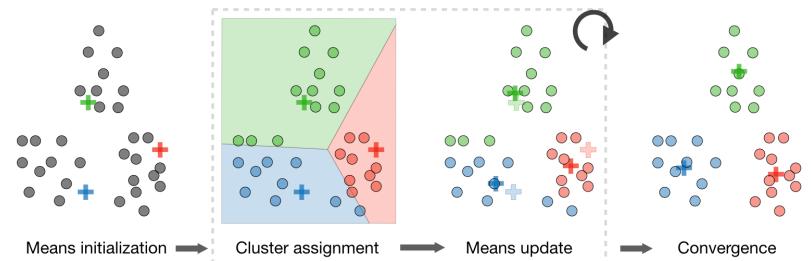
#### 2.2.2 $k$ -means clustering

We note  $c^{(i)}$  the cluster of data point  $i$  and  $\mu_j$  the center of cluster  $j$ .

**Algorithm** – After randomly initializing the cluster centroids  $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$ , the  $k$ -means algorithm repeats the following step until convergence:

$$c^{(i)} = \arg \min_j \|x^{(i)} - \mu_j\|^2$$

$$\mu_j = \frac{\sum_{i=1}^m 1_{\{c^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^m 1_{\{c^{(i)}=j\}}}$$



**Distortion function** – In order to see if the algorithm converges, we look at the distortion function defined as follows:

$$J(c, \mu) = \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

#### 2.2.3 Hierarchical clustering

**Algorithm** – It is a clustering algorithm with an agglomerative hierarchical approach that build nested clusters in a successive manner.

**Types** – There are different sorts of hierarchical clustering algorithms that aims at optimizing different objective functions, which is summed up in the table below:

Ward linkage	Average linkage	Complete linkage
Minimize within cluster distance	Minimize average distance between cluster pairs	Minimize maximum distance of between cluster pairs

#### 2.2.4 Clustering assessment metrics

In an unsupervised learning setting, it is often hard to assess the performance of a model since we don't have the ground truth labels as was the case in the supervised learning setting.

**Silhouette coefficient** – By noting  $a$  and  $b$  the mean distance between a sample and all other points in the same class, and between a sample and all other points in the next nearest cluster, the silhouette coefficient  $s$  for a single sample is defined as follows:

$$s = \frac{b - a}{\max(a, b)}$$

**Calinski-Harabaz index** – By noting  $k$  the number of clusters,  $B_k$  and  $W_k$  the between and within-clustering dispersion matrices respectively defined as

$$B_k = \sum_{j=1}^k n_{c(i)} (\mu_{c(i)} - \mu)(\mu_{c(i)} - \mu)^T, \quad W_k = \sum_{i=1}^m (x^{(i)} - \mu_{c(i)})(x^{(i)} - \mu_{c(i)})^T$$

the Calinski-Harabaz index  $s(k)$  indicates how well a clustering model defines its clusters, such that the higher the score, the more dense and well separated the clusters are. It is defined as follows:

$$s(k) = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \times \frac{N - k}{k - 1}$$

## 2.3 Dimension reduction

### 2.3.1 Principal component analysis

It is a dimension reduction technique that finds the variance maximizing directions onto which to project the data.

**Eigenvalue, eigenvector** – Given a matrix  $A \in \mathbb{R}^{n \times n}$ ,  $\lambda$  is said to be an eigenvalue of  $A$  if there exists a vector  $z \in \mathbb{R}^n \setminus \{0\}$ , called eigenvector, such that we have:

$$Az = \lambda z$$

**Spectral theorem** – Let  $A \in \mathbb{R}^{n \times n}$ . If  $A$  is symmetric, then  $A$  is diagonalizable by a real orthogonal matrix  $U \in \mathbb{R}^{n \times n}$ . By noting  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ , we have:

$$\exists \Lambda \text{ diagonal}, \quad A = U \Lambda U^T$$

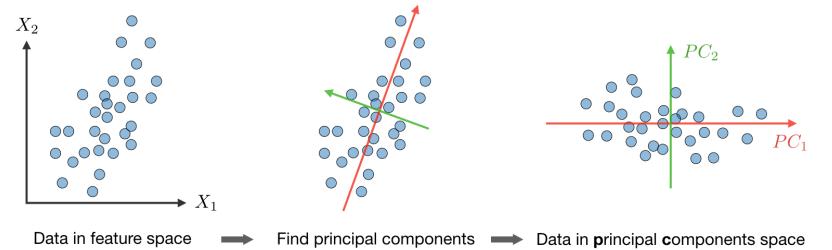
*Remark: the eigenvector associated with the largest eigenvalue is called principal eigenvector of matrix  $A$ .*

**Algorithm** – The Principal Component Analysis (PCA) procedure is a dimension reduction technique that projects the data on  $k$  dimensions by maximizing the variance of the data as follows:

- Step 1: Normalize the data to have a mean of 0 and standard deviation of 1.

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j} \quad \text{where} \quad \mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)} \quad \text{and} \quad \sigma_j^2 = \frac{1}{m} \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

- Step 2: Compute  $\Sigma = \frac{1}{m} \sum_{i=1}^m x^{(i)} x^{(i)T} \in \mathbb{R}^{n \times n}$ , which is symmetric with real eigenvalues.
- Step 3: Compute  $u_1, \dots, u_k \in \mathbb{R}^n$  the  $k$  orthogonal principal eigenvectors of  $\Sigma$ , i.e. the orthogonal eigenvectors of the  $k$  largest eigenvalues.
- Step 4: Project the data on  $\text{span}_{\mathbb{R}}(u_1, \dots, u_k)$ . This procedure maximizes the variance among all  $k$ -dimensional spaces.



### 2.3.2 Independent component analysis

It is a technique meant to find the underlying generating sources.

**Assumptions** – We assume that our data  $x$  has been generated by the  $n$ -dimensional source vector  $s = (s_1, \dots, s_n)$ , where  $s_i$  are independent random variables, via a mixing and non-singular matrix  $A$  as follows:

$$x = As$$

The goal is to find the unmixing matrix  $W = A^{-1}$  by an update rule.

**Bell and Sejnowski ICA algorithm** – This algorithm finds the unmixing matrix  $W$  by following the steps below:

- Write the probability of  $x = As = W^{-1}s$  as:

$$p(x) = \prod_{i=1}^n p_s(w_i^T x) \cdot |W|$$

- Write the log likelihood given our training data  $\{x^{(i)}, i \in [1, m]\}$  and by noting  $g$  the sigmoid function as:

$$l(W) = \sum_{i=1}^m \left( \sum_{j=1}^n \log \left( g'(w_j^T x^{(i)}) \right) + \log |W| \right)$$

Therefore, the stochastic gradient ascent learning rule is such that for each training example  $x^{(i)}$ , we update  $W$  as follows:

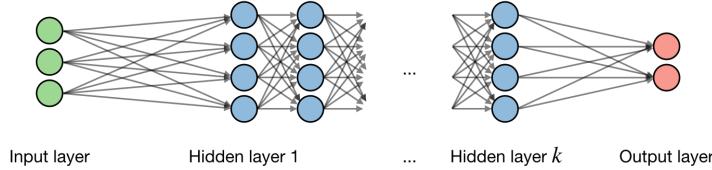
$$W \leftarrow W + \alpha \left( \begin{pmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_n^T x^{(i)}) \end{pmatrix} x^{(i)T} + (W^T)^{-1} \right)$$

### 3 Deep Learning

#### 3.1 Neural Networks

Neural networks are a class of models that are built with layers. Commonly used types of neural networks include convolutional and recurrent neural networks.

**Architecture** – The vocabulary around neural networks architectures is described in the figure below:

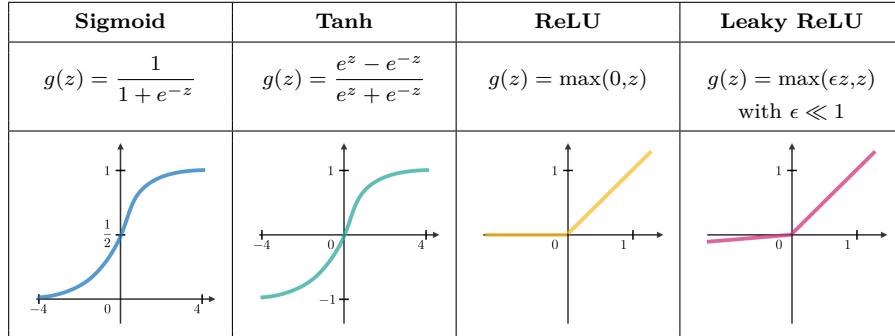


By noting  $i$  the  $i^{th}$  layer of the network and  $j$  the  $j^{th}$  hidden unit of the layer, we have:

$$z_j^{[i]} = w_j^{[i]T} x + b_j^{[i]}$$

where we note  $w$ ,  $b$ ,  $z$  the weight, bias and output respectively.

**Activation function** – Activation functions are used at the end of a hidden unit to introduce non-linear complexities to the model. Here are the most common ones:



**Cross-entropy loss** – In the context of neural networks, the cross-entropy loss  $L(z, y)$  is commonly used and is defined as follows:

$$L(z, y) = - \left[ y \log(z) + (1 - y) \log(1 - z) \right]$$

**Learning rate** – The learning rate, often noted  $\eta$ , indicates at which pace the weights get updated. This can be fixed or adaptively changed. The current most popular method is called Adam, which is a method that adapts the learning rate.

**Backpropagation** – Backpropagation is a method to update the weights in the neural network by taking into account the actual output and the desired output. The derivative with respect to weight  $w$  is computed using chain rule and is of the following form:

$$\frac{\partial L(z, y)}{\partial w} = \frac{\partial L(z, y)}{\partial a} \times \frac{\partial a}{\partial z} \times \frac{\partial z}{\partial w}$$

As a result, the weight is updated as follows:

$$w \leftarrow w - \eta \frac{\partial L(z, y)}{\partial w}$$

**Updating weights** – In a neural network, weights are updated as follows:

- Step 1: Take a batch of training data.
- Step 2: Perform forward propagation to obtain the corresponding loss.
- Step 3: Backpropagate the loss to get the gradients.
- Step 4: Use the gradients to update the weights of the network.

**Dropout** – Dropout is a technique meant at preventing overfitting the training data by dropping out units in a neural network. In practice, neurons are either dropped with probability  $p$  or kept with probability  $1 - p$ .

#### 3.2 Convolutional Neural Networks

**Convolutional layer requirement** – By noting  $W$  the input volume size,  $F$  the size of the convolutional layer neurons,  $P$  the amount of zero padding, then the number of neurons  $N$  that fit in a given volume is such that:

$$N = \frac{W - F + 2P}{S} + 1$$

**Batch normalization** – It is a step of hyperparameter  $\gamma, \beta$  that normalizes the batch  $\{x_i\}$ . By noting  $\mu_B, \sigma_B^2$  the mean and variance of that we want to correct to the batch, it is done as follows:

$$x_i \leftarrow \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

It is usually done after a fully connected/convolutional layer and before a non-linearity layer and aims at allowing higher learning rates and reducing the strong dependence on initialization.

#### 3.3 Recurrent Neural Networks

**Types of gates** – Here are the different types of gates that we encounter in a typical recurrent neural network:

Input gate	Forget gate	Output gate	Gate
Write to cell or not?	Erase a cell or not?	Reveal a cell or not?	How much writing?

**LSTM** – A long short-term memory (LSTM) network is a type of RNN model that avoids the vanishing gradient problem by adding ‘forget’ gates.

### 3.4 Reinforcement Learning and Control

The goal of reinforcement learning is for an agent to learn how to evolve in an environment.

**□ Markov decision processes** – A Markov decision process (MDP) is a 5-tuple  $(S, A, \{P_{sa}\}, \gamma, R)$  where:

- $\mathcal{S}$  is the set of states
- $\mathcal{A}$  is the set of actions
- $\{P_{sa}\}$  are the state transition probabilities for  $s \in \mathcal{S}$  and  $a \in \mathcal{A}$
- $\gamma \in [0, 1[$  is the discount factor
- $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  or  $R : \mathcal{S} \rightarrow \mathbb{R}$  is the reward function that the algorithm wants to maximize

**□ Policy** – A policy  $\pi$  is a function  $\pi : \mathcal{S} \rightarrow \mathcal{A}$  that maps states to actions.

*Remark: we say that we execute a given policy  $\pi$  if given a state  $s$  we take the action  $a = \pi(s)$ .*

**□ Value function** – For a given policy  $\pi$  and a given state  $s$ , we define the value function  $V^\pi$  as follows:

$$V^\pi(s) = E \left[ R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \dots | s_0 = s, \pi \right]$$

**□ Bellman equation** – The optimal Bellman equations characterizes the value function  $V^{\pi^*}$  of the optimal policy  $\pi^*$ :

$$V^{\pi^*}(s) = R(s) + \max_{a \in \mathcal{A}} \gamma \sum_{s' \in \mathcal{S}} P_{sa}(s') V^{\pi^*}(s')$$

*Remark: we note that the optimal policy  $\pi^*$  for a given state  $s$  is such that:*

$$\pi^*(s) = \operatorname{argmax}_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_{sa}(s') V^*(s')$$

**□ Value iteration algorithm** – The value iteration algorithm is in two steps:

- We initialize the value:

$$V_0(s) = 0$$

- We iterate the value based on the values before:

$$V_{i+1}(s) = R(s) + \max_{a \in \mathcal{A}} \left[ \sum_{s' \in \mathcal{S}} \gamma P_{sa}(s') V_i(s') \right]$$

**□ Maximum likelihood estimate** – The maximum likelihood estimates for the state transition probabilities are as follows:

$$P_{sa}(s') = \frac{\text{\#times took action } a \text{ in state } s \text{ and got to } s'}{\text{\#times took action } a \text{ in state } s}$$

**□ Q-learning** –  $Q$ -learning is a model-free estimation of  $Q$ , which is done as follows:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[ R(s, a, s') + \gamma \max_{a'} Q(s', a') - Q(s, a) \right]$$

## 4 Machine Learning Tips and Tricks

### 4.1 Metrics

Given a set of data points  $\{x^{(1)}, \dots, x^{(m)}\}$ , where each  $x^{(i)}$  has  $n$  features, associated to a set of outcomes  $\{y^{(1)}, \dots, y^{(m)}\}$ , we want to assess a given classifier that learns how to predict  $y$  from  $x$ .

#### 4.1.1 Classification

In a context of a binary classification, here are the main metrics that are important to track to assess the performance of the model.

**Confusion matrix** – The confusion matrix is used to have a more complete picture when assessing the performance of a model. It is defined as follows:

		Predicted class	
		+	-
Actual class	+	TP True Positives	FN False Negatives Type II error
	-	FP False Positives Type I error	TN True Negatives

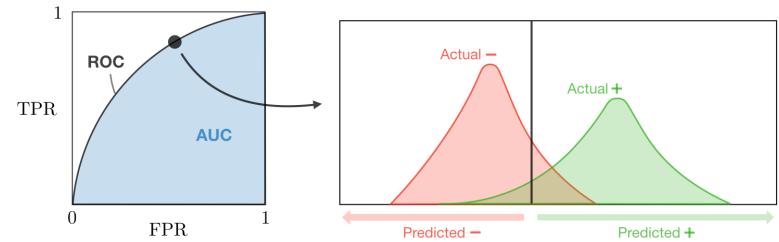
**Main metrics** – The following metrics are commonly used to assess the performance of classification models:

Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall performance of model
Precision	$\frac{TP}{TP + FP}$	How accurate the positive predictions are
Recall Sensitivity	$\frac{TP}{TP + FN}$	Coverage of actual positive sample
Specificity	$\frac{TN}{TN + FP}$	Coverage of actual negative sample
F1 score	$\frac{2TP}{2TP + FP + FN}$	Hybrid metric useful for unbalanced classes

**ROC** – The receiver operating curve, also noted ROC, is the plot of TPR versus FPR by varying the threshold. These metrics are summed up in the table below:

Metric	Formula	Equivalent
True Positive Rate TPR	$\frac{TP}{TP + FN}$	Recall, sensitivity
False Positive Rate FPR	$\frac{FP}{TN + FP}$	1-specificity

**AUC** – The area under the receiving operating curve, also noted AUC or AUROC, is the area below the ROC as shown in the following figure:



#### 4.1.2 Regression

**Basic metrics** – Given a regression model  $f$ , the following metrics are commonly used to assess the performance of the model:

Total sum of squares	Explained sum of squares	Residual sum of squares
$SS_{tot} = \sum_{i=1}^m (y_i - \bar{y})^2$	$SS_{reg} = \sum_{i=1}^m (f(x_i) - \bar{y})^2$	$SS_{res} = \sum_{i=1}^m (y_i - f(x_i))^2$

**Coefficient of determination** – The coefficient of determination, often noted  $R^2$  or  $r^2$ , provides a measure of how well the observed outcomes are replicated by the model and is defined as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

**Main metrics** – The following metrics are commonly used to assess the performance of regression models, by taking into account the number of variables  $n$  that they take into consideration:

Mallow's Cp	AIC	BIC	Adjusted $R^2$
$\frac{SS_{res} + 2(n+1)\hat{\sigma}^2}{m}$	$2[(n+2) - \log(L)]$	$\log(m)(n+2) - 2\log(L)$	$1 - \frac{(1-R^2)(m-1)}{m-n-1}$

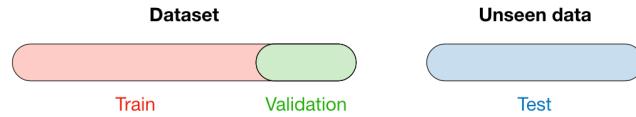
where  $L$  is the likelihood and  $\hat{\sigma}^2$  is an estimate of the variance associated with each response.

## 4.2 Model selection

**Vocabulary** – When selecting a model, we distinguish 3 different parts of the data that we have as follows:

Training set	Validation set	Testing set
- Model is trained - Usually 80% of the dataset	- Model is assessed - Usually 20% of the dataset - Also called hold-out or development set	- Model gives predictions - Unseen data

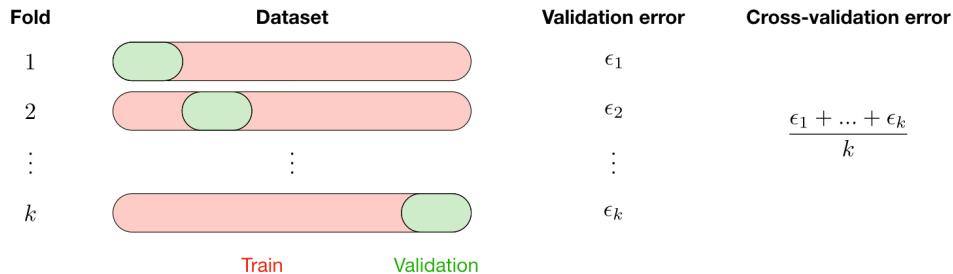
Once the model has been chosen, it is trained on the entire dataset and tested on the unseen test set. These are represented in the figure below:



**Cross-validation** – Cross-validation, also noted CV, is a method that is used to select a model that does not rely too much on the initial training set. The different types are summed up in the table below:

<i>k</i> -fold	Leave- <i>p</i> -out
- Training on $k - 1$ folds and assessment on the remaining one - Generally $k = 5$ or 10	- Training on $n - p$ observations and assessment on the $p$ remaining ones - Case $p = 1$ is called leave-one-out

The most commonly used method is called *k*-fold cross-validation and splits the training data into *k* folds to validate the model on one fold while training the model on the  $k - 1$  other folds, all of this *k* times. The error is then averaged over the *k* folds and is named cross-validation error.



**Regularization** – The regularization procedure aims at avoiding the model to overfit the data and thus deals with high variance issues. The following table sums up the different types of commonly used regularization techniques:

LASSO	Ridge	Elastic Net
- Shrinks coefficients to 0 - Good for variable selection	Makes coefficients smaller	Tradeoff between variable selection and small coefficients
$\dots + \lambda \ \theta\ _1$ $\lambda \in \mathbb{R}$	$\dots + \lambda \ \theta\ _2^2$ $\lambda \in \mathbb{R}$	$\dots + \lambda [(1-\alpha)\ \theta\ _1 + \alpha\ \theta\ _2^2]$ $\lambda \in \mathbb{R}, \alpha \in [0,1]$

**Model selection** – Train model on training set, then evaluate on the development set, then pick best performance model on the development set, and retrain all of that model on the whole training set.

## 4.3 Diagnostics

**Bias** – The bias of a model is the difference between the expected prediction and the correct model that we try to predict for given data points.

**Variance** – The variance of a model is the variability of the model prediction for given data points.

**Bias/variance tradeoff** – The simpler the model, the higher the bias, and the more complex the model, the higher the variance.

	Underfitting	Just right	Overfitting
<b>Symptoms</b>	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error - High variance	- Low training error - Training error much lower than test error - High variance
<b>Regression</b>			

Classification		
Deep learning		
Remedies	<ul style="list-style-type: none"> <li>- Complexify model</li> <li>- Add more features</li> <li>- Train longer</li> </ul>	<ul style="list-style-type: none"> <li>- Regularize</li> <li>- Get more data</li> </ul>

□ **Error analysis** – Error analysis is analyzing the root cause of the difference in performance between the current and the perfect models.

□ **Ablative analysis** – Ablative analysis is analyzing the root cause of the difference in performance between the current and the baseline models.

## 5 Refreshers

### 5.1 Probabilities and Statistics

#### 5.1.1 Introduction to Probability and Combinatorics

□ **Sample space** – The set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted by  $S$ .

□ **Event** – Any subset  $E$  of the sample space is known as an event. That is, an event is a set consisting of possible outcomes of the experiment. If the outcome of the experiment is contained in  $E$ , then we say that  $E$  has occurred.

□ **Axioms of probability** – For each event  $E$ , we denote  $P(E)$  as the probability of event  $E$  occurring. By noting  $E_1, \dots, E_n$  mutually exclusive events, we have the 3 following axioms:

$$(1) \quad 0 \leq P(E) \leq 1 \quad (2) \quad P(S) = 1 \quad (3) \quad P\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i)$$

□ **Permutation** – A permutation is an arrangement of  $r$  objects from a pool of  $n$  objects, in a given order. The number of such arrangements is given by  $P(n, r)$ , defined as:

$$P(n, r) = \frac{n!}{(n - r)!}$$

□ **Combination** – A combination is an arrangement of  $r$  objects from a pool of  $n$  objects, where the order does not matter. The number of such arrangements is given by  $C(n, r)$ , defined as:

$$C(n, r) = \frac{P(n, r)}{r!} = \frac{n!}{r!(n - r)!}$$

*Remark: we note that for  $0 \leq r \leq n$ , we have  $P(n, r) \geq C(n, r)$ .*

#### 5.1.2 Conditional Probability

□ **Bayes' rule** – For events  $A$  and  $B$  such that  $P(B) > 0$ , we have:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

*Remark: we have  $P(A \cap B) = P(A)P(B|A) = P(A|B)P(B)$ .*

□ **Partition** – Let  $\{A_i, i \in [1, n]\}$  be such that for all  $i$ ,  $A_i \neq \emptyset$ . We say that  $\{A_i\}$  is a partition if we have:

$$\forall i \neq j, A_i \cap A_j = \emptyset \quad \text{and} \quad \bigcup_{i=1}^n A_i = S$$

*Remark: for any event  $B$  in the sample space, we have  $P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$ .*

**□ Extended form of Bayes' rule** – Let  $\{A_i, i \in [1, n]\}$  be a partition of the sample space. We have:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

**□ Independence** – Two events  $A$  and  $B$  are independent if and only if we have:

$$P(A \cap B) = P(A)P(B)$$

### 5.1.3 Random Variables

**□ Random variable** – A random variable, often noted  $X$ , is a function that maps every element in a sample space to a real line.

**□ Cumulative distribution function (CDF)** – The cumulative distribution function  $F$ , which is monotonically non-decreasing and is such that  $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow +\infty} F(x) = 1$ , is defined as:

$$F(x) = P(X \leq x)$$

Remark: we have  $P(a < X \leq b) = F(b) - F(a)$ .

**□ Probability density function (PDF)** – The probability density function  $f$  is the probability that  $X$  takes on values between two adjacent realizations of the random variable.

**□ Relationships involving the PDF and CDF** – Here are the important properties to know in the discrete (D) and the continuous (C) cases.

Case	CDF $F$	PDF $f$	Properties of PDF
(D)	$F(x) = \sum_{x_i \leq x} P(X = x_i)$	$f(x_j) = P(X = x_j)$	$0 \leq f(x_j) \leq 1$ and $\sum_j f(x_j) = 1$
(C)	$F(x) = \int_{-\infty}^x f(y)dy$	$f(x) = \frac{dF}{dx}$	$f(x) \geq 0$ and $\int_{-\infty}^{+\infty} f(x)dx = 1$

**□ Variance** – The variance of a random variable, often noted  $\text{Var}(X)$  or  $\sigma^2$ , is a measure of the spread of its distribution function. It is determined as follows:

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

**□ Standard deviation** – The standard deviation of a random variable, often noted  $\sigma$ , is a measure of the spread of its distribution function which is compatible with the units of the actual random variable. It is determined as follows:

$$\sigma = \sqrt{\text{Var}(X)}$$

**□ Expectation and Moments of the Distribution** – Here are the expressions of the expected value  $E[X]$ , generalized expected value  $E[g(X)]$ ,  $k^{\text{th}}$  moment  $E[X^k]$  and characteristic function  $\psi(\omega)$  for the discrete and continuous cases:

Case	$E[X]$	$E[g(X)]$	$E[X^k]$	$\psi(\omega)$
(D)	$\sum_{i=1}^n x_i f(x_i)$	$\sum_{i=1}^n g(x_i) f(x_i)$	$\sum_{i=1}^n x_i^k f(x_i)$	$\sum_{i=1}^n f(x_i) e^{i\omega x_i}$
(C)	$\int_{-\infty}^{+\infty} xf(x)dx$	$\int_{-\infty}^{+\infty} g(x)f(x)dx$	$\int_{-\infty}^{+\infty} x^k f(x)dx$	$\int_{-\infty}^{+\infty} f(x) e^{i\omega x} dx$

Remark: we have  $e^{i\omega x} = \cos(\omega x) + i \sin(\omega x)$ .

**□ Revisiting the  $k^{\text{th}}$  moment** – The  $k^{\text{th}}$  moment can also be computed with the characteristic function as follows:

$$E[X^k] = \frac{1}{i^k} \left[ \frac{\partial^k \psi}{\partial \omega^k} \right]_{\omega=0}$$

**□ Transformation of random variables** – Let the variables  $X$  and  $Y$  be linked by some function. By noting  $f_X$  and  $f_Y$  the distribution function of  $X$  and  $Y$  respectively, we have:

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$$

**□ Leibniz integral rule** – Let  $g$  be a function of  $x$  and potentially  $c$ , and  $a, b$  boundaries that may depend on  $c$ . We have:

$$\frac{\partial}{\partial c} \left( \int_a^b g(x)dx \right) = \frac{\partial b}{\partial c} \cdot g(b) - \frac{\partial a}{\partial c} \cdot g(a) + \int_a^b \frac{\partial g}{\partial c}(x)dx$$

**□ Chebyshev's inequality** – Let  $X$  be a random variable with expected value  $\mu$  and standard deviation  $\sigma$ . For  $k, \sigma > 0$ , we have the following inequality:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

### 5.1.4 Jointly Distributed Random Variables

**□ Conditional density** – The conditional density of  $X$  with respect to  $Y$ , often noted  $f_{X|Y}$ , is defined as follows:

$$f_{X|Y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

**□ Independence** – Two random variables  $X$  and  $Y$  are said to be independent if we have:

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

□ **Marginal density and cumulative distribution** – From the joint density probability function  $f_{XY}$ , we have:

Case	Marginal density	Cumulative function
(D)	$f_X(x_i) = \sum_j f_{XY}(x_i, y_j)$	$F_{XY}(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f_{XY}(x_i, y_j)$
(C)	$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy$	$F_{XY}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{XY}(x', y') dx' dy'$

□ **Distribution of a sum of independent random variables** – Let  $Y = X_1 + \dots + X_n$  with  $X_1, \dots, X_n$  independent. We have:

$$\psi_Y(\omega) = \prod_{k=1}^n \psi_{X_k}(\omega)$$

□ **Covariance** – We define the covariance of two random variables  $X$  and  $Y$ , that we note  $\sigma_{XY}^2$  or more commonly  $\text{Cov}(X, Y)$ , as follows:

$$\text{Cov}(X, Y) \triangleq \sigma_{XY}^2 = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$$

□ **Correlation** – By noting  $\sigma_X, \sigma_Y$  the standard deviations of  $X$  and  $Y$ , we define the correlation between the random variables  $X$  and  $Y$ , noted  $\rho_{XY}$ , as follows:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Remarks: For any  $X, Y$ , we have  $\rho_{XY} \in [-1, 1]$ . If  $X$  and  $Y$  are independent, then  $\rho_{XY} = 0$ .

□ **Main distributions** – Here are the main distributions to have in mind:

Type	Distribution	PDF	$\psi(\omega)$	$E[X]$	$\text{Var}(X)$
(D)	$X \sim \mathcal{B}(n, p)$ Binomial	$P(X = x) = \binom{n}{x} p^x q^{n-x}$ $x \in \llbracket 0, n \rrbracket$	$(pe^{i\omega} + q)^n$	$np$	$npq$
	$X \sim \text{Po}(\mu)$ Poisson	$P(X = x) = \frac{\mu^x}{x!} e^{-\mu}$ $x \in \mathbb{N}$	$e^{\mu(e^{i\omega}-1)}$	$\mu$	$\mu$
(C)	$X \sim \mathcal{U}(a, b)$ Uniform	$f(x) = \frac{1}{b-a}$ $x \in [a, b]$	$\frac{e^{i\omega b} - e^{i\omega a}}{(b-a)i\omega}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	$X \sim \mathcal{N}(\mu, \sigma)$ Gaussian	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2}$ $x \in \mathbb{R}$	$e^{i\omega\mu - \frac{1}{2}\omega^2\sigma^2}$	$\mu$	$\sigma^2$
	$X \sim \text{Exp}(\lambda)$ Exponential	$f(x) = \lambda e^{-\lambda x}$ $x \in \mathbb{R}_+$	$\frac{1}{1 - \frac{i\omega}{\lambda}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

## 5.1.5 Parameter estimation

□ **Random sample** – A random sample is a collection of  $n$  random variables  $X_1, \dots, X_n$  that are independent and identically distributed with  $X$ .

□ **Estimator** – An estimator  $\hat{\theta}$  is a function of the data that is used to infer the value of an unknown parameter  $\theta$  in a statistical model.

□ **Bias** – The bias of an estimator  $\hat{\theta}$  is defined as being the difference between the expected value of the distribution of  $\hat{\theta}$  and the true value, i.e.:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Remark: an estimator is said to be unbiased when we have  $E[\hat{\theta}] = \theta$ .

□ **Sample mean and variance** – The sample mean and the sample variance of a random sample are used to estimate the true mean  $\mu$  and the true variance  $\sigma^2$  of a distribution, are noted  $\bar{X}$  and  $s^2$  respectively, and are such that:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad s^2 = \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

□ **Central Limit Theorem** – Let us have a random sample  $X_1, \dots, X_n$  following a given distribution with mean  $\mu$  and variance  $\sigma^2$ , then we have:

$$\bar{X} \underset{n \rightarrow +\infty}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

## 5.2 Linear Algebra and Calculus

### 5.2.1 General notations

□ **Vector** – We note  $x \in \mathbb{R}^n$  a vector with  $n$  entries, where  $x_i \in \mathbb{R}$  is the  $i^{th}$  entry:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$$

□ **Matrix** – We note  $A \in \mathbb{R}^{m \times n}$  a matrix with  $m$  rows and  $n$  columns, where  $A_{i,j} \in \mathbb{R}$  is the entry located in the  $i^{th}$  row and  $j^{th}$  column:

$$A = \begin{pmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Remark: the vector  $x$  defined above can be viewed as a  $n \times 1$  matrix and is more particularly called a column-vector.

□ **Identity matrix** – The identity matrix  $I \in \mathbb{R}^{n \times n}$  is a square matrix with ones in its diagonal and zero everywhere else:

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

Remark: for all matrices  $A \in \mathbb{R}^{n \times n}$ , we have  $A \times I = I \times A = A$ .

**□ Diagonal matrix** – A diagonal matrix  $D \in \mathbb{R}^{n \times n}$  is a square matrix with nonzero values in its diagonal and zero everywhere else:

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{pmatrix}$$

Remark: we also note  $D$  as  $\text{diag}(d_1, \dots, d_n)$ .

## 5.2.2 Matrix operations

**□ Vector-vector multiplication** – There are two types of vector-vector products:

- inner product: for  $x, y \in \mathbb{R}^n$ , we have:

$$x^T y = \sum_{i=1}^n x_i y_i \in \mathbb{R}$$

- outer product: for  $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ , we have:

$$xy^T = \begin{pmatrix} x_1 y_1 & \cdots & x_1 y_n \\ \vdots & & \vdots \\ x_m y_1 & \cdots & x_m y_n \end{pmatrix} \in \mathbb{R}^{m \times n}$$

**□ Matrix-vector multiplication** – The product of matrix  $A \in \mathbb{R}^{m \times n}$  and vector  $x \in \mathbb{R}^n$  is a vector of size  $\mathbb{R}^m$ , such that:

$$Ax = \begin{pmatrix} a_{r,1}^T x \\ \vdots \\ a_{r,m}^T x \end{pmatrix} = \sum_{i=1}^n a_{c,i} x_i \in \mathbb{R}^m$$

where  $a_{r,i}^T$  are the vector rows and  $a_{c,j}$  are the vector columns of  $A$ , and  $x_i$  are the entries of  $x$ .

**□ Matrix-matrix multiplication** – The product of matrices  $A \in \mathbb{R}^{m \times n}$  and  $B \in \mathbb{R}^{n \times p}$  is a matrix of size  $\mathbb{R}^{m \times p}$ , such that:

$$AB = \begin{pmatrix} a_{r,1}^T b_{c,1} & \cdots & a_{r,1}^T b_{c,p} \\ \vdots & & \vdots \\ a_{r,m}^T b_{c,1} & \cdots & a_{r,m}^T b_{c,p} \end{pmatrix} = \sum_{i=1}^n a_{c,i} b_{r,i}^T \in \mathbb{R}^{m \times p}$$

where  $a_{r,i}^T, b_{r,i}^T$  are the vector rows and  $a_{c,j}, b_{c,j}$  are the vector columns of  $A$  and  $B$  respectively.

**□ Transpose** – The transpose of a matrix  $A \in \mathbb{R}^{m \times n}$ , noted  $A^T$ , is such that its entries are flipped:

$$\forall i, j, \quad A_{i,j}^T = A_{j,i}$$

Remark: for matrices  $A, B$ , we have  $(AB)^T = B^T A^T$ .

**□ Inverse** – The inverse of an invertible square matrix  $A$  is noted  $A^{-1}$  and is the only matrix such that:

$$AA^{-1} = A^{-1}A = I$$

Remark: not all square matrices are invertible. Also, for matrices  $A, B$ , we have  $(AB)^{-1} = B^{-1}A^{-1}$

**□ Trace** – The trace of a square matrix  $A$ , noted  $\text{tr}(A)$ , is the sum of its diagonal entries:

$$\text{tr}(A) = \sum_{i=1}^n A_{i,i}$$

Remark: for matrices  $A, B$ , we have  $\text{tr}(A^T) = \text{tr}(A)$  and  $\text{tr}(AB) = \text{tr}(BA)$

**□ Determinant** – The determinant of a square matrix  $A \in \mathbb{R}^{n \times n}$ , noted  $|A|$  or  $\det(A)$  is expressed recursively in terms of  $A_{\setminus i, \setminus j}$ , which is the matrix  $A$  without its  $i^{\text{th}}$  row and  $j^{\text{th}}$  column, as follows:

$$\det(A) = |A| = \sum_{j=1}^n (-1)^{i+j} A_{i,j} |A_{\setminus i, \setminus j}|$$

Remark:  $A$  is invertible if and only if  $|A| \neq 0$ . Also,  $|AB| = |A||B|$  and  $|A^T| = |A|$ .

## 5.2.3 Matrix properties

**□ Symmetric decomposition** – A given matrix  $A$  can be expressed in terms of its symmetric and antisymmetric parts as follows:

$$A = \underbrace{\frac{A + A^T}{2}}_{\text{Symmetric}} + \underbrace{\frac{A - A^T}{2}}_{\text{Antisymmetric}}$$

**□ Norm** – A norm is a function  $N : V \rightarrow [0, +\infty]$  where  $V$  is a vector space, and such that for all  $x, y \in V$ , we have:

- $N(x + y) \leq N(x) + N(y)$
- $N(ax) = |a|N(x)$  for  $a$  scalar
- if  $N(x) = 0$ , then  $x = 0$

For  $x \in V$ , the most commonly used norms are summed up in the table below:

Norm	Notation	Definition	Use case
Manhattan, $L^1$	$\ x\ _1$	$\sum_{i=1}^n  x_i $	LASSO regularization
Euclidean, $L^2$	$\ x\ _2$	$\sqrt{\sum_{i=1}^n x_i^2}$	Ridge regularization
$p$ -norm, $L^p$	$\ x\ _p$	$\left(\sum_{i=1}^n x_i^p\right)^{\frac{1}{p}}$	Hölder inequality
Infinity, $L^\infty$	$\ x\ _\infty$	$\max_i  x_i $	Uniform convergence

□ **Linearly dependence** – A set of vectors is said to be linearly dependent if one of the vectors in the set can be defined as a linear combination of the others.

*Remark: if no vector can be written this way, then the vectors are said to be linearly independent.*

□ **Matrix rank** – The rank of a given matrix  $A$  is noted  $\text{rank}(A)$  and is the dimension of the vector space generated by its columns. This is equivalent to the maximum number of linearly independent columns of  $A$ .

□ **Positive semi-definite matrix** – A matrix  $A \in \mathbb{R}^{n \times n}$  is positive semi-definite (PSD) and is noted  $A \succeq 0$  if we have:

$$A = A^T \quad \text{and} \quad \forall x \in \mathbb{R}^n, \quad x^T Ax \geq 0$$

*Remark: similarly, a matrix  $A$  is said to be positive definite, and is noted  $A \succ 0$ , if it is a PSD matrix which satisfies for all non-zero vector  $x$ ,  $x^T Ax > 0$ .*

□ **Eigenvalue, eigenvector** – Given a matrix  $A \in \mathbb{R}^{n \times n}$ ,  $\lambda$  is said to be an eigenvalue of  $A$  if there exists a vector  $z \in \mathbb{R}^n \setminus \{0\}$ , called eigenvector, such that we have:

$$Az = \lambda z$$

□ **Spectral theorem** – Let  $A \in \mathbb{R}^{n \times n}$ . If  $A$  is symmetric, then  $A$  is diagonalizable by a real orthogonal matrix  $U \in \mathbb{R}^{n \times n}$ . By noting  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ , we have:

$$\exists \Lambda \text{ diagonal}, \quad A = U \Lambda U^T$$

□ **Singular-value decomposition** – For a given matrix  $A$  of dimensions  $m \times n$ , the singular-value decomposition (SVD) is a factorization technique that guarantees the existence of  $U$   $m \times m$  unitary,  $\Sigma$   $m \times n$  diagonal and  $V$   $n \times n$  unitary matrices, such that:

$$A = U \Sigma V^T$$

## 5.2.4 Matrix calculus

□ **Gradient** – Let  $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$  be a function and  $A \in \mathbb{R}^{m \times n}$  be a matrix. The gradient of  $f$  with respect to  $A$  is a  $m \times n$  matrix, noted  $\nabla_A f(A)$ , such that:

$$\left( \nabla_A f(A) \right)_{i,j} = \frac{\partial f(A)}{\partial A_{i,j}}$$

*Remark: the gradient of  $f$  is only defined when  $f$  is a function that returns a scalar.*

□ **Hessian** – Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function and  $x \in \mathbb{R}^n$  be a vector. The hessian of  $f$  with respect to  $x$  is a  $n \times n$  symmetric matrix, noted  $\nabla_x^2 f(x)$ , such that:

$$\left( \nabla_x^2 f(x) \right)_{i,j} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

*Remark: the hessian of  $f$  is only defined when  $f$  is a function that returns a scalar.*

□ **Gradient operations** – For matrices  $A, B, C$ , the following gradient properties are worth having in mind:

$$\nabla_A \text{tr}(AB) = B^T$$

$$\nabla_A f(A) = (\nabla_A f(A))^T$$

$$\nabla_A \text{tr}(ABA^T C) = CAB + C^T AB^T$$

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