

MACHINE LEARNING

Science is like sex: sometimes something useful
come out, but that is not the reason we are doing it

– Richard Feynman

Marco Gori

*Dipartimento di Ingegneria dell'Informazione e
Scienze Matematiche*

Università di Siena

Machine Learning, University of Siena
2018-2019

Machine Learning

A CONSTRAINT-BASED APPROACH



MK
MORGAN KAUFMANN

Marco Gori

Contents

Preface	xiii
Notes on the Exercises	xix
CHAPTER 1 The Big Picture	2
1.1 Why Do Machines Need to Learn?	3
1.1.1 Learning Tasks	4
1.1.2 Symbolic and Subsymbolic Representations of the Environment	9
1.1.3 Biological and Artificial Neural Networks	11
1.1.4 Protocols of Learning	13
1.1.5 Constraint-Based Learning	19
1.2 Principles and Practice	28
1.2.1 The Puzzling Nature of Induction	28
1.2.2 Learning Principles	34
1.2.3 The Role of Time in Learning Processes	34
1.2.4 Focus of Attention	35
1.3 Hands-on Experience	38
1.3.1 Measuring the Success of Experiments	39
1.3.2 Handwritten Character Recognition	40
1.3.3 Setting up a Machine Learning Experiment	42
1.3.4 Test and Experimental Remarks	45
1.4 Challenges in Machine Learning	50
1.4.1 Learning to See	50
1.4.2 Speech Understanding	51
1.4.3 Agents Living in Their Own Environment	52
1.5 Scholia	54
CHAPTER 2 Learning Principles	60
2.1 Environmental Constraints	61
2.1.1 Loss and Risk Functions	61
2.1.2 Ill-Position of Constraint-Induced Risk Functions	69
2.1.3 Risk Minimization	71
2.1.4 The Bias–Variance Dilemma	75
2.2 Statistical Learning	83
2.2.1 Maximum Likelihood Estimation	83
2.2.2 Bayesian Inference	86
2.2.3 Bayesian Learning	88
2.2.4 Graphical Models	89
2.2.5 Frequentist and Bayesian Approach	92
2.3 Information-Based Learning	95
2.3.1 A Motivating Example	95

2.3.2	Principle of Maximum Entropy	97
2.3.3	Maximum Mutual Information	99
2.4	Learning Under the Parsimony Principle	104
2.4.1	The Parsimony Principle	104
2.4.2	Minimum Description Length	104
2.4.3	MDL and Regularization	110
2.4.4	Statistical Interpretation of Regularization	113
2.5	Scholia	115
CHAPTER 3	Linear Threshold Machines	122
3.1	Linear Machines	123
3.1.1	Normal Equations	128
3.1.2	Undetermined Problems and Pseudoinversion	129
3.1.3	Ridge Regression	132
3.1.4	Primal and Dual Representations	134
3.2	Linear Machines With Threshold Units	141
3.2.1	Predicate-Order and Representational Issues	142
3.2.2	Optimality for Linearly-Separable Examples	149
3.2.3	Failing to Separate	151
3.3	Statistical View	155
3.3.1	Bayesian Decision and Linear Discrimination	155
3.3.2	Logistic Regression	156
3.3.3	The Parsimony Principle Meets the Bayesian Decision	158
3.3.4	LMS in the Statistical Framework	159
3.4	Algorithmic Issues	162
3.4.1	Gradient Descent	162
3.4.2	Stochastic Gradient Descent	164
3.4.3	The Perceptron Algorithm	165
3.4.4	Complexity Issues	169
3.5	Scholia	175
CHAPTER 4	Kernel Machines	186
4.1	Feature Space	187
4.1.1	Polynomial Preprocessing	187
4.1.2	Boolean Enrichment	188
4.1.3	Invariant Feature Maps	189
4.1.4	Linear-Separability in High-Dimensional Spaces	190
4.2	Maximum Margin Problem	194
4.2.1	Classification Under Linear-Separability	194
4.2.2	Dealing With Soft-Constraints	198
4.2.3	Regression	201
4.3	Kernel Functions	207
4.3.1	Similarity and Kernel Trick	207
4.3.2	Characterization of Kernels	208

4.3.3	The Reproducing Kernel Map	212
4.3.4	Types of Kernels	214
4.4	Regularization	220
4.4.1	Regularized Risks	220
4.4.2	Regularization in RKHS	222
4.4.3	Minimization of Regularized Risks	223
4.4.4	Regularization Operators	224
4.5	Scholia	230
CHAPTER 5	Deep Architectures	236
5.1	Architectural Issues	237
5.1.1	Digraphs and Feedforward Networks	238
5.1.2	Deep Paths	240
5.1.3	From Deep to Relaxation-Based Architectures	243
5.1.4	Classifiers, Regressors, and Auto-Encoders	244
5.2	Realization of Boolean Functions	247
5.2.1	Canonical Realizations by and-or Gates	247
5.2.2	Universal nand Realization	251
5.2.3	Shallow vs Deep Realizations	251
5.2.4	LTU-Based Realizations and Complexity Issues	254
5.3	Realization of Real-Valued Functions	265
5.3.1	Computational Geometry-Based Realizations	265
5.3.2	Universal Approximation	268
5.3.3	Solution Space and Separation Surfaces	271
5.3.4	Deep Networks and Representational Issues	276
5.4	Convolutional Networks	280
5.4.1	Kernels, Convolutions, and Receptive Fields	280
5.4.2	Incorporating Invariance	288
5.4.3	Deep Convolutional Networks	293
5.5	Learning in Feedforward Networks	298
5.5.1	Supervised Learning	298
5.5.2	Backpropagation	298
5.5.3	Symbolic and Automatic Differentiation	306
5.5.4	Regularization Issues	308
5.6	Complexity Issues	313
5.6.1	On the Problem of Local Minima	313
5.6.2	Facing Saturation	319
5.6.3	Complexity and Numerical Issues	323
5.7	Scholia	326
CHAPTER 6	Learning and Reasoning With Constraints	340
6.1	Constraint Machines	343
6.1.1	Walking Through Learning and Inference	343
6.1.2	A Unified View of Constrained Environments	352

6.1.3	Functional Representation of Learning Tasks	359
6.1.4	Reasoning With Constraints	364
6.2	Logic Constraints in the Environment	373
6.2.1	Formal Logic and Complexity of Reasoning	373
6.2.2	Environments With Symbols and Subsymbols	376
6.2.3	T-Norms	384
6.2.4	Łukasiewicz Propositional Logic	388
6.3	Diffusion Machines	392
6.3.1	Data Models	393
6.3.2	Diffusion in Spatiotemporal Environments	399
6.3.3	Recurrent Neural Networks	400
6.4	Algorithmic Issues	404
6.4.1	Pointwise Content-Based Constraints	405
6.4.2	Propositional Constraints in the Input Space	408
6.4.3	Supervised Learning With Linear Constraints	413
6.4.4	Learning Under Diffusion Constraints	416
6.5	Life-Long Learning Agents	424
6.5.1	Cognitive Action and Temporal Manifolds	425
6.5.2	Energy Balance	430
6.5.3	Focus of Attention, Teaching, and Active Learning	431
6.5.4	Developmental Learning	433
6.6	Scholia	437
CHAPTER 7	Epilogue	446
CHAPTER 8	Answers to Exercises	452
	Section 1.1	453
	Section 1.2	454
	Section 1.3	455
	Section 2.1	455
	Section 2.2	459
	Section 3.1	465
	Section 3.2	468
	Section 3.3	471
	Section 3.4	472
	Section 4.1	473
	Section 4.2	475
	Section 4.3	479
	Section 4.4	486
	Section 5.1	487
	Section 5.2	489
	Section 5.3	490
	Section 5.4	492
	Section 5.5	494

Section 5.7	495
Section 6.1	497
Section 6.2	500
Section 6.3	502
Section 6.4	504
Appendix A Constrained Optimization in Finite Dimensions	508
Appendix B Regularization Operators	512
Appendix C Calculus of Variations	518
C.1 Functionals and Variations	518
C.2 Basic Notion on Variations	520
C.3 Euler–Lagrange Equations	523
C.4 Variational Problems With Subsidiary Conditions	526
Appendix D Index to Notation	530
Bibliography	534
Index	552

CHAPTER

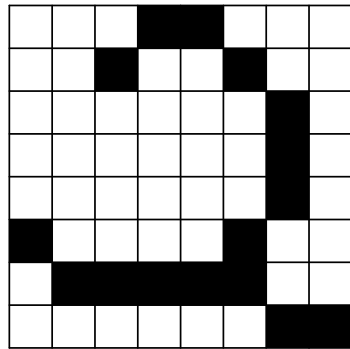
The Big Picture

1

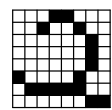
Let's start!



Why do machines need to learn?



*Handwritten
characters: The 2^d
warning!*



~ 0001100000100100000000100000001000000010100001000111110000000011.

Patterns and segmentation

*Segmentation might
be as difficult as
recognition!*

signal. Unfortunately, those analyses are doomed to fail. The sentence “computers are attacking the secret of intelligence”, **quickly pronounced, would likely**

com / pu / tersarea / tta / ckingthesecre / tofin / telligence.

In vision

Region Segmentation



Learning tasks

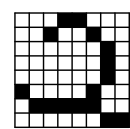
Agent: $\chi : \mathcal{E} \rightarrow \mathcal{D}$.

$$\pi : \mathcal{E} \rightarrow \mathcal{X}$$

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

$$h : \mathcal{Y} \rightarrow \mathcal{O}$$

$\chi = h \circ f \circ \pi$,
where π is the input
encoding, f is the
learning function,
and h is the output
encoding.


$$\begin{array}{l} \xrightarrow{\pi} (0, 0, 0, 1, 1, 0, 0, 0, \dots, 0, 0, 0, 0, 0, 0, 1, 1)' \\ \xrightarrow{f} (0, 0, 1, 0, 0, 0, 0, 0, 0)' \xrightarrow{h} 2. \end{array}$$

Regression and classification

Agent: $\chi : \mathcal{E} \rightarrow \mathcal{D}$.

$$\mathcal{O} \subset \mathbb{N}$$

$$|\mathcal{O}| = 10.$$

$$\mathcal{O} = \mathbb{R}$$

Structured representations

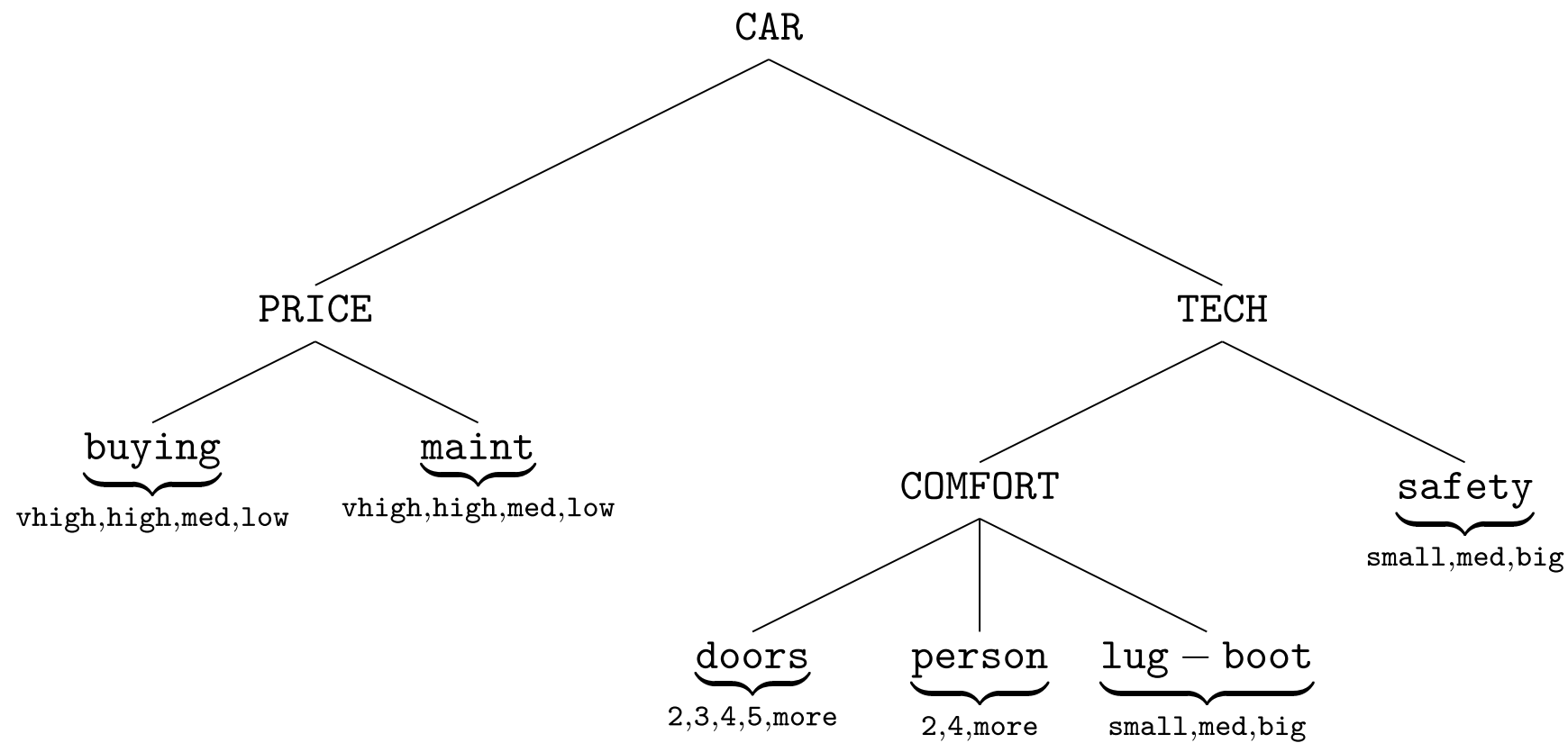


FIGURE 1.1

This learning task is presented in the UCI Machine Learning repository
<https://archive.ics.uci.edu/ml/datasets/Car+Evaluation>.

Structured representations (con't)

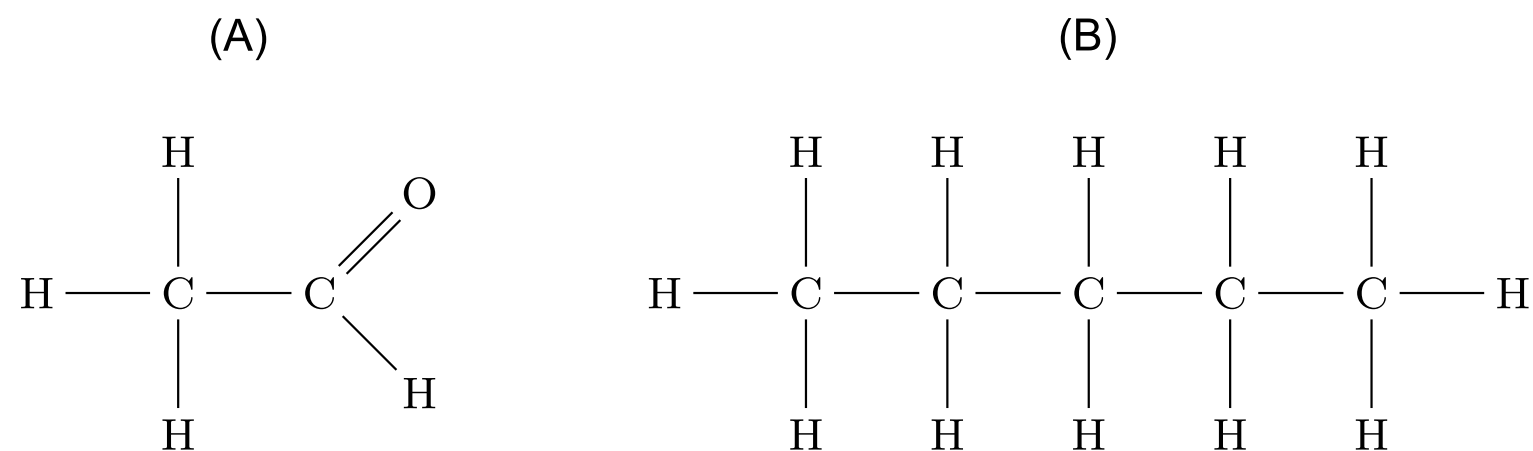


FIGURE 1.2

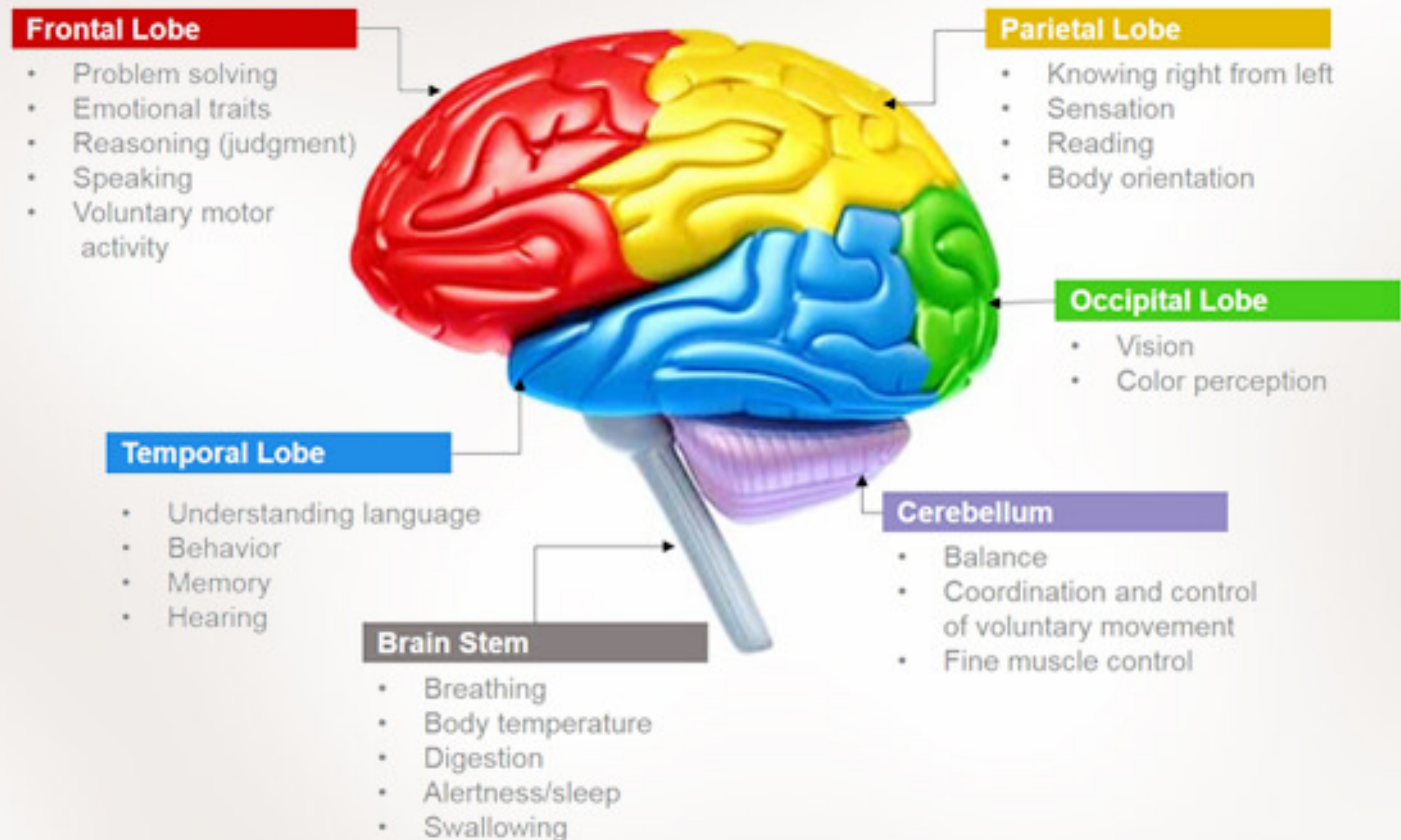
Two chemical formulas: (A) acetaldehyde with formula CH_3CHO , (B) N-heptane with the chemical formula $\text{H}_3\text{C}(\text{CH}_2)_5\text{CH}_3$.

Biological neurons



about 100 billion, 7,000 synaptic connections each

Brain and localization of functionalities



Artificial neurons

$$a_i = b_i + \sum_{j=1}^d w_{i,j} x_j,$$

$$y_i = \sigma(a_i) = 1/(1 + e^{-a_i}).$$

Feedforward neural network

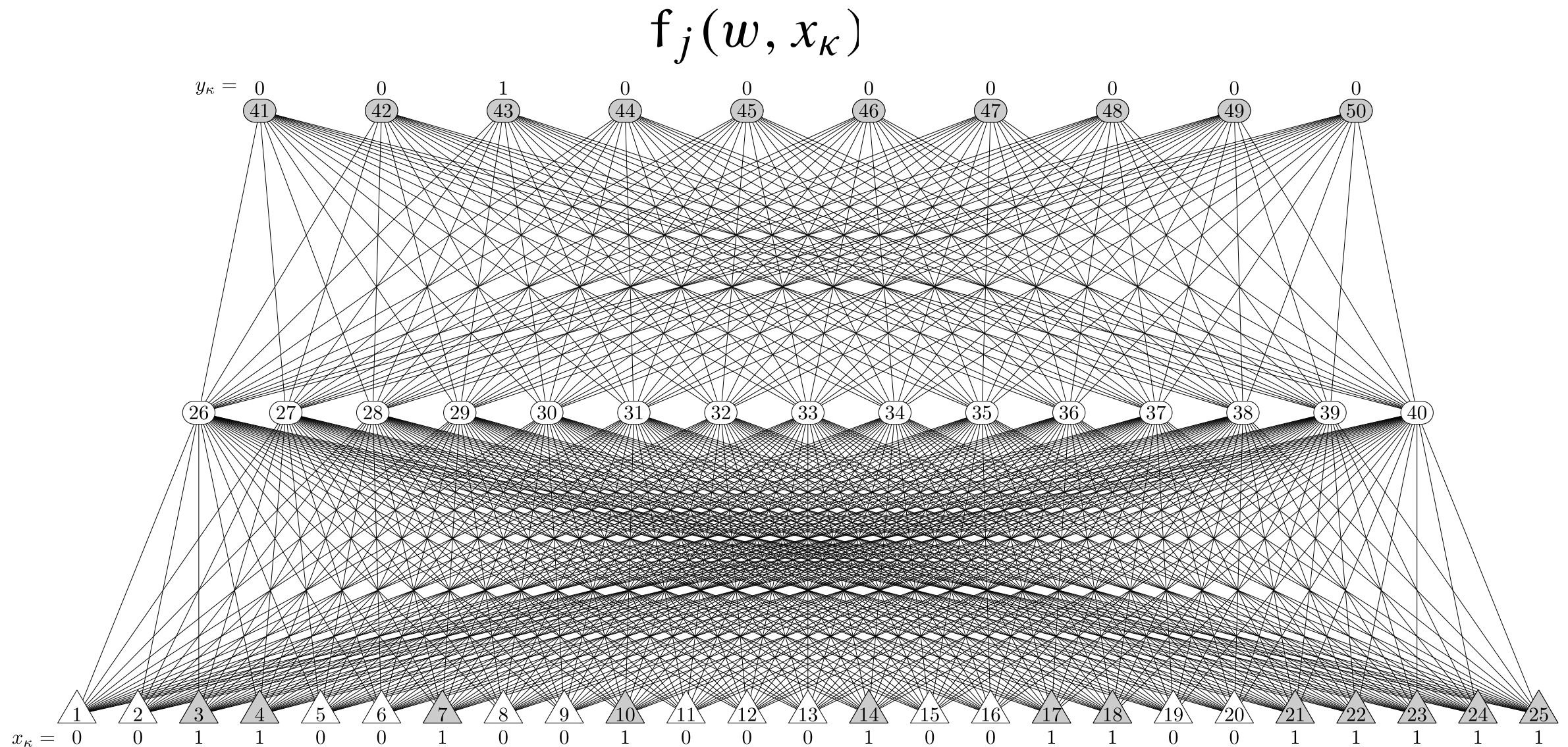


FIGURE 1.3

Recognition of handwritten chars. The incoming pattern $x = \pi(\text{grid})$ is processed by the feedforward neural network, whose target consists of firing only neuron 43. This corresponds with the *one-hot* encoding of class “2”.

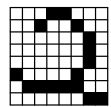
LEARNING PROTOCOLS

Supervised learning

$$\mathcal{L} = \{(e_1, o_1), \dots, (e_\ell, o_\ell)\}$$

2

$$\{(x_\kappa, y_\kappa), \kappa = 1, \dots, \ell\}$$



~ 0001100000100100000000100000001000000010100001000111110000000011.

Error function

$$(\mathcal{N}, \mathcal{L}) \rightsquigarrow E(\cdot)$$

$$E(w) = \sum_{\kappa=1}^{\ell} \sum_{j=1}^n (1 - y_{\kappa j} f_j(w, x_{\kappa}))_+$$

Unsupervised learning

$$x \in \mathcal{X} \subset \mathbb{R}^d$$

$$\bar{x} \in \mathcal{X}$$

$$\|x - \bar{x}\| < \rho$$

$$\mathcal{N}_\rho = \{x \in \mathcal{X} \mid \|x - \bar{x}\| < \rho\}$$

$$\text{vol}(\mathcal{N}_\rho) = \frac{(\sqrt{\pi})^d}{\Gamma(1 + \frac{d}{2})} \rho^d$$

Everything is in the peel!

Space oddities at high dimensions.

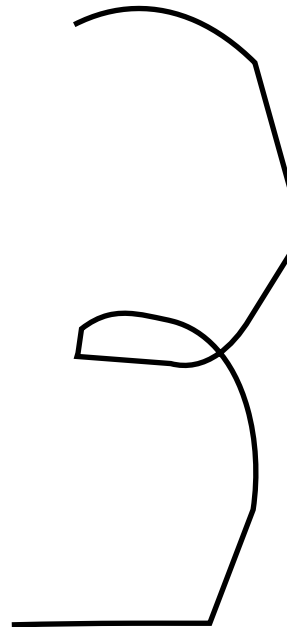
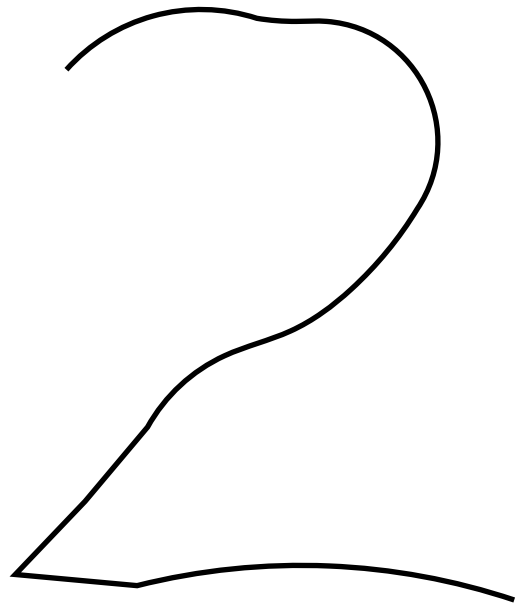
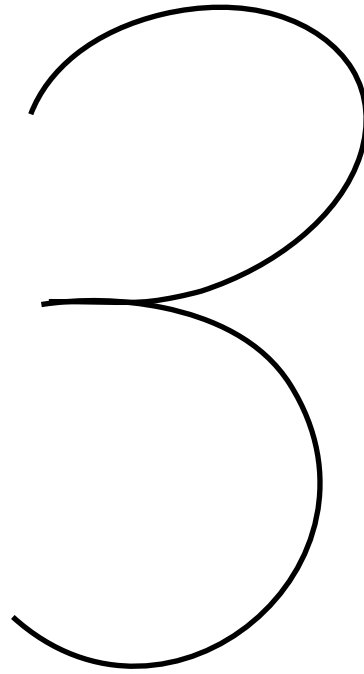
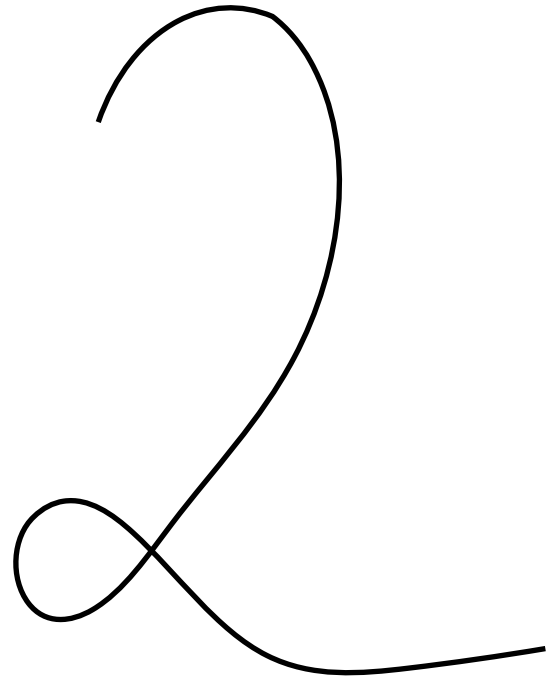
$$\mathcal{P}_\epsilon = \{x \in \mathcal{X} \mid \|x - \bar{x}\| < \rho \quad \text{and} \quad \|x - \bar{x}\| > \rho - \epsilon\}$$

As $d \rightarrow \infty$, the orange collapses to its peel. Hence, no thresholding criterion can discriminate the patterns.

$$\begin{aligned} \text{vol}(\mathcal{P}_\epsilon) &= \lim_{d \rightarrow \infty} \text{vol}(\mathcal{N}_\rho) \left(1 - \frac{\text{vol}(\mathcal{N}_\epsilon)}{\text{vol}(\mathcal{N}_\rho)} \right) \\ &= \text{vol}(\mathcal{N}_\rho) \left(1 - \lim_{d \rightarrow \infty} \left(\frac{\rho - \epsilon}{\rho} \right)^d \right) = \text{vol}(\mathcal{N}_\rho) \end{aligned}$$

A nice exercise ...

Compute the into-char Euclidean distance in MNIST!
You'll learn a lot about space oddities ...



Pattern auto-encoding

$$f(w, x_K) \simeq x_K$$

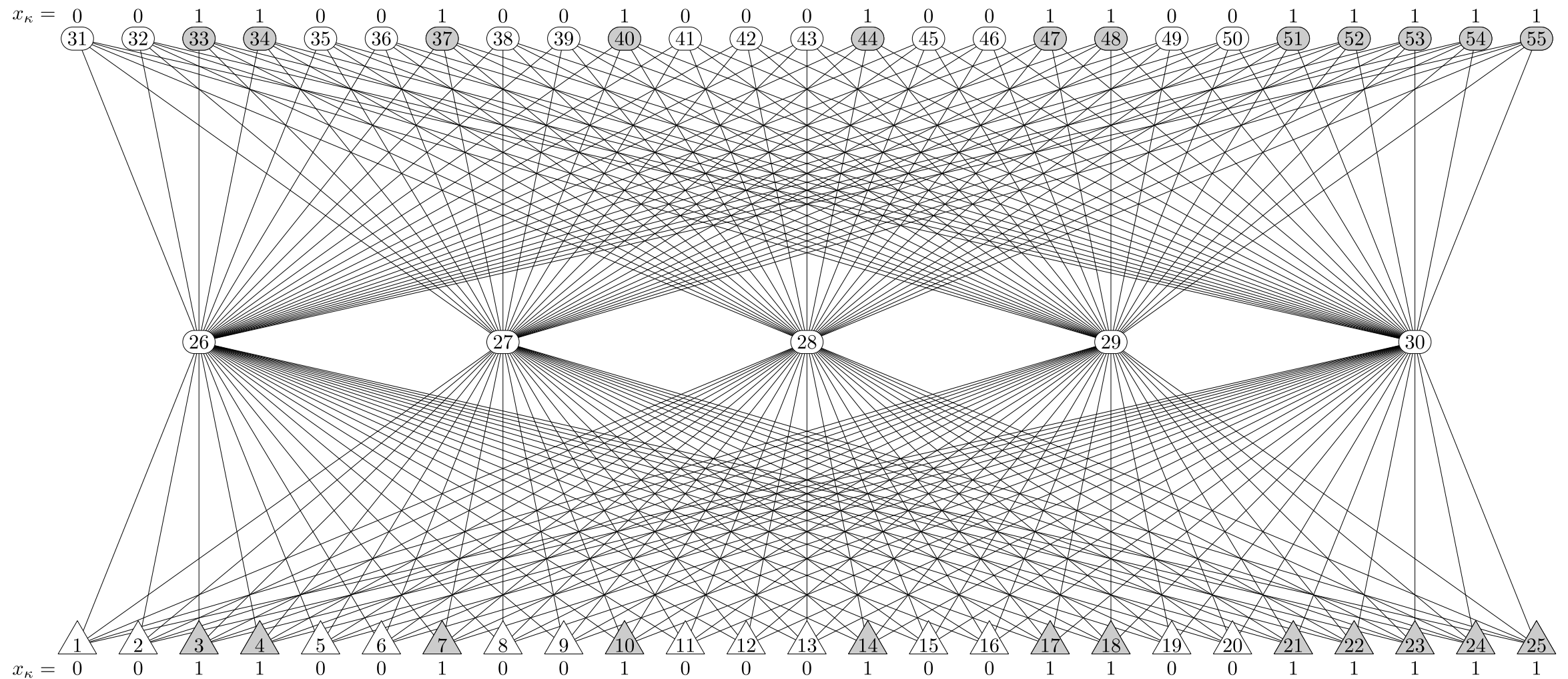


FIGURE 1.4

Pattern auto-encoding by an MLP. The neural net is supervised in such a way to reproduce the input to the output. The hidden layer yields a compressed pattern representation.

Pattern auto-encoding (con't)

$$\mathcal{D} = \{x_1, \dots, x_\ell\} \subset \mathcal{X}^\ell$$

$$w^\star = \arg \min_w \sum_{x_k \in \mathcal{D}} \|f(w, x_k) - x_k\|^2$$

$$s_{\mathcal{D}}(x) := \|x - f(w^\star, x)\|$$

$$\mathcal{X}_{\mathcal{D}}^\rho := \{x \in \mathcal{X} \mid s_{\mathcal{D}}(x) < \rho\}$$

Other protocols of learning

- semi-supervised learning
- transductive learning
- reinforcement learning
- active learning