Chapter 1. Introduction

* 1. Who Should Read This Book?
  2. Historical Trends in Deep Learning
     1. The Many Names and Changing Fortunes of Neural Networks

Three waves :

1. cybernetics (1940~1960) :

* motive : theory of biological learning / single neuron level
* ADALINE(adaptive linear element) or linear model
* SGD(stochastic gradient descent)
* backlash because linear model cannot implement XOR function ( later, in 2010, rectified linear unit solved this problem )

1. connectionism or parallel distributed processing (1980s~1990s) :

* neural network / 1~2 hidden layer
* motive : “a large number of networked simple computational units 🡪 intelligent”
* “distributed representation” : “each input to a system should be represented by many features, and each feature should be involved in the representation of many possible inputs”
* Back-propagation / LSTM ( used in natural language processing tasks at Google )
* CIFAR NCAP : Canadian Institute for Advanced Research / Neural Computation and Adaptive Perceptioin : Geoffrey Hinton, Yoshua Bengio, Yann LeCun

1. deep learning(2006~) :

* greedy layer-wise pretraining
* able to train deeper / theoretical importance of depth
  + 1. Increasing Dataset Sizes
    2. Increasing Model Sizes
    3. Increasing Accuracy, Complexity and Real-World Impact

Part 1. Applied Math and Machine Learning Basics

Chapter2. Linear Algebra

2.1 Scalars, Vectors, Matrices and Tensors

2.2 Multiplying Matrices and Vectors

2.3 Identity and Inverse Matrices

2.4 Linear Dependence and Span

- span of a set of vectors : the set of all points obtainable by linear combination of the original vectors.

2.5 Norms : the size of a vector

- general case :

- max norm :

2.6 Special Kinds of Matrices and Vectors

- Diagonal matrices : diag(v)

2.7 Eigendecomposition

- For real symmetric matrix, can be expressed as ( where Q is an orthogonal matrix composed of eigenvectors of A, is a diagonal matrix )

2.8 Singular Value Decomposition(SVD)

- more generally applicable. Every real matrix has SVD. Useful for non-square matrices

- when A is m\*n matrix, U,D,V are m\*m, m\*n, n\*n and D is a diagonal matrix

2.9 The Moore-Penrose Pseudoinverse

- for a non-square matrix, pseudoinverse matrix might be ( sometimes does not exist )

2.10 The Trace Operation

- Tr(AB) = Tr(BA)

2.11 The Determinant

2.12 Example : Principal Components Analysis (PCA)

- Lossy compression method, famous for ML

Chapter 3. Probability and Information Theory

3.1 Why Probability?

3.2 Random Variables

3.3 Probability Distributions

3.3.1 Discrete Variables and Probability Mass Functions

- PMF(probability mass function)

3.3.2 Continuous Variables and Probability Density Functions

- PDF(probability density function)

3.4 Marginal Probability

- marginal probability distribution : when it depends on more than one variable

3.5 Conditional Probability

3.6 The Chain Rule of Conditional Probabilities

3.7 Independence and Conditional Independence

3.8 Expectation, Variance and Covariance

- Covariance : how much two values are linearly related to each other and also how much they change

- correlation : normalize the contribution of each vector, so only how much related

- Independent 🡪 Covariance=0 (True), but Covariance=0 🡪 Independent (False)

3.9 Common Probability Distributions

3.9.1 Bernoulli Distribution

3.9.2 Multinoulli Distribution : multiple events of Bernoulli event

3.9.3 Gaussian Distribution ( the normal distribution )

- most common over real numbers ( central limit theorem )

- maximum uncertainty over all PDF with same variance. ( worst case modeling )

3.9.4 Exponential and Laplace Distributions

- sharp dense point at x = 0 🡪 Exponential Distribution

- sharp dense point at x = u 🡪 Laplace Distribution

3.9.5 The Dirac Distribution and Empirical Distribution

- Empirical Distribution : combinations of Dirac Distribution. Be used when we choose a dataset of training examples. And also easiest modeling for training.

3.9.6 Mixtures of Distributions

3.10 Useful Properties of Common Functions

- logistic sigmoid : smooth version of step function

- softplus function : smooth version of max( 0, x )

3.11 Bayes’ Rule

3.12 Technical Details of Continuous Variables

- Measure theory : useful to define magnitude of subsets in abstract domain or things

Length, volume, area are easy example of measure theory

PDF also use Measure theory to define the actual probability.

3.13 Information Theory

- theory that describes magnitude of Information. ( as Entropy )

- “Less likely events should have higher information content”

- definition of self-information of an event X = x : I(x) = -log(P(x))

- unit : if use e as base 🡪 nats

If use 2 as base 🡪 bits or shannons

- Shannon entropy : Quantification of uncertainty in an entire probability distribution

- H(x) = E[I(x) = -E[logP(x)]

- called as the differential entropy when x is continuous

3.14 Structured Probabilistic Models

- Describing probability distribution with more simple functions by distinguish independent variables

- e.g : when a and c are independent with b, p(a,b,c) = p(a) p(b|a) p(c|b)

- These factorizations can be described w/ graphs.