A kao textbook draft

Machine Learning Fundamentals

An introduction to the basic principles and methods of machine learning

Hui Jiang

July 16, 2020

Cambridge University Press

Detailed Contents

Preface v

Detailed Contents xi

1 Introduction 1

- 1.1 **What is Machine Learning?** 1
- 1.2 **Basic Concepts in Machine Learning** 4
  - Classification vs. Regression 4
  - Supervised vs. Unsupervised Learning 5
  - Simple vs. Complex Models 5
  - Parametric vs. Non-parametric Models 7
  - Over-fitting vs. Under-fitting 8
  - Bias-Variance Tradeoff 10
- 1.3 **General Principles in Machine Learning** 11
  - Occam's Razor 11
  - No Free Lunch Theorem 11
  - Law of the Smooth World 12
  - Curse of Dimensionality 14
- 1.4 **Advanced Topics in Machine Learning** 15
  - Reinforcement Learning 15
  - Meta-Learning 16
  - Causal Inference 16
  - Other Advanced Topics 16
- Exercises 18

2 Mathematical Foundation 19

- 2.1 **Linear Algebra** 19
  - Vectors and Matrices 19
  - Linear Transformation as Matrix Multiplication 20
  - Basic Matrix Operations 21
  - Eigenvalues and Eigenvectors 23
  - Matrix Calculus 25
- 2.2 **Probability and Statistics** 27
  - Random Variables and Distributions 27
  - Expectation: Mean, Variance and Moments 28
  - Joint, Marginal and Conditional Distributions 30
  - Common Probability Distributions 33
  - Transformation of Random Variables 40

2.3 **Information Theory** . . . 41
Information and Entropy . . . 41
Mutual Information . . . 43
KL Divergence . . . 46
2.4 **Mathematical Optimization** . . . 48
General Formulation . . . 49
Optimality Conditions . . . 50
Numerical Optimization Methods . . . 59
Exercises . . . 64

3 Supervised Machine Learning (in a nutshell) **67**
3.1 **Overview** . . . 67
3.2 **Case Studies** . . . 72

4 Feature Extraction **77**
4.1 **Feature Extraction: Concepts** . . . 77
Feature Engineering . . . 77
Feature Selection . . . 78
Dimensionality Reduction . . . 79
4.2 **Linear Dimension Reduction** . . . 79
Principal Component Analysis . . . 80
Linear Discriminant Analysis . . . 84
4.3 **Nonlinear Dimension Reduction (I): Manifold Learning** . . . 86
Locally Linear Embedding . . . 87
Multidimensional Scaling . . . 88
Stochastic Neighborhood Embedding . . . 89
4.4 **Nonlinear Dimension Reduction (II): Neural Networks** . . . 90
Autoencoder . . . 90
Bottleneck Features . . . 91
Lab Project I . . . 92
Exercises . . . 93

DISCRIMINATIVE MODELS **95**

5 Statistical Learning Theory **97**
5.1 **Formulation of Discriminative Models** . . . 97
5.2 **Learnability** . . . 99
5.3 **Generalization Bounds** . . . 100
Finite Model Space: $|\mathbb{H}|$ . . . 100
Infinite Model Space: VC Dimension . . . 102
Exercises . . . 105

6 Linear Models **107**
6.1 **Perceptron** 108
6.2 **Linear Regression** 112
6.3 **Minimum Classification Error** 113
6.4 **Logistic Regression** 114
6.5 **Support Vector Machines (SVM)** 116
Linear SVM 116
Soft SVM 121
Nonlinear SVM: the kernel trick 123
Solving Quadratic Programming 125
Multi-Class SVM 127
Lab Project II 129
Exercises 130

7 Learning Discriminative Models in General **133**
7.1 **A General Framework to Learn Discriminative Models** 133
Common Loss Functions in ML 135
Regularization based on L_p norm 136
7.2 **Ridge Regression and LASSO** 139
7.3 **Matrix Factorization** 140
7.4 **Dictionary Learning** 145
Lab Project III 149
Exercises 150

8 Neural Networks **151**
8.1 **Artificial Neural Networks (ANN)** 152
Basic Formulation of ANNs 152
Mathematical Justification: Universal Approximator 154
8.2 **Neural Network Structures** 156
Basic Building Blocks to Connect Layers 156
Case Study (I): Fully-Connected Deep Neural Networks 165
Case Study (II): Convolutional Neural Networks (CNN) 166
Case Study (III): Recurrent Neural Networks (RNN) 170
Case Study (IV): Transformer 171
8.3 **Learning Algorithms for Neural Networks** 174
Loss Function 175
Automatic Differentiation 176
Optimization Using Mini-batch SGD 187
8.4 **Heuristics and Tricks for Optimization** 188
Other SGD-variant Optimization Methods: ADAM 190
Regularization 193
Fine-tuning Tricks 194

8.5 **End-to-End Learning** . . . 195
Sequence-to-Sequence Learning . . . 196
Lab Project IV . . . 198
Exercises . . . 199

9 Ensemble Learning **201**
9.1 **Formulation of Ensemble Learning** . . . 201
Decision trees . . . 203
9.2 **Bagging** . . . 205
Random Forests . . . 206
9.3 **Boosting** . . . 207
Gradient Boosting . . . 208
AdaBoost . . . 209
Gradient Tree Boosting . . . 212
Lab Project V . . . 214
Exercises . . . 215

GENERATIVE MODELS **217**

10 Overview of Generative Models **219**
10.1 **Formulation of Generative Models** . . . 219
10.2 **Bayesian Decision Theory** . . . 220
Generative Models for Classification . . . 221
Generative Models for Regression . . . 225
10.3 **Statistical Data Modelling** . . . 226
Plug-in MAP Decision Rule . . . 227
10.4 **Density Estimation** . . . 229
Maximum Likelihood Estimation . . . 229
Maximum Likelihood Classifier . . . 231
10.5 **Generative Models (in a nutshell)** . . . 232
Generative vs. Discriminative Models . . . 233
Exercises . . . 235

11 Unimodal Models **237**
11.1 **Gaussian Models** . . . 238
11.2 **Multinomial Models** . . . 241
11.3 **Markov Chain Models** . . . 243
11.4 **Generalized Linear Models** . . . 248
Probit Regression . . . 250
Poisson Regression . . . 250
Log-linear Models . . . 251
Exercises . . . 254

12 Mixture Models **255**
12.1 **Formulation of Mixture Models** 255
Exponential Family (e-family) 257
Formal Definition of Mixture Models 259
12.2 **Expectation-Maximization (EM) Method** 259
Auxiliary Function: eliminating log-sum 260
Expectation-Maximization Algorithm 263
12.3 **Gaussian Mixture Models** 266
K-means Clustering for Initialization 268
12.4 **Hidden Markov Models (HMMs)** 269
HMMs: mixture models for sequences 270
Evaluation Problem: Forward-Backward Algorithm 274
Decoding Problem: Viterbi Algorithm 277
Training Problem: Baum-Welch Algorithm 278
Lab Project VI 285
Exercises 286

13 Entangled Models **289**
13.1 **Formulation of Entangled Models** 289
Framework of Entangled Models 290
Learning of Entangled Models in General 292
13.2 **Linear Gaussian Models** 294
Probabilistic PCA 294
Factor Analysis 296
13.3 **Non-Gaussian Models** 298
Independent Component Analysis 298
Independent Factor Analysis 299
Hybrid Orthogonal Projection and Estimation 300
13.4 **Deep Generative Models** 301
Variational Autoencoders 301
Generative Adversarial Nets 305
Exercises 307

14 Bayesian Learning **309**
14.1 **Formulation of Bayesian Learning** 309
Bayesian Inference 311
Maximum a Posterior Estimation 312
Sequential Bayesian Learning 313
14.2 **Conjugate Priors** 316
Maximum Marginal Likelihood Estimation 321
14.3 **Approximate Inference** 322
Laplace's Method 322

Variational Bayesian Methods . . . 324
14.4 **Gaussian Processes** . . . 330
Gaussian Processes as Non-parametric Priors . . . 331
Gaussian Processes for Regression . . . 333
Gaussian Processes for Classification . . . 336
Exercises . . . 338

15 Graphical Models **341**
15.1 **Concepts of Graphical Models** . . . 341
15.2 **Bayesian Networks** . . . 344
Conditional Independence . . . 344
Representing Generative Models as Bayesian Networks . . . 349
Learning Bayesian Networks . . . 351
Inference Algorithms . . . 353
Case Study (I): Naive Bayes Classifier . . . 359
Case Study (II): Latent Dirichlet Allocation . . . 360
15.3 **Markov Random Fields** . . . 363
Formulation: Potential and Partition Functions . . . 363
Case Study (III): Conditional Random Fields . . . 366
Case Study (IV): Restricted Boltzmann Machines . . . 367
Exercises . . . 370

APPENDIX **373**

A Other Probability Distributions **375**
A.1 Uniform Distribution . . . 375
A.2 Poisson Distribution . . . 375
A.3 Gamma Distribution . . . 376
A.4 Inverse-Wishart Distribution . . . 376
A.5 von Mises–Fisher distribution . . . 377

Bibliography **379**

Notation **397**

Index **399**