

Introduction to Machine Learning

67577

Course Book

INTRODUCTION TO MACHINE LEARNING, THE HEBREW UNIVERSITY, JERUSALEM
ADDRESS TO CODE EXAMPLES AND LABS REPOSITORY
Written by and by.

First release, October 2020



Contents

0.1 Preface	9
0.1.1 Notation	9
0.1.2 Data Sets Used in Book, Labs and Examples	9
1 Mathematical Basis	11
 1.1 Linear Algebra	11
1.1.1 Hyperplanes	11
1.1.2 Projecting Matrices	11
1.1.3 Matrix Decomposition	11
 1.2 Calculus	11
1.2.1 High Order Derivatives	11
1.2.2 Convexity	11
 1.3 Probabilities Theory	11
1.3.1 Distributions and Random Variables	11
1.3.2 Multi-variate Distributions	11
1.3.3 Joint- and Marginal Distributions	11
1.3.4 PDF and CDF of Distributions	11
1.3.5 Measurements Of Concentration	11
2 Introduction & Linear Regression	13
 2.1 Introduction to Statistical Learning	13
2.1.1 Estimation Theory	13
2.1.2 Risk & Loss Functions	13
2.1.3 Learning Principles	13
2.1.4 Lab: Python Data Analysis - First Steps	13

2.1.5 Lab: Data Simulation and Sampling	13
2.2 Linear Regression	13
2.2.1 Ordinary Least Squares	13
2.2.2 Weighted Least Squares	13
2.2.3 Geometric Interpretation	13
2.2.4 Categorical Variables	13
2.2.5 Lab: Linear Regression	13
2.3 Beyond Linearity	13
2.3.1 Polynomial Fitting	13
2.3.2 Poisson Regression	13
2.3.3 Lab: Polynomial Fitting	13
3 Classification	15
3.1 Classification Overview	16
3.1.1 Loss Function	16
3.1.2 Type-I and Type-II Errors	16
3.1.3 Statistical Measures of Performance	16
3.2 Logistic Regression	16
3.2.1 A Probabilistic Model For Noisy Labels	16
3.2.2 Computational Implementation	16
3.2.3 Interpretability	16
3.2.4 ROC Curve	16
3.2.5 Lab: Logistic Regression	16
3.3 Half-Space Classifier	16
3.3.1 Learning Linearly Separable Data Via ERM	16
3.3.2 Computational Implementation	16
3.3.3 The Perceptron Algorithm	16
3.4 Support Vector Machines	16
3.4.1 Maximum Margin Learning Principal	16
3.4.2 Hard-SVM	16
3.4.3 Soft-SVM	16
3.5 Nearest Neighbors	16
3.5.1 Graph-Based Approach For Learning	16
3.5.2 Classification & Regression Using k -NN	16
3.5.3 Computational Implementation	16
3.5.4 Selecting Value of k Hyper-Parameter	16
3.5.5 Variants of Nearest Neighbors	16
3.6 Decision Trees	16
3.6.1 Axis-Parallel Partitioning of \mathbb{R}^d	16
3.6.2 Classification & Regression Trees	16
3.6.3 Growing a Classification Tree	16
3.6.4 NP-Hardness and CART Heuristic	16
3.6.5 Pruning a Decision Tree	16

3.7 Bayes Classifiers	16
3.7.1 Bayes Optimal Classifier	16
3.7.2 Naive Bayes	16
3.7.3 Linear Discriminant Analysis	16
3.7.4 Quadratic Discriminant Analysis	16
3.7.5 Lab: Maximum Likelihood Estimation	16
4 PAC Theory of Statistical Learning	17
4.1 Theoretical Framework For Learning	18
4.1.1 Data-Generation Model	18
4.1.2 The Realizability Assumption	18
4.1.3 Learning As A Game - First Attempt	18
4.1.4 Probably- and Approximately Correct Learners	18
4.2 No Free Lunch and Hypothesis Classes	18
4.2.1 No Free Lunch!	18
4.2.2 Restricting for Hypothesis Classes	18
4.2.3 Learning As A Game - Final Attempt	18
4.2.4 Example: Threshold Functions	18
4.3 PAC Learnability of Finite Hypothesis Classes	18
4.4 VC-Dimension	18
4.4.1 Formal Definition	18
4.4.2 VC-Dimension of Finite Hypothesis Classes	18
4.4.3 Example: Axis Aligned Rectangles	18
4.4.4 Example: Half-Spaces	18
4.5 Agnostic PAC: Extending Framework	18
4.5.1 Data-Generation Model Over $\mathcal{X} \times \mathcal{Y}$	18
4.5.2 Relaxing Realizability Assumption	18
4.5.3 Introducing General Loss Functions	18
4.5.4 Agnostic PAC Learnability	18
4.6 Uniform Convergence Property	18
4.6.1 ε -Representative Datasets	18
4.6.2 Achieving Uniformity In \mathcal{H} and \mathcal{D}	18
4.7 The Fundamental Theorem of Statistical Learning	18
5 Ensemble Methods	19
5.1 Bias-Variance Trade-off	19
5.1.1 Generalization Error Decomposition	19
5.1.2 Lab: Bias-Variance Via Decision Trees	19
5.1.3 Lab: Bias-Variance Via Polynomial Fitting	19

5.2 Ensemble/Committee Methods	19
5.2.1 Weak-Learnability	19
5.2.2 Uncorrelated Predictors	19
5.2.3 Correlated Predictors	19
5.2.4 Committee Methods In Machine Learning	19
5.3 Boosting Weak-Learners	19
5.3.1 AdaBoost Algorithm	19
5.3.2 Gradient Boosting Algorithm	19
5.3.3 Lab: Boosting - Image Classification	19
5.4 Bagging	19
5.4.1 Bootstrapping	19
5.4.2 Bagging Reduces Variance	19
5.4.3 Random Forests Bagging and De-correlating Decision Trees	19
6 Regularization, Model Selection and Model Evaluation	21
6.1 Regularization	21
6.1.1 Best Subset Selection	21
6.1.2 L_q Nrom Regularizes	21
6.1.3 Ridge Regularization	21
6.1.4 Convexity vs. Sparsity	21
6.1.5 Lasso Regularization	21
6.1.6 Lab: Regularized Logistic Regression	21
6.2 Model Selection and -Evaluation	21
6.2.1 Cross Validation	21
6.2.2 Bootstrap	21
6.2.3 Common Model Selection Mistakes	21
6.2.4 Lab: Selecting Regularized Model	21
7 Unsupervised Learning	23
7.1 Dimensionality Reduction	23
7.1.1 Preserved Data Properties	23
7.1.2 Principal Component Analysis	23
7.1.3 Lab: PCA	23
7.2 Clustering	23
7.2.1 K-Means	23
7.2.2 Mixture of Gaussians	23
7.2.3 Spectral Clustering	23
7.2.4 Lab: K-Means++	23
7.2.5 Lab: Parameters Estimation In MoG	23
8 Convex Optimization and Gradient Descent	25
8.0.1 Gradient Descent Learning Principal	25
8.0.2 Utilizing Sub-gradients For GD	25

8.0.3 Stochastic Gradient Descent	25
8.0.4 Variants Of Gradient Descent	25
8.0.5 Initialization Conditions	25
8.0.6 Tuning Learning Rates	25
9 Online- and Reinforcement Learning	27
10 Deep Learning	29

0.1 Preface**0.1.1 Notation**

Table 1: Notation Summary

Symbol	Meaning
n	Number of samples
d, k	Number of features
\mathcal{X}	Domain set
\mathcal{Y}	Response set
\mathcal{Z}	$(\mathcal{X} \times \mathcal{Y})$ The product space of domain set and response set
$S = x_1, \dots, x_n$	A sequence of samples from domain set
$S = z_1, \dots, z_n$	A sequence of samples and corresponding responses from product space \mathcal{Z}

0.1.2 Data Sets Used in Book, Labs and Examples



1. Mathematical Basis

1.1 Linear Algebra

1.1.1 Hyperplanes

1.1.2 Projecting Matrices

1.1.3 Matrix Decomposition

Eigenvalues Decomposition

Singular Values Decomposition

1.2 Calculus

1.2.1 High Order Derivatives

1.2.2 Convexity

1.3 Probabilities Theory

1.3.1 Distributions and Random Variables

1.3.2 Multi-variate Distributions

1.3.3 Joint- and Marginal Distributions

1.3.4 PDF and CDF of Distributions

1.3.5 Measurements Of Concentration



2. Introduction & Linear Regression

2.1 Introduction to Statistical Learning

2.1.1 Estimation Theory

Estimators

2.1.2 Risk & Loss Functions

2.1.3 Learning Principles

Empirical Risk Minimization

Maximum Likelihood

2.1.4 Lab: Python Data Analysis - First Steps

2.1.5 Lab: Data Simulation and Sampling

2.2 Linear Regression

2.2.1 Ordinary Least Squares

2.2.2 Weighted Least Squares

2.2.3 Geometric Interpretation

2.2.4 Categorical Variables

2.2.5 Lab: Linear Regression

2.3 Beyond Linearity

2.3.1 Polynomial Fitting

2.3.2 Poisson Regression

2.3.3 Lab: Polynomial Fitting



3. Classification

3.1 Classification Overview

3.1.1 Loss Function

3.1.2 Type-I and Type-II Errors

3.1.3 Statistical Measures of Performance

3.2 Logistic Regression

3.2.1 A Probabilistic Model For Noisy Labels

The Hypothesis Class

Learning Via Maximum Likelihood

3.2.2 Computational Implementation

3.2.3 Interpretability

3.2.4 ROC Curve

3.2.5 Lab: Logistic Regression

3.3 Half-Space Classifier

3.3.1 Learning Linearly Separable Data Via ERM

3.3.2 Computational Implementation

3.3.3 The Perceptron Algorithm

3.4 Support Vector Machines

3.4.1 Maximum Margin Learning Principal

3.4.2 Hard-SVM

3.4.3 Soft-SVM

The Kernel Trick

3.5 Nearest Neighbors

3.5.1 Graph-Based Approach For Learning

3.5.2 Classification & Regression Using k -NN

3.5.3 Computational Implementation

3.5.4 Selecting Value of k Hyper-Parameter

3.5.5 Variants of Nearest Neighbors

3.6 Decision Trees

3.6.1 Axis-Parallel Partitioning of \mathbb{R}^d



4. PAC Theory of Statistical Learning

4.1 Theoretical Framework For Learning

4.1.1 Data-Generation Model

4.1.2 The Realizability Assumption

4.1.3 Learning As A Game - First Attempt

4.1.4 Probably- and Approximately Correct Learners

4.2 No Free Lunch and Hypothesis Classes

4.2.1 No Free Lunch!

4.2.2 Restricting for Hypothesis Classes

4.2.3 Learning As A Game - Final Attempt

4.2.4 Example: Threshold Functions

4.3 PAC Learnability of Finite Hypothesis Classes

4.4 VC-Dimension

4.4.1 Formal Definition

4.4.2 VC-Dimension of Finite Hypothesis Classes

4.4.3 Example: Axis Aligned Rectangles

4.4.4 Example: Half-Spaces

4.5 Agnostic PAC: Extending Framework

4.5.1 Data-Generation Model Over $\mathcal{X} \times \mathcal{Y}$

4.5.2 Relaxing Realizability Assumption

4.5.3 Introducing General Loss Functions

4.5.4 Agnostic PAC Learnability

4.6 Uniform Convergence Property

4.6.1 ε -Representative Datasets

4.6.2 Achieving Uniformity In \mathcal{H} and \mathcal{D}

The Case Of Finite \mathcal{H}

The General Case - Infinite \mathcal{H}

4.7 The Fundamental Theorem of Statistical Learning



5. Ensemble Methods

5.1 Bias-Variance Trade-off

5.1.1 Generalization Error Decomposition

5.1.2 Lab: Bias-Variance Via Decision Trees

5.1.3 Lab: Bias-Variance Via Polynomial Fitting

5.2 Ensemble/Committee Methods

5.2.1 Weak-Learnability

5.2.2 Uncorrelated Predictors

5.2.3 Correlated Predictors

5.2.4 Committee Methods In Machine Learning

5.3 Boosting Weak-Learners

5.3.1 AdaBoost Algorithm

5.3.2 Gradient Boosting Algorithm

5.3.3 Lab: Boosting - Image Classification

5.4 Bagging

5.4.1 Bootstrapping

5.4.2 Bagging Reduces Variance

5.4.3 Random Forests Bagging and De-correlating Decision Trees



6. Regularization, Model Selection and Model

6.1 Regularization

- 6.1.1 Best Subset Selection
- 6.1.2 L_q Norm Regularizes
- 6.1.3 Ridge Regularization
- 6.1.4 Convexity vs. Sparsity
- 6.1.5 Lasso Regularization
- 6.1.6 Lab: Regularized Logistic Regression

6.2 Model Selection and -Evaluation

- 6.2.1 Cross Validation
- 6.2.2 Bootstrap
- 6.2.3 Common Model Selection Mistakes
 - Over-estimating Generalization Error
 - Under-estimating Generalization Error
- 6.2.4 Lab: Selecting Regularized Model



7. Unsupervised Learning

7.1 Dimensionality Reduction

7.1.1 Preserved Data Properties

7.1.2 Principal Component Analysis

Closest Subspace Interpretation

Generalized Linear Regression Interpretation

Maximum Retained Variance Interpretation

Projection- vs. Coordinates of Data-Points

Variants of PCA

Euclidean Embedding

7.1.3 Lab: PCA

7.2 Clustering

7.2.1 K-Means

K-Means++

7.2.2 Mixture of Gaussians

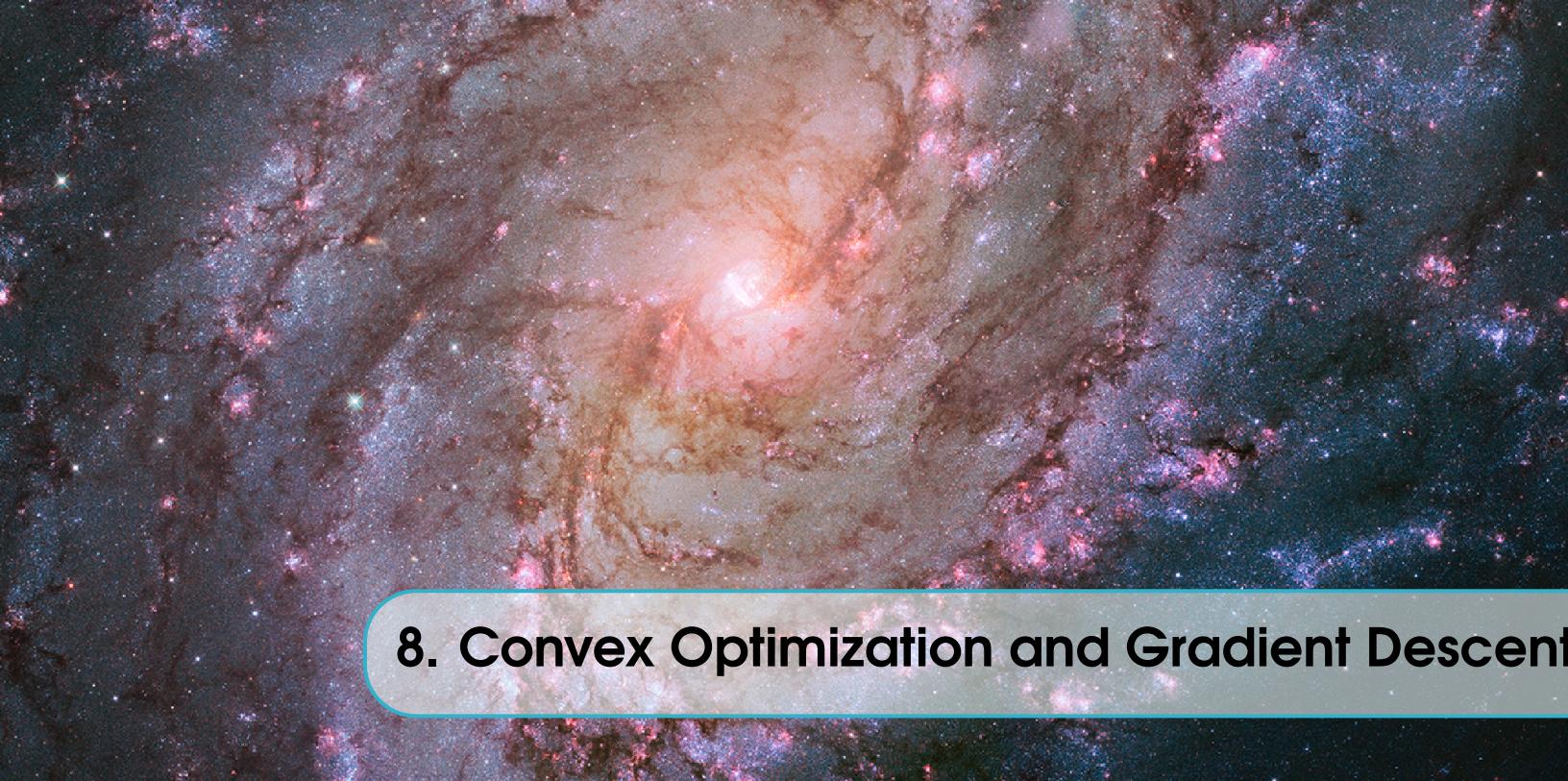
Expectation Minimization Learning Principle

Estimating Model Parameters

7.2.3 Spectral Clustering

7.2.4 Lab: K-Means++

7.2.5 Lab: Parameters Estimation In MoG



8. Convex Optimization and Gradient Descent

- 8.0.1 Gradient Descent Learning Principal
- 8.0.2 Utilizing Sub-gradients For GD
- 8.0.3 Stochastic Gradient Descent
- 8.0.4 Variants Of Gradient Descent
- 8.0.5 Initialization Conditions
- 8.0.6 Tuning Learning Rates



9. Online- and Reinforcement Learning



10. Deep Learning