Machine Learning Handbook

Xinhe Liu

2018-2-28

Contents

Part I

High-level Views

Chapter 1

Math Review

1.1 Linear Algebra

Concepts:

- scalar, vector, matrix, tensor(n-rank tensor, matrix is a rank 2 tensor)
- Gaussian Elimination, rank
- p-norm
$$|X|_p = (\sum_i |x_i|^p)^{\frac{1}{p}}$$
- inner product$< x_i, y_i >$, outer product
- orthogonaldimension, basis, orthogonal basis
- linear transformation $Ax = y$
- eigenvalue, eigenvector $Ax = \lambda x$ (transformation and speed)
- vector space, linear space(with summation, scalar production), inner product space(inner product space)

1.2 Probability

Concepts:

- Classic Probability Model: Frequentist
- Bayesian Probability Theory
- Random variable, continuous RV, discrete RV, probability mass function, probability density function, cumulative density function
- Bernoulli distribution, Binomial distribution(n,p)

  $$P(X = k) = \binom{N}{k} p^k (1-p)^{(n-k)}$$

  , Poisson distribution

  $$P(X = k) = \lambda^k \frac{e^{-\lambda}}{k!}$$

- uniform distribution, exponential distribution

  $$e^{-\frac{x}{\theta}}\theta, P(x > s + t|X > s) = P(x > t)$$

  , normal distribution, t-distribution

- expectation, moments, variance, covariance, correlation coefficient

Theorems:

- Law of Total Probability
- Bayesian Theorem

  $$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

  P(H)-prior probability, P(D—H)-likelihood, P(H—D)-posterior probability,

1.3 Statistics

Concepts:

- data, sample, statistics, population
- point estimation, interval estimation, confidence interval(intepreation??)
- method of moments: $E(X^k)$ based on LOLN.
- maximum likelihood estimation. Multiply p.m.f/p.d.f since every sample is independent. Maximize the likelihood of finding samples.
- Estimation criteria
  - unbiased
  - efficient
  - coherent
- Hypotheis test, type-I error(wrongly reject), type-II error(wrongly accept)

Theorems:

- Law of Large Number
- Central Limit Theorem
- Bias/Variance decomposition (error = bias + variance + noise)

$$E[(Y-\hat{\mu}(X))^2] == E[(Y-f(x)+f(x)-\hat{\mu}(X))^2]$$

$$= E[(Y-f(x)]^2 + 2E[(Y-f(x))(f(x)-\hat{\mu}(X))] + E[(f(x)-\hat{\mu}(X))^2]$$

$$= \sigma_x^2 + Bias(\hat{\mu}(X))^2 + Var(\hat{\mu}(X))$$

1.4 Optimization Theory

- Objective function/Evaluation function, constrained/unconstrained optimizationFeasible Set, Optimal Solution, Optimal Value, Binding Constraints, Shadow Price, Infeasible Price, Infeasibility, Unboundedness
- Linear Programming
- Lagrange Multiplier

$$L(x, y, \lambda) = f(x, y) + \lambda\varphi(x, y)$$

- Convex Set, Convex Function $f : S \to R$ is convex if and only if $\bigtriangledown^2 f(\mathbf{x})$ is positive semidefinite

Optimization Methods:

- Linear Search Method: Direction First, Step Size second
  - Gradient Descent: Batch Processing(Use all samples) vs Stochastic Gradient Descent(Use one sample)
  - Newton's Method: Use Curvature Information
- Trust Region: Step first, direction second. Find optimal direction of second-order approximation. If the descent size is too small, make step size smaller.
- Heuristics Method
  - Genetic Algorithm
  - Simulated Annealing
  - Partical Swarming/Ant Colony Algorithm

Theorems:

-

1.5 Information Theory

Concepts:

- Information
$$h(A) = -log_2 p(A)$$
(bit)
- (Information Source) Entropy
$$H(X) = -\sum_{i=1}^{n} p(a_i) log_2 p(a_i) \leq log_2 n$$
Maximize under equal probability
- Conditional Entropy
$$H(Y|X) = -\sum_{i=1}^{n} p(x_i) H(Y|X = x_i) = -\sum_{i=1}^{n} p(x_i) \sum_{j=1}^{n} p(y_j|x_i) log_2 p(y_j|x_i)$$
$$= \sum_{i=1}^{n} \sum_{j=1}^{n} p(x_i, y_j) log_2 p(y_j|x_i)$$
- Mutual Information/Information Gain
$$I(X;Y) = H(Y) - H(Y|X)$$
- Kullback-Leibler Divergence (K-L) Divergence
$$D_{KL}(P||Q) = \sum_{i=1}^{n} p(x_i) log_2 \frac{p(x_i)}{q(x_i)} \neq D_{KL}(Q||P)$$
Measures the Distance of two distributions. The optimal encoding of information has the same bits as the entropy. Measures the extra bits if the real distribution is q rather than p. (Using P to approximate Q)

Theorems:

- The Maximum Entropy Principle. Without extra assumption, max entropy/equal probability has the minimum prediction risk.

1.6 Formal Logic

Concepts

- Generative Expert System: Rule+Facts+Deduction Engine
- Godel's incompleteness theorems

Chapter 2

Computational Learning Theory

Part II

Supervised Learning Models

Chapter 3

Regression

3.1 Linear Regressions

3.1.1 Assumptions

Classic Assumptions for Statistics:

-

3.1.2 Intepratation

$$f(x) = \mathbf{w}^T \mathbf{x} = \sum_{i=1}^{n} w_i x_i$$

$$\mathbf{w}* = (\mathbf{X^T X})^{-1} \mathbf{X^T y}$$

RSS Approach:

MLE Approach

Assuming noise is normal, maximize

$$p(\mathbf{x_1}, \mathbf{x_2} ... \mathbf{x_n} | \mathbf{w}) = \prod_k \frac{1}{\sqrt{2\pi}\sigma} exp[-\frac{1}{2\sigma^2}(y_k - \mathbf{w_t x_k})^2]$$

3.1.3 Lasso-Least Absolute Shrinkage and Selection Operator

$$min||y_k - \mathbf{w^T x}_k||^2 + \lambda||\mathbf{w}||_1$$

Chapter 4

Logistic Regression and General Linear Model

Chapter 5

Naive Bayesian

Chapter 6

Tree Models and Ensemble Learning

Part III

Unsupervised Learning Models

Chapter 7

Clustering

Chapter 8

Dimension Reduction

Part IV

Deep Learning and Enhanced Learning Theory

Chapter 9

Multi-layer Perceptron

Chapter 10

Multi-layer Perceptron

Chapter 11

Multi-layer Perceptron

Chapter 12

Multi-layer Perceptron

Chapter 13

Multi-layer Perceptron

Chapter 14

Multi-layer Perceptron