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,

1.2.1 Important Distributions, Moment Generating Functions

1. Normal Distribution, See next chapter
2. Bernoulli Distribution
3. Exponential Distribution $f_x(x, \theta) = \theta e^{-\theta x}$
4. Poisson Distribution

1.3 Statistics

1.3.1 Concepts

- parameter(constant for probability model), statistic (model of sample data), data, sample, population
- point estimation, interval estimation, Confidence Interval($P(L \leq \theta \leq U)$, notice: θ is not random, L, U is random! (We repeat constructing confidence interval a n times, α percent of the times, it will contain $theta$.
- Hypothesis test, type-I error(wrongly reject), type-II error(wrongly accept)

Estimator and Estimation

- Method of Moments: $E(X^k)$ based on LOLN.
  If We have p parameters, we can use p moments to form a system of equations to solve $\theta_1, ...\theta_p$

  $$\sum_{i=1}^{n} X_i^j = E(X^j)$$

  , for j = 1,...,p
- Maximum Likelihood Estimation. Multiply p.m.f/p.d.f since every sample is independent. Maximize the likelihood of finding samples.

If $X_1, ...X_n \overset{i.i.d}{\sim} f_x(x, \theta)$,

$$l(\theta) = \prod_{i=1}^{n} f_{X_i} f_{x_i}(x_i; \theta), L(\theta) = logl(\theta)$$

$$\hat{\theta}_{MLE} = argmax_{\theta} f_x(x; \theta) = argmax_{\theta} L(\theta)$$

Analytical or Numerically solved.

$$\frac{\partial}{\partial \theta}[logL(\theta)] = 0, \frac{\partial^2}{\partial \theta^2}[logL(\theta)] < 0$$

, for multiple parameters, we need the Hessian matrix to be negative definite $x^t H x < 0, \forall x$

- Properties of MLE

  1. Invariance $\hat{\theta}$ is MLE of θ, then $g(\hat{\theta})$ is MLE of $g(\theta)$
  2. Consistency
     $$P(\hat{\theta} - \theta) \to 0$$
     as $n \to 0, \forall \epsilon > 0$ Under the conditions
     (a) $X_1, ...X_n \overset{i.i.d}{\sim} f_x(x|\theta)$
     (b) parameters are identifiable, $\theta \neq \theta', f_x(x|\theta) \neq f_x(x|\theta')$
     (c) densities $f_x(x|\theta)$ has common support(set of x with positive density/probability), $f_x(x|\theta)$ is differentiable at θ
     (d) parameter space Ω contains open set ω where true θ_0 is an interior point
  3. Asymptotic Normality
     $$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \to N(0, I^{-1}(\theta_0))$$
     $$I(\theta_0) = E(-(\frac{\partial}{\partial \theta}[logf(x, \theta)])^2) = E(-\frac{\partial^2}{\partial \theta^2}[logf(x, \theta)])$$
     called the Fisher Information
     $$\hat{\theta}_{MLE} \approx N(\theta_0, \frac{1}{nI(\theta_0)})$$
     $$nI(\theta_0) = E(-\frac{\partial^2}{\partial \theta^2}logL(\theta))$$

So the Variance of MLE($1/E(-\frac{\partial^2}{\partial\theta^2} logL(\theta)))$ is the reciprocal of amount of curvature at MLE.
Usually, We can just use the *observed Fisher Information* (curvature near θ_{MLE}) instead. $(I(\theta_{MLE}))$
$\frac{1}{nI(\theta_0)}$ is called Cramer-Rao Lower Bound.
Under Multi-dimensional Case,

$$I(\theta_0)_{ij} == E(-\frac{\partial^2}{\partial\theta_i\partial\theta_j}[logf(x,\theta)])$$

$Hessian \approx nI(\theta_0)$ $Hessian^{-1} \approx nI(\theta_0)$ when we use numerical approach.

Under the above four conditions plus

(a) $\forall x \in \chi$, $f_x(x|\theta)$ is three times differentiable with respect to θ, and third derivative is continuous at θ, and $\int f_x(x|\theta)dx$ can be differentiated three times under integral sign

(b) $\forall\theta \in \Omega, \exists c, M(x)$ (both depends on θ_0) such that

$$\frac{\partial^3}{\partial\theta^3}[logf(x,\theta)] \leq M(x), \forall x \in \chi, \theta_0 - c < \theta < \theta_0 + c, E_{\theta_0}[M(x)] < \infty$$

- Δ -Method: $g(th\hat{e}ta_{MLE})$ is approximately

$$N(g(\theta),(g'(\theta))^2\frac{1}{nI(\theta)})$$

  if asymptotic normality is satisfied.

- Estimation criteria
  - Unbiased $E(\hat{\theta}) = \theta$
  - Minimum Variance (MVUE, minimum variance unbiased estimator) $Var(\hat{\theta}) < Var(\theta')$
  - Efficient
  - Coherent

1.3.2 Important Distributions

1. Normal Distribution, $X_1, ...X_n \sim N(\mu, \sigma^2)$ then

   (a) $\bar{X}$ and s^2 are independent

(b) $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

(c) $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

(d) $\frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\frac{(n-1)s^2}{\sigma^2}\frac{1}{\sqrt{n-1}}} \sim t_{n-1}$

2. Multi-variate normal distribution

$$f_x(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

(a) $X_1, ...X_n$ normal $\Leftarrow (X_1, ...X_n)$ is multivariate normal. (Not equivalent)

(b) $E(X) = \mu, Var(X) = \Sigma$

(c) Linear transformations $AX + b \sim N(A\mu + b, A\Sigma A^T)$ remain multivariate normal

(d) Marginals are multivariate normal, each sub-vector is multivariate normal, the parameters are just sub-matrices.

(e) All conditionals are multivariate normal

3. t-distribution: like normal distribution, but heavier tails

(a) $Z \sim N(0,1), Y \sim \chi^2_\nu$, Z, Y independent,

$$X = Z/\sqrt{Y/\nu} \sim t_\nu$$

(b) pdf has polynomial tails (decays much slower than exponential ones)

(c) $\nu = 1$, it is the **Cauchy Distribution**, with very heavy tails (no expectation)

(d) The MCF not exist. $E(|X|^k) < \infty$ for $k < \nu$, $E(|X|^k) = \infty$ for $k > \nu$

(e) $X \sim t_\nu, E(X) = 0, Var(X) = \frac{\nu}{\nu-2}$

$$f_X(x) = \frac{1}{\pi(1+x^2)}$$

4. χ^2 distribution

$$f_x(x) = \frac{1}{(2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, x \in [0,\infty) \sim Gamma(\frac{k}{2}, \frac{1}{2})$$

(a) $E(X) = k, Var(X) = 2k, M_X(t) = (\frac{1}{1-2^t})^{k/2}$

(b) $X \sim N(0,1) \Rightarrow X^2 \sim \chi^2$, $X_1, ...X_n \sim N(0,1) i.i.d \Rightarrow \sum X_i^2 \sim \chi^2$,

$$f_X(x) = \frac{1}{\pi(1+x^2)}$$

1.3.3 Theorems

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

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

1.3.4 Practice/Examples

1. sample mean($\bar{X}$) is unbiased. Sample variance ($\frac{1}{n-1}\sum_{i=1}^{n} x_i^n$) is unbiased. But sample std is not unbiased. $SE(\bar{X}) = \frac{\sigma^2}{n}$
2. $\hat{Cov}(X.Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{Y})$ is unbiased
3. Distributions with Expectation not exist? (Cauchy)
4. Common Confidence Intervals:
   μ: $P(-t_{\alpha/2,n-1} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq t_{\alpha/2,n-1}) = 1 - \alpha$,
   σ: $P(a \leq \frac{(n-1)s^2}{\sigma^2} \leq b) = 1 - \alpha$
5. Solve MLE/MOM for beta, exponential ($n/\sum X_i$, normal
6. * prove Asymptotic Normality of MLE(hint: using Taylor Expansion for $\theta, \hat{\theta}$)
7. * Use t^{th} quantile to approximate c.d.f, what's the distribution? ($Y_n = \frac{1}{n}\sum I(X_i < x)$, a Bernoulli distribution with $p = F_x(x)$, $\sqrt{n}[Y_n(x) - F_x(x)] \sim N(0, F(x)(1 - F(x))$.

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

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

3.1.3 Lasso-Least Absolute Shrinkage and Selection Operator

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

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