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
- Bayes' Rule

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

  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 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

$$D_{KL}(P||Q) = \sum_{i=1}^{n} p(x_i) log_2 \frac{p(x_i)}{q(x_i)} \neq D_{KL}(Q||P)$$

$$D_{KL}(f, \hat{f})) = \int_{-\infty}^{\infty} log(\frac{f_X(x)}{\hat{f(x)}}) f_X(x) dx$$

  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) K-L divergence plays an important role in both information theory and MLE theory. MLE $\hat{\theta}$ is actually finding the closest K-L Distance approximation of $f(x;\theta)$ to sample distribution.

Theorems:

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

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 Formal Logic

Concepts

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

Chapter 2

Statistics

2.1 Concepts

2.1.1 Basic

- 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*.

2.1.2 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 $x^t Hx < 0, \forall x$

- 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: $g(\hat{theta}_{MLE})$ is approximately

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

if asymptotic normality is satisfied.
In Multivariate Case:

$$\hat{\theta} \sim N(\theta,\Sigma/n), \theta, \hat{\theta} \in R^p$$

$$g : R^p \to R^m$$

$$g(\hat{\theta}) \sim N(g(\theta), G\Sigma G^T/n)$$

$$G = \begin{pmatrix} \frac{\partial g_1(\theta)}{\partial\theta_1} & \cdots & \frac{\partial g_1(\theta)}{\partial\theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m(\theta)}{\partial\theta_1} & \cdots & \frac{\partial g_m(\theta)}{\partial\theta_p} \end{pmatrix}$$

- Estimation criteria
  - Unbiased $E(\hat{\theta}) = \theta$

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

2.1.3 Model Selection

AIC - Akaike Information Criterion

By K-L Distance

$$D_{KL}(f, \hat{f})) = \int_{-\infty}^{\infty} log(\frac{f_X(x)}{f(\hat{x})}) f_X(x) dx$$

$$= const + \frac{1}{2} \int (-2log\hat{f}(x)) f(x) dx = const + AIC$$

$$A(f, \hat{f}) = -2logL(\theta) + 2p(\frac{n}{n-p+1})$$

2.1.4 Hypothesis Testing

- Hypothese, Test Statistic(T), Rejection Region
- p-value (chance of rejecting, largest choice of α that we would fail to reject H_0)
- type-I error(wrongly reject), type-II error(wrongly accept)

Hypothesis Testings (Based on the distribution of $\hat{\theta}$)

- Wald Test

$$T = \frac{\hat{\theta} - \theta_0}{Se(\hat{\theta})}$$

$$\hat{\theta}_{MLE} \approx N(\theta_0, \frac{1}{nI(\theta_0)})$$

$$T = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{1}{nI(\theta_0)})}}$$

- Likelihood Ratio Test
- Score Test

*Computation-based hypothesis testing approach

- Permutation tests:
  Test $X_1, ..X_n \sim F, Y_1, ...Y_n \sim G, if F = G$. Use
  $T = Mean(X_i) - Mean(Y_i)$, each time scramble X and V labels and should not not change the distributions of vectors $X_1, ...X_n, Y_1..., Y_n$
- Bootstrapping:
  $X_1, ...X_n \sim F$ with $T = T(X_1, ..., X_n)$, to get the distribution of T, **sample with replacement.** The belief is $(\hat{\theta} - \theta)$ should behave the same as $(\theta * -th\hat{e}ta)$. The first quantity can be treated like a pivot. (use $(\theta *_1 -\hat{\theta}_1), ...(\theta *_n -\hat{\theta}_n)$ to test.

Multiple Testing

- Family-wise Error Rate(FWER) the probability of rejecting at least one of at least one null hypothesis
  Under independence, the probability of making mistake when all null are true: (P(any type I mistake) = 1-P(no type I mistake for all) = $1 - (1 - \alpha)^M = \beta$)
- Bonferroni correction, assuming independence
  $P(\bigcup_{i=1}^n typeI_mistake) \leq \sum_{i=1}^n P(typeI_mistake) \leq M\alpha$, control at $\alpha = \frac{\alpha}{M}$
  α being to small will impact power of the individual tests!
- False Discovery Rate(FDR): bound the fraction of type-I errors. R be the total number of hypotheses rejected. V be the number of rejected hypotheses that were actually null. Let FDR = V/max(R,1), control $E(FDR) \leq \alpha$.

2.2 Theorems

- Law of Large Number

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

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

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

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

5. F-Distribution

More Generalized Distributions

1. Generalized Error Distribution (symmetric)
2. Non-standard t-distribution (shift and scaling, heavy tailed, symmetric)
3. Theodossious skewed t-distribution
4. Theodossious skewed t-distribution plus shift

2.4 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 $p = F_x(x)$,
$\sqrt{n}[Y_n(x) - F_x(x)] \sim N(0, F(x)(1 - F(x))$.

8. $X_1, ...X_n \sim Binomial(n, p)$, What's the MLE for p and Fisher Information? ($\hat{p} = \frac{x_i}{n}, I(p) = 1/p(1-p), var(p) = \frac{p(1-p)}{n}$)

9. $(x_i, y_i) \sim N(\mu_i, \sigma^2)$, find MLE for σ ($\frac{1}{4N}\sum(x_i - y_i)$)

10. How can you get N(0,1) random variables from U[0,1]? (Method1: Inverse Transformation, Method2; Use
$SumZ_i^2 \sim \chi_k^2, k = 2, F^{-1}(u) = -2log(1-u)$,
$R^2 \sim \chi^2, Z_1 = Rcos\theta, Z_2 = Rsin\theta, \theta \in [0, 2\pi]$

11. (Permutation test) how can you test $X_1, ..., X_n \sim F$, how can you test if F is symmetric? (Multiply -1 on all two form two sample groups)

12. Draw a bootstrap sample, what fraction of original data points appear in this sample on average?
Define I be the indicator is it is in the sample.
$E(\frac{1}{n}\sum I_i = E(I_i) = P(\text{ith point shows up}) = 1 - (1 - \frac{1}{n})^n$

Chapter 3

Bayesian Statistical Theory

Bayes' Rule

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x,y)}{f_X(x)} = \frac{f_{(X,Y)}(x,y)}{\int f(x|y)f(y)dy}$$

Byesian Inference:
All parameters are random variables,

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

$$\pi(\theta|x) \sim f(x|\theta)\pi(\theta)$$

$\pi(\theta)$ is the prior distribution, $\pi(\theta|x)$ is the posterior distribution for θ given x.

Bayes Estimator

$$\hat{\theta}_{Bayes} = E(\theta|X) = \int \theta\pi(\theta|X)d\theta$$

Conjugate Distribution: $f(x), \pi$ is called conjugate distributions if model $\pi(\theta|x), \pi(\theta)$ follows the same Distribution

eg. Bernoulli(θ) and Beta(α, β), ($\pi(\theta|x) \sim$ Beta($\alpha + \sum X_i, \beta + n - \sum X_i$) ($f(x|\theta) = \prod_{i=1}^n f_{X_i}(X_i|\theta)$)

$$\hat{\theta}_{Bayes} = E(\theta|X) = \frac{\alpha + \sum X_i}{\alpha + \beta + n}$$

$$= \frac{\sum X_i}{n} \frac{n}{\alpha + \beta + n} + \frac{\alpha}{\alpha + \beta} \frac{\alpha + \beta}{\alpha + \beta + n}$$

The prior mean (second term) influences less as n grows.

Poisson(θ) and $Gamma(\alpha + \sum X_i, \beta + n)$

3.1 Bayesian Decision Theory

In State ω we take action $a \in A$, incurr Loss $L(\omega, a)$, how to choose a?
Risk:

$$R(a|x) = \sum_{j=1}^{k} L(\omega_j, a) P(\omega_j|x)$$

Decision Rule $d \in A$

$$d^*(x) = \arg\min_{a \in A} R(a|x)$$

Chapter 4

Computational Learning Theory

Part II

Supervised Learning Models

Chapter 5

Regression Overview and Linear Regression

5.1 Overview

5.1.1 Type of Models

All Basic Models begins with **Linear Regression** Because

- Linear relationship is the simplest relationship other than constant relationship or "null" model (average)
- It's a global model
- Data Invariance: Simple linear model don't do any pre-processing or transformation on the covariants.
- Very Explainable, limited interpretation power.

So, the alternation/improvements also focuses on these aspects

- Nonlinear features-Introduction of basis function
  - Polynomial Regression

  - Spline Models(eg. Cubic Spline, Smoothing Spline)

- Nonlinear parameters: Parameters Self-adjusting.(activation function is an example of basis function as well)
  - Neutral-Network
- global nonlinear: global nonlinear on both parameters and features achieved by linkage function, extends regression models to classification.
  - Generalized Linear Model
- Change the global model to a local model
  - Local Regression (Regression + KNN)
  - Nonparametric Regression
  - Kernel Function
  - Distance Based Learning
- Data Preprocessing (Transformation) and Dimension Reduction
  - PCA
  - LDA
  - Manifold Learning
- Improve Generalization Capability from outside (not from inside the model)
  - Regularization Methods(eg. Ridge, Lasso)
  - Ensemble Learning(Stacking, Aggregating): Random Forest, Boosting(GBDT), Deep Learning...

5.1.2 The Key Questions

- What assumptions are the model making
- How will we access the validity of those assumptions
- How can we be confident about out-of-sample fitting (overfitting problem)
- How do we make predictions and quantify the uncertainty in models?

5.2 Linear Regression

Common Terms

1. Independent Variable, Features, Covariates, Predictors
2. Dependent Variable, Response, Output (variable)
3. Scaling - transform a variable to have mean zero and variance one

5.2.1 Assumptions

Classic Assumptions for Statistics:

1. Linear Relationship between covariates and dependent variable
2. $E(\varepsilon) = 0$
3. $Var(\varepsilon) = \sigma^2$: Homoscedasticity
4. ε is independent with covariates
5. x is observed without error (and no perfect multicollinearity in multivariate case)
6. (optional, Gauss-Markov Theorem) ε is normal - when it is, OLS and MLE agrees and to be BLUE(Best Linear Unbiased Estimator)

Testing the Assumptions of Linear Regression

- Scatter Plot
  Linear Relationship and Outliers
- Residual Analysis $\hat{\varepsilon} = y - \hat{y}$
  Diagnostic Plots:
  1. Plot of Residuals vs. Fitted Values
  2. Normal Probability Plot
  3. Plot Residuals versus time (see any trend of fit)

- Cook's Distance
$$D_j = \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_{i(-j)})^2}{(p+1)\hat{\sigma}^2}$$
Test Against $F_{(p+1),(n-p-1}$ degrees of freedom, over 50th percentile will definitely become a problem

Resolutions of Assumption Violations

- Verify the Linear Relationships again. (non-linear regression, generalized linear models)
- Transformations (for outliers, heteroskesticities, etc)
- Use different models on different periods/data
- Weighted **Least Squares regression**, **Robust Regression** (for outliers, heteroskesticity)

5.2.2 Intepretaion

Under Normal Condition, we have

$$y \sim N(\beta_0 + \beta x_i, \sigma^2), L(\theta) = (\frac{1}{\sqrt{2\pi}\sigma)}^n exp(-\frac{\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2})$$

Equivalent to minimize

$$RSS(\theta) = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

$$\partial_{\beta_i} RSS = 0, i = 1, 2$$

, we get
$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \beta_1 = r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}, \beta_0 = \bar{y} - \hat{\beta}\bar{x}$$

In Multi-variate Case:

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

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

Assuming noise is normal, maximize

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

Another matrix representation

$$f(\beta) = min(Y - X\beta)^T(Y - X\beta), f'(\beta) = 2X^T(Y - X\hat{\beta}) = 0$$

to solve $\hat{\beta}$

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

Model Selection and Overfitting Resolve

By AIC or BIC.

Variance Error In Prediction

$$V(\hat{y^*} - y^*) = \sigma^2 + \sigma^2[\frac{1}{n} + \frac{x^* - \bar{x})^2}{(n-1)s_x^2}]$$

$$= V(E(y^*) - y^*) + V(\hat{y^*} - E(y^*)) + 2cov(\hat{y^*} - y^*, \hat{y^*} - y^*)$$

The cross term is zero, the first term is variance with ε^*, second term is variance in β.

The confidence interval is $\hat{y^*} \pm t_{\alpha/2,n-2}SE(\hat{y^*})$.

5.3 Multiple Linear Regression

5.3.1 Lasso-Least Absolute Shrinkage and Selection Operator

5.4 Regularization, Ridge and Lasso

5.5 Practice/Examples

1. What is Anscombe's Quartet

Chapter 6

Logistic Regression and Generalized Linear Model

Chapter 7

Neutral-Network

Chapter 8

Distance and Kth Nearest Neighbors

Chapter 9

Naive Bayesian

Chapter 10

Tree Models and Ensemble Learning

Chapter 11

Classification Overview and Support Vector Machine

Stacking Aggregating: Random Forest, Boosting(GBDT), Deep Learning...

Part III

Unsupervised Learning Models

Chapter 12

Clustering

Chapter 13

Dimension Reduction

13.1 PCA

13.2 LDA

Part IV

Deep Learning and Enhanced Learning Theory

Markov Chain Hidden Markov Chain Markov Blanket Bayesian Network
EM Algorithm

Chapter 14

Multi-layer Perceptron