Chapter 1

Linear models

1.1 Problem Description

Input data **x** N*M matrix, N data points, each has M dimension, or M features
Target value **t**
predict value **y(x)**
Sometimes we use input transformation, to replace x with $\phi(x)$ (basis function), a N*D matrix, also called design matrix Φ. Design matrix is often fixed non-linear functions of input variables, since it would be not necessary if it's linear(parameter vector w can cover this condition).

Candidate of basis function

- powers of x, x^j
- Gaussian basis function, $\phi(x) = exp\{-\frac{(x-\mu_j)^2}{2s^2}\}$, where μ_j governs the location, and s governs the space scale.
- logistic sigmoidal $\phi_j(x) = \sigma(\frac{x-\mu_j}{s})$, where $\sigma(a) = \frac{1}{1+exp(-a)}$
- tanh $tanh(x) = \frac{e^x-e^{-x}}{e^x+e^{-x}} = 2\sigma(2x) - 1$
- fourier basis (warelet)

1.2 Least squares

model

$$y = w^T x$$

error function (squared error)

$$C(x) = (w^T\Phi - t)^T(w^T\Phi - t)$$

gradient w.r.t w

$$\nabla C = \Phi^T(\Phi w - t)$$

let it be 0, we get

$$w = (\Phi^T\Phi)^{-1}\Phi^T t$$

(we need to investigate 2nd order derivative matrix to make sure this zero point is not a saddle point.)

to predict a test data

$$y' = w^T x'$$

1.3 Maximum likelihood

all data points are assumed i.i.d Gaussian, $t = y(x, w) + \epsilon$ where ϵ is the zero mean Gaussian additive noise.

likelihood function

$$p(t|X, w, \beta) = \prod_n N(t_n|w^t\phi(x_n), \beta^{-1})$$

maximize the likelihood, or the logarithm of the likelihood, is equal to minimize the sum of squared error, and at last get the same result as the least squares.

if assume other density distribution, we get different model. e.g. for laplacian noise, $p(\epsilon) = \frac{1}{2a}e^{-\frac{|z|}{a}}$, we get model $\hat{x} = argmin_x|ax - y|_1$

1.4 LDA

It's said, that LDA is equivalent to liner regression, for two classes problems.

1.5 Softmax

1.6 Bayesian

According to Bayesian formula:

$$p(w|X, t) = \frac{p(t|X, w, \beta) * p(w)}{\sum_\beta p(t|X, w, \beta) * p(w)}$$

Given that $p(w|\alpha) = N(0, \alpha I)$ (isotropic Gaussian with zero mean), maximize the above function is same as minimize $|y - t|^2 + \alpha w^T w$. This is squared errors plus ridge regularization

Q: how to interpret lasso regression with bayesian theory?

1.7 Kernel methods

regularized sum of squares,

$$J(w) = \frac{1}{2}(\Phi w - y)^T(\Phi w - y) + \lambda w^T w$$

gradient

$$\nabla J = \Phi^T(\Phi w - y) + \lambda w$$

Let the gradient be 0, we get

$$w = \Phi^T a$$

where

$$a = -\frac{1}{\lambda}(\Phi w - y)$$

replace $\Phi^T a$ into J(w), we get

$$J(w) = \frac{1}{2}(w^T\Phi^T\Phi w - y^T\Phi w - w^T\Phi^T y + y^T Y) + \lambda w^T w$$

$$J(a) = \frac{1}{2}(a^T\Phi\Phi^T\Phi\Phi^T a - y^T\Phi\Phi^T a - a^T\ Phi\Phi^T y + y^T y) + \lambda a^T\Phi\Phi^T a$$

let

$$K = \Phi\Phi^T$$

, we get

$$J(a) = \frac{1}{2}(a^T KKa - y^T Ka - a^T K^T y + y^T y) + \lambda a^T Ka$$

$$\nabla J(a) = KKa - Ky + 2\lambda Ka$$

let it be 0, we get

$$a = (K + 2\lambda I)^{-1} y$$

then to predict for x, we use

$$\hat{y}(x) = \phi(x)\Phi^T a = \phi(x)\Phi^T(K + 2\lambda I)^{-1} y = k(x)^T(K + 2\lambda I)^{-1} y$$

where $\phi(x)$ is a row vector, and K is called kernel matrix.

Although $k(x) = \phi(x)\Phi$ and $K = \Phi\Phi^T$, but in computation, we don't really need Φ, we can design kernel functions directly, e.g. $k(x_m, x_n) = f(x_m, x_n) = \phi(x_m)^T\phi(x_n)$.

The function must be a valid kernel, or, in other words, it must correspond to a scalar product in some feature space. Sometimes we can construct kernel functions corresponding to infinite dimensional feature spaces.

A commonly used kernel is Gaussian kernel:

$$k(x, x') = exp(\frac{-|x - x'|^2}{2\sigma^2})$$

Problem:

- how to train this model? or not to train at all?

  A: reasoning about the process of inducing the prediction formula, we get the value of a by minimizing J(a), so we don't need to train the model again, it's already the best. But still we can adjust λ or kernel functions, and test them with validation sets.

  Also in regularized linear regression model, these is a penalty parameter λ, it's chosen too, not trained.

- the prediction formula uses only one unknown variable λ, how to select this value?

  A: ref previous problem.

1.7.1 Gaussian Process

Assume w is isotropic Gaussian, $y = \Phi w, t = y + \epsilon$,y is linear combination of w, ϵ is also Gaussian independent of w, then y is also Gaussian. let

$$P(w) = N(0, \alpha I)$$

$$P(\epsilon) = N(0, \beta)$$

then

$$E(y) = \Phi E(w) = 0$$
$$Cov(y) = E(yy^T) = E(\Phi w w^T \Phi^T) = \Phi E(ww^T)\Phi^T = \alpha \Phi \Phi^T = K$$

$$P(y) = N(0, K)$$
$$P(t|y) = N(0, \beta^{-1} I)$$
$$P(t) = N(t|0, C) = N(t|0, K + \beta^{-1})$$

the co-variance matrix of P(t) can be calculated by kernel functions, without using design matrix Φ, this make it possible to handle infinite dimensional feature space.

Problems:

- How to predict with Gaussian process model?

  A:use conditional Gaussian distribution formula:

- How to train a Gaussian Process model? what parameters do we need to find?

- It's said that GP is modeling directly on functions space, which function?

Chapter 2

Algorithms

Chapter 3

Statistical Learning

3.1 Statistical concept

- population regression line
  best linear approximation to the true model.
- least square line
- bias vs variance
  $MSE = BIAS^2 + VARIANCE$
- formula of standard error(SE)

$$Var(\mu) = SE(\mu)^2 = \frac{\sigma^2}{n}$$

- t-statistic, t-distribution (TBD)
- degrees of freedom - number of free (independent) variables
- f-statistic
- F1-score
- RSE - Residual Standard Error

$$RSE = \sqrt{\frac{1}{n-2}RSS}$$

- R^2

$$R^2 = \frac{TSS - RSS}{TSS}$$

  TSS total sum of squares.

$$TSS = \Sigma_i(y_i - E(y))^2$$

  RSS residual sum of squares.

$$RSS = \Sigma(y_i - y)^2$$

- p-value
- null hypothesis (H) - assume null hypothesis, compute p-value, which shows the probability that the data is observed given that H is true. this is to say, p = P(D—H), the posterior probability? refer to chapter 37 of [1] and chapter 3 of [2]
- alternative hypothesis
- LDA

$$P(Y = k|X = x) = \frac{P(X = x|y = k) * P(y = k)}{\Sigma_l P(X = x|y = l) * P(y = l)}$$

prior probability: $P(Y = k)$ suppose $P(X = x|y = l)$ has normal distribution, and has equal variance for all l. then, to get the largest value of $P(Y = k|X = x)$, we need only to get the largest nominator in the right part. let $\pi_k = P(y = k)$ the formula above have same denominator for each class k, we need only compute the nominator. plug in the normal distribution formula, we get

$$d_k(x) = \pi_k \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2}(x - \mu_k)^2)$$

taking the log we get:

$$log(d_k(x)) = log(\pi_k) - \frac{1}{2\sigma^2}(x - \mu_k)^2 - log(\sqrt{2\pi}\sigma)$$

the last item is the same for all k, so we can ignore it also. and for given x, $\frac{x^2}{2\sigma^2}$ is also a constant, at last we get:

$$log(\pi_k) + \frac{\mu_k x}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

we can assign x the class k for which the above formula is largest.

3.2 neural networks

- sigmoid function $f(z) = \frac{1}{1+e^{-z}}$ why do we use -z other than z? Maybe to make f(z) a ascending function.

3.3 spline functions

- make a comparison between Gradient descent and Neuton method
- gradient descent, if we adjust parameters each time after one sample training, it's called stochastic gradient descent, and if we adjust parameters once after a batch training, it's called gradient descent. + stochastic gradient descent, on-line learning + gradient descent, batch learning. problem: why it's called stochastic? problem2: some books call on-line learning those methods which test a few examples, not just limited to only one.

- check if the lasso regularization method finds the same model as SVD or PCA. 3Mar2006: svd and pca find potential dimensions (linearly combinations of existing ones), they do base-transformation, so apparently get different answer from lasso? and if so, check if other dimension-reducing methods get same answer as lasso.

3.4 Bootstrap

Original data points $X = \{x_1, x_2, ..., x_N\}$

Create new data set by drawing N points at random from X, with replacement (some x may be replicated, and others may be absent). this process can be repeated L times, and we got L data sets.

3.5 Perceptron

ESL,4.5 compute a linear combination of input features and return the sign, is called perceptron.

It's linear, so the separator contains 0 point. $y = sign(w^T x)$

How to train?

For mis-classified sample x_i, let $w \leftarrow w + y_i \eta x_i$. Repeat it until converged.

3.6 Random forest

3.7 Extra trees

3.8 Gradient Boost machine

3.9 Kernel methods

3.10 Gaussian process

Linear regression in GP's view
Linear classification in GP's view
In normal Gaussian process, the index set is time series. but in machine learning context, the index set is often the index of input data.

- how GP model make a prediction?

3.11 Ensembling learning

3.11.1 Ensembing methods

- vote
- weighted vote
- averaging

- geometric mean (maybe better than averaging)
- rank averaging

  Rank the predictions first, then averaging ranks.
- stacked generalization

  Use a pool of base classifiers, and then use another classifier to combine their predictions, aiming to reduce generalization error.
- blending

  almost same as stacked generalization.

3.11.2 reference

kaggle ensembling guide, mlwave.com

3.12 model metrics

- logloss
- AUC, area under curve (ROC curve)
- accuracy
- precision and recall

  precision: fraction of detections that were reported by the model that were correct.

  recall: fraction of true events that were detected.
- F1 score

$$F1 = \frac{2pr}{p+r}$$

- coverage

  fractions of examples that a machine learning system can produce a response. (the rest of which need a human being to decide.)

3.13 time series

for single variable time series,

weak stationary, means the auto-covariance depends on the separation of x_s and x_t, say, $|s-t|$, and not on where the points are located in time, and the mean value function is constant.

strictly stationary, means the probabilistic behavior of every collection of values $\{x_{t1}, x_{t2}, ..., x_{tn}\}$ is identical to a time shift of the values $\{x_{t1+h}, x_{t2+h}, ..., x_{tn+h}\}$

for k-dimensional time series

weakly stationary if $E(z_t) = \mu$, a K-dimensional constant vector, and $Cov(z_t) = E((z_t - \mu)(z_t - \mu)^T) = \Sigma_z$, a constant $K \times K$ positive-definite matrix. in other words, the first two moments of Zt is time-invariant.

strictly stationary has the same meaning as for single-variable time series. the joint probability distribution of m collection, (z_{t1}, z_{tm}), is the same as that of (z_{t1+j}, z_{tm+j})

3.14 Mixture Gaussian

3.15 EM

3.16 references

[1] information theory, inference, and learning algorithms by David J.C. MacKay
[2] statistical foundation of machine learning by Gianluca Bontempi

Chapter 4

Discretization

4.1 Bayesian Discretization

The basic idea is to use K-means algorithm to cluster values of an attribute, while the user has to specify the K value.

4.2 ID3

use information entropy to choose attributes and cutpoint.

$$H(S) = -\sum_{x \in S} p(x) log_2 p(x)$$

4.3 C4.5

4.4 Maximum Marginal Entropy

4.5 Ent-MDLP (minimum description length principle)

basic idea: a cutpoint for a set of points is accepted if the cost or length of the message required to send after partition is less than the cost or length of the message required to send before the partition.

4.6 χ^2

Chapter 5

Optimization

Newton method

BFGS

Q:why in convex optimization problem the contraints functions need to be all convex?

5.1 Line Search

$$x_{k+1} = x_k + t_k \eta_k$$

where η_k is the search direction, and t_k is the step size.

- Scale Invariance

  if a direction and step length are selected base on an algorithm that is not sersitive to x, it's called scale invariant.

- Armijo condition

  $$f(x_k + \alpha p_k) <= f(x_k) + c_1 \alpha \triangledown f_k^T p_k$$

  where $\triangledown f_k^T p_k$ is the directional derivative. Armijo condition make sure that the decrease of f is proportional to both a and directional derivative. $c_1 \in (0, 1)$. But Armijo condition can't assure the sufficiency of decrease since for small α near x_k the condition is almost always met.

- Wolfe conditions

  $$\triangledown f(x_k + \alpha_k p_k)^T p_k >= c_2 \triangledown f_k^T p_k$$

  Wolfe conditions makes sure that the step length is not too small, as an additional condition to Armijo condition. It means the directional derivative at α_k is not too low, since low value means in this direction the value of f will decrease quickly and we may not need to stop here.

- Strong Wolfe conditions

- Goldstein conditions

- Backtracking line search

  just give a big stop lenght, and shrink it slowly. then we can avoid problem of Armijo condition.

Chapter 6

statistics

- point estimation
- set estimation
- confidence interval
- confidence set
- sufficient statistics

6.1 parameter inference

6.1.1 methods

- moments. why called moments?
- Maximum likelihood
- bayesian inference

6.1.2 check assumption

goodness-of-fit test

6.2 hypothesis testing

Null hypothesis, Alternative hypothesis

6.3 Monte Carlo statistics

Null hypothesis, Alternative hypothesis

Chapter 7

Math

7.1 SVD

SVD(singular value decomposition) has a relation to Lagrange multipliers theorem. Find the details.

- they both uses λ
- SVD can be proved using Lagrange Multipliers theorem, and Weierstrass theorem. (*topics in matrix analysis*, Horn Johnson).
- Horn Johnson also discussed the related history.
- *Vector calculus, linear algebra, and differential forms* also gives a brief history, in which the author says SVD is first proved by Lagrange, and the **eigenvalues are in fact Lagrange multipliers**. (but it seems that the author deleted this content in new edition of the book)

7.1.1 Numerical methods to compute SVD

- common methods
- fast methods
- methods used by industry

papers to check:
Large-scale Parallel Collaborative Filtering for the Netflix Prize
Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions?
Golub's and Van Loan's matrix computations
Å. Björck's Numerical Methods for Least Squares Problems
Nela Bosner Fast Methods for Large Scale Singular Value Decomposition Doctoral Thesis
On Parallelizing Matrix Multiplication by the Column-Row Method? Andrea Campagna? Konstantin Kutzkov? Rasmus Pagh§
links:
https://en.wikipedia.org/wiki/Principal_component_analysis

software:
Eigen, Armadillo and Trilinos

7.2 PCA

what's the difference between SVD and PCA? If we remove the mean from features before doing PCA, then PCA is just SVD.

(prove it, using PCA's correlation matrix)

7.3 linear algebra

for matrix A, f(x) = Ax, what is the max(f(x))? does it relate to eigen-value/eigen-vector of A? if A has rank(A) eigen-vectors, what's the relationship between these eigen-vectors and f(x)?

anti-symmetric matrix: $-A = A^T$, for such matrix, $x^T Ax = 0$

7.3.1 Cholesky decomposition

- Freobenius norm

  sum of square of each elements

7.4 Probability Theory

7.4.1 Geometric distribution

$$P(k) = (1-p)^{k-1} * p$$

the name maybe has something to do with geometric sequence: a sequence of numbers where each term after the first one is found by multiplying the previous one by a fixed, non-zero number called the common ratio.(wikipedia) also named ???? in contrast, the arithmetic sequence: a sequence of numbers where the difference between consecutive ones is constant (common difference). ????
Let P(0) = p, P(1) = 1-p, then P(k) means (k-1) 1's before a 0 occures.

7.4.2 Binomial distribution

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

Let P(0) = p, P(1) = 1-p, then P(k) means k out of n times experiment outcome is 0.

7.4.3 Gaussian distribution (normal distribution)

$$N(x|\mu,\sigma) = P(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

$\beta = 1/\sigma^2$ is called precision

the sum of a random variables(**N terms**), which is itself a random variable, is Gaussian, as long as the **number of terms(N) in the sum** is large enough. (central limit theorem)

7.4.4 random variable

Entropy

$$H[x] = -\sum_x p(x) log_2 p(x)$$

Differential Entropy

$$-\int p(x) lnp(x) dx$$

The distribution that maximizes the differential entropy is the Gaussian (prove it)

Cross Entropy

$$H(p,q) = -\int_x p(x) log(q(x))$$

7.4.5 Random vectors

Probability density of a random vector, is the joint probability density of its components.

7.4.6 Gaussian random vectors

If $(x_1, x_2, ..., x_n)$ is a Gaussian random vector, then each element x_i is a Gaussian random variable, and each subset from a Gaussian random vector, too. However, this relationship is not reversible, that is, if $x_1, x_2, ..., x_n$ are all Gaussian random variables, the vector $(x_1, x_2, ..., x_n)$ is not necessarily a Gaussian Random vector.

The properties of Gaussian random vectors follow largely from the jointly Gaussian property rather than merely the property of being individually Gaussian.

Problem: derive the joint probability density function of a Gaussian random vector.

$p(x|m,\Sigma) = (2\pi)^{-D/2} * |\Sigma|^{1/2} * exp(-\frac{1}{2}(x-m)^T \Sigma^- 1(x-m))$

Mode of Gaussian distribution

7.4.7 Concetps

- central limit theorem
- likelihood

  Likelihood of parameter w, is the conditional probability of output given w, $p(t|w)$

- Mode of distribution

  The mode is a value that appears most often in a set of data, or the most probable value. For Gaussian, the mode equals to mean.

- What is a Moment Generating Function(MGF)?

  In probability theory and statistics, MGF is another way to describe a random variable besides probability density function or cumulative distribution function. However, notice that not all random variables have a MGF.

  MGF is defined as following:

  $$M_x(X) = E[e^{tX}], t \in R$$

  whenever this expectation exists.

- Moment from wikipedia: the points represent probability density, the zeroth moment is the total probability(one), the first moment is the mean, the second central moment is the variance, the third is the skewness, and fourth kurtosis.

7.5 Information Theory

- conditional entropy of X give Y:

$$\begin{aligned} H(X|Y) &= \sum_y H(X|Y=y) \\ &= \sum_y P(y) \left[\sum_x P(x|y) \log \frac{1}{p(x|y)} \right] \\ &= \sum_{x,y} P(x,y) \log \frac{1}{p(x|y)} \end{aligned}$$

- relative entropy, or Kullback-leibler divergence:

$$D_{KL}(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)}$$

  where P and Q are probablities over the same alphabet Ax.

- Gibb's inequality

$$D_{KL}(P||Q) >= 0$$

7.6 Miscs

- *Moore-Penrose pseudo-inverse* of a matrix
  $(\phi^T \phi)^{-1} \phi^T$

- Gram matrix
  $\phi \phi^T$

7.7 Multi-variable functions

- interior - x is interior to set A if there is some neighborhood U of x, such that $U \subset A$
- open set - every point of A is interior, then A is open.
- closed set - complement A^c is open, then A is closed.
- compact set - in E^n, closed and bounded set. more general definition: a subset S of a topological space S_0 is compact if every open covering of S contains a finite sub-covering.

7.7.1 Problems

- why topology define compact set in such a wield way?
- what are the problems that measure theory solved while other old theories can't solve?

7.8 Vector Calculus

- gradient
- derivative of vector formulas
  $\frac{d(x^T x)}{dx} = 2x$
  $\frac{d(x^T Mx)}{dx} = (M + M^T)x$
  $\frac{d(Mx)^T(Mx)}{dx} = \frac{d(x^T M^T Mx)}{dx} = 2(M^T M)x$
  $\frac{d(Mx-y)^T(Mx-y)}{dx}$
  $= \frac{d(X^T M^T Mx - x^T M^T y - y^T Mx - y^T y)}{dx}$
  $= 2M^T Mx - M^T y - (y^T M)^T$
  $= 2M^T(Mx - y)$

7.9 Analysis

- L^1 space: integrable functions on R^d with the norm defined as

$$||f(x)|| = \int_{R^d} |f(x)|dx$$

- L^2 space: integrable functions on R^d with the norm defined as

$$||f(x)|| = \left(\int_{R^d} |f(x)|^2 dx\right)^{\frac{1}{2}}$$

  inner product defined as

$$< f, g >= \int f(x)\overline{g(x)}dx$$

metric defined as

$$d(f,g) = ||f-g||$$

it's a hilbert space.

- Quadratic forms

  Quadratic forms are polynormials, all of whose terms are of degree 2.

  quadratic forms as a sum of squares: all quadratic forms on R^n can be decomposed into sums of m linearly independant linear functions.

  $$Q(x) = (\alpha_1(x))^2 + ... + (\alpha_k(x))^2 - (\alpha_{k+1}(x))^2 ... - (\alpha_{k+l}(x))^2$$

  where $x \in R^n$

  and the number k and l are independant of specifically choosen linear functions, they depend only on Q(x).

  (k,l) is the so-called signature of the quadratic form $Q(x)$

  Q:how to use properties of quadratic forms to decide the type of a critical point? local mininum/maximum/saddle?

- Differential Forms

- inverse function theorem

  invertablity of derivative of f means invertablity of f locally, because derivative D(f) is very good local approximation for f.

  $F(u_0 + h) = F(u_0) + D_{u_0}h + r(h)$

  where $r(h) = o(||h||)$ as $h \to 0$

- Banach fixed point theorem

- Questions

  The determinant of a matrix is the volume of the parallegram spanned by the vectors of the matrix, prove it.

7.10 Topology

- first-countable space

  A space X is said to be first-countable if each point has a countable neighbourhood basis (local base).

- second-countable space

  A space is said to be second-countable if its topology has a countable base. More explicitly, this means that a topological space T is second countable if there exists some countable collection $U = \{U_i\}_{i=1}^{\infty}$ of open subsets of T such that any open subset of T can be written as a union of elements of some subfamily of U.

- hausdorff space

  A space X where all distinct points of X are pairwise neighborhood-seperable. If x, y are distinct points of X, there exists a neighbor U of x and a neighbor V of y such that $U \cap V = \emptyset$

- open

  how is open defined? or, is 'open' a predefined concept in topology, in other words, 'open' is not defined inside topology?

- closed

7.11 Matrix Manifold

- chart

  A bijection ϕ of a subset U of M onto an open subset of R^d is called a d-dimentional chart of set M, denoted by (U, ϕ)

- atlas An atlas of M into R^d is a collections of charts (U_α, ϕ_α) of M such that

  1) $\bigcup_\alpha U_\alpha = M$ 2) the elements of an atlas overlap smoothly. if $U_\alpha \cup U_\beta \neq \emptyset$, the sets $\phi_\alpha(U_\alpha \cap U_\beta)$ and $\phi_\beta(U_\alpha \cap U_\beta)$ are open sets of R^d and the changes of coordinates $\phi_\beta \circ \phi_\alpha^{-1} : R^d \to R^d$ is smooth.

- manifold

  a (d-dimensional) manifold is a couple (M, A^+), where M is a set and A^+ is a maximal atlas of M into R^d, such that the topology induced by A^+ is hausdorff and second-countable.

  lie: this definition is different from the one defined in analysis fields. In those literatures a manifold is a topological space that locally resembles euclidean space near each point. Maybe the only defference is the set M.

- embedded submanifolds

- quotient manifolds

- generalized eigen-value problem

  finding eigen pairs of matrix pencil is known as the generalized eigen-value problem. $Av = \lambda Bv$ where (A, B) is a matrix pencil.

- Stiefel manifold

  $\{M : R^{n*p}\}$ where all columns of M are linearly independant, and $p \leq n$, is called noncompact Stiefel manifold of full-rand n*p matrices.

  $\{X \in R^{n*p}, X^T X = I_p\}$ is compact Stiefel manifold.

- quotient manifold

  the set of equivalent classes of a relation r of M is called the quotient of M by , denoted as M/r.

- problems

  is it possible to use category theory to interpret some topics in matrix manifold?

Chapter 8

Problems

Generative model: $p(y|x) = p(x|y)p(y)/p(x)$, models both input and output distributions explicitly or implicitly.

Discriminative model: $p(y|x)$ models posterior distribution directly.

Chapter 9

NLP

9.1 Word Vectors

9.1.1 Skip Gram model

use centor word to predict context words, maximize the average log probability $\frac{1}{T\sum}$

$$p(w_o|w_c) = \frac{exp(< w_o, w_c >)}{\sum_{j \in V} exp(< w_j, w_c >)})$$

cost function:

- full softmax
- hierarchical softmax
- NCE(noise contrastive estimation)
- NEG(negative samping)

$$log(\sigma < wo, wi >) + \sum E_{w_i p_n(w)} log(\sigma(- < wo, wi >))$$

9.1.2 CBOW model

CBOW - use context words to predict center word, with the context words represented as a summary of word vectors.

9.1.3 GloVe

GloVe