Linear regression — minimize w.r.t. mean-square error

Error: $\hat{R}(w) = \|Xw - y\|_2^2$ — Assuming X^TX invertible

closed form: $w^* = (X^TX)^{-1}X^Ty$ $\Leftrightarrow$ full rank $\Leftrightarrow$ $n > p$

Gradient: $\nabla_w = 2X^T(Xw - y)$

Ridge Regression

Error: $\hat{R}(w) = \sum_i^n (y_i - w^Tx_i)^2 + \lambda\|w\|_2^2$

closed form: $w^* = (X^TX + \lambda I)^{-1}X^Ty$

Gradient: $\nabla_w = -2\sum_i^n (y_i - w^Tx_i)x_i + 2\lambda w$

Lasso L^1 penalty — penalize β — gradient

$\hat{\beta} = \sum_{i=1}^n (y_i - x^T\beta)^2 + \lambda\sum_{j=1}^d |\beta_j|$ — $\lambda\,\mathrm{sign}(\beta)$

$= (y - X\beta)^T(y - X\beta) + \lambda\|\beta\|_1$ — 当 $\beta \neq (0,0,\dots 0)$

No closed form solution.

Bayesian Linear Regression

Set a prior over β

Ridge Regression is equivalent as setting a Gaussian

Prior $\pi(\beta) \sim N(\beta|0, \frac{\sigma^2}{\lambda}I) \propto \exp\{-\frac{\lambda}{2\sigma^2}\beta^T\beta\}$

Finding posterior

Properties Involving linear Regression

Ridge	Lined	Lasso
biased	Unbiased	biased
Consistent	Consistent	Consistent

smallest Var among unbiased

Gaussian Process Regression

Rewrite joint dist.

$P\begin{bmatrix} y \\ y_* \end{bmatrix} = N\left(y \middle| 0, \begin{bmatrix} C_n & k \\ k^T & c \end{bmatrix}\right)$

$C_n = K + \sigma^2 I_n$, $c = k(x_{n+1}, x_{n+1}) + \sigma^2$, $k = k(x_{n+1}, X)$, $K = k(X, X)$

with k being the kernel function

Prediction with GP

$p(y_*|x_*, X, y) = N(y_*|\mu_*, \sigma_*^2)$

$C_n = K + \sigma^2 I_n$, $\mu_{y_*} = k^T C_n^{-1} y$, $\sigma_*^2 = c - k^T C_n^{-1} k$

$c = k(x_*, x_*) + \sigma^2$

predict new values from

$p(y_*|x_*, X, y) = N(y_*|\mu_y', \sigma_{y_*}^2)$

$\mu_{y_*} = k^T(K + \sigma^2 I)^{-1} y$

Kernel: · Can map to infinite dimension
· computationally cheaper

If K_1, K_2 are valid: ① 相乘 ② 相加 ③ Scaler ④ $K_1(h(\cdot), h(\cdot))$

$h: \mathcal{X} \to \mathcal{X}$ ⑤ $h(K)$ when h is poly / exp

Polynomial Kernel $(x^Tx' + 1)^d$ | $\tanh(\alpha x^Tx' + c)$ is not

RBF $\exp(-\frac{\|x - x'\|^2}{h^2})$ | $[1, -1], [-1, 2]$

To dis-prove P.s.d.

找一对 $(\cdot,\cdot)(\cdot,\cdot)$ $\det[K] < 0$

$\binom{n+d}{d}$

Logistics $R(w) = \sum_{i=1}^n \log(1 + \exp(-y_i w^T x_i))$

$P[y|x] = \frac{1}{1 + \exp(-yw^Tx)}$

MAP

$\arg\max_w P(w|D) = \arg\max_w P(D|w)\pi(w)$

MLE = MAP when $n \to \infty$ or the prior dist. is uniform

Laplace – L1
Gaussian – L2

Bias–Variance tradeoff — 有一个升有一个降

$\mathrm{Bias}(\hat{f}) = \mathbb{E}[\hat{f}] - f$

$\mathrm{Var}(\hat{f}) = \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2]$

Square Error Decomposition

$\mathbb{E}_D\mathbb{E}_{X,Y}[(\hat{f}(x) - Y)^2] =$

$\mathbb{E}_{X,Y}[(\mathbb{E}_{Y|X}[Y] - Y)^2]$ — noise²

$+ \mathbb{E}_X\mathbb{E}_D[(\hat{f}_D(x) - \mathbb{E}_D[\hat{f}(x)])^2]$ — Var

$+ \mathbb{E}_X[(\mathbb{E}_D\hat{f}_D(x)) - \mathbb{E}_{Y|X}[Y])^2]$ — bias²

Var ↑ overfit — bias ↑ underfit

Fit to every pt. — All is const.

↑ tends to underfit / high variance.

Cross validation / LOO 是对 $\mathbb{E}_D[\hat{R}(x)]$ 的 unbiased prediction.

把 Dataset 分成 K 份, 留一份 Test

LOO K=N (high variance, unbiased)

Bootstrapping. 抽 w/ replacement

The chance of a sample in bootstrap $1 - (1 - \frac{1}{n})^n = 0.632$

Therefore: ERM on Z would have 63% accuracy by memorization ⟹ over-confident

Jackknife. A way to de-biased estimator

Estimate of an $\hat{S}_n$'s bias — 估值偏差

$\hat{S}^{JK} = \hat{S}_n - \mathrm{bias}^{JK}$ — JK Estimator (去掉偏差)

$\mathrm{bias}^{JK} = (n-1)(\tilde{S}_n - \hat{S}_n)$

$\tilde{S}_n = \frac{1}{n}\sum \hat{S}_{n-1}^{-i}$ is the avg. LOO Estimator

It works for any distribution.

Bootstrap debiased. $\bar{S} = 2\hat{S} - \frac{1}{B}\sum_b S^*(b)$

Cramer Rao LB

$\mathrm{Var}[\hat{\theta}] \geq I_n(\theta)^{-1}$ — $\hat{\theta}$ is unbiased

$I_n(\theta) = -\mathbb{E}[\frac{\partial^2 \log p(x|\theta)}{\partial\theta\partial\theta}]$

AIC $= -2\log P(x|\theta) + 2k$

BIC $= -2\log P(x|\theta) + k\log n$

Bayes Optimal Classifier For 0-1 loss

$\hat{c}(x) = \begin{cases} y & \mathbb{P}[y|x] \geq 1-d \;\exists y \\ D & \mathbb{P}[y|x] < 1-d \;\forall y \end{cases}$

选哪个? 最高 a posterior probability 的那个.

SVD For linear regression

$X = UDV^T$ — $U \in \mathbb{R}^{n \times d}$ — $V \in \mathbb{R}^{d \times d}$

$X\hat{\beta}_{LS} = X(X^TX)^{-1}X^Ty = UDV^T(VDU^TUDV^T)^{-1}VDU^Ty$

$U^TU = I_{d \times d}$ — $= UDV^T(VD^2V^T)^{-1}VDU^Ty$

$VV^T = I_{d \times d}$ — $= UU^Ty$

Classification

GD: $a(k+1) = a(k) - \eta_k \sum \nabla$

SGD — 从 $\nabla(w_t, x_i, y_i)$ 随机

Newton optimization

$a(k+1) = a(k) - H^{-1}\nabla J$ — $H = \frac{\partial^2 J}{\partial a_i \partial a_j}$

Linear — $\mathrm{sign}(a^Tx)$ for binary classification

Perceptron Loss: $J_p(a) = \sum_{mis}(-a^Tz)$ — $-ya^Tx$

⟹ $a(k+1) = a(k) + y(k)\nabla J(a)$

Converges if data is separable.

· 不 maximize margin — 不看 分错的 mis

LDA 要 maximize 两个 class 在 DB 上 的距离

Fisher's Linear Discriminant Analysis

Maximize dist. of the means of the projected classes to find projection plane separating them best.

Projected mean for class α $\tilde{\mu}_\alpha = \frac{1}{n_\alpha}\sum w^Tx = w^T\mu_\alpha$

Dist. of project means for 2 classes

$|w^T(\mu_1 - \mu_2)|$

Fisher's Criterion: $J(w) = \frac{\|\tilde{\mu}_1 - \tilde{\mu}_2\|^2}{\tilde{\Sigma}_1 + \tilde{\Sigma}_2} = \frac{w^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^Tw}{w^T(\Sigma_1 + \Sigma_2)w}$

Nearest neighbour

选最近的 sample 的 label, 或根据 k-NN 的 K 值来取

LDA for multiple-class: One vs-rest

SVM:

Primal constrained:

$\min_w w^Tw + C\sum_{i=1}^n \xi_i$ st. $yw^Tx \geq 1 - \xi_i$, $\xi \geq 0$

Primal unconstraind

$\min_w w^Tw + C\sum_{i=1}^n \max(0, 1 - y \cdot w^Tx)$

Dual

$\max \sum\alpha_i - \frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j x_i^Tx_j$ st $0 \leq \alpha_i \leq C$

Lagrangian's Multiplier $\min f(w)$ st $g(w) = 0$ $i \leq m$

$L(w, \mu) = f(w) + \sum \mu g(w)$ ⟸ $\nabla L = 0$

General $\min f(w)$ s.t. $g(w) = 0$ $h(w) \leq 0$

$L(w, \mu, \alpha) = f(w) + \sum\mu_i g_i + \sum\alpha_i h_i$

Strong dual Dual 的 $\mathrm{Opt}(d) = \max L(\ )$

Kernel SVM

$\max_\alpha \sum\alpha_i - \frac{1}{2}\sum\alpha_i\alpha_j y_i y_j k(\ )$ st $0 \leq \alpha_i \leq C$

$y = \mathrm{sign}(\sum_{i=1}^n \alpha y_i k(\ ))$

Logistic

$P(y = y|x, w) = \mathrm{Ber}(y; \sigma(w^Tx)) = \frac{1}{1 + e^{-yw^Tx}}$

Multiclass Softmax

$P(y = i|x, w) = \frac{\exp(w_i^Tx)}{\sum \exp(w_j^Tx)} = p_i$

Multiclass SVM still a hard margin.

$\min_{w, \xi \geq 0} \frac{1}{2}w^Tw + C\sum\xi_i$ st. $\forall y_i \in Y$

$(w^Ty_i) - w_x(w^Ty) \geq 1 - \xi_i$

Structural SVM: $\hat{z} = \arg\max(w_z^Ty + w_{z,0})$

$\min \frac{1}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \xi_i$ $\xi_i \geq H_i$ $\forall i$

where $H(i) = \max_{y \in Y} L(y_i, y) - w^T(\phi(x_i, y_i) - \phi(x_i, y))$

XOR ⟹ $k(x_1, y_1) = (1 + x_1y_1 + x_2y_2)^2$

$\frac{\partial L}{\partial w} = w - \sum\alpha z_i y_i = 0$ — $\min \frac{1}{2}w^Tw + C\sum_{i=1}^n \xi_i$

$\frac{\partial L}{\partial \xi} = C - \alpha_i - \beta_i = 0$ — $z_i(w^Ty_i + w_0) \geq 1 - \xi_i$, $\xi_i \geq 0$

$\frac{\partial L}{\partial w_0} = -\sum\alpha_i z_i = 0$

$L(\ ) = \frac{1}{2}w^Tw + C\sum\xi_i - \sum\beta_i\xi_i - \sum\alpha_i(\ \cdots\ )$

$= \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum\sum\alpha_i\alpha_j z_i z_j y_i^T y_j + \sum_{i=1}^n (C - \alpha_i - \beta_i)\xi_i$

$= \sum\alpha_i - \frac{1}{2}\sum\sum\alpha_i\alpha_j z_i z_j y_i^T y_j$

Dual max $W(\alpha) =$ subject to $C \geq \alpha_i \geq 0$

$\alpha_i \leq C$ 是因为 $C - \alpha_i - \beta_i = 0$ $\beta_i \geq 0$

Primal for SSVM

$\min_w \frac{1}{2}w^Tw + C\sum_i^n \xi_i$

st. $w^T\psi(z_i, y_i) - w^T\psi(z, y_i) \geq \Delta(z, z_i) - \xi_i$ ($\alpha_{i,j}$)

$\forall i\ \forall z \neq z_i$

($\xi_i \geq 0$) β_i

Largian:

$L(\ ) = \frac{1}{2}w^Tw - \sum^n \alpha_i[z_i(w^Ty_i + w_0) - 1] -$

$\frac{\partial L}{\partial w} = 0$ $w = \sum\alpha_i z_i y_i$; $\frac{\partial L}{\partial w_0} = 0$ ⟹ $\sum\alpha_i z_i = 0$

Dual's problem:

$\max W(\alpha) = \sum^n\alpha_i - \sum\sum\alpha_i\alpha_j z_i z_j y_i^T y_j$

st $\forall i\ \alpha_i \geq 0$ and $\sum\alpha_i z_i = 0$

$w^* = \sum\alpha^* z_i y_i$; $w_0^* = -\frac{1}{2}(\min w^* y_i + \max w^* y_i)$

Soft margin

Primal: $\min_w \frac{1}{2}w^Tw + C\sum_{i=1}^n \xi_i$

st. $z_i(w^Ty_i + w_0) \geq 1 - \xi_i$, $\xi_i \geq 0$

$L(\alpha, \beta) = \frac{1}{2}w^Tw + C\sum^n\xi_i - \sum^n\alpha_i[z_i(w^Ty_i + w_0) - 1 + \xi_i] - \sum^n\beta_i\xi_i$

$\frac{\partial L}{\partial \xi_i} = 0$ ⟹ $C - \alpha_i - \beta_i = 0$

Dual is the same for this. But $C \geq \alpha_i \geq 0\ \forall i$

Multiclass SVM w/ slacks $\xi_i \geq 0$

$\min_{w, \xi} \frac{1}{2}w^Tw + C\sum_i\xi_i$ st. $(\ ) - \max(\ ) \geq 1 - \xi_i$ $\forall y_i$

Ensemble Method

Usually combine weak learner ($\varepsilon < 0.5$) to be great in classification.

Bagging: Train weak learners to bootstrap sets w/ equal weights

Boosting: All data. but the classifiers are re-weighted in a sequential manner for learners.

Decision Trees

Stumps: 垂直于 axis 的分割 $h(x) = \text{sign}(a x_i - t)$

Decision trees: Recursive trees of stumps.

Adaboost

最小化 min $f^*(x) = \arg\min \sum \exp(-y_i f(x_i))$

1. Initialize 所有权重为 $1/n$
2. Fit a weak classifier $c_b(x)$ $\quad \alpha_b = \log\left(\frac{1}{\varepsilon_b} - 1\right)$
3. Compute Error $\varepsilon_b = \sum w_i \mathbb{1}\{错\} / \sum w_i^{(b)}$
4. Update $w_i^{(b+1)} = w_i^{(b)} \exp\{\alpha_b \mathbb{1}\{错\}\}$

Return $\hat{C}(x) = \text{sign}(\sum^B \alpha_b C_b x)$

一定要是多个的 weak classifier 组

WINNOW

Performs better when many dimensions are irrelevant.

Search for 2 weight vector $a^+ a^-$ (for each class)

If a pt. is misclassified:

$a_i^+ \leftarrow \alpha^{+x_{ki}} a_i^+$;

$a_i^- \leftarrow \alpha^{-x_{ki}} a_i^-$ (if $y_k = +1$)

$a_i^+ \leftarrow \alpha^{-x_{ki}} a_i^+$;

$a_i^- \leftarrow \alpha^{+x_{ki}} a_i^-$ (if $y_k = -1$)

Exponential update of weights.

PSD for a matrix

$\det(A) \geq 0$ eigenvalue $\lambda_i \geq 0$

$x^T A x \geq 0 \ \forall x$

matrix derivative

$\frac{d a^T x}{dx} = \frac{d x^T a}{dx} = a \qquad \frac{\partial}{\partial x} a^T x b = a b^T \qquad \frac{\partial a^T x^T b}{\partial x} = b a^T$

$\frac{dK^{-1}}{dK} = -K^{-1} K' K^{-1} \qquad \frac{\partial x^T a x}{\partial a} = a^T(x + x^T)$

$\frac{d f(x)^T g(x)}{dx} = f(x)^T \frac{dg(x)}{dx} + g(x)^T \frac{df(x)}{dx}$

OLS

unbiasedness

$\text{Var}(\hat\beta) = \text{Var}((X^TX)^{-1}X^T(X\beta + \varepsilon)) = \text{Var}[(X^TX)^{-1}X^T\varepsilon]$
$= \sigma^2 (X^TX)^{-1} X^T X (X^TX)^{-1}$
$= \sigma^2 (X^TX)^{-1}$

sigmoid $\sigma(x) = \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{1+\exp(x)}$

$\nabla\sigma = \sigma(x)(1-\sigma(x))$

$\nabla \tanh x = 1 - \tanh^2(x)$

Some examples in VC dimension

Function	VC dim
Threshold $\mathbb{1}(-\infty, a)$	1
Intervals (a, b)	2
union of 2/3 intervals	4/6
half planes in $\mathbb{R}^2$	3
Convex polygons $\mathbb{R}^2$	∞
convex polygons in $\mathbb{R}^2$	$2k+1$

- ERM 不一定有效, Reactangle 有个例子
- 证明 Threshold function PAC learnable 的例子.
- Def Shattering. A Hypothesis class $\mathcal{H}$ shatters a finite set $C \subset \mathcal{X}$ if $|\mathcal{H}_C| = 2^{|C|}$
  - Proof of a threshold function class is 1
- Axis-Aligned Rectangles

$h(a_1, a_2, b_1, b_2) = \begin{cases} 1, & 在正方形内 \\ 0 & \text{otherwise} \end{cases}$

$VC\dim(\ ) = 4$. $\therefore$ 如果外面的点 (1,1,1,1,0) 则 Reactangles contains 点, 0 which is labelled as 0.

Argument for threshold functions

$\mathcal{H}$ threshold function in $\mathbb{R}$. $|\mathcal{H}_C|$ is the set of all functions from $\{C_1\} \mapsto \{0,1\}$ $\quad C = \{C_1\}$ $\mathcal{H}$ shatters $\{C_1\}$

$\Rightarrow \mathcal{C} = \{C_1, C_2\}$ $C_1 \leq C_2$ 没有 function $h \in \mathcal{H}$ 能把 $C_1 = 0, C_2 = 1$ 这个 label map 出来.

Since $h \in \mathcal{H}$, $h(C_1) = 0$ and $h(C_2) = 1 \Rightarrow$ Not all functions from $\{C_1, C_2\}$ to $\{0,1\}$, therefore $\mathcal{H}$ does NOT shatter. $VC = 1$

如果是 $VC < \infty$ 则一定 PAC-learnable.

If $VC\dim(\mathcal{H}) = d$, let $C \subset \mathcal{X}$, its effective size is only $O(|C|^d) \Rightarrow$ the size of $\mathcal{H}_C$ grows polynomially rather expo. in d

let $\tau_{\mathcal{H}}(m) = \max_{|C| = m} |\mathcal{H}_C|$, no. of different functions in $\mathcal{H}$ from $C \mapsto \{0,1\}$

Sauer's Thm: $VC\dim(\mathcal{H}) = d < \infty$

Then $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^d \binom{m}{i}$, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$

SVD for ridge

$\hat\beta = (X^TX + \lambda I)^{-1} X^T y = UD(D^2 + \lambda I)^{-1} D U^T y = \sum_{j=1} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y$

Optimal learning rate

$H = \frac{\partial J}{\partial a_i \partial a_j}$, using Taylor Expansion

$J(a(k+1)) \approx J(a(k)) - \eta(k) \nabla J^T \nabla J + \frac{1}{2}\eta^2(k) \nabla J^T H \nabla J$ (rate)

把这项 $= 0 \Rightarrow -\|\nabla J\|^2 + \eta \nabla J^T H \nabla J = 0$

就是梯度不再下降

$\eta_{opt} = \frac{\|\nabla J\|^2}{\nabla J^T H \nabla J}$ optimal

PAC learn

Efficient: 在 $n > m(\frac{1}{\varepsilon}, \frac{1}{\delta})$ 是有 poly. 的情况下

PAC learnable

$\mathbb{P}[\ R(\hat c) \leq \varepsilon\ ] \geq 1 - \delta$ $\quad$ For $n \geq m(\varepsilon, \delta)$, $\delta, \varepsilon \in (0,1)$

Classical Proof:

The PAC learnable of the threshold function

Efficient PAC learnable

If a algorithm A receive Z of size $n \geq \text{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta})$ then $\hat A$ such that $\mathbb{P}[R(\hat c) \leq \varepsilon] \geq 1 - \delta$

When $\hat R(\hat c) = 0$ and $|\mathcal{H}| < \infty$

$n > \frac{1}{\varepsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta})$, $\mathbb{P}[R(\hat c) \leq \varepsilon] \geq 1 - \delta$

Proof

$\mathbb{P}[\hat R(\hat c) = 0 \wedge R(\hat c) > \varepsilon] \leq \mathbb{P}[\exists h, \hat R(h) = 0 \wedge R(h) > \varepsilon]$

$\leq \sum_{h \in \mathcal{H}} \mathbb{P}[\hat R_n(h) = 0 \wedge R(h) > \varepsilon]$

$= \sum \mathbb{P}[\hat R_n(h) = 0 \mid R(h) > \varepsilon]\ \mathbb{P}[R(h) > \varepsilon]$

$\leq$ 一一 [illegible] $\leq \sum_h \prod_{j \leq n} \mathbb{P}[x_j \in h \Delta c \mid \mathbb{P}[h \Delta c]]$

$\leq |\mathcal{H}|(1-\varepsilon)^n$

Hoeffding's inequality

a. $\mathbb{P}[X \geq \varepsilon] \leq \frac{\mathbb{E}[X]}{\varepsilon}$ when $X \geq 0$

b. $\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$ $\quad X \in [a,b]$, $s > 0$

$\mathbb{P}[S_n - \mathbb{E}S_n \geq t] \leq \exp\left[-\frac{2t^2}{\sum_i^n (b_i - a_i)^2}\right]$

Hoeffding's Bound

$\mathbb{P}[\hat S_n - \mathbb{E}[\hat S_n] \geq \varepsilon] \leq \exp\left[\frac{-2n\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right]$

Union Bound: $\mathbb{P}[\max X_i > \varepsilon] \leq \sum \mathbb{P}[X_i > \varepsilon]$

ERM for hyperplanes: $\mathbb{P}[R(\hat c) - \inf R(c) > \varepsilon] \leq$

$(1 + 2\binom{n}{d}) e^{2d\varepsilon} e^{-\frac{n\varepsilon^2}{2}}$ if 当 $R(c) = 0$

$\mathbb{P}[R(\hat c) > \varepsilon] \leq 2\binom{n}{d} \exp[-\varepsilon(n-d)]$

Newton's Algo

(condition: choose $a(k+1)$ optimally to minimize the 2nd order Taylor expansion of $J(a(k+1))$)

$J(a(k+1)) \approx J(a(k)) + \nabla J^T (a(k+1) - a(k))$ $\quad$ J 是 general cost function

$+ \frac{1}{2}(a(k+1) - a(k))^T H (a(k+1) - a(k))$

chose a optimal weight vector $a(k+1)$!

$\frac{\partial}{\partial a(k+1)} J(a(k+1)) = \nabla J + H(a(k+1) - a(k)) = 0$

Rule $\Rightarrow a(k+1) = a(k) - H^{-1} \nabla J$

证明 [illegible] $= \frac{1}{2} w^T w$ 是 convex 要想一下

VC-dim

classical example. Threshold $\quad VC = 1$

Interval $C = 2$

要证明 a hypothesis class 有 VC dimension, 先证明 for $i = 1, \ldots, d$, $C \subset \mathcal{X}$, $|C| = d$ which is 可以包涵所有 都可以被 $\mathcal{H}$ 给 shattered 住 $\{c_1 \ldots c_d\} \to \{0,1\}^d$ 的组合, 然后再证明 $d+1$ 不可行.

Nonparametric Bayesian Method

having infinitely many parameters.

ADP, $DP(\alpha, H)$ is a dist. over probability space on a space Θ

α concentration parameter $\quad H$ is the base measure

For every partition of Θ and $GP(\alpha, H)$ $\alpha \uparrow$ 方差变小

$G(T_1) \ldots G(T_k) \sim \text{Dir}(\alpha H(T_1) \ldots \alpha H(T_k))$

Stick-Breaking Problem

Sampling $(p_1 \ldots p_k) \sim \text{Dir}(\alpha_1 \ldots \alpha_k)$ is the same as Beta

Fix $\alpha \Rightarrow \beta_i \sim \text{Beta}(1, \alpha)$

$\beta_1 \sim \text{Beta}(1, \alpha)$ $\beta_2 \sim \text{Beta}(\frac{1}{2}, \alpha)$ $p_2 = \beta_2(1 - \beta_1)$

$\beta_k \sim \text{Beta}(1 - \alpha)$ $p_k = \beta_k(1 - \sum p_i)$

Chinese Restaurant Problem

对于新来的顾客

p_C $n+1$ join table $|P\rangle = \begin{cases} \frac{|T|}{\alpha + n} \\ \frac{\alpha}{\alpha + n} \end{cases}$

第一个新的为 $\frac{\alpha}{\alpha}$.

Gibbs sampling for cluster assignment

先确定有 k 个 cluster (k 非常容易) 然后判断数据属于哪个 cluster.

$p(z_i = k \mid z_{-i}, x, \alpha, H) \propto p(z_i = k \mid z_{-i}, \alpha, \mu)\ p(x \mid z_i = k, z_{-i}, \alpha, \mu)$

$\propto p(z_i = k \mid z_{-i}, \alpha)\ p(x_i \mid x_{-i}, z_i = k, z_{-i}, \mu)$.

$\propto \underbrace{p(z_i = k \mid z_{-i}, \alpha)}_{\text{prior}}\ \underbrace{p(x_i \mid x_{-i}, z_i = k, z_{-i}, \mu)}_{\text{likelihood}}$ $\quad p(x_i \mid z_i, \mu)$

If H is normal

$p(z_i = k \mid z_{-i}, x, \alpha, \mu) = \begin{cases} \frac{N_{k,-i}}{\alpha + N - 1} p(x_i \mid x_{-i,k}, \mu) \\ \frac{\alpha}{\alpha + N - 1} p(x_i \mid \mu) & \text{otherwise 新的 cluster} \end{cases}$

Pseudo-code:

```
for i=1 to N in random order do
  Remove xi's sufficient statistics from old cluster zi;
  for k=1 to K do
    compute Pk(xi) = Pk(xi | x-i,k)
    set Nk,-i = |x-i,k|
    compute p(zi=k | z-i, X) = Nk,-i / (α+N-1)
  end for
  Compute P*(xi) = p(xi | μ)
  Compute p(zi = * | z-i, x)
  Normalize p(zi | ·);
  Sample zi ~ p(zi | ·)
  Add xi's sufficient statistics to new cluster zi
  If any cluster is empty, remove it and decrease K
End For.
```