NO
DATE

(1.1) · 算法在训练数据D上训练后的预测结果 $f(\vec{x};D)$

· 算法在足够多数据上训练后的预测结果 $\bar{f}(\vec{x}) = \mathbb{E}_D[f(\vec{x};D)]$

· 数据集上的标签 y_D

· 真实标签 $y = \mathbb{E}_D[y_D]$

$$f(\vec{x};D) \overset{\text{~~bias~~ var.}}{\sim} \bar{f}(\vec{x}) \overset{\text{var.}}{\sim} y_D \overset{\text{Noise}}{\sim} y_D$$

variance bias noise

$$\mathbb{E}_D[(f(\vec{x};D)-y_D)^2]$$
$$= \mathbb{E}[(f(\vec{x};D)-\bar{f}(\vec{x}))^2] + \mathbb{E}[(\bar{f}(\vec{x})-y_D)^2] + 2\mathbb{E}[(f(\vec{x};D)-\bar{f}(\vec{x}))(\bar{f}(\vec{x})-y_D)]$$
$$= \mathbb{E}[(f(\vec{x};D)-\bar{f}(\vec{x}))^2] + \mathbb{E}[(\bar{f}(\vec{x})-y)^2] + \mathbb{E}[(y-y_D)^2] + 2\mathbb{E}[(y-y_D)(\bar{f}(\vec{x})-y)]$$
$$= \mathbb{E}[(f(\vec{x};D)-\bar{f}(\vec{x}))^2] + \mathbb{E}[(\bar{f}(\vec{x})-y)^2] + \mathbb{E}[(y-y_D)^2]$$
$$= var(\vec{x}) + bias^2(\vec{x}) + noise$$

(1.2) k-fold 中 k 越大

· 每次训练的样本越接近最后用于预测的样本，因此 bias 越小

· 重复次数更多，取平均，variance 更小

(1.3) 100

· 坏处在于计算量巨大.

· 好处在于和最后用于预测的模型接近，准确性高（因为）

(2.1) $E = \sum_{i=1}^{m}(y_i - wx_i - b)^2$

$$\frac{\partial E}{\partial w} = \sum_{i=1}^{m} 2(y_i - wx_i - b)(-x_i)$$
$$= 2\left(w\sum_{i=1}^{m}x_i^2 - \sum_{i=1}^{m}(y_i-b)x_i\right) = 0$$
$$\Rightarrow w\sum_{i=1}^{m}x_i^2 = \sum_{i=1}^{m}(y_i-b)x_i$$
$$\frac{\partial E}{\partial b} = \sum_{i=1}^{m} 2(y_i - wx_i - b)(-1) = 0$$
$$\Rightarrow b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i)$$
$$\Rightarrow w\left(\sum x_i^2 - \frac{1}{m}\sum_{ij}x_ix_j\right) = \sum y_ix_i - \sum_i y_i\left(\frac{1}{m}\sum_j x_j\right)$$
$$\Rightarrow w = \frac{\sum y_i(x_i - \bar{x})}{\sum x_i^2 - (\sum x_i)^2}$$
$$\begin{cases} w = \dfrac{\sum y_i(x_i-\bar{x})}{\sum x_i^2 - (\sum x_i)^2} \\ b = \frac{1}{m}\sum_i(y_i - wx_i). \end{cases}$$

(2.2) $E = (y - Wx)^T(y - Wx)$

$$= (y^T - x^TW^T)(y - Wx)$$
$$= y^Ty - y^TWx - x^TW^Ty + x^TW^TWx$$
$$\frac{\partial E}{\partial W} = -X^Ty - X^Ty + X^TXW + X^TXW = 0$$
$$\Rightarrow X^TXW = X^Ty$$
$$\Rightarrow W = (X^TX)^{-1}X^Ty \quad (X^TX^{-1})\text{ 满秩时}$$

(2.3) $\frac{\partial E}{\partial W} = 2(X^TXW - X^Ty)$

$$\frac{\partial E}{\partial W^2} = 2X^TX$$
$$\Rightarrow W \leftarrow W - \left(\frac{\partial^2 E}{\partial W^2}\right)^{-1}\left(\frac{\partial E}{\partial W}\right)$$
$$= W - (X^TX)^{-1}(X^TXW - X^Ty)$$
$$= W - W + (X^TX)^{-1}X^Ty$$
$$= (X^TX)^{-1}X^Ty$$

由 扫描全能王 扫描创建

NO
DATE

(2.4)

$$P(y_i|w,x_i)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-wx)^2}{2\sigma^2}\right)$$

$$\log P(y_i|w,x_i)=-\frac{(y_i-wx_i)^2}{2\sigma^2}+\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)$$

$$\Rightarrow L(w)=\sum_{i=1}^m \log P(y_i|w,x_i)$$

$$=\sum_{i=1}^m -\frac{(y_i-wx_i)^2}{2\sigma^2}+m\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)$$

$$\frac{\partial L(w)}{\partial w}=\sum_{i=1}^m \frac{2(y_i-wx_i)x_i}{2\sigma^2}=0$$

$$\Rightarrow w=\frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

(与线性回归结果一致).

(3.1)

$$P(y_i|w,x_i)=\left(\frac{e^{wx_i}}{1+e^{wx_i}}\right)^{y_i}+\left(\frac{1}{1+e^{wx_i}}\right)^{1-y_i}$$

$$\log P(y_i|w,x_i)=y_i\log\left(\frac{e^{wx_i}}{1+e^{wx_i}}\right)-(1-y_i)\log(1+e^{wx_i})$$

$$L(w)=\sum_{i=1}^m \log P(y_i|w,x_i)$$

(3.2)

$$\frac{\partial L(w)}{\partial w}=\sum_{i=1}^m y_i\frac{1+e^{wx_i}}{e^{wx_i}}\cdot\frac{x_ie^{wx_i}(1+e^{wx_i})}{(1+e^{wx_i})^2}-\sum_{i=1}^m(1-y_i)\frac{1}{1+e^{wx_i}}e^{wx_i}x_i$$

$$=\sum_{i=1}^m\frac{x_iy_i}{1+e^{wx_i}}-\sum_{i=1}^m\frac{(1-y_i)x_ie^{wx_i}}{1+e^{wx_i}}$$

$$=0$$

$$\cancel{\Rightarrow\sum_{i=1}^m x_i(1-y_i)e^{wx_i}=\sum_{i=1}^m x_iy_i}\ \times$$

令 $g(x)=\frac{1}{1+e^x}$

$$L(w)=\sum_{i=1}^m y_i\log(1-g(wx_i))+(1-y_i)\log(g(wx_i))$$

$$\frac{\partial L(w)}{\partial w}=\sum_{i=1}^m\left[y_i\frac{(-1)}{1-g(wx_i)}+(1-y_i)\frac{1}{g(wx_i)}\right]\frac{\partial g(wx_i)}{\partial w}$$

$$=\sum_{i=1}^m\frac{1-g(wx_i)-y_i}{g(wx_i)(1-g(wx_i))}g'(wx_i)(1-g(wx_i))x_i$$

$$=\sum_{i=1}^m\left(\frac{e^{wx_i}}{1+e^{wx_i}}-y_i\right)x_i$$

梯度下降

$$w\leftarrow w-\alpha\sum_{i=1}^m\left(\frac{e^{w^Tx_i}}{1+e^{w^Tx_i}}-y_i\right)x_i$$

(3.3) (4)

信息增益公式

$$H(D)=-\sum_{k=1}^{|y|}p_k\log p_k$$

$$H^{split}(D)=\sum_i\frac{|D_i|}{|D|}H(D_i)$$

增益 $=H^{split}(D)-H(D)$.

(4.2) additive model.

$$\hat y^{(t)}=\hat y^{(t-1)}+f_t(x_i)$$

$$obj^{(t)}=\sum_{i=1}^m l(y_i,\hat y^{(t-1)}+f_t(x_i))+\Omega(f_t)$$

$$\simeq\sum_{i=1}^m -2y_i(\hat y^{(t-1)}+f_t(x_i))+(\hat y^{(t-1)}+f_t(x_i))^2+\Omega$$

$$\simeq\sum_{i=1}^m\left(2(\hat y^{(t-1)}-y_i)f_t(x_i)+f_t^2(x_i)\right)+\Omega(f_t)$$

when we consider MSE

$$\simeq\sum_{i=1}^m\left(l(y_i,\hat y^{(t-1)})+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\right)+\Omega$$

in general form

$$obj^{(t)}\simeq\sum_{i=1}^m\left(g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\right)+\gamma T+\frac{\lambda}{2}\sum_{j=1}^T w_j^2$$

$$=\sum_{i=1}^m\left(g_iw_{q(x_i)}+\frac{1}{2}h_iw_{q(x_i)}^2\right)+\gamma T+\frac{\lambda}{2}\sum_{j=1}^T w_j^2$$

$$=\sum_{j=1}^T\left(\sum_{i\in I_j}g_i\right)w_j+\frac{1}{2}\left(\left(\sum_{i\in I_j}h_i\right)+\lambda\right)w_j^2+\gamma T$$

$$=\sum_{j=1}^T\left[\left(\sum_{i\in I_j}g_i\right)w_j+\frac{1}{2}\left[\left(\sum_{i\in I_j}h_i\right)+\lambda\right]w_j^2\right]+\gamma T$$

$$=\sum_{j=1}^T\left[G_jw_j+\frac{1}{2}(H_j+\lambda)w_j^2\right]+\gamma T$$

$$w^*=-\frac{G_j}{H_j+\lambda}\qquad obj^*=-\frac{1}{2}\sum_{j=1}^T\frac{G_j^2}{H_j+\lambda}$$

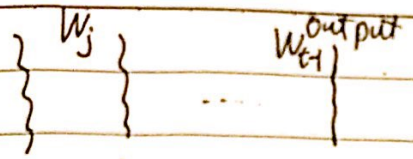
split 增益

$$gain=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma$$

由 扫描全能王 扫描创建

NO
DATE

⑤



层 j $j+1$ t $\hat{y}^{(i)} = x_t^{(i)}$

维度 d_j d_{j+1} d_t

$$E = \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

$$\frac{\partial E}{\partial W_j} = \sum_{i=1}^{m} \frac{\partial E}{\partial \hat{y}_{(i)}} \cdot \frac{\partial \hat{y}^{(i)}}{\partial x_{j+1}^{(i)}} \circ \frac{\partial x_{j+1}^{(i)}}{\partial W_j x_j} \otimes \frac{\partial W_j x_j}{\partial W_j}$$

其中 $\frac{\partial \hat{y}^{(i)}}{\partial x_{j+1}^{(i)}} = \prod_{s=j+1}^{s-1} \frac{\partial x_{s+1}^{(i)}}{\partial x_s^{(i)}}$

$$= \prod_{s=j+1}^{s-1} \left[\left(x_{s+1}^{(i)} \circ (1 - x_{s+1}^{(i)}) \right) \otimes I \right] \circ W_s$$ (逆序).

$$\Rightarrow \frac{\partial E}{\partial W_j} = \sum_{i=1}^{m} 2(\hat{y}^{(i)} - y^{(i)}) \cdot \frac{\partial \hat{y}^{(i)}}{\partial x_{j+1}^{(i)}} \circ x_{j+1}^{(i)} \circ (1 - x_{j+1}^{(i)}) \otimes x_j$$

其中 $\cdot$ 为矩阵乘法, $\prod$ 为矩阵连乘

$\circ$ 为 element wise product

$\otimes$ 为 Cartesian product.

⑥.1 原问题: $\min \frac{1}{2}\|w\|^2$

s.t. $y_i(w^T x_i + b) \geq 1 \quad \forall i \in [m]$

拉格朗日函数

$$L(w, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i (w^T x_i + b) - 1 \right)$$

KKT条件:

$$\frac{\partial L}{\partial w} = \vec{w} - \sum_i \alpha_i y_i \vec{x}_i = 0$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{m} \alpha_i y_i$$

$$\Rightarrow L(w, \alpha) = -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j - b \sum_{i=1}^{m} \alpha_i y_i + \sum_{i=1}^{m} \alpha_i$$

$\Rightarrow$ 对偶问题

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j$$

s.t. $\alpha_i \geq 0 \quad \sum_i \alpha_i y_i = 0 \quad \forall i \in [m]$

⑥.2 原问题 $\min \frac{1}{2}\|w\|^2 + C \sum \xi_i$

s.t. $y_i (w^T x_i + b) \geq 1 - \xi_i$

$\xi_i \geq 0$

$$L(\alpha, \beta, w, b, \xi) = \frac{1}{2}\|w\|^2 + C \sum \xi_i - \sum \beta_i \xi_i - \sum \alpha_i \left(y_i (w^T x_i + b) - 1 + \xi_i \right)$$

KKT condition:

$$\frac{\partial L}{\partial w} = \vec{w} - \sum \alpha_i y_i \vec{x}_i = 0$$

$$\frac{\partial L}{\partial b} = \sum \alpha_i y_i = 0$$

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

$\Rightarrow$ 对偶问题:

$$\max_\alpha -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j + \sum \alpha_i$$

s.t. $0 \leq \alpha_i \leq C$

$\sum \alpha_i y_i = 0$

当 $\alpha_i = 0$ 时 $\xi_i = 0$ 不是support vec.

当 $0 < \alpha_i < C$ 时 $\xi_i = 0$ 是support vec. 在margin上.

当 $\alpha_i = C$, $\xi_i > 0$ & $\xi_i \leq 1$ 时 是support vec 在margin内, 但划分依然正确

当 $\alpha_i = C$, $\xi_i > 1$ 时, 是support vec. 处于错误划分的一侧.

由 扫描全能王 扫描创建

NO
DATE

(6.3) SMO. 考虑每次变动2个α. α_1, α_2

约束 $\sum \alpha_i y_i = 0 \Rightarrow \alpha_1 y_1 + \alpha_2 y_2 = \xi$

$\Rightarrow \alpha_1 = (\xi - \alpha_2 y_2)/y_1 = y_1(\xi - \alpha_2 y_2)$

目标 $W(\alpha_1, \alpha_2, \alpha_3 \cdots)$

$= W((\xi - \alpha_2 y_2) y_1, \alpha_2, \alpha_3 \cdots)$

$= A\alpha_2^2 + B\alpha_2 + C.$

$\Rightarrow$ find minimizer α_2^*,

考虑约束:

$$\begin{cases} 0 \le y_1(\xi - \alpha_2 y_2) \le C \\ 0 \le \alpha_2 \le C \end{cases} \Rightarrow \alpha_2 \in [L, H]$$

$\Rightarrow \alpha_2^{*new} = clip(\alpha_2^*, [L, H]).$

$\alpha_1^{*new} = y_1(\xi - \alpha_2^{*new} y_2)$

(6.4) simply change $\vec{x}_1 \cdot \vec{x}_2 \to K(\vec{x}_1, \vec{x}_2)$

$f(\vec{x}) = \sum \alpha_i y_i K(\vec{x}_i, \vec{x}) + b$

Ⓐ in some cases it can be easily written as $K(\vec{x}_1, \vec{x}_2) = \phi(\vec{x}_1) \cdot \phi(\vec{x}_2)$

Ⓑ $\forall x_i \quad i \in [n]$

$$K = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & & \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{bmatrix}$$

is a 半正定矩阵.

(7.1) 参见2014年题目 9 & 11题.

(7.2) full Bayesian

training: $P(\theta|S) = \frac{P(S|\theta)P(\theta)}{P(S)}$

$\propto P(\theta) \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta)$

testing: $P(y|x, S) = \int_\theta p(y|x, \theta) p(\theta|S) d\theta$

full Bayesian method is computational expensive.

in contrast, MAP takes

$\theta = \arg\max_\theta \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta) p(\theta)$

MLE takes

$\theta = \arg\max_\theta \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta)$

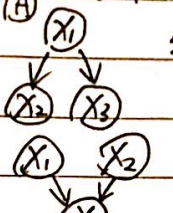
to avoid integration in prediction step.

(8.1) 可以写出联合概率分布,然后求解.

$P_B(x_1, \cdots x_d) = \prod_{t=1}^{d} P_B(x_i | \pi_i)$

π_i 表示结点 x_i 的父结点

(8.2) Ⓐ 几个基本结论.



给定 x_1, x_2, x_3 的结点

$x_2 \perp x_3 | x_1$

$x_1 \perp\!\!\!\perp x_2$

给定 x_3 时, x_1, x_2 不独立

$x \to y \to z \quad x \perp z | y.$

Ⓑ 写出公式,观察 $p(x, y) \overset{?}{=} p(x)p(y)$

Ⓒ 把V结构的父结点连边.

有向边变为无向边

给定的 r.v. 从图中去掉,如果剩下的两个变量分属不同连通子图则 $x_1, x_2 | x$.

(8.3) 随机产生一个与证据变量E一致的情形

每次循环选取一个除证据变量以外的变量

使用其 Markov blank 中其它变量的取值

计算 $P(Q|MB)$, 依此概率采样得到新值

$P(Q=q|E=e) \simeq \frac{n_q}{T}$ ⟶ $Q=q$ 的步数.

↳ 总采样次数

由 扫描全能王 扫描创建

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

NO

DATE

(9.1) $J(c,\mu)=\sum_{i=1}^N \|x^{(i)}-\mu_{c^{(i)}}\|^2$

E-step is to cal.

$$c^{(i)}=\arg\min_c J(c,\mu)$$
$$=\arg\min_c \|x^{(i)}-\mu_c\|^2 \quad ①$$

M-step is to cal.

$$\mu_j=\arg\min_{\mu_j} J(c,\mu).$$

$$\frac{\partial J}{\partial \mu_j}=\sum_{i=1}^N 2(x^{(i)}-\mu_j)\,\mathbb{I}(c^{(i)}=\mu_j)=0$$

$$\Rightarrow \mu_j=\frac{1}{N}\sum_{i=1}^N \mathbb{I}(c^{(i)}=\mu_j)x^{(i)} \quad ②$$

① and ② are essentially EM algorithm for k-means

(9.2) log-likelihood.

$$\mathcal{L}=\log p(D|\pi,\mu,\Sigma)$$
$$=\sum_{i=1}^N \log p(x^{(i)}|\pi,\mu,\Sigma)$$
$$=\sum_{i=1}^N \log\left(\sum_{k=1}^K \pi_k \mathcal{N}(x^{(i)}|\mu_k,\Sigma_k)\right)$$

(A) $$\frac{\partial \mathcal{L}}{\partial \mu_k}=\sum_{i=1}^N \frac{\pi_k \mathcal{N}(x^{(i)}|\mu_k,\Sigma_k)}{\sum_{k=1}^K \pi_k \mathcal{N}(x^{(i)}|\mu_k,\Sigma_k)}\Sigma_k^{-1}(x^{(i)}-\mu_k)$$

$$=\sum_{i=1}^N \gamma(z_k^{(i)})\Sigma_k^{-1}(x^{(i)}-\mu_k)=0$$

$$\Rightarrow \mu_k=\frac{1}{\sum_{i=1}^N \gamma_{ik}}\sum_{i=1}^N \gamma_{ik}x^{(i)}$$

(B) $$\frac{\partial \mathcal{L}}{\partial \Sigma_k}=\sum_{i=1}^N \frac{\pi_k \mathcal{N}(x^{(i)}|\mu_k,\Sigma_k)}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(x^{(i)}|\mu_{k'},\Sigma_{k'})}\left(-\frac{1}{2}\frac{1}{|\Sigma|^2}\right)\left(\Sigma-\frac{(x-\mu)(x-\mu)^T}{\cancel{\Sigma}}\right)=0$$

$$\Rightarrow \Sigma_k=\frac{1}{\sum_{i=1}^N \gamma_{ik}}\sum_{i=1}^N \gamma_{ik}(x-\mu)(x-\mu)^T$$

(C) consider $\sum_{k=1}^K \pi_k=1$

$$L=\mathcal{L}(\mu,\Sigma,\pi)+\lambda\left(\sum_{k=1}^K \pi_k-1\right)$$

$$\frac{\partial L}{\partial \pi_k}=\sum_{i=1}^N \frac{\mathcal{N}(x^{(i)}|\mu_k,\Sigma_k)}{\sum_{k=1}^K \pi_k \mathcal{N}(x^{(i)}|\mu_k,\Sigma_k)}+\lambda=0$$

$$\Rightarrow \lambda=-\sum_{i=1}^N \gamma_{ik} \quad \forall k\in[K]$$

$$\Rightarrow \pi_k=\frac{1}{N}\sum_{i=1}^N \mathcal{N}(x^{(i)}|\mu_k,\Sigma_k)$$

(D) finally E step

$$\gamma_{ik}=\frac{\pi_k \mathcal{N}(x^{(i)}|\mu_k,\Sigma_k)}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(x^{(i)}|\mu_{k'},\Sigma_{k'})}$$

in M step.

$$\begin{cases}\mu_k=\frac{1}{\sum_{i=1}^N \gamma_{ik}}\sum_{i=1}^N \gamma_{ik}x^{(i)}\\ \Sigma_k=\frac{1}{\sum_{i=1}^N \gamma_{ik}}\sum_{i=1}^N \gamma_{ik}(x^{(i)}-\mu)(x^{(i)}-\mu)^T\\ \pi_k=\frac{1}{N}\sum_{i=1}^N \mathcal{N}(x^{(i)}|\mu_k,\Sigma_k)\end{cases}$$

(E) 当 GMM 中 $\Sigma=\sigma I$ $\sigma\to 0$ 时.

$$\gamma_{ik}=\frac{\pi_k \mathcal{N}(x^{(i)}|\mu_k,\Sigma_k)}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(x^{(i)}|\mu_{k'},\Sigma_{k'})}$$

$$=\frac{\pi_k \exp\left(-\frac{1}{2\sigma}\|x^{(i)}-\mu_k\|_2^2\right)}{\sum_{k'=1}^K \pi_{k'}\exp\left(-\frac{1}{2\sigma}\|x^{(i)}-\mu_{k'}\|_2^2\right)}$$

$$=\mathbb{I}\left(k=\arg\min_{k'}\|x^{(i)}-\mu_{k'}\|_2^2\right)$$

GMM $\Rightarrow$ k-means.

(9.3) 对一般问题:

$$\log(D|\Theta)=\sum_{i=1}^m \log\left(\sum_z p(x^{(i)},z|\Theta)\right)$$
$$=\sum_{i=1}^m \log\left(\sum_z q(z)\frac{p(x^{(i)},z|\Theta)}{q(z)}\right)$$
$$=\sum_{i=1}^m \log \mathbb{E}_{q(z)}\left(\frac{p(x^{(i)},z|\Theta)}{q(z)}\right)$$
$$\geq \sum_{i=1}^m \mathbb{E}_{q(z)}\log\left(\frac{p(x^{(i)},z|\Theta)}{q(z)}\right) \quad \text{(Jesen's)}$$
$$=\sum_{i=1}^m \sum_z q(z)\log\left(\frac{p(x^{(i)},z|\Theta)}{q(z)}\right)\cancel{\equiv \mathcal{L}(\Theta)}$$
$$\equiv \mathcal{L}(q(z)|\Theta)$$

由 扫描全能王 扫描创建

回忆

$$\mathcal{L}(\Theta|q(z)) = \sum_i \left(\sum_z q(z)\log P(x^{(i)},z|\Theta) - q(z)\log q(z)\right)$$

$$= \sum_i \left(\sum_z q(z)\log\frac{P(x^{(i)},z|\Theta)}{P(x^{(i)}|\Theta)} + \sum_z q(z)\log P(x^{(i)}|\Theta) - \sum_z q(z)\log q(z)\right)$$

$$= \sum_i \log P(x^{(i)}|\Theta) + \sum_i\sum_z q(z)\log P(z|x^{(i)},\Theta) - \sum_i\sum_z q(z)\log q(z)$$

$$= \log P(D|\Theta) + \sum_z q(z)\log P(z|D,\Theta) - \sum_z q(z)\log q(z)$$

$$= \log P(D|\Theta) + \sum_z q(z)\log\frac{P(z|D,\Theta)}{q(z)}$$

$$= \log P(D|\Theta) + KL(q(z)\,\|\,P(z|D))$$

i.e. $\log P(D|\Theta) = \mathcal{L}(q(z)|\Theta) + KL(q(z)\|P(z|D))$

当 $q(z)=P(z|D)$ 时等号成立.

$$\Rightarrow \begin{cases} \text{E step}: & q(z)=P(z|D) \\ \text{M step}: & \Theta = \arg\max_\Theta \sum_{i=1}^m \sum_z q(z)\log\frac{P(x^{(i)},z|\Theta)}{q(z)} \end{cases}$$

(9.4) 证明(9.6)即证 $\mathcal{L}(\Theta^{(t+1)}|q^{(t+1)}(z))$ 不小于 $\mathcal{L}(\Theta^{(t)}|q^{(t)}(z))$

$$\mathcal{L}(\Theta^{(t+1)}|q^{(t+1)}(z)) = \sum_{i=1}^m\sum_z q^{(t+1)}(z)\log\frac{P(x^{(i)},z|\Theta^{(t+1)})}{q^{(t+1)}(z)}$$

$$\geq \sum_i\sum_z q^{(t)}(z)\log\frac{P(x^{(i)},z|\Theta^{(t+1)})}{q^{(t)}(z)} \quad \text{(9.3 结论)}$$

$$\geq \sum_i\sum_z q^{(t)}(z)\log\frac{P(x^{(i)},z|\Theta^{(t)})}{q^{(t)}(z)} \quad (\arg\max)$$

$$= \mathcal{L}(\Theta^{(t)}|q^{(t)}(z))$$

(10.1) 假设每个分类器的正确率为 p, 且相互独立, 由 Heoffding ineq.

$P(H(n)\leq k) = \sum_{i=0}^k C_n^i p^i (1-p)^{n-i}$

为 majority vote 的错误率.

$P(H(n)\leq k) < \exp(-2(p-\frac{1}{2})^2 n)$

因此 majority vote 的错误率随分类器个数指数下降.

(10.2) (A) additive model is model of the form

$f(x) = \sum_{m=1}^M \beta_m G_m(x)$

i.e. the weighted sum of several base models

(B) in each step, we choose β_m, G_m to minimize

$$L(\beta_m, G_m) = \mathbb{E}_{(x,y)\sim D}\left[\exp(-y_i(f_{m-1}(x_i)+\beta_m G_m(x_i)))\right]$$

$$L(\beta_m, G_m) = \sum_{i=1}^N \exp(-y_i f_{m-1}(x_i))\exp(-y_i\beta_m G_m(x_i))$$

令 $w_i^{(m)} = \exp(-y_i f_{m-1}(x_i))$

$$L = \sum_{i=1}^N w_i^{(m)}\exp(-y_i\beta_m G_m(x_i))$$

考虑 $y_i, G_m(x_i) \in \{+1,-1\}$.

$$L = \sum_{i:y_i=G_m(x_i)} w_i^{(m)} e^{-\beta_m} + \sum_{i:y_i\neq G_m(x_i)} w_i^{(m)} e^{+\beta_m}$$

$$= \sum_{i=1}^N w_i^{(m)} e^{-\beta_m} + \sum_{i=1}^N w_i^{(m)}(e^{+\beta_m}-e^{-\beta_m})\mathbb{I}(y_i\neq G_m(x_i))$$

$$G_m^* = \arg\min \sum_{i=1}^N w_i^{(m)}\mathbb{I}(y_i\neq G_m(x_i))$$

i.e. a classifier trained with weight $w_i^{(m)}$

$$\frac{\partial L}{\partial\beta_m} = \sum_{i=1}^N -w_i^{(m)}e^{-\beta_m} + \sum_{i=1}^N w_i^{(m)}(e^{+\beta_m}+e^{-\beta_m})\mathbb{I}(\neq) = 0$$

由 扫描全能王 扫描创建

NO
DATE

$e^{-\beta_m}\sum_i w_i^{(m)} = (e^{-\beta_m}+e^{+\beta_m})\sum_i w_i^{(m)} \mathbb{I}(y_i \neq G_m)$

$\Rightarrow \beta_m^* = \frac{1}{2}\log\left(\frac{\sum w_i^{(m)}}{\sum w_i^{(m)}\mathbb{I}(\neq)} - 1\right)$

let $err_m = \frac{\sum \mathbb{I}(\neq) w_i^{(m)}}{\sum w_i^{(m)}}$

$\Rightarrow \beta_m^* = \frac{1}{2}\log\left(\frac{1-err_m}{err_m}\right)$

权重更新 $w_i^{(m+1)}$

$w_i^{(m+1)} = w_i^{(m)} \exp(-y_i \beta_m G_m(x_i))$

其中 $\exp(-y_i\beta_m G_m(x_i))$

$= \mathbb{I}(y_i = G_m(x_i))e^{-\beta_m} + \mathbb{I}(y_i \neq G_m(x_i))e^{+\beta_m}$

$= e^{-\beta_m}[\mathbb{I}(=) + \mathbb{I}(\neq)e^{+2\beta_m}]$

各样本权重都scale常数不影响结果.

$\Rightarrow w_i^{(m+1)} = w_i^{(m)} \exp(\alpha_m \mathbb{I}(y_i \neq G_m(x_i)))$

其中 $\alpha_m = 2\beta_m$

总结:

$$\begin{cases} G_m^* = \arg\min \sum_{i=1}^{N} w_i^{(m)} \mathbb{I}(y_i \neq G_m(x_i)) \\ \alpha_m = \log\left(\frac{1-err_m}{err_m}\right) \\ w_i^{(m+1)} = w_i^{(m)} \exp(\alpha_m \mathbb{I}(y_i \neq G_m(x_i))) \end{cases}$$

(10.3) 假设单个错误率 $\varepsilon_i = f_i - h$

f_i: 预测结果 h: 真实结果

假设 $\begin{cases} \text{独立 } \mathbb{E}[\varepsilon_i \varepsilon_j] = \sigma^2 \delta_{ij} \\ \text{无偏 } \mathbb{E}[\varepsilon_i] = 0 \end{cases}$

平均偏方误差

$E_{avg} = \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}[\varepsilon_i^2] = \sigma^2$

bagging 平方误差

$E_{bag} = \mathbb{E}\left[\left(\frac{1}{M}\sum_{i=1}^{M}\varepsilon_i\right)^2\right]$

$= \frac{1}{M}\sigma^2$

$\Rightarrow$ bagging 能够减小 variance

random forest:

repeat

- randomly sample from D (bootstrap)
- randomly sample features
- use sampled D & features to grow tree

(11) 参考14年 13题.

(12) KNN 的错误率

$P(err) = 1 - \sum_{c\in\mathcal{Y}} P(c|x)P(c|z)$ ← 最近邻

贝叶斯分类器

$c^* = \arg\max_{c\in\mathcal{Y}} P(c|x)$

错误率 $1 - P(c^*|x)$

$P(err) = 1 - \sum_{c\in\mathcal{Y}} P(c|x)P(c|z)$

$\simeq 1 - \sum_{c\in\mathcal{Y}} P(c|x)^2$ $\quad x$ close to z

$\leq 1 - P^2(c^*|x)$

$= (1 - P(c^*|x))(1 + P(c^*|x))$

$\leq (1 - P(c^*|x)) \times 2$

(13.1) 使用狄拉克符号.

一个样本点 $|x_n\rangle$ 一组正交单位基 $\{|\mu_i\rangle\}$.

$\langle\mu_i|\mu_j\rangle = \delta_{ij}$

$|x_n\rangle = \sum_{i=1}^{p} |\mu_i\rangle\langle\mu_i|x_n\rangle$

$= \left(\sum_{i=1}^{d} |\mu_i\rangle\langle\mu_i| + \sum_{i=d+1}^{p} |\mu_i\rangle\langle\mu_i|\right)|x_n\rangle$

$\simeq \sum_{i=1}^{d} |\mu_i\rangle\langle\mu_i|x_n\rangle + \sum_{i=d+1}^{p} |\mu_i\rangle\langle\mu_i|\bar{x}\rangle$

$= |x_n^*\rangle$

重构误差 $= J = \frac{1}{N}\sum_{n=1}^{N} \| |x_n\rangle - |x_n^*\rangle \|^2$

$= \frac{1}{N}\sum_{n=1}^{N}\sum_{i=d+1}^{p} \| (\langle\mu_i|x_n\rangle - \langle\mu_i|\bar{x}\rangle)|\mu_i\rangle \|^2$

$= \frac{1}{N}\sum_{n=1}^{N}\sum_{i=d+1}^{p} (\langle\mu_i|x_n\rangle - \langle\mu_i|\bar{x}\rangle)^2$

低维空间为 d 维.

由 扫描全能王 扫描创建

NO

DATE

$= \frac{1}{N}\sum_{i=d+1}^{p} \langle \mu_i | \sum_{n=1}^{N} (|x_n\rangle - |\bar{x}\rangle)(\langle x_n| - \langle \bar{x}|) | \mu_i \rangle$

$= \frac{1}{N}\sum_{i=d+1}^{p} \langle \mu_i | S | \mu_i \rangle$

其中 S 为 covariance matrix

$\Rightarrow \min \sum_{i=d+1}^{p} \langle \mu_i | S | \mu_i \rangle$

s.t. $\langle \mu_i | \mu_j \rangle = \delta_{ij}$

$\Rightarrow$ 拉格朗日乘子法

$L = \sum_i \langle \mu_i | S | \mu_i \rangle - \sum_i \lambda_i (\langle \mu_i | \mu_i \rangle - 1)$

$\frac{\partial L}{\partial \mu_i} = 2S|\mu_i\rangle - 2\lambda_i |\mu_i\rangle$

$\Rightarrow S|\mu_i\rangle = \lambda_i |\mu_i\rangle$

即为特征值方程 $J = \sum_i \langle \mu_i | S | \mu_i \rangle$

$= \sum \lambda_i$

因此应选择 d 个最大的特征值做为 $\lambda_1 \sim \lambda_d$，剩 p-d 个较小的做为 $\lambda_{d+1} \sim \lambda_p$

(13.2) 一组正交基 $|\mu_i\rangle$ 张成其中 d 个张成自子低维空间，对每个基 $|\mu_i\rangle$ 在其方向上的样本方差为

$Var = \frac{1}{N}\sum_{n=1}^{N} (\langle x_n | \mu_i \rangle - \langle \bar{x} | \mu_i \rangle)^2$

$= \frac{1}{N} \langle \mu_i | \sum_{n=1}^{N} (|x_n\rangle - |\bar{x}\rangle)$

$(\langle x_n| - \langle \bar{x}|) | \mu_i \rangle$

$= \frac{1}{N} \langle \mu_i | S | \mu_i \rangle$

同样的方法 $\begin{cases} \max Var \\ \text{s.t. } \langle \mu_i | \mu_i \rangle = 1 \end{cases}$

可求得特征值方程，并且应取最大的 d 个特征值.

(14) 解

经验误差 $\hat{\epsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} |h(x_i) - f(x_i)|$

泛化误差 $\epsilon(h) = \mathbb{E}_{x\sim D}[h(x) - f(x)]$

Chernoff bound:

$P(|\hat{\epsilon}(h) - \epsilon(h)| \geq \gamma) \leq 2\exp(-2\gamma^2 m)$ #

$P(\exists h \in \mathcal{H}, |\hat{\epsilon}(h) - \epsilon(h)| \geq \gamma) = P(\exists h \in \mathcal{H}, A(h))$

$= P(A(h_1) \vee A(h_2) \vee \cdots)$

$\leq k\, P(A(h)) \leq 2k\exp(-2\gamma^2 m)$ union bound

$\Rightarrow P(\forall h \in \mathcal{H}, |\hat{\epsilon}(h) - \epsilon(h)| \geq \gamma)$

$\geq 1 - 2k\exp(-2\gamma^2 m)$

(14.2) we write

$P(|\epsilon(\hat{h}) - \hat{\epsilon}(\hat{h})| \leq \gamma) \geq 1 - 2\exp(-2\gamma^2 m) = 1-\delta$

as. $|\epsilon(\hat{h}) - \epsilon(\hat{h})| \leq \gamma$ w.p. $1-\delta$

denote $\begin{cases} \hat{h} = \arg\min_{h\in\mathcal{H}} \hat{\epsilon}(h) \\ h^* = \arg\min_{h\in\mathcal{H}} \epsilon(h) \end{cases}$

w.p. at least $1-\delta$

$\epsilon(\hat{h}) \leq \hat{\epsilon}(\hat{h}) + \gamma \leq \hat{\epsilon}(h^*) + \gamma \leq \epsilon(h^*) + 2\gamma$

i.e. $P(|\epsilon(\hat{h}) - \epsilon(h^*)| \leq 2\gamma) \geq 1-\delta$

(14.3) (A) $h(x) = \mathbb{1}\{a < x\}$ $VC = 1$

$\begin{cases} \text{it can shatter point } x=0, \text{ chose } a = \pm 1 \\ \text{it can't shatter any two point } x_1, x_2 \\ \text{since } x_1 = 0 \quad x_2 = 0 \quad (x_1 < x_2) \text{ can't be realized.} \end{cases}$

(B) $h(x) = \mathbb{1}\{a < x < b\}$ $VC = 2$

point $x_1 = 0$ $x_2 = 1$

由 扫描全能王 扫描创建

NO
DATE

it can't shatter any 3 points $x_1<x_2<x_3$ since $x_1=1$ $x_2=0$ $x_3=1$ can't be realized.

(C) $h(x)=\mathbb{1}\{a\sin x>0\}$. $VC=1$

point $x_1=1$ can be shattered by choosing $a=\pm1$

any two point cannot be shattered if $x_1=x_2$ can be achieved, then $x_1\neq x_2$ cannot be achived, vice versa.

(D) $h(x)=\mathbb{1}\{\sin(x+a)>0\}$. $VC=2$

point $(x_1,x_2)=(\frac{\pi}{2},\frac{3\pi}{2})$ can be shattered by choosing $a=0$ or $a=\pi$.

any three points x_1,x_2,x_3

define $x_i'=x_i \bmod 2\pi$, ...

if $x_1'<x_2'<x_3'$

$x_1=x_3=1$ $x_2=0$ cannot be achieved

(15.1) HMM joint distribution

$P(x_1,y_1,x_2,y_2\dots)=P(y_1)P(x_1|y_1)\prod_{i=2}^{n}P(y_i|y_{i-1})P(x_i|y_i)$

(15.2) 参见14年题目16题

$P(\vec{x})=\frac{1}{Z}\prod_{Q\in C}\psi_Q(x_Q)$

$\psi_Q(x_Q)=e^{-H_Q(x_Q)}$

C: set of all cliques.

条件概率用贝叶斯公式来算.

(16.1) BFS / DFS / A^* ... 略

(16.2) proof.

由 $h(n)\le c(n,n')+h(n')$

$g(n)+h(n)\le g(n)+c(n,n')+h(n')$

$=g(n')+h(n')$

⇒ 顺着路径走 $f(n)$ 越来越大

由此每扩展一个结点的时候，其最优路径就被找到了，不然越往后走 $f(n)$ 越大，$h(n)$ 相同，则 $g(n)$ 越大，不可能比第一次扩展时更优.

(16.3) SA repeat

random select $x_{t+1}\in N(x_t)$

if $f(x_{t+1})\ge f(x_t)$

$x_{t+1}\to x$

else:

$x_{t+1}\to x$ w.p. $e^{\Delta E/T}$

decrease T.

GA repeat

next_gen ← random select via fitness

mutate(next_gen)

crossover(next_gen)

由 扫描全能王 扫描创建

令$N(\mu,\Sigma)$表示$P(x_i|\mu,\Sigma)$

(17.1) $L=\log P(D|\Phi,\mu,\Sigma)$

$=\sum_{i=1}^m \log P(x_i|\Phi,\mu,\Sigma)$

$=\sum_{i=1}^m \log P(x_i|y_i,\mu,\Sigma)+\log P(y_i|\Phi)$

$=\sum_{i=1}^m\sum_{j=0,1}\mathbb{I}(y_i=j)\log N(\mu_j,\Sigma)$

$+\sum_{i=1}^m y_i\log\Phi+(1-y_i)\log(1-\Phi)$

$\frac{\partial L}{\partial\Phi}=\sum_{i=1}^m y_i\frac{1}{\Phi}+\sum_{i=1}^m(1-y_i)\frac{-1}{1-\Phi}=0$

$\Rightarrow\Phi m-\Phi\sum_{i=1}^m y_i=\sum_{i=1}^m y_i-\Phi\sum_{i=1}^m y_i$

$\Rightarrow\Phi=\frac{1}{m}\sum_{i=1}^m y_i$

$\frac{\partial L}{\partial\mu_j}=\sum_{i=1}^m\mathbb{I}(y_i=j)\frac{\partial\log N(\mu_j,\Sigma)}{\partial\mu_j}$

$=\sum_{i=1}^m\mathbb{I}(y_i=j)\Sigma^{-1}(x_i-\mu_j)=0$

$\Rightarrow\mu_j=\sum_{i=1}^m\mathbb{I}(y_i=j)x_i/\sum_{i=1}^m\mathbb{I}(y_i=j)$

$\frac{\partial L}{\partial\Sigma}=\sum_{i=1}^m\sum_j\mathbb{I}(y_i=j)\frac{\partial\log N(\mu_j,\Sigma)}{\partial\Sigma}$

$=\sum_{i=1}^m\sum_j\mathbb{I}(y_i=j)\frac{1}{2}\left[\Sigma^{-1}(x_i-\mu_j)(x_i-\mu_j)^T\Sigma^{-1}-\Sigma^{-1}\right]=0$

$\Rightarrow\sum_{i=1}^m\sum_j\mathbb{I}(y_i=j)\left((x_i-\mu_j)(x_i-\mu_j)^T-\Sigma\right)=0$

$\Rightarrow\Sigma=\frac{\sum_{i=1}^m\sum_j\mathbb{I}(y_i=j)(x_i-\mu_j)(x_i-\mu_j)^T}{m}$

(18.1) y 某篇文章的 spam label.

x_j 文章出现第j个词

Naive Bayes:

$P(x_1,\dots x_V|y)=\prod_{j=1}^V P(x_j|y)$

一共m篇文章，对数似然函数

$L=\sum_{i=1}^m\left[\sum_{j=1}^V\log(P(x_j^{(i)}|y^{(i)}))+\log P(y^{(i)})\right]$

待估计参数

$\Phi=P(y^{(i)}=1)$

$\Phi_{j0}=P(x_j^{=1}|y=0)$

$\Phi_{j1}=P(x_j^{=1}|y=1)$

$L(\Phi,\Phi_{j0},\Phi_{j1})$

$=\sum_{i=1}^m\sum_{j=1}^V\mathbb{I}(y^{(i)}=0)\log\Phi_{j0}\,\mathbb{I}(x^{(i)}=1)$

$+\sum_{i=1}^m\sum_{j=1}^V\mathbb{I}(y^{(i)}=1)\log\Phi_{j1}\,\mathbb{I}(x^{(i)}=1)$

$+\sum_{i=1}^m\log\Phi\,\mathbb{I}(y^{(i)}=1)+\log(1-\Phi)\mathbb{I}(y^{(i)}=0)$

(18.2) $\frac{\partial L}{\partial\Phi}=\frac{1}{\Phi}\sum_{i=1}^m\mathbb{I}(y^{(i)}=1)-\frac{1}{1-\Phi}\sum_{i=1}^m\mathbb{I}(y^{(i)}=0)=$

$\Rightarrow(1-\Phi)\sum_{i=1}^m\mathbb{I}(y^{(i)}=1)=\Phi\sum_{i=1}^m\mathbb{I}(y^{(i)}=0)$

$\Rightarrow\Phi=\frac{1}{m}\sum_{i=1}^m\mathbb{I}(y^{(i)}=1)$

$\frac{\partial L}{\partial\Phi_{j0}}=\sum_{i=1}^m\frac{1}{\Phi_{j0}}\mathbb{I}(y^{(i)}=0)\mathbb{I}(x^{(i)}=1)-\lambda_0=0$

$\Rightarrow\Phi_{j0}=\frac{1}{\lambda_0}\sum_{i=1}^m\mathbb{I}(y^{(i)}=0)\mathbb{I}(x_j^{(i)}=1)$

同理 $\Rightarrow\Phi_{j1}=\frac{1}{\lambda_1}\sum_{i=1}^m\mathbb{I}(y^{(i)}=1)\mathbb{I}(x_j^{(i)}=1)$

其中 $L=L-\lambda_0(\sum_{j=1}^V\Phi_{j0}-1)-\lambda_1(\sum_{j=1}^V\Phi_{j1}-1)$

λ_0为归一化因子

$\lambda_0=\sum_{j=1}^V\sum_{i=1}^m\mathbb{I}(y^{(i)}=0)\mathbb{I}(x_j^{(i)}=1)$

= total # words in $y^{(i)}=0$ email

i.e.

$\Phi=\frac{1}{m}\sum_{i=1}^m\mathbb{I}(y^{(i)}=1)$

$$\begin{cases}\Phi_{j0}=\dfrac{\sum_{i=1}^m\mathbb{I}(y^{(i)}=0\wedge x_j^{(i)}=1)}{\sum_{i=1}^m\sum_{j=1}^V\mathbb{I}(y^{(i)}=0\wedge x_j^{(i)}=1)}\\[2ex]\Phi_{j1}=\dfrac{\sum_{i=1}^m\mathbb{I}(y^{(i)}=1\wedge x_j^{(i)}=1)}{\sum_{i=1}^m\sum_{j=1}^V\mathbb{I}(y^{(i)}=1\wedge x_j^{(i)}=1)}\end{cases}$$

由 扫描全能王 扫描创建

NO
DATE

(19) perceptron learning

$h_\theta(x) = \text{sign}(\theta^T x)$

$\theta \leftarrow \theta + y^{(i)} x^{(i)}$ if wrong on $(x^{(i)}, y^{(i)})$

THM. $\|x^{(i)}\| \le D$

$\exists u$ s.t. $\|u\|^2 = 1$ $y^{(i)}(u^T x^{(i)}) \ge \gamma$

pf. k-th error made on $\theta^{(k)}$ $(x^{(i)}, y^{(i)})$

$\theta^{(k+1)} = \theta^{(k)} + y^{(i)} x^{(i)}$

$\theta^{(k+1)} u = \theta^{(k)} u + y^{(i)} x^{(i)} u \ge \theta^{(k)} u + \gamma$

$\Rightarrow \theta^{(k+1)} \cdot u \ge k\gamma \Rightarrow \|\theta^{(k+1)}\|^2 \ge k^2\gamma^2$

$\|\theta^{(k+1)}\|^2 = \|\theta^{(k)} + y^{(i)} x^{(i)}\|^2$

$= \|\theta^{(k)}\|^2 + \|y^{(i)} x^{(i)}\|^2 + 2 y^{(i)} x^{(i)} \cdot \theta^{(k)}$ wrong on $\theta^{(k)}$

$\le \|\theta^{(k)}\|^2 + \|x^{(i)}\|^2$

$\le \|\theta^{(k)}\|^2 + D^2$

$\Rightarrow \|\theta^{(k+1)}\|^2 \le kD^2$

$\Rightarrow k^2\gamma^2 \le kD^2 \Rightarrow k \le \frac{D^2}{\gamma^2}$

由 扫描全能王 扫描创建