$\hat{Y} = \hat{f}(x)$ $Y = f(x) + \varepsilon_i$

$E(Y-\hat{Y})^2 = E(f(x) + \varepsilon_i - \hat{f}(x))^2$

$E((f(x)-\hat{f}(x))^2 + 2\varepsilon_i(f(x)-\hat{f}(x)) + \varepsilon_i^2)$

$= E[f(x)-\hat{f}(x)]^2 + Var(\varepsilon_i)$

$+2E[\varepsilon_i(f(x)-\hat{f}(x))]$

Pattern Recognition and machine learning

Inference need stronger assumptions.

parametric model: assump 函数的形式

non-___ : 不是没有参数, 而是没有预先假设的参数形式

$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda\int g''(t)^2 dt$ g(t)没有假设

$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$

$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda\int g''(t)^2 dt$ non-parameters

smoothing spline (non perfect) (only for low dimensions)

根据不同λ下的 testing error 选择模型 ⇒ MSE is difficult ⇐ no test data

⇒ Cross Validation

Bias-Variance Trade off

如果换成Y, 要再加上 Var(ε)

$E_D[f(x_0) - \hat{f}(x_0; D)]^2$ 不管哪个 data 去拟合, 不是 test 集

$= E_D[f(x_0) - E_D(\hat{f}) + E_D(\hat{f}) - \hat{f}]^2$

没有 randomness, 可将E直接去掉 ⇒ $bias^2$

$= E_D[f - E_D\hat{f}]^2 + E_D[(E_D\hat{f} - \hat{f})^2]$

$+2E_D[(f(x_0) - E_D\hat{f})(E_D\hat{f} - \hat{f})] = 0$

$= bias^2 + Var(E_D\hat{f} - \hat{f}) + (E_D[E_D\hat{f} - \hat{f}])^2$

$= bias^2 + Var(E_D\hat{f} - \hat{f})$

→ non training error or testing error.

$\frac{1}{n}\sum(y_i - \hat{g}(x_i))^2$ empirical testing error.

$E_{Y,X\sim T}[(Y - \hat{g}(x))^2]$ expected testing error

Condition on the training data D

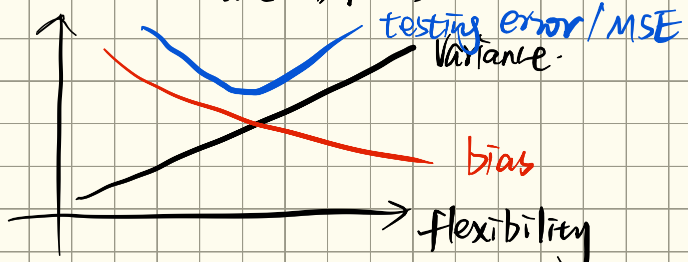
but $E_D = E_D E_{Y,X\sim T}[(Y - \hat{g}(x))^2 | D] = bias^2 + variance + noise$

unconditional one   conditional one.

In practice diff 很小

in CV. un conditional

In AIC. BIC conditional



overfit: small bias & large variance.

why ML flexible but small error?

signal >> noise   overfit will not be a big issue.

ML 仍然会 overfit eg. k-lines. noise very big.

事实上. 对 bias 来说我们要知道 true fun. f(x).

⇒ simulate.

interaction: effect of x_1 on Y depends on x_2. (tree model)

$Y = \beta_0 + \beta_1 x$

$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ $\beta_0 = \bar{y} - \hat{\beta}_1\bar{x}$

$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

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

$R^2 = \frac{TSS - RSS}{TSS}$ $TSS = \sum (y_i - \bar{y})^2$

more data. $R^2 \downarrow$

TSS: $\min \ell (y_i - \beta_0)^2$ $\beta_0 = \frac{1}{n}\sum y_i = \bar{y}$

$TSS = \sum_1 (y_i - \hat{\beta}_0)^2 = \sum (y_i - \bar{y})^2$ ($\hat{\beta}_0 = \bar{y}$)

TSS > RSS. 对于 training dataset

⇓ 只适用于 simple. $y = \beta_0 + \beta_1 x_i$

对 traing $\in [0,1]$ 对 test may < 0

⇒ overfit

Multiple linear regression

$\hat{\beta} = (X^T X)^{-1} X^T y$

why need T/F.

而不是直接回归 test minimized?

Variable selection:

different goal.

回归 + test: pick model best prediction.

T/F: control type 1 error.

zero but think it's non zero. 拒真

eg. drug. 我认为有用 ⇒ Big Issue.

may be inconsistent results.

if many data ⇒ converge

potential problems (related to inference)

1. non-linearity
2. correlation of error: $Var(\hat{\beta}) \neq \hat{\sigma}^2 (X^T X)^{-1}$
3. error variance is not constant (log transform)
4. outlier
5. leveraged points $\frac{\partial \hat{y}_i}{\partial y_i}$ 对 y_i 敏感的点. ↑ observed data
6. collinearity.

if test 1 para ⇒ F ≈ T

but assump ε normal

more para. ⇒ F.

if data change. model 不变 ⇒ flexibility = 0

flexible. 对data非常 sensetive ⇒ flexibility = sensetivity

$x_{new}^T \beta$ (check uncertainty) variance of y relys on variance of β

$df = \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$

对 linear model

$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$

$df = \sum_{i=1}^{n} h_{ii}$ 1 or 0

revise

randomness ⇒ make it preise take expectation.

$E\left[\sum \frac{\partial \hat{y}_i}{\partial y_i}\right]$

$= tr(H)$

why not $E[\sum |\frac{\partial \hat{y}_i}{\partial y_i}|]$ ⇐ $h_{ii} \in [0,1]$ ⇐ $H^2 = H = H^T H$

$= tr(X^T X (X^T X)^{-1}) = tr(I_{p+1}) = p+1$

① model paras

② how search paras

X 是 $n \times (p+1)$ $X^T X = (p+1) \times (p+1) = p+1$

= number of parameters be upper bounded by number of data.

可能有 constrains.

extension: additive model (no interaction) (stable)

eg. XGBoost: depth of tree = 1

⇒ additive

interaction: 与 correlation 无关. effect of X_1 on Y depends on X_2

additive 与 interaction. XGBoost: when depth of tree ≠ 1 ⇒ may some interaction ⇒ not improve

Solution: ① L2 平方会导致对 outlier 更敏感 (把 outlier 认为不是)

② L1 + 迭代 (MM). 更 robust

change w_i: (whether converge?)

Step: fit L2 + L1 version.

看是否 different 用 L1 identify outlier.

{ why need L2? robust & efficiency (var($\hat{\beta}$))

if no outlier. L2 smallest variance.

or $\min_{\beta} \sum |y_i - x_i^T \beta|$

let $t_i = y_i - x_i^T \beta$

$\min_{\beta} \sum t_i$

st. $-t_i \leq t_i \leq t_i$ linear programming

K-nearest neighbors: non-parametric

k=1 training error = 0

MM algorithm

non-differentiable

要找相切于 t^k, 且大于的 curve. 找最小值 t^{k+1}. $t^k \leftarrow t^{k+1}$

why not subgradient? don't converge

Lecture 3 Classification

why not linear? ① $x_{new}^T \hat{\beta}$ 不是在 [0,1]

② Qualitative 不能划分为 quantitative

③ classes 之间的差异不一致

⇒ $p(y_i = 1 | x_i) = \frac{1}{1 + \exp(-x_i^T \beta)}$ sigmoid function

softmax active function

单调 + [0,1]

也可以替 cdf.

assumption:

① ② y_i independent

logistic regression: $\log \frac{P(y_i = 1 | x_i)}{P(y_i = 0 | x_i)} = \log \frac{\frac{1}{1 + \exp(-x_i^T \beta)}}{\frac{\exp(-x_i^T \beta)}{1 + \exp(-x_i^T \beta)}} = \log \exp(x_i^T \beta) = x_i^T \beta$

(linear)

① nearly unbiased ⇐ linear

② easy interpretation

$\prod_{i=1}^{n} P(y_i | x_i, \beta) = \prod_{y_i = 1} \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \prod_{y_i = 0} \frac{1}{1 + \exp(x_i^T \beta)}$

$= \prod_{i=1}^{n} \frac{\exp(y_i x_i^T \beta)}{1 + \exp(x_i^T \beta)}$

↳ $\max(\log P(D|\theta))$

$= \max(\sum_{i=1}^{n} \log P(y_i | x_i, \beta))$

$= \max\left(\sum_{i=1}^{n}[y_i x_i^T\beta - \log(1+\exp(x_i^T\beta))]\right)$ 对β求导

$\sum y_i x_i - \frac{\exp(x_i^T\beta)x_i}{1+\exp(x_i^T\beta)}$ $(p+1)\times n$ $n\times 1$

$\sum x_i(y_i - p_i) = 0 = x^T(y-p)$

再求二阶导. $\frac{\partial}{\partial\beta} = \sum_i \left(\frac{x_i \exp(-x_i^T\beta)(-x_i^T)}{(1+\exp(-x_i^T\beta))^2} \right)$

$H = \frac{\partial^2 L}{\partial\beta\partial\beta^T}$ not compatible

$= \sum_{i=1}^{n} p_i(1-p_i)x_i x_i^T$

$= -x^T W x$

$W = \mathrm{diag}[p_i(1-p_i)]$

⇒ Newton Method

$\beta^{new} = \beta^{old} - H^{-1}\nabla L|_{\beta^{old}}$

$\min f(x)$. $g(x) = f'(x) = 0$

$= g(x_0) + g'(x_0)(x-x_0) = 0$

$x = -g'(x_0)^{-1}g(x_0) + x_0$

⇒ 应用. $L = \sum_{i=1}^{n}[y_i\beta^T x_i - \log(1+\exp(\beta^T x_i))]$

$L = \sum_{i=1}^{n}(y_i - p_i)x_i = x^T(y-p)$

$H = -x^T W x$

$\Rightarrow \beta^{new} = \beta^{old} - (x^T W x)^{-1}x^T(y-p)$

reweighted LS. $\min \sum_{i=1}^{n} w_i(y_i - x_i^T\beta)^2$

$= (y - x\beta)^T W(y - x\beta)$

$\hat\beta = (x^T W x)^{-1}x^T W y$

Another. 伯努利 distribution

$P(y_i) = p_i^{y_i}(1-p_i)^{(1-y_i)}$

$P(D|\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)}$

$\log P(D|\beta) = \sum_{i=1}^{n} y_i\log p_i + (1-y_i)\log(1-p_i)$ — cross entropy

伯努利写成 multiclass 容易 normalize

$P(y_i) = y_1^{p_1} y_2^{p_2} y_3^{p_3}$ (不相互影响) (垂直)

β: joint effect and margin effect. non causual effect

generative model: can generate data without anything given parameters

但 logistic 必须有 X

discriminative model (generative model) 本质上是估计 μ_k Σ_k $(f_k(x))$

Linear: $X|Y=k \sim N(\mu_k, \Sigma)$ 利用 μ_k, Σ_k 估计 $p(y|x)$, 挑选 max p.

求 $\delta(x)$ $\log\frac{p(Y=1|x)}{p(Y=2|x)} = (\mu_1^T\Sigma^{-1} - \mu_2^T\Sigma^{-1})x + \beta_0$ 实现 classification

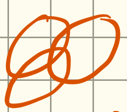
linear $\delta(x) = \mu_k^T\Sigma^{-1}x - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k$ + $\log\pi_k$. $\propto P_k(x) = P(Y|X)$

π_k, μ_k. how to estimate?

μ_k. center
Σ. 方向和 size

$\mu_k = \frac{1}{n_k}\sum_{y_i=k} x_i$ (用小组内数据)

$\hat\Sigma = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{y_i=k}(x_i - \hat\mu_k)(x_i - \hat\mu_k)^T$



so 方向大小都相同

(用了所有数据)

$\frac{1}{1+\exp} \to 1$

LDA 没有 numerical issue. $\exp \to 0$ $-x_i\beta \to \infty$ ⇒ numerical issue

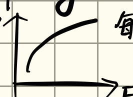
假设 newton's method (separable dataset), optimal = ∞. ⇒ not converge ⇒ logistic Reg 的问题

但 LDA 只需计算即可

different: ① LDA assume normal error distribution
② LDA no numerical issue

sensitivity: TP specificity: TN.

AUC ROC Curve



每个点对应一个 thershold

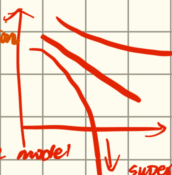
TP, FP = 1 − TN specificity

every 入有1个 ROC

inbalance data.

difference: AUC. not assume model

loss fun: testing data comes from distribution (eg. cross entropy)



sublinear: subgradient method

linear: gradient method $\beta^{new} = \beta^{old} - s \cdot g|_{\beta^{old}}$

+ strongly convex

super linear: Newton's method $\beta^{new} = \beta^{old} - H^{-1}\nabla L|_{\beta^{old}}$

time-consuming

H^{-1} 给出了 optimal s so [step] quicker than Gradient

QDA. (μ_k, Σ_k) 界限是曲线

哪个模型好取决于 bias & variance trade off. 不取决于 assumption 是否正确.

if enough data ⇒ variance is not important.

Validation set approach: 人为划分比例. { 多少合适? 怎样选择数据?

Cross validation {
leave one out: variance 还大. no randomness
correlation 很大, 导致 variance 很大. although bias 小 nearly small bias

k-folders CV

CV 低估了 variance (因为我们以为 uncorrelated), confidence interval 偏小 mean 准确 confidence interval 不准确
}

用回归选取 variable selection: training data: CV + selection 防止 data snopping

但 PCA 没有用. ∴ PCA 不影响

naive Classification: $X|Y=k \sim N(\mu_k, \Sigma_k)$ Σ_k is diagnal ⇒ independent

LOOCV 很麻烦. 但可以直接计算. (Ridge 也可以) $[e_i] = \frac{e_i}{1-h_{ii}}$ if h_{ii} 大 ⇒ flexibility 大. $[e_i]$ 大

只要 $\hat y = Sy$ ⇒ similar tricks (Woodbury lemma)

variable selection: best subset. forward. backward.

AIC = −2 loglikelihood $(\hat\theta)$ + 2d

BIC = −2 loglikelihood $(\hat\theta)$ + 2 log(n) d

for better prediction AIC > BIC

if enough data. BIC > AIC both use approximation, less time

CV 更好

CV don't care about which para but the number of paras.

最后给 all data 选出来的 p 个 paras

bootstrap: mean is different, variance structure is similar

best subset approach: L_0 norm; forward: approximation for L_0 norm.

Regularization: explicit: Ridge / Lasso

implicit: forward stepwise · variable selection.

/ gradient descent / XGBoost · / dropout ≈ Ridge

Ridge: L_2 standardize is important. ① 标准化 ② 入型 rescale

$(X^TX+\lambda I)^{-1}X^Ty \Leftarrow U(D+\lambda I)^{-1}U^TX^Ty$ 奇异问题 df ≠ number of parameters

Lasso: L_1 variable selection · coordinate descent

convex + differentiable

or convex + non-diff + seperatable.

soft shrinkage operator $sign(\tilde\beta)(|\tilde\beta|-\lambda)_+$

revised: ① inner product, track $y_i \Rightarrow r_i$

② active set

③ warm start + active set

SNR小时, 效果比较差

$\hat\beta_\lambda^{relax}(\gamma) = \gamma\cdot\hat\beta_\lambda + (1-\lambda)\hat\beta^{LS}$

$df = \frac{1}{\sigma^2}\sum_{i=1}^n Cov(y_i, \hat y_i)$

$S(\tilde\beta,\lambda) = \arg\min \frac{1}{2}(\tilde\beta-\beta)^2+\lambda|\beta|$

$\frac{\partial}{\partial\beta} = -(\tilde\beta-\beta)+\lambda\, sign(\beta)=0$

$\beta>0$: $-\tilde\beta+\beta+\lambda=0$

$\beta = (\tilde\beta-\lambda)_+$ $\tilde\beta>0$

$\beta<0$: $-\tilde\beta+\beta-\lambda=0$

$\beta = \lambda-|\tilde\beta|$

Lasso: $\frac{\partial R}{\partial\beta_j} = \sum_i\left(-x_{ij}(y_i-\tilde y_i^{(j)}-x_{ij}\beta_j)+\lambda\, sign(\beta_j)\right)=0$

$\Rightarrow \beta_j = S\left(\sum_i x_{ij}(y_i-\tilde y_i^{(j)}),\lambda\right)$.

必须 standardize.

RF: if classification: voting

if regression: mean

out of Bagging error: ith data 没被选中.

$\frac{1}{n}\sum(y_i-\hat f_i)^2 \simeq$ LOOCV

advantage: when training, already know LOOCV test error.

bagging: correlated.

RF: $\frac{m}{p} = 0.1, 0.5, 1$. more search, df ↑.

search p paras

search less.

search m paras.

eg. best subset · most search

不是比 bagging 好 signal/noise ratio

能够用 bagging / random to forward stepwise.

variance 换 bias ↑ approximation to best subset.

uncorrelated its variance is not that small.

to ↓ variance

Boost

(1) 对于每个数据给一个 w_i

(2) (a) 迭代 m 棵树，对每棵树 $f_i(x)$

loss function $= \sum_{i=1}^{n} w_i^{(m)} I(f_m(x_i) \neq y_i)$ 分类错误为 0，分类错误为 1

(b) $\varepsilon_m = \dfrac{\sum_{i=1}^{n} w_i^{(m)} I(f_m(x_i) \neq y_i)}{\sum_{i=1}^{n} w_i^{(m)}}$

$\alpha_m = \log\left[\dfrac{1-\varepsilon_m}{\varepsilon_m}\right]$

$w_i^{(m+1)} = w_i^{(m)} \exp\left[\alpha_m I(f_m(x_i) \neq y_i)\right]$

(3) 最终预测：$F_M(x) = \operatorname{sign}\left(\sum_{m=1}^{M} \alpha_m f_m(x)\right)$

负取决于符号。

$L(y, F) = \exp(-y_i F(x))$ $F_M(x) = \frac{1}{2}\sum_{i=1}^{M} \alpha_m f_m(x)$

在 M 次迭代时构建的树。

$F(x)$ 代表之前所有的 α 和 f，虽然不能这么看

所以用归纳法，假设已知 $\alpha_1 \dots \alpha_{m-1}$，$F_1 \dots F_{m-1}$

只取决于 α_m 和 F_m

$$L(y, F) = \sum \exp(-y_i F_m(x))$$
$$= \sum \exp\left[-y_i\left(F_{m-1}(x) + \frac{1}{2}\alpha_m f_m(x)\right)\right]$$
$$= \sum \exp\left(-y_i F_{m-1}(x) - \frac{1}{2} y_i \alpha_m f_m(x)\right) \quad w_i^{(m)} = \exp(-y_i F_{m-1}(x_i))$$
$$= \sum w_i^{(m)} \exp\left(-\frac{1}{2} y_i \alpha_m f_m(x)\right)$$

化简 $y_i f_m(x) =$

one fact: $\sum w_i^{(m)} = \sum_{i \in I} w_i^{(m)} + \sum_{i \in M} w_i^{(m)}$

$$L(y, F) = \sum_{i \in I} w_i^{(m)} \exp\left(-\frac{1}{2}\alpha_m\right) + \sum_{i \in M} w_i^{(m)} \exp\left(\frac{1}{2}\alpha_m\right)$$
$$= \exp\left(-\frac{1}{2}\alpha_m\right)\left(\sum_{i=1}^{M} w_i^{(m)} - \sum_{i \in M} w_i^{(m)}\right) + \sum_{i \in M} w_i^{(m)} \exp\left(\frac{1}{2}\alpha_m\right)$$
$$= \exp\left(-\frac{1}{2}\alpha_m\right)\sum_{i=1}^{M} w_i^{(m)} + \sum_{i \in M} w_i^{(m)}\left(\exp\left(\frac{1}{2}\alpha_m\right) - \exp\left(-\frac{1}{2}\alpha_m\right)\right)$$

令 $L(y, F)$ 最小

对 α_m 求导 $\dfrac{\partial}{\partial \alpha_m} = \sum_{i=1}^{M} w_i^{(m)} \exp\left(-\frac{1}{2}\alpha_m\right) \cdot \left(-\frac{1}{2}\right)$

$$+ \sum_{i \in M} w_i^{(m)}\left(\frac{1}{2}\exp\left(\frac{1}{2}\alpha_m\right) + \frac{1}{2}\exp\left(-\frac{1}{2}\alpha_m\right)\right) = 0$$
$$-\sum w_i^{(m)} + \sum_{i \in M} w_i^{(m)}\left(\exp(\alpha_m) + 1\right) = 0$$
$$\alpha_m = \log\left(\frac{\sum_i w_i^{(m)}}{\sum_{i \in M} w_i^{(m)}} - 1\right)$$
$$= \log \frac{\sum w_i^{(m)}}{\sum_i w_i^{(m)} - \sum_{i \in M} w_i^{(m)}}$$

$$\frac{1-\varepsilon_m}{\varepsilon_m} = \frac{\sum_{i=1}^{n} w_i^{(m)} I(f_m(x_i) \neq y_i)}{\sum_{i=1}^{n} w_i^{(m)} I(f_m(x_i) \neq y_i)} = \frac{\sum w_i^{(m)} - \sum w_i^{(m)} I(f_m(x_i) \neq y_i)}{\sum w_i^{(m)} I(f_m(x_i) \neq y_i)}$$

这个 I 是分错的时候 1，对 1 求和，是 y_i 和 f 分 1 和 1

$\therefore \alpha_m = \log \dfrac{1-\varepsilon_m}{\varepsilon_m}$

指数来证明 Adaboost 其实就是 logistic regression.

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

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

$= \exp(-y_i[F_{m-1}(x_i) - \frac{1}{2}\alpha_m y_i f_m(x_i)])$

$= w_i^{(m)} \exp(-\frac{1}{2}\alpha_m y_i f_m(x_i))$ $\quad y_i f_m(x_i) = 1 - 2I(f_m(x_i) \neq y_i)$

$= w_i^{(m)} \exp(-\frac{1}{2}\alpha_m(1 - 2I(f_m(x_i) \neq y_i)))$

$= w_i^{(m)} \exp(-\frac{1}{2}\alpha_m + \alpha_m I(f_m(x_i) \neq y_i))$

$= w_i^{(m)} \exp(-\frac{\alpha_m}{2}) \cdot \exp(\alpha_m I(f_m(x_i) \neq y_i))$

同个 m，对所有的 w 来说，都是个常数，可以忽略

$\Rightarrow w_i^{(m+1)} = w_i^{(m)} \exp(\alpha_m I(f_m(x_i) \neq y_i))$ 跟未归一化的权重公式一样

为什么是 logistic？

对总体的 loss

$E_{x,y}[\exp(-yF(x))] = \sum_y \int \exp(-yf(x)) p(y|x) p(x) dx$

对 F 求导 $\frac{\partial}{\partial F(x)} = \sum_y y \exp[-yF(x)] p(y|x) p(x)$

$= \int \exp F(x) p(y=-1|x) - \exp -F(x) p(y=1|x)$

$\Rightarrow F(x) = \frac{1}{2} \log \frac{P(y=1|x)}{P(y=-1|x)}$ ⇒ logistic regression.

Forward Stagewise Additive Model

$f(x) = \sum_{m=1}^{M} \beta_m b(x, \gamma_m)$ 如果 b 是一棵树，那么 γ_m 包含所有的 parameter（包含从哪开始到哪结束）。M 个 estimators 求和

若是 square loss：

$L(y, f(x)) = (y - f(x))^2$

$L(y_i, f_{m-1}(x) + \beta b(x_i, \gamma)) = (y_i - f_{m-1}(x) - \beta b(x_i, \gamma))^2$

$= (\underbrace{r_{im}}_{\text{current residual}} - \beta b(x_i, \gamma))^2$

$\beta b(x, \gamma) = \beta \sum c_i I(x \in S_i) = \sum \beta c_i I(x \in S_i)$

所以不用考虑 β 的问题

initial: $\hat{f}_0(x) = \arg\min_{\rho_0} \sum_{i=1}^{n} L(y_i, \rho_0)$ Friedman's Gradient Boosting

iteration:

① $r_i = -\frac{\partial}{\partial F_{m-1}(x_i)} L(y_i, F_{m-1}(x_i))\Big|_{F_{m-1}(x_i) = \hat{f}_{m-1}(x_i)}$ 相当于之前的 r_i

② fit $g(x)$ 用为来预测 r_i 但我们以前的 square loss 的 residual 是 $y_i - \hat{y}_i$ 但不知道其它的 loss function 的 r

③ $\tilde{\rho} = \arg\min_{\rho} \sum_{i=1}^{n} L(y_i, \hat{f}_{m-1}(x_i) + \rho g(x_i))$

④ $\hat{f}_m(x) = \hat{f}_{m-1}(x) + \tilde{\rho} g(x)$

if L = squared loss function.

$L(y_i, F) = \frac{1}{2}(y_i - F(x_i))^2$

$\frac{\partial L}{\partial F} = y_i - F(x_i)\Big|_{F_{m-1}(x_i) = \hat{F}_{m-1}(x_i)} = y_i - \hat{F}_{m-1}(x_i)$ 还是 current residual

$\beta^{new} = \beta^{old} + s(-g)|_{\beta^{old}}$ 类似于 gradient descent

initial: $f_0(x)$ constant. $\hat{f}_0(x) = \arg\min_{\rho_0} \sum_{i=1}^{n} L(y_i, \rho_0)$ Stochastic Gradient Boosting

① $r_i = -\frac{\partial}{\partial f_{m-1}(x_i)} L(y_i, f_{m-1}(x_i))\Big|_{f_{m-1}(x) = \hat{f}_{m-1}(x_i)}$

② 随机抽选 p 的 x cases data $= A$

③ 用 A 训练一个 regression tree $A \to r_i$（有 k 个 terminal nodes）

④ 对每个 terminal node 都有属于它的 S_k

$\tilde{\rho}_k = \arg\min_{\rho} \sum_{x_i \in S_k} L(y_i, \hat{f}_{m-1}(x_i) + \rho)$

⑤ $\hat{f}_m(x) = \hat{f}_{m-1}(x) + \nu g(x)$

$g(x) = \sum c_k I(x \in S_k)$

如何得到 c_k？$\to \rho_k$. $g(x) = \sum \rho_k I(x \in S_k)$

对落到 S_k 区域的点自己的 $\rho_k = \arg\min_{\rho} \sum_{x_i \in S_k} L(y_i, \hat{f}_{m-1}(x_i) + \rho)$

不断循环 ⇒ $g(x) = \sum \rho_k I(x \in S_k)$

举例子：loss $= -\frac{2}{n} \sum_{i=1}^{n} [y_i f(x_i) - \log(1 + \exp(f(x_i)))]$

initial: $f_0 = \log \frac{\sum y_i}{\sum (1 - y_i)}$ p_i

① $r_i = -\frac{\partial}{\partial f_{m-1}} = -\frac{2}{n}\sum_i \left[y_i - \frac{\exp f(x_i)}{1+\exp f(x_i)} \right] \sum \left(y_i - \frac{1}{1+\exp(-f(x_i))} \right) \left[\frac{2}{n} \right]$

② $\theta_{min} = \text{argmin} \sum (-g_{im} - T(x_i;\theta))^2$

用 θ 划分的 S_k 视作 ρ_k.

$\rho_k = \frac{\sum_{i \in S_k}(y_i - p_i)}{\sum_{i \in S_k} p_i(1-p_i)}$ ← why?

// 用 1-step Newton $\gamma = \frac{\sum}{n}(y_i - p_i)$

$H = \frac{\partial^2 L}{\partial p^2} = -\frac{2}{n}\sum p_i(1-p_i)$

$\rho = 0 - H^{-1}g = \frac{\sum_{i \in S_k}(y_i - p_i)}{\sum_{i \in S_k} p_i(1-p_i)}$

$\frac{\partial L}{\partial f} = -\frac{2}{n}\sum \left(y_i - \frac{1}{1+\exp(-f(x_i))} \right)$

$\frac{\partial L^2}{\partial f^2} = \frac{2}{n}\sum \left(+ \frac{+\exp(-f(x_i))}{(1+\exp(-f(x_i)))^2} \right)$

$= -\frac{2}{n}\sum p_i(1-p_i)$

λ: learning rate. Shrinkage

Lasso and Boosting.

(又用 square loss $\min_b \sum_i (r_i - \varepsilon b(x_i))^2 = \min_j \sum (r_i^2 - 2x_{ij}r_i + x_{ij}^2)$

r_i 相同 $\sum x_{ij}^2 = 1 \Rightarrow \max_j \{\sum_i x_{ij} r_i\}$

$x_j^T r$ $x^T(y - x\beta^{k-1})$

inner product with x_j and r.

(又用 Lasso L1 regularization

EM

E-step: $E_z[\log P(x, z|\theta)]$

M-step: $Q(\theta, \theta^{old}) = E_{z|x,\theta^{old}}[\ln p(x,z|\theta)] = \sum_z \ln p(x,z|\theta) \cdot p(z|x,\theta^{old})$

and $\max Q(\theta, \theta^{old})$

$\theta^{old} \leftarrow \theta^{new}$

首先，the missing data 指的不是 nan，而是不能直接观察到的数据

因此我们引入 z 我们假设其为 indicator.

假设 z 有 3 个值. $P(z=k) = \pi_k$

$P(x|z=1) \sim N(\mu_1, \Sigma_1)$ $P(x|z=2) \sim N(\mu_2, \Sigma_2)$, $P(x|z=3) \sim N(\mu_3, \Sigma_3)$

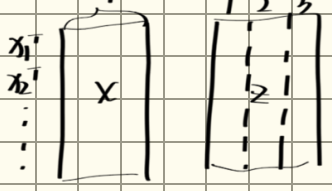
$p(x_n) = \sum_z p(x_n, z) = \sum_z p(x_n|z)p(z) = \sum_k p(x_n|\mu_k, \Sigma_k)\cdot\pi_k$.

$P(x) = \prod_{i=1}^n \left(\sum_k P(x_n|\mu_k, \Sigma_k)\cdot\pi_k \right)$

$\log P(x) = \sum_{i=1}^n \log \left(\sum_k P(x_n|\mu_k, \Sigma_k)\pi_k \right)$ π_k, μ_k, Σ_k 都是 θ^{old}.

Problem: ① 直接对 μ_k, Σ_k, π_k 求导很难有 closed form.

② 直接对 θ 迭代，如何保证 update 之后 $\sum \pi_k = 1$.



z_1^T $z_1^T = (1,0,0)$ 代表属于第 1 个类别

So $\sum_k z_{nk} = 1$

$\pi_1 = p(z_{n1} = 1)$

$\pi_2 = p(z_{n2} = 1)$

$\pi_3 = p(z_{n3} = 1)$

$p(z_n) = \prod_{k=1}^3 \pi_k^{z_{nk}}$

So $p(x_n|z_n) = \prod_{k=1}^3 p(x_n|\mu_k, \Sigma_k)^{z_{nk}}$

$\log P(x_n|z_n) = \sum_{k=1}^3 z_{nk} \log P(x_n|\mu_k, \Sigma_k)$

$p(x|z,\theta) = \prod_{i=1}^n \prod_{k=1}^3 p(x_n|\mu_k, \Sigma_k)^{z_{nk}}$

$p(x, z|\theta) = \prod_{i=1}^n \prod_{k=1}^3 p(x_n|\mu_k, \Sigma_k)^{z_{nk}} \prod_{k=1}^3 \pi_k^{z_{nk}}$

$= \prod_{i=1}^n \prod_{k=1}^3 \left(p(x_n|\mu_k, \Sigma_k)\pi_k \right)^{z_{nk}}$

$E_z[\log P(X, z|\theta)]$

$p(z|\theta) = p(z|z)p(z) = \prod_{k=1}^{3} p(z_n|\theta_k)^{z_{nk}} = \prod_k \pi_k^{z_{nk}}$

$\log = \sum z_{nk}\left(\log \pi_k + \log p(x_n|\theta_k)\right)$

$E_z[\log P(X, z|\theta)]$

$= E_z\left[\sum_{i=1}^{n}\sum_{k=1}^{3} z_{nk}\left(\log \pi_k + \log p(x_n|\mu_k, \Sigma_k)\right)\right]$

因此 E 时为常数 (θ^{old} 与 z_{nk})

$= E_z\left[\sum_{i=1}^{n}\sum_{k=1}^{3} z_{nk}\left(\log \pi_k + \log P(x_n|\mu_k, \Sigma_k)\right)\right]$

$= E_z\left[\sum_{i=1}^{n}\sum_{k=1}^{3} P(z_{nk}=1|x, \theta^{old})\left(\log \pi_k + \log P(x_n|\mu_k, \Sigma_k)\right)\right]$ 与 z_{nk} 无关, 取决 θ

可以理解为是 posterior mean

$P(z_{nk}=1|x, \theta^{old}) = \frac{P(z_{nk}=1, x_n|\theta^{old})}{p(x_n|\theta^{old})}$ how to calculate

$= \frac{p(x_n|z_{nk}=1, \theta^{old})p(z_{nk}=1)}{\sum_{j=1}^{k} p(x_n, z_{nj}=1|\theta^{old})}$

$= \frac{p(x_n|\mu_k, \Sigma_k)\cdot \pi_k}{\sum_{j=1}^{k} p(x_n|\mu_j, \Sigma_j)\pi_j}$ all θ^{old}, can get

⇒ 求得 E, 完成 E-step

let $E[z_{nk}|x, \theta] = \gamma(z_{nk})$

$\sum_{n=1}^{N}\sum_{k=1}^{k}\gamma(z_{nk}) = \sum_{n=1}^{N}\frac{\sum_{k=1}^{k} p(x_n|\mu_k, \Sigma_k)\pi_k}{\sum_{j=1}^{k} p(x_n|\mu_j, \Sigma_j)\pi_j}$

$= \sum_{n=1}^{N} 1 = N.$

取 $\sum_{k=1}^{k}\gamma(z_{nk}) = \sum_{k=1}^{k} E[z_{nk}|x, \theta] = 1$

即对之一有 γ 的期望为 1, 即为 x_n 落在 3 类中的概率为 1. 合法.

然后为 M-step

M-step: $Q(\theta, \theta^{old}) = E_{z|x,\theta^{old}}[\ln p(x, z|\theta)] = \sum_z \ln p(x, z|\theta)\cdot p(z|x, \theta^{old})$

and max $Q(\theta, \theta^{old})$

$\theta^{old} \to \theta^{new}$

$Q(\theta, \theta^{old}) = \sum_{i=1}^{n}\sum_{k=1}^{3}\underbrace{\gamma(z_{nk})}_{\theta^{old}}\left(\log \pi_k + \log P(x_n|\mu_k, \Sigma_k)\right)$ θ.

max $Q(\theta, \theta^{old})$ with constrains $\sum_k \pi_k = 1$

⇒ lagrange

① 对 π_k 求导: $\sum_{i=1}^{N}\gamma(z_{nk})\frac{1}{\pi_k} + \lambda = 0$. λ 未知. 对应的项 k

由于 $\sum_k \gamma(z_{nk}) = 1$

⇒ 对 k 项的同求 $\sum_{k=1}^{3}$

⇒ $N + \sum_k \lambda\pi_k = 0$. $\sum \pi_k = 1$ $\boxed{\lambda = N}$

$\sum_{i=1}^{N}\gamma(z_{nk}) - N\pi_k = 0$ $\pi_k = \frac{\sum_{i=1}^{N}\gamma(z_{nk})}{N}$.

同样. 对 μ_k 和 Σ_k 也使用同样的方法. 但不把 π_k 视为

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\underbrace{(\log\pi_k}_{\text{fixed}}+\log P(x_n|\mu_k,\Sigma_k))$$

$$\gamma(z_{nk})\cdot\log P(x_n|\mu_k,\Sigma_k)$$

根据 $x_i\sim N(\mu,\Sigma)\Rightarrow \hat{\mu}=\frac{\sum x_i}{n}\quad \Sigma=\frac{\sum x_i}{n}$

$$\Rightarrow \mu_k=\frac{\sum x_n\gamma(z_{nk})}{\sum\gamma(z_{nk})}\qquad \Sigma_k=\frac{\sum\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T}{\sum\gamma(z_{nk})}$$

一个很好用性质: $\ln p(x|\pi,\mu,\Sigma)=\ln\left(\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\cdot p(x_n|\mu_k,\Sigma_k)\right)$ 这里不再展开了.

$$=\sum_{i=1}^{N}\ln\sum_{k=1}^{K}\pi_k p(x_n|\mu_k,\Sigma_k)$$

↑ 单调递增. ⇒ 使得 EM. stable

why?

$\ln p(x|\pi,\mu,\Sigma)=L+KL$

$$L(q,\theta)=\sum_z q(z)\log P(x,z|\theta)-\sum_z q(z)\log q(z)$$

$$KL(q\|p)=-\sum_z q(z)\log p(z|x,\theta)+\sum_z q(z)\log q(z)$$

$$L+KL=\sum_z q(z)\left(\log P(x,z|\theta)-\log(p(z|x,\theta))\right)$$

$$=\sum_z q(z)\log p(x|\theta)$$

$$=\log P(x|\theta)\sum_z q(z)=1$$

$$=\log P(x|\theta)$$ 分解完成.

下证 $KL\geq 0$. Jensen's equality: $f(E(x))\leq E(f(x))$ for convex function

$$KL=-\sum_z q(z)\log P(z|x,\theta)+\sum_z q(z)\log q(z)\geq 0$$

$f=-\ln x\qquad x=\frac{p(z)}{q(z)}$

$$=-\sum_z q(z)\log\frac{p(z)}{q(z)}\qquad =E[f(x)]$$

$$\geq -\ln\cdot\left(\frac{p(z)}{q(z)}\cdot q(z)\right)=0$$

∴ L 是 $\log p(x|\theta)$ 的 lower bound.

- at E step $\ln p(x|\theta^{old})=L(q,\theta^{old})\Leftrightarrow q(z)=p(z|x,\theta^{old})$
- at M step. $L(q,\theta^{old})\leq L(q,\theta^{new})=\ln p(x|\theta^{new})$

递增.