Representations

Conditional expected risk:
$R(f,X) = \int_Y Q(Y, f(x)) P(Y|x)\, dY$

Total expected risk:
$R(f) = \int_Y \int_X Q(Y, f(X)) P(x, Y)\, dx\, dY$.

Empirical risk: $\hat{R}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} Q(y_i, \hat{f}(x_i))$

Density Estimation in Regression

Bayes Rule: $P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$

Cramer-Rao bound: $\mathbb{E}[(\hat{\theta}-\theta_0)^2] \geq \frac{1}{I_n(\theta_0)}$

Fisher info: $I_n(\theta_0) = n V_{\theta_0}\left[\frac{\partial}{\partial\theta}\log P(Y|\theta)\right]_{\theta_0}$

Efficiency: $\mathbb{E}[(\hat{\theta}_{eff}-\theta_0)^2] = \frac{1}{I_n(\theta_0)}$

$\hat{\theta}_{ML}$ is asymp. efficient (not eff with finite).

Summary on MLE:

① Consistency: $\hat{\theta}_n^{ML} \xrightarrow{P} \theta_0$

② Equivariance: $g(\hat{\theta}_{ML})$ is MLE of $g(\theta)$ (if g is invertable)

③ Asymp. normality: $\sqrt{n}(\theta-\hat{\theta}_n) \rightsquigarrow N(0, J^{-1})$;
$V(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log P(x|\theta)}{\partial\theta\,\partial\theta^T}\right]$ $I(\theta) = V\left[\frac{\partial \log P(x|\theta)}{\partial\theta}\right]$

④ Asymp. efficiency: $\lim_{n\to\infty}(V[\hat{\theta}^{ML}(x_1,\dots x_n)]\, I(\theta)) = 1$

Bayesian Learning:

$P(X=x|\chi) = \int p(x-\theta|\chi)\, d\theta = \int p(x|\theta)p(\theta|\chi)\, d\theta$

Assume: $p(x|\mu) \sim N(\mu, \sigma^2)$, $p(\mu) \sim N(\mu_0, \sigma_0^2)$

$p(\mu|\chi) \sim N(\mu_n, \sigma_n^2)$, $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$

$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}$, $\frac{\sigma^2}{\sigma_0^2}$: effective #sample

$p(x|\chi) \sim N(\mu_n, \sigma^2+\sigma_n^2)$

Multivariate case:
$\Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_0^{-1}$ $\Sigma_n^{-1}\mu_n = n\Sigma^{-1}\hat{\mu}_0 + \Sigma_0^{-1}\mu_0$

Recursive: $p(\theta|\chi^n) = \frac{p(x_n|\theta)p(\theta|\chi^{n-1})}{\int p(x_n|\theta)p(\theta|\chi^{n-1})\, d\theta}$

Regression

$\hat{\beta}_{RSS} = (X^TX)^{-1}X^Ty \sim N(\beta, (X^TX)^{-1}\sigma^2)$
(smallest var among unbiased estimators)

Bias-Var-Tradeoff

$|D| \downarrow |F| \uparrow \rightarrow$ overfit $\rightarrow$ var $\uparrow$ bias $\downarrow$

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

Avoid Overfitting:

Regularization, CV, Ensembles

Ridge Regression:

$RSS(\beta;\lambda) = (y-X\beta)^T(y-X\beta) + \lambda\beta^T\beta$

Gaussian prior: $\beta \sim N(0, \sigma^2/\lambda I)$.

LASSO:

$RSS(\beta;\lambda) = (y-X\beta)^T(y-X\beta) + \lambda\|\beta\|_1$

Laplace prior: $p(\beta) = \frac{\lambda}{4\sigma^2}\exp(-|\beta|\frac{\lambda}{2\sigma^2})$

SVD:

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

$X\hat{\beta}^{ls} = UDV^T(VD^2V^T)^{-1}VDU^Ty = UU^Ty$

$X\hat{\beta}^{ridge} = UD(D^2+\lambda I)^{-1}DU^Ty = \sum_{j=1}^{d} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^Ty$

Gaussian Process for Regression

Bayes linear regression

$p(\beta|\lambda) \sim N(\beta|0, \Lambda^{-1})$, $p(\beta|X,y,\Lambda) \sim N(\mu_\beta, \Sigma_\beta)$

$\mu_\beta = (X^TX+\sigma^2\Lambda)^{-1}X^Ty$, $\Sigma_\beta = \sigma^2(X^TX+\sigma^2\Lambda)^{-1}$

$\Lambda = \lambda I$, $\sigma = 1 \Rightarrow$ ridge

Kernel functions & Kernel engineering

linear: X^TX'. polynomial: $(X^TX'+1)^p$

RBF: $\exp(-\|x-x'\|^2/h^2)$

sigmoid (tanh): $\tanh kX^TX' - b$

Valid kernels: $ck_1(x,x')$ $(c>0)$;
$f(x)k_1(x,x')f(x')$; $q(k_1(x,x'))$ (q is poly with non-negative coeff); $\exp(k_1(x,x'))$;
x^TAx' (A is symmetric PSD).

Prediction by Gaussian Process

$p\left(\begin{bmatrix} y \\ y_{n+1}\end{bmatrix}\middle| x_{n+1}, X, \sigma\right) = N\left(\begin{bmatrix} y \\ y_{n+1}\end{bmatrix}\middle| 0, \begin{bmatrix} C_n & k \\ k^T & c\end{bmatrix}\right)$

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

$p(y_{n+1}|x_{n+1}, X, y) = N(\mu_{y_{n+1}}, \sigma^2_{y_{n+1}})$

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

$\sigma^2_{y_{n+1}} = c - k^TC_n^{-1}k$

$p(y_{n+1}|x_{n+1}, X, y) = \frac{p([y^T, y_{n+1}]^T|x_{n+1}, X)}{\int p([y^T, y_{n+1}]^T|x_{n+1}, X)\, dy_{n+1}}$

Conditional Gaussian distr.

$p(a_2|a_1=a) = N(a_2|\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(a-\mu_1), \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$.

Numerical Estimation Techniques

select $\hat{f}(x) \in F$ with min pred err:
CV / Bootstrap / Jackknife.

Naive Bootstrap:

$\hat{R}^* = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n} l(y_i, \hat{f}^{*b}(x_i))$

$\forall b$ $P\{z_i \in Z^{*b}\} = 1-(1-\frac{1}{n})^n \longrightarrow 1-\frac{1}{e} = 0.632$

LOO Bootstrap:

$\hat{R}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} l(y_i, \hat{f}^{*b}(x_i))$

$C^{-i} \subset \{1\dots B\}$ contains z^{*b} without (x_i, y_i).

0.632 bootstrap: $\hat{R}^{(.632)} = 0.368\hat{R}_{train} + 0.632\hat{R}^{(1)}$

Jackknife:

loo estimator $\hat{S}_{n-1}^{(-i)}$ to est. bias

$\mathbb{E}[\hat{S}_n] - S = \frac{a_1}{n} + \frac{a_2}{n^2} + \dots$

$\mathbb{E}[\hat{S}_{n-1}^{(-i)}] - S = \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \dots$

$bias^{JK} = (n-1)(\tilde{S}_n - \hat{S}_n)$ $\tilde{S}_n = \frac{1}{n}\sum_{i=1}^{n}\hat{S}_{n-1}^{(-i)}$

$\mathbb{E}[bias^{JK}] = \frac{a_1}{n} + \frac{(2n-1)a_2}{(n-1)n^2} + O(n^{-2})$

JK estimator: $\hat{S}^{JK} = \hat{S}_n - bias^{JK}$

Linear Discriminant Functions

Gradient Descent: $a(k+1) = a(k) - \eta(k)\nabla J(a(k))$

Optimal lr:

$H = \frac{\partial^2 J}{\partial a_i \partial a_j}$, 2nd Taylor exp.

$\frac{\partial}{\partial\eta}J(a(k+1)) = \frac{\partial}{\partial\eta}(J(a(k)) - \eta(k)\nabla J^T\nabla J + \frac{1}{2}\eta^2(k)\nabla J^TH\nabla J)$

$\eta^{opt} = \frac{\|\nabla J\|^2}{\nabla J^TH\nabla J}$ ($J(a) = J(a(k)) + \nabla J^T(a-a(k)) + \frac{1}{2}(a-a(k))^TH(a-a(k))$).

Newton's Algorithm:

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

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

Perceptron

$J_p(a) = \sum_{x\in\chi^{mc}}(-a^Tx)$ (criterion)

$a(k+1) = a(k) + \eta(k)\sum_{x\in\chi^{mc}}x$ (rule).

Support Vector Machines

KKT Conditions:

① (Feasible) $g_i(w^*) = 0$; ② (Feasible) h

③ (Complementary slackness) $\alpha_j^* h_j(w^*)$

④ (Minimizes Lagrangian) $\frac{\partial}{\partial w}L(w, \lambda^*, \alpha^*$

Hard Margin SVM:

minimize $\|w$

minimize $T(w) = \frac{1}{2}w^Tw$

subject to $\forall i$: $z_i(w^Ty_i + w_0$

maximize $W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}z_i$

subject to $\forall i$: $\alpha_i \geq 0 \wedge \sum_{i=1}^{n} z_i 0$

$w^* = \sum_{i=1}^{n}\alpha_i^* z_i y_i$. $w_0^* = \frac{1}{2}(\min_{i:z_i=1} w^*y_i$

$g^*(y) = \sum_{i\in SV} z_i\alpha_i^* y_i^T y + w_0^*$.

Optimal margin: $w^Tw = \sum_{i\in SV}\alpha_i^*$

Soft Margin SVM:

minimize $T(w,\xi) = \frac{1}{2}w^Tw + C$

subject to. $\forall i$: $z_i(w^Ty_i + w_0) \geq$

$L(w, w_0, \xi, \alpha, \beta) = \frac{1}{2}w^Tw + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\beta$
$-\sum_{i=1}^{n}\alpha_i(z_i(w^Ty_i + w_0$

maximize $W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}z_i$

subject to. $\sum_{j=1}^{n} z_j\alpha_j = 0 \wedge \forall i$ $C \geq$

Multiclass SVMs:

$\min_{w,\xi\geq 0}\frac{1}{2}w^Tw + C$

s.t. $(w_{z_i}^Ty_i + w_{z_i,0}) - \max_{z\neq z_i}(w_z^Ty_i$

Structural SVMs: (joint feature ma

hard functional margin SSVMs:

$\min_w \frac{1}{2}w^Tw$ "margin"

s.t. $w^T\psi(z_i, y_i) - \max_{z\neq z_i} w^T\psi(z_i, y_i')$

$\min_{w,\xi\geq 0}\frac{1}{2}w^Tw + C\sum_{i=1}^{n}\xi_i$

s.t. $w^T\psi(z_i, y_i) - w^T\psi(z, y_i) \geq$

$\psi_i(z) := \psi(z_i, y_i) - \psi(z, y_i)$ Δ_i

$L(\vec{w}, \xi, \alpha, \beta) = \frac{1}{2}w^Tw + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\sum$

$(w^T\psi_i(z_j) - \Delta_i(z_j) + \xi_i) - \sum_{j=1}^{n}\beta_j$

maximize $-\frac{1}{2}\|\sum_{i=1}^{n}\sum_{z\neq z_i}\alpha_{ij}\psi_i(z_j)\|$

subject to. $0 \leq \sum_{z\neq z_i}\alpha_{ij} \leq C$; $0 \leq$

Ensemble Methods

Bayesian inference by model averag
(weight by posterior).

ERM: $\hat{C}^* \leftarrow \operatorname{argmin}\hat{R}_D(C) = \operatorname{argmin}\frac{1}{n}\#\{$

Bagging: for $b=1$ to B do

Z^{*b} = b-th bootstrap sample from Z

Construct classifier C_b based on Z^{*b}.

end for

return $\hat{C}_B(x) = \text{sgn}(\sum_{i=1}^{B} C_i(x))$.

Covariance ↓, Variance similar. Bias –

AdaBoost initialize weight $w_i = \frac{1}{n}$.

for $b=1$ to B do

Fit $C_b(x)$ to the training data with w_i

$\epsilon_b \leftarrow \sum_{i=1}^{n} w_i^{(b)} \mathbb{I}\{C_b(x_i) \neq y_i\} / \sum_{i=1}^{n} w_i^{(b)}$

$\alpha_b \leftarrow \log \frac{1-\epsilon_b}{\epsilon_b}$

Set $\forall i$: $w_i \leftarrow w_i \exp(\alpha_b \mathbb{I}\{y_i \neq C_b(x_i)\})$.

end for. $\hat{C}_B(x) = \text{sgn}(\sum_{b=1}^{B} \alpha_b C_b(x))$

PAC Learning

learnable: $\exists \mathcal{A}$ can learn any concept in C.

if $\mathcal{A}$ receives sample $n \geq poly(1/\epsilon, 1/\delta, \dim(X))$, $P(R(\hat{c}) - \inf_{c \in C} R(c) \leq \epsilon) \geq 1-\delta$

Efficient PAC learnable: $\mathcal{A}$ runs in time $poly(1/\epsilon, 1/\delta)$. unions of k interval: $2k$

VC dimension: convex polygons in $\mathbb{R}^2$ with at most k vertices: $2k+1$

VC ineq $P\{R(\hat{C}_n^*) - \inf_{c \in C} R(c) > \epsilon\} \leq P\{\sup_{c \in C} |\hat{R}_n(c) - R(c)| > \frac{\epsilon}{2}\}$

Markov ineq. $P\{X \geq \epsilon\} \leq \frac{\mathbb{E}[X]}{\epsilon}$ (nonneg. X).

Hoeffding's Lemma: $\mathbb{E}[\exp(sX)] \leq \exp(s^2(b-a)^2/8)$

Hoeff's Theorem: $P\{S_n - \mathbb{E}[S_n] \geq t\} \leq \exp(-\frac{2t^2}{\sum_i (b_i-a_i)^2})$.

Normalized: $P\{\tilde{S}_n - \mathbb{E}[\tilde{S}_n] \geq \epsilon\} \leq \exp(-\frac{2n^2\epsilon^2}{\sum_i (b_i-a_i)^2/n})$

Union Bound: $P\{\cup_{i=1}^{\infty} A_i\} \leq \sum_{i=1}^{\infty} P\{A_i\}$.

Error Bound: $P\{\sup_{c \in C} |\hat{R}_n(c) - R(c)| > \epsilon\} \leq 2N\exp(-2n\epsilon^2)$

(N is cardinality of C).

Nonparametric Bayesian Methods

Dirichlet distr. $Dir(x|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{n} x_k^{\alpha_k - 1}$

$B(\alpha) = \frac{\prod_{k=1}^{n} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{n} \alpha_k)}$, $\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$

Stick breaking: $\beta_i \sim Beta(1,\alpha)$. $p_k = \beta_k(1 - \sum_{i=1}^{k-1} p_i)$

$p \sim GEM(\alpha)$, $p = \{p_k\}_{k=1}^{\infty}$

$\mathbb{E}[\#\text{clusters}] = \sum_{i=1}^{N} \frac{\alpha}{\alpha+i} \sim O(\alpha \log(N))$.

DP Mixture Model:

① Prob. of clusters: $p = (p_1, p_2 \cdots) \sim GEM(\alpha)$.

② Centers of clusters: $\mu_k \sim N(\mu_0, \sigma_0)$ $k=1,2\cdots$

③ Assignment: $z_i \sim Categorical(p), i=1 \cdots N$.

④ Coord. of data: $x_i \sim N(\mu_{z_i}, \sigma), i=1 \cdots N$

Gibbs sampling $p(z_i=k|z_{-i}, x, \alpha, \mu) \propto$

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

$= \begin{cases} \frac{N_{k,-i}}{\alpha+N-1} p(x_i|x_{-i,k}, \mu) & \text{for exist } k \\ \frac{\alpha}{\alpha+N-1} p(x_i|\mu) & \text{otherwise} \end{cases}$

Basics

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

$N(\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$.

$f(x) = f(a) + (x-a)f'(a) + \frac{1}{2}(x-a)^2 f''(a) + \cdots$

Binomial: $Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

$X \sim N(\mu,\Sigma)$, $Y = A + BX \Rightarrow Y \sim N(A+B\mu, B\Sigma B^T)$.

Moments $Var[X] = \int_x (x-\mu)^2 p(x) dx$

$Var[X] = \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

$Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$

$Cov[X,Y] = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$

$Cov[aX, bY] = ab\, Cov[X,Y]$

$K_{XY} = Cov[X,Y] = \mathbb{E}[XY^T] - \mathbb{E}[X]\mathbb{E}[Y^T]$

Matrix Identities $\frac{\partial}{\partial x}(b^T x) = \frac{\partial}{\partial x}(x^T b) = b$

$\frac{\partial}{\partial x}(x^T A x) = (A^T + A)x$; $\frac{\partial}{\partial x}(b^T A x) = A^T b$;

$\frac{\partial}{\partial X}(c^T X b) = cb^T$; $\frac{\partial}{\partial X}(c^T X^T b) = bc^T$;

$\frac{\partial}{\partial x}(\|x-b\|_2) = \frac{x-b}{\|x-b\|_2}$; $\frac{\partial}{\partial x}(\|x\|_2^2) = 2x$;

$x^T A x = Tr(x^T A x) = Tr(x x^T A) = Tr(A x x^T)$

$\frac{\partial}{\partial A} Tr(AB) = B^T$; $\frac{\partial}{\partial A} \log|A| = A^{-T}$

functions $\sigma(x) = 1/(1+e^{-x})$

$\sigma'(x) = \sigma(x)(1-\sigma(x))$; $\tanh = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

$\tanh'(x) = 1 - \tanh^2(x)$; $disc(x) = \text{sgn}(\log \frac{P(y=1|x)}{P(y=-1|x)})$

Inequalities Jensen: r.v. X & convex f

$\Rightarrow f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$; $1 - z \leq e^{-z}$;

Cauchy-Schwarz: $(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$

Extra

Naive Bayes (X iid).

$p(y|x) = \frac{P(y)P(x|y)}{\sum_y P(y)P(x|y)}$; $y = \text{argmax}_{y'} P(y'|x)$

EM likelihood $p(X|\theta) = \prod_{i=1}^{n} \sum_{c=1}^{k} \pi_c P(x_i|c, \mu_c, \Sigma_c)$

$p(X, M|\theta) = \prod_{i=1}^{n} \prod_{c=1}^{k} (\pi_c P(x_i, c, \mu_c, \Sigma_c))^{M_{ic}}$

E-step: $\gamma_{ic} = \mathbb{E}_{M|X,\theta^0}[M_{ic}] = P(c|x_i, \theta^0)$

$= \frac{\pi_c P(x_i|c, \mu_c^0, \Sigma_c^0)}{\sum_{v=1}^{k} \pi_v P(x_i|v, \mu_v^0, \Sigma_v^0)}$

M-step: $Q(\theta) = \mathbb{E}_{M|X,\theta^0}[L(X, M|\theta)]$

$= \sum_{i=1}^{n} \sum_{c=1}^{k} \gamma_{ic} \log(\pi_c P(x_i|c, \mu_c, \Sigma_c))$; $\mu_c = \frac{\sum_{i=1}^{n} \gamma_{ic} x_i}{\sum_{i=1}^{n} \gamma_{ic}}$

$\Sigma_c = \frac{\sum_{i=1}^{n} \gamma_{ic}(x_i - \mu_c)(x_i - \mu_c)^T}{\sum_{i=1}^{n} \gamma_{ic}}$; $\pi_c = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ic}$;

Proofs & Sample Problems

Axis-aligned Rectangle PAC learnable

Goal: $P(R(\hat{R}) \leq \epsilon) \geq P(\hat{R}IG) \geq 1 - 4e^{-\frac{n\epsilon}{4}}$

① $\hat{R}IG \Rightarrow R \setminus \hat{R} \subset T^\epsilon = \cup T_i^\epsilon$

$Pr(R \setminus \hat{R}) = Pr(T^\epsilon) \leq \sum_i Pr(T_i^\epsilon) = \epsilon$

② $Pr(\neg \hat{R}IG) = Pr(\hat{R}$ doesn't intersect T^ϵ some $i)$

$\leq \sum_i Pr(\hat{R} \cap T_i^\epsilon = \emptyset) = \sum_i Pr(x \notin T_i^\epsilon)$

$= \sum_i \prod_{j \leq n} Pr(x_j \notin T_i^\epsilon) = \sum_i \prod_{j \leq n} (1 - \frac{\epsilon}{4})$

$= 4(1-\frac{\epsilon}{4})^n \leq 4e^{-\frac{\epsilon n}{4}}$. $(n \geq \frac{4}{\epsilon} \times \frac{4}{8} \geq \frac{4}{\epsilon}\ln\frac{4}{8})$

Cramer-Rao $V = \frac{\partial}{\partial\theta} \log P(x_1 \cdots x_n|\theta)_{\theta_0}$

$= \frac{\frac{\partial}{\partial\theta} P(x_1 \cdots x_n|\theta)|_{\theta_0}}{P(x_1 \cdots x_n|\theta_0)}$. $\mathbb{E}[V] = 0$. $Var[V] = I_n(\theta_0)$

$\mathbb{E}[V(\hat{\theta} - \theta_0)] = \mathbb{E}[V\hat{\theta}] = \frac{\partial}{\partial\theta}\mathbb{E}[\hat{\theta}] = 1$

$\mathbb{E}[V^2]\mathbb{E}[(\hat{\theta}-\theta_0)^2] \geq (\mathbb{E}[V(\hat{\theta}-\theta_0)])^2 = 1$.

Optimal margin of dual SVM

$w^* = \sum_{i \in SV} \alpha_i^* z_i y_i$; $\sum_{i \in SV} \alpha_i^* z_i = 0$; $z_i(w^T y_i + w_0) = 1$

$w^T w = \sum_{i \in SV} \alpha_i^* z_i w^T y_i = \sum_{i \in SV} \alpha_i^* (1 - z_i w_0) = \sum_{i \in SV} \alpha_i^*$

Hoeffding's Lemma

$\mathbb{E}_x e^{sx} \leq \mathbb{E}_x e^{sb} \frac{x-a}{x-b} + \mathbb{E}_x e^{sa} \frac{b-x}{b-a} = \frac{b}{b-a} e^{sa}$

$- \frac{a}{b-a} e^{sb} = e^{\phi(u)}$ (*); for $p = \frac{-a}{b-a}$, $u = s(b-a)$

$\phi(u) = -pu + \log(1-p+pe^u)$; $\phi'(u) = -p + \frac{p}{p+(1-p)e^{-u}}$

$\phi''(u) = \frac{p(1-p)e^{-u}}{(p-(1-p)e^{-u})^2 + 4p(1-p)e^{-u}} \leq \frac{1}{4}$;

$\phi(u) = \phi(0) + \phi'(0)u + \frac{1}{2}\phi''(0)u^2 \leq 0 + 0u + \frac{1}{8}s^2(b-a)^2$

AdaBoost $c(x) = \text{Sgn}(F(x)) = \text{sgn}(\sum_{b=1}^{B} \alpha_b C_b(x))$;

$l(F(x), y) = \exp(-yF(x))$.

① Exp. loss: $F^*(x) \in \text{argmin}_{F(x)} \mathbb{E}_{Y|X}[e^{-YF(x)}]$.

$\Rightarrow F^*(x) = \frac{1}{2}\log\frac{P(y=1|x)}{P(y=-1|x)}$; $p(y=1|x) = \frac{1}{1+e^{-2F^*(x)}}$

② Stagewise reg. $F(x) \leftarrow F(x) + \alpha c(x)$

$J(F(x)) = \mathbb{E}_{Y|X}[\exp(-YF(x) - \alpha c(x) \cdots$

$= \mathbb{E}_{Y|X}[e^{-YF(x)}(1 - \alpha c(x)y + \frac{1}{2}\alpha^2 c^2 \cdots$

$C^*(x) = \text{argmin}_{c(x)} J(F(x)) = \text{argmax}_{c(x)} \mathbb{E}_{Y|X} \cdots$

$= \{+1$ if $\mathbb{E}_{Y|X}[e^{-YF(x)} \cdot y] > 0$; $-1 \cdots$

③ optimal α for classifier $c(x)$.

$\alpha^* = \text{argmin}_\alpha \mathbb{E}_{Y|X}[e^{-YF(x)} e^{-\alpha Y c(x)} \cdots$

$= \text{argmin}_\alpha [e^\alpha \mathbb{E}_{Y|X} e^{-YF(x)} \mathbb{I}\{c(x) \neq y\} \cdots$

$e^{-YF(x)}(1 - \mathbb{I}\{c(x) \neq y\})] \Rightarrow \alpha^* = \frac{1}{2}\log \cdots$

$err = \frac{\mathbb{E}_{Y|X}[e^{-YF(x)} \mathbb{I}\{c(x) \neq y\}]}{\mathbb{E}_{Y|X}[e^{-YF(x)}]}$.

Adaboost update:

$F(x) \leftarrow F(x) + \alpha^* c(x) = F(x) + \frac{1}{2}\log(\cdots$

$w(x,y) \leftarrow w(x,y)\exp(-\alpha^* y\, c(x))$

$= w(x,y)\exp(\log(\frac{1-err}{err})\mathbb{I}\{y \cdots$

Universal concept class is not PAC

$Pr[R(\hat{c}) > \epsilon] = Pr[\hat{R}_n(\hat{c}) = 0 \,\&\, R(\hat{c}) > \epsilon \cdots$

$\leq Pr[\exists h \in H: \hat{R}_n(h) = 0 \,\&\, R(h) > \epsilon]$

$\leq \sum_{h \in H} Pr[\hat{R}_n(h) = 0 \,\&\, R(h) > \epsilon]$

$= \sum_{h \in H} Pr[\hat{R}_n(h) = 0 | R(h) > \epsilon] Pr[R(h \cdots$

$\leq \sum_{h \in H} Pr[\hat{R}_n(h) = 0 | R(h) > \epsilon]$ $Pr(h(x \cdots$

$x \in Z$ is in $h \Delta c$ ⟶ ⊕ $Pr(h \cdots$

$\leq \sum_{h \in H} \prod_{j \leq n} R(X_j \notin h \Delta c | Pr(h \Delta c) > \epsilon) \leq \cdots$

Gibbs Sampling

for $i = 1$ to N in random order do

Remove x_i's sufficient statistics f…

for $k = 1$ to K do

$p_k(x_i) = p_k(x_i | X_{-i,k})$

$N_{k,-i} = |X_{-i,k}|$

$p(z_i = k | z_{-i}, X) = \frac{N_{k,-i}}{\alpha + N - 1}$

end for

$p_*(x_i) = p(x_i|\mu)$

$p(z_i = * | z_{-i}, X)$

Normalize $p(z_i|\cdot)$

Sample $z_i \sim p(z_i|\cdot)$

Add x_i's sufficient statistics to…

if any is empty, remove it and…

end for.