Max Entropy Priors (17)

Max ignorant priors: prior w/ largest entropy

$p^* = \arg\max_{p \in \Delta_\mathcal{X}} H(p) = \arg\min_{p \in \Delta_\mathcal{X}} D(p \| U)$ where U is uniform distn.

$\Rightarrow p^*(y) = \exp(\lambda^T t(y) - \alpha(\lambda))$ s.t. $E_{p^*}[t_i(y)] = \overline{t_i}$

Infinite Alphabet case: same as above

Conjugate Priors (18)

permutation invariance

Conditionally i.i.d Model: $P_{Y_1 \cdots Y_N | X}(y_1 \cdots y_N | x) = \prod_{n=1}^N P_{Y|X}(y_n | x)$

Exchangable: $(y_1 \cdots y_N)$ exchangable if $\forall$ perm π,

$P_{Y_1 \cdots Y_N}(y_1 \cdots y_N) = P_{Y_1 \cdots Y_N}(y_{\pi(1)} \cdots y_{\pi(N)})$

Theorem1: $Y_1, Y_2, \cdots$ is inf. exchangable if $\forall N$ $\exists \mathcal{X}$, cond iid model $P_{Y|X}$ s.t. $p(y_1 \cdots y_N) = \int P_{Y_1 \cdots Y_N | X}(y_1 \cdots y_N | x) p_X(x) dx$

Conjugate Prior Family: $\mathcal{Q} = \{q(\cdot;\theta) \in \mathcal{P}^\mathcal{X} : \theta \in \Theta \subseteq \mathbb{R}^k\}$ s.t. mapping from $\Theta \mapsto q$ cont. & invertible. $\mathcal{Q}$ is C.P. Fam if $\forall y \in \mathcal{Y}$, we have $P_X(\cdot) \in \mathcal{Q} \Rightarrow P_{X|y}(\cdot | y) \in \mathcal{Q}$.

↳ equivalent to $p_X(\cdot) \in \mathcal{Q} \Rightarrow P_{X|y_1 \cdots y_N}(\cdot | y_1 \cdots y_N) \in \mathcal{Q}$

Natural Conjugate Prior Family: Natural if $\forall y$, $\exists$ param $\theta(y)$ s.t. $q(\cdot;\theta(y)) \propto P_{Y|X}(y|\cdot)$

Theorem2: For cond. iid models s.t. $\mathcal{Y} \subseteq \mathbb{R}$ region, $P_{Y|X}(\cdot|x)$ cont. func, if conj. prior fam exists, then for every $N \geq 1$ $\exists$ cont. func $t_N(\cdot, \cdots, \cdot)$ s.t. $t_N(y_1 \cdots y_N)$ is a sufficient stat wrt inf. abt x, and t_N finite dim ind. of N.

Conjugate Prior of Exp Families: If $p_{Y|X} = \exp(\lambda(x)^T t(y) - \alpha(x) + \beta(y))$ then $\mathcal{Q} = \{q(\cdot;t,N) = \exp\{t^T \lambda(x) - N\alpha(x) - \gamma(t,N)\}\}$ is natural conj. prior fam.

Info Geometry of ML & EM (19)

$\hat{x}_{ML}(y) = \arg\max_a P_Y(y;a)$

Fact: $\frac{1}{N}\sum_{n=1}^N f(u_n) = \mathbb{E}_{\hat{p}(\cdot|u)}[f(U)]$

Fact: $\hat{x}_{ML}(y) = \arg\min_a D(\hat{p}_y(\cdot|y) \| P_Y(\cdot|a))$

↳ M-projection (minimizing second term of KL divergence).

Direct EM Algo:

E-step: Compute $U(x;\hat{x}^{(k-1)})$, where

$U(x;x') = \sum_{z \in \mathcal{Z}^N} p_{Z|Y}(z|y;x') \log p_Z(z;x)$

M-step: Compute $\hat{x}^{(k)} = \arg\max_x U(x;\hat{x}^{(k-1)})$.

If $p_{Z|Y}(z|y;x) = \prod_{n=1}^N p_{Z_n|Y_n}(z_n|y_n;x)$, $y_n = y(z_n)$, then E-step simplifies to

$U(x;x') = \sum_{n=1}^N \sum_{c \in \mathcal{Z}} p_{Z|Y}(c|y_n;x') \log p_Z(c;x)$.

Note $p_{Z|Y}(z|y;x) = \frac{p_Z(z;x)}{p_Y(y;x)} 1_{z \in y^{-1}(y)}$, $p_Y(y;x) = \sum_{z \in y^{-1}(y)} p_Z(z;x)$

$\rightarrow$ want $x = \arg\min_x D(p_{\hat{Z}}(\cdot;x') \| p_Z(\cdot;x))$.

Let $p^*_{\hat{Z}}(\cdot;x) = \arg\min_{\hat{p}_Z \in \hat{\mathcal{P}}^Z(y)} D(\hat{p}_Z(\cdot) \| p_Z(\cdot;x))$

$= \frac{p_Z(z;x)\, \hat{p}_Z(y(z))}{p_Y(y(z);x)}$

$\rightarrow$ set $\frac{1}{N}U(x;x') = -D(\hat{p}^{x'}_Z(\cdot;x') \| p_Z(\cdot;x)) - H(\hat{p}^{x'}_Z(z|y;x'))$

New EM: E-step: $\hat{p}^{x^{(k-1)}}_Z(\cdot;\hat{x}^{(k-1)}) = \arg\min_{\hat{p}_Z \in \hat{\mathcal{P}}^Z(y)} D(\hat{p}_Z(\cdot) \| p_Z(\cdot;\hat{x}^{(k-1)}))$

M-step: $\hat{x}^{(k)} = \arg\min_x D(\hat{p}^{x^{(k-1)}}_Z(\cdot;\hat{x}^{(k-1)}) \| p_Z(\cdot;x))$

Data Processing Inequality (DPI): $g: \mathcal{Z} \to \mathcal{Y}$, p_Y, q_Y induced by p_Z, q_Z. Then $D(p_Z \| q_Z) \geq D(p_Y \| q_Y)$ w/ equality iff $\frac{p_{Z|Y}(z)}{q_{Z|Y}(z)} = \frac{p_Y(y(z))}{q_Y(y(z))}$

Theorem.1: $p_Y(y;x) = \exp\{\sum_k x_k t_k(y) - \alpha(x) + \beta(y)\}$, then $\hat{x}_{ML}$ satisfies $\frac{1}{N}\sum_{n=1}^N t_k(y_n) = \mathbb{E}_{\hat{p}(\cdot;\hat{x}_{ML})}(t_k(y)) = \mathbb{E}_{\hat{p}_Y(\cdot;\hat{x})}[t_k(y)]$

Typical Seqs, Large Deviations (21)

WLLN: If $u_1 \cdots u_N$ i.i.d r.v. w/ mean μ, $E[|u_n|] < \infty$, then

$\lim_{N\to\infty} \mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^N u_n - \mu\right| > \varepsilon\right) = 0$ (conv. in prob.)

Typical Set: Note $\hat{L}_p(y) = \frac{1}{N}\sum \log p(y_n) \to -H(p)$

↳ Defn: seq y is ε-typical wrt p if $|\hat{L}_p(y) + H(p)| \leq \varepsilon$

$T^\varepsilon_N(p) = \{y \in \mathcal{Y}^N : |\hat{L}_p(y) + H(p)| \leq \varepsilon\}$

Theorem (Asymptotic Equipartition Property): If $T^\varepsilon_N(p)$ is ε-typical set, then $\lim_{N\to\infty} P\{T^\varepsilon_N(p)\} = 1$, $2^{-N(H(p)+\varepsilon)} \leq p^N(y) \leq 2^{-N(H(p)-\varepsilon)}$ and $(1-\varepsilon)2^{N(H(p)-\varepsilon)} \leq |T^\varepsilon_N(p)| \leq 2^{N(H(p)+\varepsilon)}$ for large enough N.

Divergence ε-typical wrt p relative to q: $T^\varepsilon_N(p\|q) = \{y : |\hat{L}_{p\|q}(y) - D(p\|q)| \leq \varepsilon\}$

$- \lim_{N\to\infty} P\{T^\varepsilon_N(p\|q)\} = 1$

Theorem: $(1-\varepsilon)2^{-N(D(p\|q)+\varepsilon)} \leq Q\{T^\varepsilon_N(p\|q)\} \leq 2^{-N(D(p\|q)-\varepsilon)}$

Cramer's Theorem: If $y = (y_1 \cdots y_N)$ iid from q, $t: \mathcal{Y} \to \mathbb{R}$ s.t. $\kappa < \mathbb{E}_q[t(y)] < \infty$, then for any $\gamma > \mu$, have:

$\lim_{N\to\infty} -\frac{1}{N}\log P\left(\frac{1}{N}\sum_{n=1}^N t(y_n) \geq \gamma\right) = E_C(\gamma)$

where $E_C[\gamma] \triangleq D(p(\cdot;x)\|q)$ w/ $p(y;x) = q(y)e^{x t(y) - \alpha(x)}$, and x s.t. $\mathbb{E}_{p(\cdot;x)}[t(y)] = \gamma$.

Furthermore, we can express $E_C(\gamma) = (\log e)(x\gamma - \alpha(x))$

Can calculate: $\frac{1}{\log e} \cdot \frac{x}{dx}\left(D(p(\cdot;x)\|q)\right) = x\ddot{\alpha}(x) > 0$ $\forall x > 0$

Method of Types + Sanov (22)

Type: prob distn. $\hat{p}(b;y) = \frac{1}{N}\sum_{n=1}^N 1_{b=y_n} = \frac{N_b(y)}{N}$

Set of Types: $\hat{\mathcal{P}}^\mathcal{Y}_N$ = set of all possible types

Type class: $T^\mathcal{Y}_N(p) = \{y \in \mathcal{Y}^N : \hat{p}(\cdot;y) = p(\cdot)\}$

Exponential Rate Notation: $f(N) \doteq 2^{N\alpha}$ means $\lim_{N\to\infty} \frac{\log_2 f(N)}{N} = \alpha$.

Properties: $|\hat{\mathcal{P}}^\mathcal{Y}_N| \leq (N+1)^{|\mathcal{Y}|}$

Lemma: If $q \in \mathcal{P}^\mathcal{Y}$ defined over y, then $q^N(y) = 2^{-N(D(\hat{p}(\cdot;y)\|q) + H(\hat{p}(\cdot;y)))}$ (general inf. theory version)

Lemma: $|T^\mathcal{Y}_N(p)| \doteq 2^{NH(p)}$, $cN^{-|\mathcal{Y}|}2^{NH(p)} \leq |T^\mathcal{Y}_N(p)| \leq 2^{NH(p)}$

Theorem.1: $cN^{-|\mathcal{Y}|}2^{-ND(p\|q)} \leq Q\{T^\mathcal{Y}_N(p)\} \leq 2^{-ND(p\|q)}$

Sanov's Theorem: $S \subset \mathcal{P}^\mathcal{Y}$ arbitrary, $q \in \mathcal{P}^\mathcal{Y}$ arbitrary, then $Q\{S \cap \hat{\mathcal{P}}^\mathcal{Y}_N\} \leq (N+1)^{|\mathcal{Y}|} 2^{-ND(p_0\|q)} \doteq 2^{-ND(p_0\|q)}$

where $p^* = \arg\min_{p \in cl(S)} D(p\|q)$ = I-proj of q onto $cl(S)$.

If $cl(S) = cl(int(S))$, then $Q\{S \cap \hat{\mathcal{P}}^\mathcal{Y}_N\} \doteq 2^{-ND(p_0\|q)}$

Pf sketch: bound size $|S|$ w/ $(N+1)^{|\mathcal{Y}|}$ and $Q\{T_N(p_N)\} \geq cN^{-|\mathcal{Y}|}2^{-ND(p\|q)}$

Conditional Limit Thm: If $S \subset \mathcal{P}^\mathcal{Y}$ closed & convex, $y \in \mathcal{Y}^N$ iid wrt $q \notin S$, then for any $\varepsilon > 0$ we have $\lim_{N\to\infty} P(|\hat{p}(b;y) - p_*(b)| > \varepsilon \mid \hat{p}(\cdot;y) \in S) = 0$

where $p_* = I$-proj from q to S; the type $\to p_*$

Pinsker ineq: for any $p, q \in \mathcal{P}^\mathcal{Y}$, we have $\|p - q\|_1 = \sum_{b \in \mathcal{Y}} |p(b) - q(b)| \leq \sqrt{2\ln 2 \cdot D(p\|q)}$

Asymptotics of Hypo Testing (23)

Setup: Binary hypo test $H_0: P_Y(y) = \prod_{n=1}^N P_0(y_n)$, $H_1: P_Y(y) = \prod_{n=1}^N P_1(y_n)$

→ Log. likelihood ratio: $\hat{L}(y) = \frac{1}{N}\sum_{n=1}^N \log \frac{P_1(y_n)}{P_0(y_n)}$

Typically biased rule: $g(y) = g_{A,\gamma}(H_1|y) = \begin{cases} 0 & \text{if } y \in J^N_\gamma(p_1|p_0) \setminus J^N_\gamma(p_1|p_0) \\ 1 & \text{if } y \in J^N_\gamma(p_1|p_0) \setminus J^N_\gamma(p_1|p_0) \\ \frac{1}{2} & \text{otherwise} \end{cases}$

↳ get $P_F \doteq \varepsilon + 2^{-ND(p_1\|p_0)}$, $P_M \doteq \varepsilon + 2^{-ND(p_0\|p_1)}$

Theorem: For log-likelihood ratio test w/ threshold γ $\left(\hat{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma\right)$ with $\hat{L}(y) = \frac{1}{N}\sum_{n=1}^N \log \frac{P_1(y_n)}{P_0(y_n)}$, we have:

$P_F \doteq 2^{-ND(p_\gamma\|p_0)}$, $P_M \doteq 2^{-ND(p_\gamma\|p_1)}$ where

$p_\gamma = p(y; x_\gamma) = \frac{1}{Z(x_\gamma)} p_0^{1-x_\gamma} p_1^{x_\gamma}$ w/ x_γ s.t. $D(p_\gamma\|p_0) - D(p_\gamma\|p_1) = \gamma$.

NP-hypo testing: Constrain P_F, minimize P_M

→ If P_F doesn't decay exponentially, set $\gamma = -D(p_0\|p_1)$, $P_F = P_0 - P_M \doteq 2^{-ND(p_0\|p_1)}$

Bayesian Hypo testing: $E[L] = C_{10}P_0P_F + C_{01}P_1P_M \doteq 2^{-N\min(D(p_\gamma\|p_0), D(p_\gamma\|p_1))}$

→ satisfied at $\gamma = 0$, $E_C = D(p_*\|p_1) = D(p_*\|p_0)$

Asymptotics of Param Est. (25)

Identifiability: $x \in \mathcal{X}$ s.t. $D(p_Y(\cdot;x)\|p_Y(\cdot;x')) = 0 \Leftrightarrow x = x'$

Consistency: $\hat{x}_N(y^N) \xrightarrow{p} x$ as $N \to \infty$ (weak) / $\xrightarrow{a.s.}$ (strong)

Asymptotic Normality: $\sqrt{N}(\hat{x}_N - x) \xrightarrow{d} N(0, \sigma^2(x))$ for $\sigma^2(x) \geq 0$

Asymptotic Efficiency: Asymptotic Normal and $\sigma^2(x) = 1/J_Y(x)$ (1/Fisher info)

Theorem1: If $|\mathcal{Y}|, |\mathcal{X}| < \infty$, $y_1 \cdots y_N$ iid $\sim p_Y(\cdot;x)$ for identifiable x, and $p_Y(y;x) > 0$ $\forall y \in \mathcal{Y}, x \in \mathcal{X}$, then ML estimate $\hat{x}_{ML}(y^N)$ weakly consistent.

Theorem2: If $p_Y(y;x) = \exp(a(x)t(y) - \alpha(x) + \beta(y))$, $y^N \sim p_Y(\cdot;x)$ id $\ddot{\alpha}(x) = J_Y(x) > 0$, then $\hat{x}_N(y^N)$ is strongly consistent.

Theorem3: same as above, now $\hat{x}_N(y^N)$ is asymptotically efficient.

Folk thm: $\hat{x}_N(y^N) \sim N\left(x, \frac{1}{N J_Y(x)}\right)$

Theorem 6: If $\mathcal{P} = \{p_Y(\cdot;x) : x \in \mathcal{X}\}$, $|\mathcal{X}|, |\mathcal{Y}| < \infty$, $y_1 \cdots y_N$ iid according to $q_Y \notin \mathcal{P}$, $q_Y(y) > 0$, $p_Y(y;x) > 0$ $\forall y, x$, and $x_* = \arg\min_{x \in \mathcal{X}} D(q_Y \| p_{Y|X}(\cdot|x))$ unique. Then $\hat{x}_N(y_1 \cdots y_N) \to x_*$.

Laplace Method: Approximate $\int_a^b e^{N g(z)} dz$ if g twice diff., unique max of g $z_0 \in (a,b)$ s.t. $\dot{g}(z_0) = 0$, then

$\frac{\int_a^b e^{N g(z)} dz}{e^{N g(z_0)} \sqrt{\frac{2\pi}{N|\ddot{g}(z_0)|}}} \to 1$ as $N \to \infty$

We want to approximate posteriors $P_{X|y^N}(x|y^N) = \frac{\int p_X(a) \prod_{n=1}^N P_{Y|X}(y_n|a) da}{\int \cdots}$

If $\mathcal{X}$ cont. alphabet, often $p_{X|y^N}(\cdot|y^N) \approx N\left(\hat{x}_N(y^N), \frac{1}{N J_Y(x)}\right)$.

Theorem8: If $\tilde{x} = \sqrt{N}(x - \hat{x}_N(y^N))$, y^N iid from $p_{Y|X}(\cdot|x)$, $x \sim p_X$, $\ddot{\alpha}(x) = J_Y(x)$, and $p_Y(y;x) = \exp(a(x)y - \alpha(x) + \beta(y))$, then $\ln(p_{\tilde{X}|y^N}(\tilde{x}|y^N)) \xrightarrow{a.s.} \frac{1}{2}J_Y(x)\tilde{x}^2$ as $N \to \infty$.

Asymptotics of Inf. & Universality (26)

Universal inference: After enough data, can make predictions abt future data.

Defn Universal Predictor: $S_N = \{p_{Y^N|X}(\cdot;x) \in \mathcal{P}^{\mathcal{Y}^N}, x \in \mathcal{X}\}$ family of models, $q_1, q_2, \cdots q_N \in \mathcal{P}^\mathcal{Y}$ seq. of predictors, $q_N(\cdot;y^{N-1})$ est. of dist given y^{N-1}, then seq is universal if $\lim_{N\to\infty} \max_{x \in \mathcal{X}} \bar{p}_N(x) = 0$, where

$\bar{p}_N(x) = \frac{1}{N}\sum_{n=1}^N p_n(x)$ w/ $p_n(x) = \mathbb{E}_{p_{Y^n}(\cdot;x)}\left[\log \frac{p_{Y_n|Y^{n-1}}(y_n|y^{n-1};x)}{q_n(y_n|y^{n-1})}\right]$

and $p_{Y_n|Y^{n-1}}(y_n|y^{n-1};x) = \frac{p_{Y^n}(y^n;x)}{p_{Y^{n-1}}(y^{n-1};x)}$.

↳ Also: $p_n(x) = \mathbb{E}_{p_{Y^{n-1}}(\cdot;x)}\left[D(p_{Y_n|Y^{n-1}}(\cdot|Y^{n-1};x) \| q_n(\cdot;Y^{n-1}))\right]$

↳ $\bar{p}_N(x) = \frac{1}{N} D(p_{Y^N|X}(\cdot|x) \| q_{Y^N}(\cdot))$, where $q_{Y^N}(y^N) = \prod_{n=1}^N q_n(y_n|y^{n-1})$

Theorem1: If model capacity of model from S_N is C_N, $y^N = (y_1 \cdots y_N)$ from $p \in S_N$, then universal inf possible iff $\lim_{N\to\infty} \frac{C_N}{N} = 0$.

Asymptotic Least informative prior: $p^{in}_X \in \mathcal{P}^\mathcal{X}$ for fam w/ capacity C_N if $I(x;y^N)$ associated w/ $p_{Y^N}(y^N;x) p^{in}_X(x)$ satisfies $C_N - I(x;y^N) = o(1)$ as $N \to \infty$.

↳ Construct predictor $q_{Y_n|Y^{n-1}}(y_n|y^{n-1}) = \frac{q(y^n)}{q(y^{n-1})}$ w/ p^{in}_X prior.

Theorem2: Model fam S_α $|\mathcal{Y}|, |\mathcal{X}| < \infty$, $y^N \sim p_Y(\cdot;x)$ where $x \sim p_X$, $p(x|a) > 0$ $\forall x, a$, then:

$D(p_{Y^N|x}(\cdot|x) \| p_{Y^N}) = -\log p_X(x) + o(1)$ and $I(x;y^N) = H(p_X) + o(1)$

and $C_N = \log|\mathcal{X}| + o(1)$ as $N \to \infty$, achieved by unif prior $p^{in}_X = \frac{1}{|\mathcal{X}|}$.

Theorem 4: $S = \{\exp(a(x)t(y) - \alpha(x) + \beta(y))\}$, $y^N \sim p_{Y^N|X}(\cdot|x)$, then $D(p_{Y^N|X}(\cdot|x) \| p_{Y^N}) = \frac{1}{2}\log\frac{N J_Y(x)}{2\pi e} - \log p_X(x) + o(N)$, $N \to \infty$.

(cont. alphabet)

Corr: If $\int_\mathcal{X} \sqrt{J_Y(x)}\, dx < \infty$ then $I(x;y^N) = \frac{1}{2}\log\frac{N}{2\pi e} - D(p_X \| q^*) + \log\int_\mathcal{X} \sqrt{J_Y(x)} dx + o(1)$

where $p^*_X = \sqrt{J_Y(x)} / \int_\mathcal{X} \sqrt{J_Y(x)} dx$ ← Jeffreys's prior

Theorem 3: For $S = \{\exp(a(x)t(y) - \alpha(x) + \beta(y))\}$, $\tilde{x} = \sqrt{N}(x - \hat{x}_N(y^N))$, we have $D(p_{\tilde{X}|y^N}(\cdot|y^N) \| N(0, 1/J_Y(x))) \xrightarrow{a.s.} 0$ (achieves model capacity)

Theorem 7: If $y^N \in \mathcal{Y}^N$ generated iid wrt $q_Y \notin \mathcal{P}$, then no predictor is universal. Asymptotic loss $\bar{p}_N(y_N) \to \frac{1}{N} D(q_{Y^N} \| p_{Y^N}) \geq D_{min} > 0$

where $q_{Y^N}(y^N) = \prod q_Y(y_i)$, $D_{min} = \min_{x \in \mathcal{X}} D(q_Y \| p_Y(\cdot;x))$.

Tradeoff between class size – large ⇒ C_N large, small ⇒ not in class