12. Maximum Entropy Models

Notions of Information Theory

Elements of Information Theory Cover & Thomas 2006.

Entropy 熵

def (entropy) discrete r.v. X p.m.f. $p(x) = \Pr(X=x)$

entropy of X: $\boxed{H(X) = -\mathbb{E}[\log p(x)] = -\sum_{x\in\mathcal{X}} p(x)\cdot\log p(x)}$ ☆

property: $H(X) \geq 0$

底数(base): e, 2.

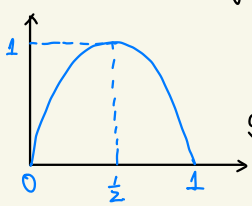
X: Bernoulli r.v.

$H(X_p) = -p\cdot\log p - (1-p)\cdot\log(1-p)$

↳ binary entropy function

$-\lg p(x)$: # of bits to represent $p(x)$

$H(X)$: average # of bits to represent X.



$0\log 0 = 0$

generally, $|\mathcal{X}| = N \geq 1$

$$H(X) = \mathbb{E}\left[\log\frac{1}{p(X)}\right] \leq \log\mathbb{E}\left[\log\frac{1}{p(X)}\right] = \log\left(\sum_{x\in\mathcal{X}}\frac{p(x)}{p(x)}\right) = \log N.$$

Upper Bound of entropy: $\log N$

衡量体系的混乱程度

entropy: lower bound on lossless data compression

Relative Entropy 相对熵 (KL-Divergence: KL散度.) ☆☆☆

Kullback Leibler

def $D(p\|q) = \mathbb{E}_p\left[\log\frac{p(X)}{q(X)}\right] = \sum_{x\in\mathcal{X}} p(x)\cdot\log\frac{p(x)}{q(x)}$

① coding length

$0\log 0 = 0$

$0\log\frac{0}{0} = 0$

$a\log\frac{a}{0} = +\infty,\ a>0$

P编码 ⇒ 最优编码的平均长度为 $H(p)$

$$D(p\|q) = \underbrace{\mathbb{E}_p\left[\log\frac{1}{q(x)}\right]}_{\text{用q的最优编码来编码P所需的平均长度}} - H(p) \quad : \text{\#additional bits}$$

② difference between distribution "距离"

Proposition. (Non-negativity) $\forall p, q.\ D(p\|q) \geq 0$. & $D(p\|q)=0$ iff $p=q$.

proof: $$-D(p\|q) = \sum_{x:p(x)>0} p(x)\cdot\log\frac{q(x)}{p(x)} \leq \log\left(\sum_{x:p(x)>0} p(x)\cdot\frac{q(x)}{p(x)}\right)$$

$$= \log\left(\sum_{x:p(x)>0} q(x)\right) \leq \log 1 = 0$$

$\Rightarrow D(p\|q) \geq 0$

△取等: $\frac{q(x)}{p(x)} = \alpha = 1$ □.

p和q相近 ⇒ $D(p\|q)$ 小.

p和q相差大 ⇒ $D(p\|q)$ 大

Relative Entropy is not a distance!

$$D(p\|q) = \sum_x p(x)\log\frac{p(x)}{q(x)}$$ asymmetric! $D(p\|q) \neq D(q\|p)$

* Jenson-Shannon divergence: $$D_{JS}(p\|q) = \frac{1}{2}D_{KL}\left(p\,\|\,\frac{p+q}{2}\right) + \frac{1}{2}D_{KL}\left(q\,\|\,\frac{p+q}{2}\right)$$

symmetric.

Cor $\boxed{\text{(log-sum inequality)}}$ $a_1,\cdots,a_n,b_1,\cdots,b_n \geq 0$, $$\boxed{\sum_{i=1}^n a_i\log\frac{a_i}{b_i} \geq \sum_{i=1}^n a_i\cdot\log\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}}$$

Proof: $\sum_{i=1}^n a_i \neq 0$. $\sum_{i=1}^n b_i \neq 0$.

scaling → $\sum_{i=1}^n a_i = \sum_{i=1}^n b_i = 1$.

LHS: $D(a\|b)$

RHS: 0

$a_i \to c a_i$: $\displaystyle\sum_{i=1}^n c a_i \cdot \log \frac{c a_i}{b_i} \geq \sum_{i=1}^n c a_i \cdot \log \frac{\sum_{i=1}^n c a_i}{\sum_{i=1}^n b_i}$

$\Rightarrow \cancel{c \sum_{i=1}^n a_i \log c} + c \sum_{i=1}^n a_i \log a_i - c \sum_{i=1}^n a_i \log b_i \geq \cancel{c \sum_{i=1}^n a_i \log c} + c \sum_{i=1}^n a_i \log \sum_{i=1}^n a_i - c \sum_{i=1}^n a_i \log \sum_{i=1}^n b_i$

$b \to c b_i$: $\displaystyle\sum_{i=1}^n a_i \log \frac{a_i}{c b_i} \geq \sum_{i=1}^n a_i \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n c b_i}$ 消 $-\sum_{i=1}^n a_i \log c$. 即可 □

Cor (Joint convexity of relative entropy) $(p, q) \mapsto D(p \| q)$

$f(x, y)$: $\alpha f(x_1, y_1) + (1-\alpha) f(x_2, y_2) \geq f(\alpha x_1 + (1-\alpha) x_2, \alpha y_1 + (1-\alpha) y_2)$

Proof: $\forall x$: $\displaystyle(\alpha p_1(x) + (1-\alpha) p_2(x)) \cdot \log \frac{\alpha p_1(x) + (1-\alpha) p_2(x)}{\alpha q_1(x) + (1-\alpha) q_2(x)}$

$\displaystyle\leq \alpha p_1(x) \cdot \log \frac{\alpha p_1(x)}{\alpha q_1(x)} + (1-\alpha) \cdot p_2(x) \cdot \log \frac{(1-\alpha) p_2(x)}{(1-\alpha) q_2(x)}$, (log-sum ineq.)

$\Rightarrow$ sum over x: $D(\alpha p_1(x) + (1-\alpha) p_2(x) \| \alpha q_1(x) + (1-\alpha) q_2(x))$

$\leq \alpha D(p_1 \| q_1) + (1-\alpha) \cdot D(p_2 \| q_2)$ □.

Cor (Concavity of entropy) $p \mapsto H(p)$

Proof: fixed p_0. $\displaystyle D(p \| p_0) = \sum_{x \in \mathcal{X}} p(x) \cdot \log p(x) - \sum_{x \in \mathcal{X}} p(x) \log p_0(x)$

$\Rightarrow H(p) = \underbrace{-D(p \| p_0)}_{\text{convex}} - \underbrace{\sum_{x \in \mathcal{X}} p(x) \log p_0(x)}_{\text{linear (affine)}}$ concave. □

Proposition (Pinsker's inequality)

$D(p \| q) \geq \frac{1}{2} \|p - q\|_1^2$

(TV) total variance distance 全变差距离

NORM

$\left(|p(x_1) - q(x_1)| + |p(x_2) - q(x_2)| + \cdots + |p(x_{|\mathcal{X}|}) - q(x_{|\mathcal{X}|})| \right)$

Proof: $A=\{a_0, a_1\}$ $\quad p_0 = p(a_0)$. $\quad q_0 = q(a_0)$

fix $p_0 \in [0,1]$.

$$f(q_0) = p_0 \cdot \log\frac{p_0}{q_0} + (1-p_0)\cdot\log\frac{1-p_0}{1-q_0} - 2(p_0-q_0)^2.$$

$$\Rightarrow f(p_0) = 0 \quad q_0 \in (0,1)$$

$$f'(q_0) = p_0 \cdot \frac{-\frac{p_0}{q_0^2}}{\frac{p_0}{q_0}} + (1-p_0)\cdot\frac{+\frac{1-p_0}{(1-q_0)^2}}{\frac{1-p_0}{1-q_0}} + 4(p_0-q_0) = -\frac{p_0}{q_0} + \frac{1-p_0}{1-q_0} + 4(p_0-q_0)$$

$$= \frac{q_0-p_0}{q_0(1-q_0)} + 4(p_0-q_0) = (q_0-p_0)\cdot\left(\frac{1}{q_0(1-q_0)} - 4\right)$$

$$q_0(1-q_0) \le \frac{1}{4} \Rightarrow \begin{cases} f'(q_0) \ge 0. & q_0 \ge p_0 \\ f'(q_0) < 0. & q_0 \le p_0 \end{cases} \Rightarrow f(q_0) \ge f(p_0) = 0.$$

$$f(q_0) = D(p\|q) - 2(p_0-q_0)^2$$

$$= D(p\|q) - \frac{1}{2}\left(|p_0-q_0|^2 + |(1-p_0)-(1-q_0)|^2\right)$$

$$= D(p\|q) - \frac{1}{2}\|p-q\|_1^2 \ge 0$$

$\underline{p-q}$. $\quad \tilde{p}(a_0) = \sum_{x\in a_0} p(x)$. $\quad \tilde{q}(a_0) = \sum_{x\in a_0} q(x)$

$$a_0 = \{x \in \mathcal{X} : p(x) \ge q(x)\}$$

$$a_1 = \{x \in \mathcal{X} : p(x) < q(x)\}.$$

多 → 二

$$D(p\|q) = \sum_{x\in a_0} p(x)\log\frac{p(x)}{q(x)} + \sum_{x\in a_1} p(x)\cdot\log\frac{p(x)}{q(x)}$$

$$\ge p(a_0)\cdot\log\frac{p(a_0)}{q(a_0)} + p(a_1)\cdot\log\frac{p(a_1)}{q(a_1)} = D(\tilde{p}\|\tilde{q})$$

$$\|\tilde{p}-\tilde{q}\|_1 = (p(a_0)-q(a_0)) + (q(a_1)-p(a_1))$$

$$= \sum_{x\in\mathcal{X}} |p(x)-q(x)| = \|p-q\|_1.$$

$$\Rightarrow D(p\|q) \ge D(\tilde{p}\|\tilde{q}) \ge \frac{1}{2}\|\tilde{p}-\tilde{q}\|_1 \ge \frac{1}{2}\|p-q\|_1$$

□

Coupling 耦合

def (Conditional Relative Entropy) $p, q : \mathcal{X} \times \mathcal{Y}$ $\quad r : \mathcal{X}$

joint $\longrightarrow$ marginal

$$\mathbb{E}_{X\sim r}\left[D\big(p(\cdot|X) \,\|\, q(\cdot|X)\big)\right] = \sum_{x\in\mathcal{X}} r(x) \cdot \sum_{y\in\mathcal{Y}} p(y|x) \cdot \log \frac{p(y|x)}{q(y|x)}$$

$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} r(x)\cdot p(y|x) \cdot \log \frac{r(x)\cdot p(y|x)}{r(x)\cdot q(y|x)} = D(\tilde{p} \,|\, \tilde{q})$$

$$\tilde{p}(x,y) = r(x)\cdot p(y|x)$$

$$\tilde{q}(x,y) = r(x)\cdot q(y|x)$$

Mutual Information 互信息

X, Y

def (MI) $I(X,Y) = D\big(\underbrace{p(x,y)}_{\text{joint}} \,\|\, \underbrace{p(x)\cdot p(y)}_{\text{marginal}}\big)$

$$= \mathbb{E}_{p(x,y)}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)\cdot \log \frac{p(x,y)}{p(x)p(y)}$$

measure of the amount of information that each r.v. can provide about the other

$I(X,Y) = 0 \iff p(x,y) = p(x)\cdot p(y) \iff$ X Y 独立

def (Normalized MI) $\tilde{I}(X,Y) = \dfrac{I(X,Y)}{\big(I(X,X)+I(Y,Y)\big)/2}$.

$\tilde{I}(X,X) = 1.$

补充: $\mathcal{X}$, input space.

$\downarrow$

unnormalized relative entropy $\tilde{D}(p\|q) = \sum_{x\in\mathcal{X}} p(x)\cdot \log \frac{p(x)}{q(x)} + (q(x) - p(x))$

P.9 $\mathcal{X}$ (Simplex) : $\tilde{D}(p\|q) = D(p\|q)$

<u>def</u> (<u>Bregman divergence</u>) F: <u>convex</u> differentiable.

defined convex (<u>open</u>) set $\mathcal{C}$ in a Hilbert Space $\mathbb{H}$:

$$B_F(x\|y) = F(x) - F(y) - \langle \nabla F(y), x-y \rangle$$

<u>examples</u>: $F(x) = \|x\|^2$. $\Rightarrow B_F(x\|y) = \|x-y\|^2$

$$F(x) = \sum_{i\in I} x(i)\log x(i) - x(i) \Rightarrow B_F(x\|y) = \tilde{D}(x\|y)$$

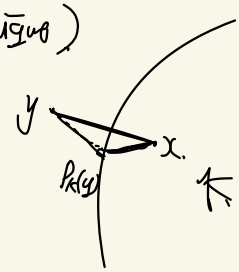
<u>Proposition</u>:

1. $B_F(x\|y) \geq 0$
2. $\langle \nabla F(x) - \nabla F(y), x-z \rangle = B_F(x\|z) + B_F(z\|x) - B_F(z\|y)$.
3. $B_F(x\|y)$ convex with its first argument
4. linearity: $B_{\alpha F + \beta G} = \alpha B_F + \beta \cdot B_G$.
5. (strictly convex) $\Rightarrow$ Pythagorean theorem:

$$B_F(x\|y) \geq B_F(x\|P_K(y)) + B_F(P_K(y)\|y)$$

$$P_K(y) = \underset{x\in K}{\arg\min}\, B_F(x\|y)$$

<u>projection</u> (unique)



熵. 相对熵 - 条件相对熵 . ....

↓ Chapter 12 ↓ Chapter 13

Fenchel Duality

duality

primal

$$\boxed{\begin{aligned} \min\ & f(x) \\ \text{s.t.}\ & g_i(x) \le 0 \\ & h_i(x) = 0 \end{aligned}}$$

convex & differentiable

$$\Rightarrow \quad L = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{n} \beta_j h_j(x)$$

$$g(\alpha, \beta) = \inf L$$

dual

$$\boxed{\begin{aligned} \max_{\alpha,\beta}\ & g(\alpha,\beta) \\ \text{s.t.}\ & \alpha \ge 0 \end{aligned}}$$

$$d^* \le p^*$$

求梯度

KKT条件
$$\begin{cases} \nabla_x L = 0 \\ \nabla_\alpha L \le 0 \\ \nabla_\beta L = 0 \\ \sum_{i=1}^{m} \alpha_i g_i(x) = 0 \end{cases}$$

constraint

f: non-differentiable

take infinite value

定义域 $\mathrm{dom}(f) = \{x \in \mathcal{X} : f(x) < +\infty\}$

def (proper function) a convex function is said to be proper if it takes in $(-\infty, +\infty]$ and if it is not uniformly equal to $+\infty$

次梯度

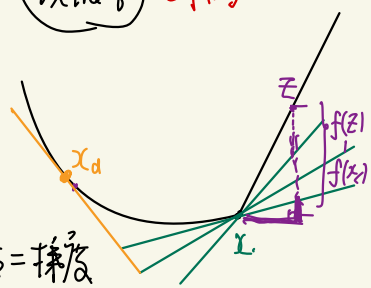
def (sub-gradient) $f: \mathcal{X} \to (-\infty, +\infty]$ proper function

a vector $g \in \mathcal{X}$ is a sub-gradient of f at point $x \in \mathrm{dom}(f)$

if $\quad f(z) \ge f(x) + \langle z - x, g \rangle \quad \forall z \in \mathcal{X}$ 次梯度

The set of all subgradients is called sub-differential 次微分 $\partial f(x)$

直线在函数图像的下方即可!



Lemma If f is differentiable at $x \in \mathrm{dom}(f)$

then $\partial f(x) = \{\nabla f(x)\}$ △ 可微: 次梯度 = 梯度

Proof: $g \in \partial f(x)$. $\forall \varepsilon \in \mathbb{R}$

$$f(x+\varepsilon\cdot(\nabla f(x)-g)) \geq f(x) + \varepsilon\cdot\langle \nabla f(x)-g, g\rangle$$

Taylor's expansion: $f(x+\varepsilon\cdot(\nabla f(x)-g)) = f(x) + \varepsilon\cdot\langle \nabla f(x), \nabla f(x)-g\rangle + o(\varepsilon\|\nabla f(x)-g\|)$

$$\Rightarrow \varepsilon\cdot\|\nabla f(x)-g\|^2 \leq o(\varepsilon\|\nabla f(x)-g\|)$$

GD $\Rightarrow$ Sub-GD

$$\Rightarrow \|\nabla f(x)-g\| = o(1) \Rightarrow \nabla f(x) = g. \quad \square$$

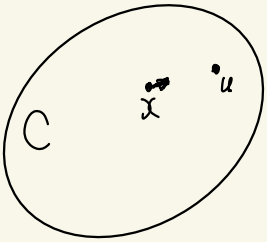
Lemma. f proper. x^* global minimizer of f. $\Leftrightarrow$ $0 \in \partial f(x^*)$

Proof: $(-\infty, +\infty]$ $m(f(x)=+\infty)=0 \Rightarrow f(x^*) \neq +\infty \Rightarrow x^* \in \mathrm{dom}(f)$

$\forall z \in \mathcal{X}$. $f(z) \geq f(x^*) + \langle z-x^*, 0\rangle \Rightarrow 0 \in \partial f(x^*)$ $\square$

Core set $C \subseteq \mathcal{X}$ "核心"

def(core) $\mathrm{Core}(C) = \{x \in C : \forall u \in \mathcal{X}. \exists \varepsilon > 0. \mid \forall t \in [0, \varepsilon]. (x+tu) \in C\}$

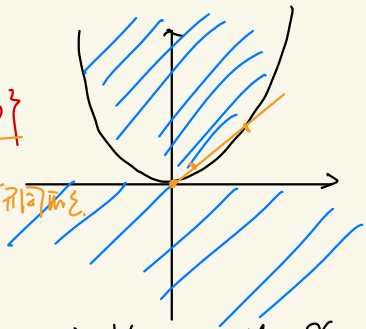


u 提供方向

interior $= \{x \in C : \exists \varepsilon > 0. B(x, \varepsilon) \subseteq C\}$

$\mathrm{interior}(C) \subseteq \mathrm{core}(C)$

$C = \{x \in \mathbb{R}^2 : x_2 \geq x_1^2 \text{ or } x_2 \leq 0\}$



$0 \in \mathrm{core}(C)$. 不同的方向可以有不同的 ε

$0 \notin \mathrm{int}(C)$ 统一的 ε

性质: (1) h convex if $\exists x_0 \in \mathrm{core}(\mathrm{dom}(h))$ s.t. $h(x_0) > -\infty \Rightarrow h(x) > -\infty$. $\forall x \in \mathcal{X}$

(2) h proper (convex) $\Rightarrow h \in (-\infty, +\infty] \Rightarrow \forall x \in \mathrm{core}(\mathrm{dom}(h))$. $\partial h(x) \neq \emptyset$

Conjugate function 共轭 (Fenchel's conjugate)

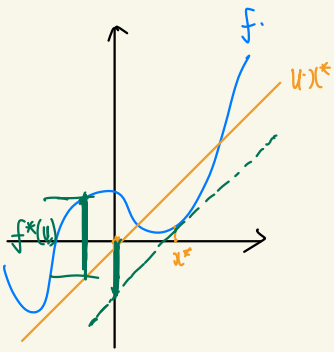
def $\mathcal{X}$ Hilbert Space. $f: \mathcal{X} \to [-\infty, +\infty]$

$$\Updownarrow$$

$$f^*: \mathcal{X} \to [-\infty, +\infty]$$

Legendre transformation

$$\boxed{f^*(u) = \sup_{x \in \mathcal{X}} \{\langle u, x\rangle - f(x)\}}$$



① f^*: convex

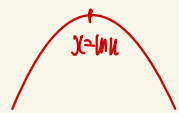
② $\exists f(x) < +\infty \Rightarrow f^* > -\infty$

③ $f \le g \Rightarrow g^* \le f^*$

④ $f^{**} = f$.

e.g.1. $f(x) = e^x$. $\quad (\langle u, x\rangle - f(x))' = (xu - e^x)' = u - e^x = 0 \Rightarrow x = \ln u$

$$f^*(u) = \begin{cases} u\ln u - u, & u > 0 \\ 0, & u = 0 \end{cases}$$



e.g.2 extended KL divergence: $\quad f(p) = \begin{cases} D(p\|p_0) & \text{if } p \in \Delta \\ +\infty & \text{o.w.} \end{cases}$ $\quad$ "simplex"

$\tilde{D}(p\|p_0)$

$f: \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$

$f^*: \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$. $\quad \forall q \in \mathbb{R}^{\mathcal{X}}$. $\quad f^*(q) = \log\left(\sum_{x \in \mathcal{X}} p_0(x) \cdot e^{q(x)}\right)$

"log-sum-exp"

softmax

机器学习

Proof: $\forall q \in \mathbb{R}^{\mathcal{X}}$:

$$\sup_{p \in \mathbb{R}^{\mathcal{X}}} \left(\langle p, q\rangle - \tilde{D}(p\|p_0)\right) = \sup_{p \in \Delta} \left(\langle p, q\rangle - D(p\|p_0)\right)$$

$$\bar{q}(x) = \frac{p_0(x) \cdot e^{q(x)}}{\sum_{x \in \mathcal{X}} p_0(x) \cdot e^{q(x)}} = \frac{p_0(x) \cdot e^{q(x)}}{\mathbb{E}_{p_0}[e^q]}$$

$\forall p \in \Delta$: $\quad \langle p, q\rangle - D(p\|p_0) = \sum_x p(x) q(x) - \sum_x p(x) \cdot \log \frac{p(x)}{p_0(x)}$

$$= \mathbb{E}_p[\log(e^q)] - \mathbb{E}_p[\log \frac{p}{p_0}]$$

$$= \mathbb{E}_p[\log \frac{p_0 \cdot e^q}{p}] = -D(p \| \bar{q}) + \log \mathbb{E}_{p_0}[e^q]$$

equality takes when: $p = \bar{q}$

$$\leq \log \mathbb{E}_{p_0}[e^q] = \text{const with } q.$$

$$\Rightarrow f^*(q) = \log \mathbb{E}_{p_0}[e^q] = \log \sum_{x \in \mathcal{X}} p_0(x) \cdot e^{q(x)}. \quad \square.$$

Lemma (Fenchel's inequality)

$$f^*(u) = \sup_x \{\langle u, x\rangle - f(x)\}$$

$$\forall x, u, \quad \boxed{f(x) + f^*(u) \geq \langle u, x\rangle}$$

Thm (Fenchel Duality Theorem) $\mathcal{X}, \mathcal{Y}$ Hilbert Spaces

$f: \mathcal{X} \to (-\infty, +\infty]$. $g: \mathcal{Y} \to (-\infty, +\infty]$ proper (convex)

$A: \mathcal{X} \to \mathcal{Y}$ bounded linear map (有界线性算子)

primal: $p^* = \inf_{x \in \mathcal{X}} \{f(x) + g(Ax)\}$

dual: $d^* = \sup_{y \in \mathcal{Y}} \{-f^*(A^*y) - g^*(-y)\}$

weak duality: $p^* \leq d^*$. ✓

strong duality: $p^* = d^*$ $\Leftarrow$ ① $0 \in \operatorname{core}(\operatorname{dom}(g) - A(\operatorname{dom}(f)))$

or

② stronger: $A(\operatorname{dom}(f)) \cap \operatorname{cont}(\operatorname{dom}(g)) \neq \emptyset$ (连续)

A^*: adjoint Operator $\langle Ax, y\rangle = \langle x, A^*y\rangle$

Proof: $f(x) + f^*(A^*y) \geq \langle A^*y, x\rangle = \langle y, Ax\rangle = -\langle -y, Ax\rangle \geq -g^*(-y) - g(Ax)$

Fenchel's inequality

$$\Rightarrow f(x) + g(Ax) \geq -f^*(A^*y) - g^*(-y) \quad \forall x, y$$

左边 take inf. 右边 take sup $\Rightarrow p^* \geq d^*$. (weak duality)

Consider $h(u) = \inf\{f(x) + g(Ax+u)\}$ $\quad h(0) = p^*$

convex

$u \in \mathrm{dom}(h) \Leftrightarrow \exists x.\ f(x) + g(Ax+u) < +\infty$

$\Leftrightarrow \begin{cases} f(x) < +\infty \Leftrightarrow x \in \mathrm{dom}(f) \\ g(Ax+u) < +\infty \Leftrightarrow Ax+u \in \mathrm{dom}(g) \end{cases}$

$\Leftrightarrow u \in \mathrm{dom}(g) - A\cdot\mathrm{dom}(f)$

$\Rightarrow \mathrm{dom}(h) = \mathrm{dom}(g) - A\cdot\mathrm{dom}(f)$

① $p^* = -\infty \Rightarrow d^* = -\infty \Rightarrow p^* = d^*$ trivially holds.

② $p^* > -\infty$. if $0 \in \mathrm{core}(\mathrm{dom}(g) - A\cdot\mathrm{dom}(f)) = \mathrm{core}(\mathrm{dom}(h))$

$\Rightarrow 0 \in \mathrm{dom}(h)$ $\quad p^* = h(0) < +\infty \quad \Rightarrow p^* \in \mathbb{R}$

由性质1, h takes value in $(-\infty, +\infty]$. $\Rightarrow$ proper

由性质2 $h(0)$ 存在次梯度. 记作 $-y$

by definition of subgradients, $\forall x \in \mathcal{X}, u \in \mathcal{U}$

$p^* = h(0) \leq h(u) + \langle y, u\rangle \leq f(x) + g(Ax+u) + \langle y, u\rangle$

adjoint

$= (f(x) - \langle A^*y, x\rangle) + (g(Ax+u) + \langle y, u\rangle + \langle A^*y, x\rangle)$

$= (f(x) - \langle A^*y, x\rangle) + (g(Ax+u) - \langle -y, Ax+u\rangle)$

inf $\quad$ inf

take infimum. $p^* = h(0) \leq -f^*(A^*y) - g^*(-y) \leq d^* \leq p^* \Rightarrow p^* = d^*$

if $(A\cdot\mathrm{dom}(f)) \cap \mathrm{cont}(\mathrm{dom}(g)) \neq \phi \Rightarrow u \in (A\,\mathrm{dom}(f)) \cap \mathrm{cont}(\mathrm{dom}(g))$

$u = Ax$. $x \in \mathrm{dom}(f)$ $\quad \Rightarrow 0 = u - Ax \in \mathrm{dom}(g) - A\cdot\mathrm{dom}(f)$

$u \in \mathrm{dom}(g)$

$u \in \mathrm{cont}(\mathrm{dom}(g)) \Rightarrow$ continuous! $\Rightarrow 0 \in \mathrm{core}(\mathrm{dom}(g) - A\,\mathrm{dom}(f)) \Rightarrow p^* = d^*$. $\square$

Density-Estimation 密度估计

$S = \{x_1, x_2, \cdots, x_m\}$ i.i.d sampling from unknown distribution $\mathcal{D}$

unsupervised learning

Goal: 通过 S 找到 $\mathcal{D}$ (from a family of possible distributions $\mathcal{P}$) Θ

① 如何选 p ② 如何选 $\mathcal{P}$

p

Maximum Likelihood Estimation 极大似然估计 (Frequentist)

$$\star \quad p^* = \underset{p \in \mathcal{P}}{\operatorname{argmax}} \prod_{i=1}^{m} p(x_i) = \underset{p \in \mathcal{P}}{\operatorname{argmax}} \sum_{i=1}^{m} \log p(x_i)$$

似然 likelihood

$p(x \mid \Theta)$

对数似然

$\hat{D}$: empirical distribution corresponding to the sample S

$$p^* = \underset{p \in \mathcal{P}}{\operatorname{argmin}} D(\hat{D} \| p)$$

$$D(p \| q) = \sum_x p(x) \cdot \log \frac{p(x)}{q(x)}$$

$$D(\hat{D} \| p) = \sum_x \hat{D}(x) \cdot \log \hat{D}(x) - \sum_x \hat{D}(x) \cdot \log p(x)$$

minimize

$$= -H(\hat{D}) - \sum_x \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{x = x_i} \right) \cdot \log p(x)$$

$$= -H(\hat{D}) - \sum_{i=1}^{m} \sum_x \frac{\mathbb{1}_{x = x_i}}{m} \cdot \log p(x)$$

$$= -H(\hat{D}) - \sum_{i=1}^{m} \frac{\log p(x)}{m} \rightarrow \text{maximize}$$

MLE: log-likelihood ↑ ⟺ relative-entropy ↓

eg $S=\{x_1,\cdots,x_m\}$ $x_i\in\{\text{head},\text{tail}\}$ $p_0\in[0,1]$

$\vec{p}=(p,1-p)$ $p\in[0,1]$

$\mathcal{P}$ 可度量的集合

MLE $\Leftrightarrow$ $D(\hat{D}\|p)=0 \Rightarrow p=\hat{D}=(\hat{p}_s,1-\hat{p}_s)=\left(\frac{n_h}{m},1-\frac{n_h}{m}\right)$

Maximum a Posteriori (MAP) 最大后验概率估计

Bayesian

$$p^*=\underset{p\in\mathcal{P}}{\operatorname{argmax}}\ \Pr(p\mid S)=\underset{p\in\mathcal{P}}{\operatorname{argmax}}\ \frac{\Pr(S\mid p)\cdot\Pr(p)}{\Pr(S)}=\underset{p\in\mathcal{P}}{\operatorname{argmax}}\ \Pr(S\mid p)\cdot\Pr(p)$$

Bayes Theorem

似然 先验 prior

$p\in\mathcal{P}$

MaxEnt Model to solve density estimation problem:

最大熵模型

Assume: learner has access to a feature mapping $\Phi: \mathcal{X}\to\mathbb{R}^N$ $\|\Phi\|_\infty\le r$

$\Rightarrow \|x\|_K^2=K(x,x)\le r^2$

Principle ☆: $\mathcal{P}$

① as agnostic as possible ☆ 不可知型

as close as possible to the uniform distribution

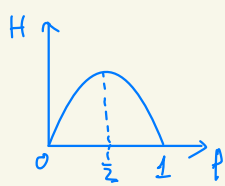
more generally: as close as possible to a prior p_0

均匀分布 概率最大

3 硬币 概率

$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

$H(p)=-\sum_x p(x)\cdot\log p(x)$



$\Rightarrow$ $\mathcal{P}$ family unknown (不需要指定分布函数族☆☆)

$\Rightarrow$ avoid bias 偏见

② $$\boxed{\left\|\mathbb{E}_{x\sim p}[\Phi(x)]-\mathbb{E}_{x\sim\hat{D}}[\Phi(x)]\right\|_\infty\le\lambda.}$$

通过p来指导采样的 x ($\Phi(x)$) 与经验分布 $\hat{D}$ 尽可能接近!

⇒ 匹配经验数据

⇒ Optimization Problem:

$$\boxed{\begin{aligned} \min_{p\in\Delta} \quad & D(p\,\|\,p_0) \\ \text{s.t.} \quad & \left\| \mathbb{E}_{x\sim p}[\Phi(x)] - \mathbb{E}_{x\sim\hat{D}}[\Phi(x)] \right\|_\infty \le \lambda \end{aligned}}$$

prior (uniform)

Δ simplex of all distributions over $\mathcal{X}$ (合法的分布函数)

$$\Delta = \left\{ \theta_0 u_0 + \cdots + \theta_k u_k \;\middle|\; \sum_{i=0}^{k} \theta_i = 1,\ \theta_i \ge 0,\ i = 0, \cdots, k \right\}$$

minimize

$$D(p\,\|\,p_0) = \sum_{x\in\mathcal{X}} p(x)\cdot\log\frac{p(x)}{p_0(x)}$$

$$= -\sum_{x\in\mathcal{X}} p(x)\log p_0(x) + \sum_{x\in\mathcal{X}} p(x)\cdot\log p(x)$$

$$= \log|\mathcal{X}| \qquad p_0(x) = \frac{1}{|\mathcal{X}|}$$

$-H(p)$ ← maximize 最大熵

~~最~~大熵分布代表信息量最少的分布!

$$\boxed{F(p) = \tilde{D}(p\,\|\,p_0) + I_C\left(\mathbb{E}_p[\Phi]\right)}$$ primal

惩罚不匹配

无约束优化问题

$$\tilde{D}(p\,\|\,p_0) = \begin{cases} D(p\,\|\,p_0) & p\in\Delta \\ +\infty & p\notin\Delta \end{cases}$$

$$I_K(x) = \begin{cases} 0, & x\in K \\ +\infty, & x\notin K \end{cases}$$

$$C = \left\{ u : \left\| u - \mathbb{E}_{\hat{D}}[\Phi] \right\|_\infty \le \lambda \right\}$$

def (Gibbs distribution)

$$\boxed{p_w[x] = \frac{p_0[x]\, e^{w\cdot\Phi(x)}}{Z(w)}}$$

← normalization factor (partition function)

$$Z(w) = \sum_{x \in \mathcal{X}} p_0[x] \cdot e^{w \cdot \Phi(x)}$$ softmax / canonical / categorical

$$\Rightarrow \boxed{G(w) = \frac{1}{m} \sum_{i=1}^{m} \log \frac{p_w[x_i]}{p_0[x_i]} - \lambda \cdot \|w\|_1}$$ dual

凸函数

Thm (Maxent duality) $\boxed{\sup_{w \in \mathbb{R}^N} G(w) = \min_p F(p)}$ ☆ strong duality

Proof: by Fenchel Duality Theorem

$$F(p) = \underbrace{\tilde{D}(p \| p_0)}_{f} + \underbrace{I_c(\mathbb{E}_p[\Phi])}_{g}$$

$$G(w) = \frac{1}{m} \sum_{i=1}^{m} \log \frac{p_w[x_i]}{p_0[x_i]} - \lambda \cdot \|w\|_1$$

模仿:

$$f(p) = \tilde{D}(p \| p_0), \quad g(u) = I_c(u)$$

$$Ap = \mathbb{E}_p[\Phi] = \sum_{x \in \mathcal{X}} p(x) \cdot \Phi(x)$$

$$p^* = \inf_x \{ f(x) + g(Ax) \}$$

$$d^* = \sup_y \{ -f^*(A^* y) - g^*(-y) \}$$

$$f^*(u) = \sup_x \{ \langle u, x \rangle - f(x) \}$$

strong duality:

① $\mathrm{cone}(\mathrm{dom}\, g - A \cdot \mathrm{dom}\, f) \ni 0$

or

② $A \cdot \mathrm{dom}\, f \cap \mathrm{cont}(\mathrm{dom}\, g) \neq \phi$

A: bounded $\quad \|Ap\| = \|p \cdot \Phi\| \leq \|p\|_1 \cdot \|\Phi\|_\infty \leq r$.

A adjoint operator A^*: $A^* w = w \cdot \Phi$ $\quad \langle Ap, w \rangle = \langle p, A^* w \rangle$

consider $u_0 \in \mathbb{R}^N$: $u_0 = \mathbb{E}_{x \sim \hat{D}}[\Phi(x)] = A\hat{D}$. $\quad \hat{D} \in \Delta = \mathrm{dom}(f)$

$\Rightarrow u_0 \in A\, \mathrm{dom}(f)$

$u_0 \in I_c \; (g(u_0) = 0) \Rightarrow u_0 \in \mathrm{cont}(\mathrm{dom}(g))$

$\Rightarrow u_0 \in A\, \mathrm{dom} f \cap \mathrm{cont}(\mathrm{dom}(g))$

$\Rightarrow A\, \mathrm{dom} f \cap \mathrm{cont}(\mathrm{dom}(g)) \neq \phi \Rightarrow$ strong duality holds!

$$f^*(q) = \log\left(\sum_{x \in \mathcal{X}} p_0(x) \cdot e^{q(x)} \right)$$

$$g^*(w) = \sup_u \left(w \cdot u - I_c(u) \right) = \sup_{u \in C} (w \cdot u) = \sup_{\|u - \mathbb{E}_p[\Phi]\|_\infty \leq \lambda} (w \cdot u)$$

$$= \sup_{\|w\|_\infty \le \lambda} \left(w\cdot\left(u + \mathbb{E}_{\hat{D}}[\Phi]\right)\right) = \sup_{\|w\|_\infty \le \lambda} (w\cdot u) + w\cdot \mathbb{E}_{\hat{D}}[\Phi]$$

$$= \lambda\cdot\|w\|_1 + \mathbb{E}_{\hat{D}}[w\cdot\Phi] \quad \text{(dual norm)}$$

$$-f^*(A^*w) - g^*(-w) = -\log\left(\sum_{x\in\mathcal{X}} p_0(x)\cdot e^{w\cdot\Phi(x)}\right) - \lambda\cdot\|w\|_1 + \mathbb{E}_{\hat{D}}[w\cdot\Phi]$$

$$= -\log Z(w) + \frac{1}{m}\sum_{i=1}^{m} w\cdot\Phi(x_i) - \lambda\cdot\|w\|_1$$

$$= \frac{1}{m}\sum_{i=1}^{m} \log\frac{e^{w\cdot\Phi(x_i)}}{Z(w)} - \lambda\|w\|_1$$

$$= \frac{1}{m}\sum_{i=1}^{m} \log\frac{p_w[x_i]}{p_0[x_i]} - \lambda\|w\|_1 = G(w)$$

$$p_w[x] = \frac{p_0[x]\, e^{w\cdot\Phi(x)}}{Z(w)}$$

by Fenchel's Duality Theorem, $\min_p F(p) = \sup_w G(w)$. $\square$

Cor $p^* = \arg\min_p F(p)$. $d^* = \sup_w G(w)$. $\forall \varepsilon > 0$

$\forall w \in \mathbb{R}^N$ s.t. $|G(w) - d^*| < \varepsilon$. then $D(p^*\|p_w) \le \varepsilon$

Proof: $G(w) - D(p^*\|p_0) + D(p^*\|p_w)$

$$= \mathbb{E}_{\hat{D}}\left[\log\frac{p_w[x]}{p_0[x]}\right] - \lambda\cdot\|w\|_1 - \mathbb{E}_{p^*}\left[\log\frac{p^*[x]}{p_0[x]}\right] + \mathbb{E}_{p^*}\left[\log\frac{p^*[x]}{p_w[x]}\right]$$

$$= -\lambda\|w\|_1 + \mathbb{E}_{\hat{D}}\left[\log\frac{p_w[x]}{p_0[x]}\right] - \mathbb{E}_{p^*}\left[\log\frac{p_w[x]}{p_0[x]}\right]$$

$$= -\lambda\cdot\|w\|_1 + w\cdot\left(\mathbb{E}_{\hat{D}}[\Phi] - \mathbb{E}_{p^*}[\Phi]\right)$$

constraints: $I_C\left(\mathbb{E}_{p^*}[\Phi]\right) = 0$

$$\Rightarrow \left\|\mathbb{E}_{\hat{D}}[\Phi] - \mathbb{E}_{p^*}[\Phi]\right\|_\infty \le \lambda$$

$$\le -\lambda\cdot\|w\|_1 + \|w\|_1\cdot\left\|\mathbb{E}_{\hat{D}}[\Phi] - \mathbb{E}_{p^*}[\Phi]\right\|_\infty \le 0$$

if w is ε-solution of dual problem,

then $D(p^*\|p_w) \le \varepsilon$

Pinsker's inequality $\Rightarrow \|p^* - p_w\|_1 \le \sqrt{2\varepsilon}$.

(p_w $\sqrt{2\varepsilon}$-close (L_1-norm) to the primal optimal)

$$\Rightarrow \boxed{\inf_{w\in\mathbb{R}^N} \lambda\|w\|_1 - \frac{1}{m}\sum_{i=1}^{m}\log p_w[x_i]}$$

SGD

Coordinate Descent

GIS (Generalized Iterative Scaling)

IIS (Improved ~)

BFGS

$\Rightarrow \forall w \in \mathbb{R}^N,\ D(p^* \| p_w) \le D(p^* \| p_0) - G(w)$

By assumption, $|G(w) - d^*| = |G(w) - \sup_w G(w)| \le \varepsilon$

$\sup_w G(w) = D(p^* \| p_0)$ (strong duality)

$\Rightarrow D(p^* \| p_w) \le \varepsilon$ □

Optimization: $\inf_w \lambda \|w\|_1 - \frac{1}{m}\sum_{i=1}^{m} \lg p_w[x_i]$

$\lambda = 0 \Rightarrow \max \sum_{i=1}^{m} \log p_w[x_i] = \max \prod_{i=1}^{m} p_w[x_i]$

(Maximum Likelihood Estimation)
极大似然估计

$\lambda \ne 0 \Rightarrow$ regularized MLE!

需要假设分布族 $\mathcal{P}$: Gibbs Distribution

神奇. 为什么? the measure of closeness!

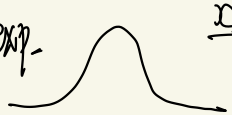
$D(p \| p_0)$ relative entropy $\longrightarrow$ Gibbs Distribution.

Gibbs Distribution $p_w[x] = \frac{p_0[x] e^{\langle w, \Phi(x) \rangle}}{Z(w)}$ Φ feature

e.g. monomial. degree ≤ 2

$x_j x_k$. x_j.

unimodel (单峰): Gaussian. Cauchy. T. χ^2. exp.



bimodel (双峰): Beta.

$p[(x_1, x_2)] = \frac{e^{-(x_1^2 + x_2^2)}}{Z}$

$p[(x_1, x_2)] = \frac{e^{-(x_1^2 + x_2^2) + x_1^2 - x_2^2}}{Z}$

Gibbs 分布是一类很丰富的分布!

extension:

① 其他 measure of closeness

KL divergave $\Rightarrow$ Bregman divergave

$\tilde{D}(p \| p_0) \Longrightarrow B_{\Phi}(p \| p_0)$

② L_2-regularization

$\min_w \lambda \cdot \|w\|_2^2 - \frac{1}{m}\sum_{i=1}^{m} \log p_w[x_i]$ 不稀疏但更精确

Generalization Bound

$$\mathcal{L}_{\mathcal{D}}(w) = \mathbb{E}_{x\sim\mathcal{D}}\left[-\log P_w[x]\right]$$

goal　　learned.

loss 越大，P_w 和 $\mathcal{D}$ 相差越大，学得越差

$$\mathcal{L}_S(w) = \mathbb{E}_{x\sim\hat{\mathcal{D}}}\left[-\log P_w[x]\right] = -\frac{1}{m}\sum_{i=1}^{m}\log P_w[x_i]$$

所以解出的 $\hat{w}$

Thm

$$\mathcal{L}_{\mathcal{D}}(\hat{w}) \le \inf_w \mathcal{L}_{\mathcal{D}}(w) + 2\|w\|_1\left(2\,\mathcal{R}_m(\mathcal{H}) + r\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right)$$

Proof:

$$\left\|\mathbb{E}_{x\sim\mathcal{D}}[\Phi(x)] - \mathbb{E}_{x\sim\hat{\mathcal{D}}}[\Phi(x)]\right\|_\infty \le 2\cdot\mathcal{R}_m(\mathcal{H}) + r\sqrt{\frac{\log\frac{1}{\delta}}{2m}} = \lambda$$

$$\mathcal{L}_{\mathcal{D}}(\hat{w}) - \mathcal{L}_S(\hat{w}) = \hat{w}\cdot\left(\mathbb{E}_{\hat{\mathcal{D}}}[\Phi] - \mathbb{E}_{\mathcal{D}}[\Phi]\right)$$

$$\le \|\hat{w}\|_1\cdot\left\|\mathbb{E}_{\hat{\mathcal{D}}}[\Phi] - \mathbb{E}_{\mathcal{D}}[\Phi]\right\|_\infty \le \lambda\cdot\|\hat{w}\|_1 \quad (\text{whatever } \hat{w})$$

$\hat{w}$: minimizer.

$$\hat{w} = \arg\min_w\left(\lambda\|w\|_1 - \frac{1}{m}\sum_{i=1}^{m}\log P_w(x_i)\right) = \arg\min_w\left(\lambda\cdot\|w\|_1 - \mathcal{L}_S(w)\right)$$

$$\mathcal{L}_{\mathcal{D}}(\hat{w}) - \mathcal{L}_{\mathcal{D}}(w) = \mathcal{L}_{\mathcal{D}}(\hat{w}) - \mathcal{L}_S(\hat{w}) + \mathcal{L}_S(\hat{w}) - \mathcal{L}_{\mathcal{D}}(w)$$

$$\le \lambda\cdot\|\hat{w}\|_1 + \mathcal{L}_S(\hat{w}) - \mathcal{L}_{\mathcal{D}}(w)$$

$$\le \lambda\cdot\|w\|_1 + \mathcal{L}_S(w) - \mathcal{L}_{\mathcal{D}}(w) \le 2\cdot\lambda\|w\|_1 \quad \forall w.$$

$$\Rightarrow \mathcal{L}_{\mathcal{D}}(\hat{w}) \le \inf_w \mathcal{L}_{\mathcal{D}}(w) + 2\|w\|_1\left(2\,\mathcal{R}_m(\mathcal{H}) + r\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right)$$

$$\mathcal{L}_{\mathcal{D}}(w^*) = \inf_w \mathcal{L}_{\mathcal{D}}(w) \qquad \mathcal{R}_m(\mathcal{H}) = \frac{1}{\sqrt{m}} \quad \Rightarrow \quad \text{w.h.p}$$

$$\mathcal{L}_{\mathcal{D}}(\hat{w}) \le \inf_w \mathcal{L}_{\mathcal{D}}(w) + O\left(\frac{\|w^*\|_1}{\sqrt{m}}\right)$$

13. Conditional Maximum Entropy Models

$$\boxed{\begin{aligned} &\min_{p\in\Delta} D(p\|p_0) && \leftarrow \text{agnostic} \\ &\text{s.t.} \ \left\| \mathbb{E}_{p}[\Phi(x)] - \mathbb{E}_{\hat{D}}[\Phi(x)] \right\|_\infty \le \lambda && \leftarrow \text{empirical} \end{aligned}}$$

density-estimation

$S = \{x_1, x_2, \cdots, x_m\}$

$\Pr(x) \sim \mathcal{D}$

Supervised learning: $S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$

i.i.d.

multi-class classification: $\mathcal{Y} = \{1, 2, \cdots, c\}$

distribution: $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$

$\mathcal{D}'$ is the marginal distribution of $\mathcal{X}$.

feature mapping $\Phi(x, y)$ $\quad \Phi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^N$. $\quad \|\Phi\|_\infty \le r$

Goal: using the training sample S to learn an accurate conditional probability $p[\cdot|x]$ for any $x \in \mathcal{X}$

$$\Rightarrow \hat{y} = \operatorname*{argmax}_{y\in\mathcal{Y}} p[y|x]$$

(Bayes Optimal Classifier)

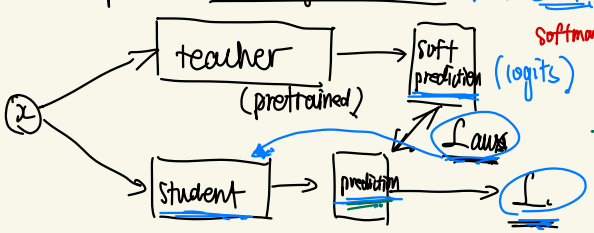
(类1, 类2, 类3)

(0.7, 0.2, 0.1) "soft prediction" 软预测

(1, 0, 0) "hard prediction"

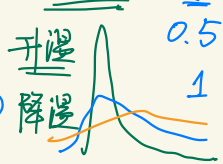
包含更多有效信息

Model Compression: Knowledge distillation 知识蒸馏



softmax

$$\frac{e^{x/T}}{\sum_i e^{x_i/T}}$$



马 ~ 驴 ← 青蛙

0.5 0.45 0.05

1 0 0

principle: ① as agnostic as possible, $p[\cdot|x] \sim \text{Uniform}$ (prior: $p_0[\cdot|x]$)

条件相对熵

② $$\left\| \mathbb{E}_{\substack{x\sim\hat{D}' \\ y\sim p[\cdot|x]}}[\Phi(x,y)] - \mathbb{E}_{(x,y)\sim\hat{D}}[\Phi(x,y)] \right\|_\infty \le \lambda$$

empirical marginal distribution

★ 条件最大熵模型

$$\min_{\bar{p}\in\Delta^{\mathcal{X}_1}} \mathbb{E}_{x\sim\hat{D}'}\left[D(p[\cdot|x] \,\|\, p_0[\cdot|x]) \right] = \sum_{x\in\mathcal{X}_1} \hat{D}'(x)\cdot D(p[\cdot|x] \,\|\, p_0[\cdot|x])$$

conditional

$p[x]\neq 0$!!!

Δ : simplex of probability distributions over $\mathcal{Y}$

$\mathcal{X}_1 = \text{supp}(\hat{D}')$

$\text{supp}(f) = \{x\in\mathcal{X} : f(x)\neq 0\}$ 非零概率的取值集合

$\bar{p}\in\Delta^{\mathcal{X}_1}$: $\bar{p} = (p[\cdot|x])_{x\in\mathcal{X}_1}$

$$\text{s.t.} \quad \left\| \mathbb{E}_{\substack{x\sim\hat{D}' \\ y\sim p[\cdot|x]}}[\Phi(x,y)] - \mathbb{E}_{(x,y)\sim\hat{D}}[\Phi(x,y)] \right\|_\infty \le \lambda$$

minimize

$$\mathbb{E}_{x\sim\hat{D}'}\left[D(p[\cdot|x] \,\|\, p_0[\cdot|x]) \right] = \mathbb{E}_{x\sim\hat{D}'}\left[\sum_{y\in\mathcal{Y}} p[y|x]\cdot\log\frac{p[y|x]}{p_0[y|x]} \right]$$

uniform $\frac{1}{C}$

$$= \mathbb{E}_{x\sim\hat{D}'}\left[+\sum_{y\in\mathcal{Y}} p[y|x]\cdot\log C + \sum_{y\in\mathcal{Y}} p[y|x]\log p[y|x] \right]$$

$$= \log C - \bar{H}(\bar{p})$$

conditional entropy $\mathbb{E}_{x\sim\hat{D}'}[-H(p[\cdot|x])]$

advantage: 不需要指定任何分布族 (no prior)

$$F(\bar{p}) = \mathbb{E}_{x\sim\hat{D}'}\left[\tilde{D}(p[\cdot|x] \,\|\, p_0[\cdot|x]) \right] + I_C\left(\mathbb{E}_{\substack{x\sim\hat{D}' \\ y\sim p[\cdot|x]}}[\Phi(x,y)] \right)$$

$$\Updownarrow$$

$$C = \left\{ u\in\mathbb{R}^N : \left\| u - \mathbb{E}_{(x,y)\sim\hat{D}}[\Phi(x,y)] \right\|_\infty \le \lambda \right\}$$

$$I_C(x) = \begin{cases} 0, & x\in C \\ +\infty, & x\notin C \end{cases}$$

$$G(w) = \frac{1}{m}\sum_{i=1}^{m} \log\frac{p_w[y_i|x_i]}{p_0[y_i|x_i]} - \lambda\cdot\|w\|_1$$

dual objective

$$p_w[y|x] = \frac{p_0[y|x]\cdot e^{w\cdot\Phi(x,y)}}{Z(w,x)} \quad ☆ \qquad Z(w,x) = \sum_{y\in\mathcal{Y}} p_0[y|x]\cdot e^{w\cdot\Phi(x,y)}$$

Thm (strong duality) $\boxed{\sup_{w\in\mathbb{R}^N} G(w) = \min_{\bar{p}\in(\mathbb{R}^{|\mathcal{Y}|})^{|\mathcal{X}_1|}} F(\bar{p})}$ ☆☆

$p^* = \inf_x \{ f(x) + g(Ax) \}$

$d^* = \sup_y \{ -f^*(A^*y) - g^*(-y) \}$

$p^* = d^*$ if $A(\mathrm{dom}(f)) \cap \mathrm{cont}(g) \neq \phi$

Cor $\bar{p}^* = \arg\min_{\bar{p}} F(\bar{p}) \quad d^* = \sup_{w\in\mathbb{R}^n} G(w)$

if $|G(w) - d^*| \le \varepsilon$ then $\mathbb{E}_{x\sim\hat{D}}[D(p^*[\cdot|x] \,\|\, p_w[\cdot|x])] \le \varepsilon$

$U_0 = \mathbb{E}_{(x,y)\sim\hat{D}}[\Phi(x,y)]$

$D(p\|q) \ge \frac{1}{2}\|p-q\|_1$

$\Rightarrow \mathbb{E}_{x\sim\hat{D}}[\,\|p^*[\cdot|x] - p_w[\cdot|x]\|_1\,] \le \sqrt{2\varepsilon}$

① $\inf_w \lambda\|w\|_1 - \frac{1}{m}\sum_{i=1}^m \log p_w[y_i|x_i]$

$\lambda = 0 \Rightarrow \max \sum_{i=1}^m \log p_w[y_i|x_i] = \max \prod_{i=1}^m p_w[y_i|x_i]$

$\lambda \neq 0 \Rightarrow$ regularization!

MLE

from Gibbs distribution 本质大概率事件

② $p[\cdot|x]$ for $x \in \mathcal{X}_1$

Since prior is KL-divergence to measure the closeness

How about $x \in \mathcal{X} \setminus \mathcal{X}_1$? (不在样本中的部分)

dual problem: $p_w[\cdot|x] = \frac{p_0[\cdot|x]\cdot e^{w\cdot\Phi(x,y)}}{Z}$ the learner already knows!

不依赖 x, 只依赖 w!

$\forall x \in \mathcal{X}$ okay!

③ $\min_{w\in\mathbb{R}^N} \lambda\|w\|_1 - \frac{1}{m}\sum_{i=1}^m \log p_w[y_i|x_i]$

$\Leftrightarrow \min_{w\in\mathbb{R}^N} \lambda\|w\|_1 + \frac{1}{m}\sum_{i=1}^m \log\left(\sum_{y\in\mathcal{Y}} e^{w\cdot\Phi(x_i,y)} - e^{w\cdot\Phi(x_i,y_i)}\right)$

$\Leftrightarrow \min_{w\in\mathbb{R}^N} \lambda\|w\|_1 - w\cdot\frac{1}{m}\sum_{i=1}^m \Phi(x_i,y_i) + \frac{1}{m}\sum_{i=1}^m \log\sum_{y\in\mathcal{Y}} e^{w\cdot\Phi(x_i,y)}$

norm　　linear　　log-sum-exp → convex

性质非常好!

SGD. Newton. GIS. IIS. Coordinate Descent

L1-regularized-AdaBoost (logistic function)

④ feature vector $\Phi(x,y)$

e.g. $T(x)$ "one-hot"

$$\Phi(x,y) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ T(x) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{真实类别}$$

$N = C$

稀疏性

二元 → 一元

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_{y-1} \\ w_y \\ w_{y+1} \\ \vdots \\ w_C \end{bmatrix}$$

$$\Rightarrow w \cdot \Phi(x,y) = w_y \cdot T(x)$$

$$\Rightarrow \min_{w \in \mathbb{R}^N} \lambda \sum_{y \in \mathcal{Y}} \|w_y\|_1 + \frac{1}{m} \sum_{i=1}^{m} \log \sum_{y \in \mathcal{Y}} e^{w_y \cdot T(x_i) - w_{y_i} \cdot T(x_i)}$$

⑤ prediction:

$$\boxed{\hat{y}(x) = \arg\max_{y \in \mathcal{Y}} P_w[y|x] = \arg\max_{y \in \mathcal{Y}} w \cdot \Phi(x,y)}$$

☆

线性模型 (log-linear model)

$y = w \cdot x + b \Rightarrow g(y) = w \cdot x + b$

$\Rightarrow y = g^{-1}(w \cdot x + b)$

用回归解决

△

$$\min_{w \in \mathbb{R}^N} \lambda \|w\|_1 - \frac{1}{m} \sum_{i=1}^{m} \log P_w[y_i|x_i]$$

区别性质 经验风险 $\mathbb{E}_D[\mathcal{L}]$ (log-loss)

$$P_w[y|x] = \frac{\cancel{P_0[y|x]} e^{w \cdot \Phi(x,y)}}{Z(w,x)}$$

Conditional Maxent Model

$$= \lambda \|w\|_1 - \frac{1}{m} \sum_{i=1}^{m} \log \frac{e^{w \cdot \Phi(x_i,y_i)}}{\sum_{y \in \mathcal{Y}} e^{w \cdot \Phi(x_i,y)}} = \boxed{\lambda \|w\|_1 + \frac{1}{m} \sum_{i=1}^{m} \log \left(\sum_{y \in \mathcal{Y}} e^{w \cdot \Phi(x_i,y) - w \cdot \Phi(x_i,y_i)} \right)}$$

二集问题 $\mathcal{Y} = \{+1, -1\}$

$$\Rightarrow \sum_{y \in \mathcal{Y}} e^{w \cdot \Phi(x_i,y) - w \cdot \Phi(x_i,y_i)}$$

$$= e^{w \cdot \Phi(x_i,+1) - w \cdot \Phi(x_i,y_i)} + e^{w \cdot \Phi(x_i,-1) - w \cdot \Phi(x_i,y_i)}$$

$$= 1 + e^{-y_i w \cdot (\Phi(x_i,+1) - \Phi(x_i,-1))}$$

$\Phi(x_i) -$

$$= 1 + e^{-y_i w \cdot \Phi(x_i)}$$

$$\min_{w\in\mathbb{R}^N} \lambda\cdot\|w\|_1 + \frac{1}{m}\sum_{i=1}^{m}\log\left(1+e^{-y_i w\cdot\Phi(x_i)}\right)$$

logistic regression

MLE 极大似然估计

$y=\{+1,-1\}$: $\frac{1}{m}\sum_{i=1}^{m}\log\left(1+e^{-y_i w\cdot(\Phi(x_i,+1)-\Phi(x_i,-1))}\right)$

逻辑斯蒂回归

(soft prediction)

$y=\{1,0\}$: $\frac{1}{m}\sum_{i=1}^{m}\log\left(1+e^{-y_i w\cdot\Phi(x_i)}\right)$

Algorithms: SGD Coordinate Descent.

Logistic Model

$$P_w[y=+1\mid x]=\frac{e^{w\cdot\Phi(x,+1)}}{Z(x)}$$

"log-odds" 对数几率

$$\log\frac{P_w[y=+1\mid x]}{P_w[y=-1\mid x]}=w\cdot\left(\Phi(x,+1)-\Phi(x,-1)\right)=w\cdot\Psi(x)$$

log-linear

prediction of logistic model is based on the sign of log-odds ratio

$$P_w[y=+1\mid x]=\frac{e^{w\cdot\Phi(x,+1)}}{e^{w\cdot\Phi(x,+1)}+e^{w\cdot\Phi(x,-1)}}=\frac{1}{1+e^{-w(\Phi(x,+1)-\Phi(x,-1))}}=\frac{1}{1+e^{-w\cdot\Phi(x)}}$$

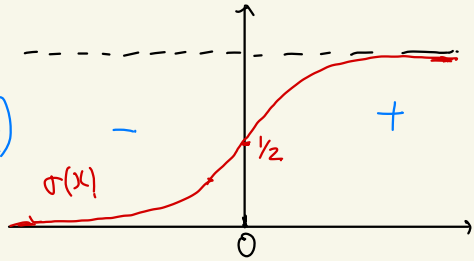
线性模型: $w\cdot\Phi(x)\longrightarrow\frac{1}{1+e^{-w\cdot\Phi(x)}}$ 逻辑斯蒂模型

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

性质超好!

我们的目标是实值(log-odds)

用回归模型拟合表



$$\sigma'(x)=\frac{+e^{-x}}{(1+e^{-x})^2}=\frac{1}{1+e^{-x}}\cdot\frac{(1+e^{-x})-1}{1+e^{-x}}=\sigma(x)(1-\sigma(x))$$

求导很方便

forward: "cache"

backward: 直接用!

sigmoid

logistic

ReLU, GELU

Generalization Bounds

$\Phi \in \mathbb{R}^N$ < N: infinity (very large) → N independent
N: finity

Thm 1 $f \in \mathcal{F} = \{ (x,y) \mapsto w \cdot \Phi(x,y) : \|w\|_1 \le 1 \}$

$$R(f) \le \underbrace{\frac{1}{m}\sum_{i=1}^{m} \log_{u_0}\left(\sum_{y \in \mathcal{Y}} e^{\frac{f(x_i,y)-f(x_i,y_i)}{\rho}} \right)}_{u_0 = 1+\frac{1}{e}} + \frac{8C}{\rho} \mathcal{R}_m(\Pi_1(\mathcal{H})) + \sqrt{\frac{\log\log_2 \frac{4r}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

$\Pi_1(\mathcal{H}) = \{ x \mapsto \Phi(x,y) \ \ \Phi \in \mathcal{H}, y \in \mathcal{Y} \}$

Proof: 多读问题 $\rho_f(x_i,y_i) = \min_{y \ne y_i} f(x_i,y_i) - f(x_i,y) = \min_{y \ne y_i} w \cdot (\Phi(x_i,y_i) - \Phi(x_i,y))$

$\forall \rho \in (0, 2r]$: $R(f) \le \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\rho_f(x_i,y_i) \le \rho} + \frac{4C}{\rho} \mathcal{R}_m(\Pi_1(\mathcal{F})) + \sqrt{\frac{\log\log_2 \frac{4r}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$

$\forall \rho > 0$

$\Pi_1(\mathcal{F}) = \{ x \mapsto f(x,y) : f \in \mathcal{F}, y \in \mathcal{Y} \}$

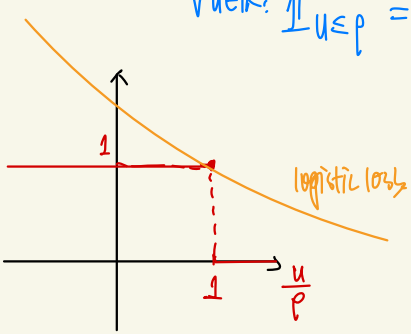
若 $\rho > 2r$. $|w \cdot \Phi(x,y)| \le \|w\|_1 \cdot \|\Phi(x,y)\|_\infty \le r < \rho$. $\Rightarrow R(f) \le 1 + \sim$

① 要证 $\mathbb{1}_{\rho_f(x_i,y_i) \le \rho} \le \log_{u_0}\left(\sum_{y \in \mathcal{Y}} e^{\frac{f(x_i,y)-f(x_i,y_i)}{\rho}} \right)$

ρ-margin loss can be upper-bounded by ρ-logistic loss

$\forall u \in \mathbb{R}$: $\mathbb{1}_{u \le \rho} = \mathbb{1}_{\frac{u}{\rho} - 1 \le 0} \le \log_{u_0}\left(1 + e^{-\frac{u}{\rho}}\right)$ $u_0 = 1 + \frac{1}{e}$

$\ll \frac{\ln\left(1+e^{-\frac{u}{\rho}}\right)}{\ln\left(1+\frac{1}{e}\right)}$



$$\mathbb{1}_{\rho_f(x_i,y_i) \le \rho} \le \log_{u_0}\left(1 + e^{-\frac{\rho_f(x_i,y_i)}{\rho}}\right)$$

$$= \log_{u_0}\left(1 + \max_{y \neq y_i} e^{\frac{f(x_i,y) - f(x_i,y_i)}{\rho}}\right)$$

$$\leq \log_{u_0}\left(1 + \sum_{y \neq y_i} e^{\frac{f(x_i,y) - f(x_i,y_i)}{\rho}}\right) = \log_{u_0} \sum_{y \in \mathcal{Y}} e^{\frac{f(x_i,y) - f(x_i,y_i)}{\rho}}$$

② $\mathcal{R}_m(\Pi_1(\mathcal{F})) \leq 2\,\mathcal{R}_m(\Pi_1(\mathcal{H}))$

$\mathcal{F}$: 由 $w\Phi$ 组成的假设空间

$\mathcal{H}$: 由 Φ 组成 ~

$$\hat{\mathcal{R}}_S(\Pi_1(\mathcal{F})) = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\substack{\|w\|_1 \leq 1 \\ y \in \mathcal{Y}}} \sum_{i=1}^{m} \sigma_i \sum_{j=1}^{N} w_j \Phi_j(x_i,y)\right]$$

$$= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\substack{\|w\|_1 \leq 1 \\ y \in \mathcal{Y}}} \sum_{j=1}^{N} w_j \sum_{i=1}^{m} \sigma_i \Phi_j(x_i,y)\right]$$

$$= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\substack{j \in [N] \\ y \in \mathcal{Y}}} \left|\sum_{i=1}^{m} \sigma_i \Phi_j(x_i,y)\right|\right]$$

dual norm.

$$\leq \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\substack{\Phi \in \mathcal{H} \\ y \in \mathcal{Y}}} \left|\sum_{i=1}^{m} \sigma_i \Phi(x_i,y)\right|\right] \leq 2\,\hat{\mathcal{R}}_S(\Pi_1(\mathcal{H}))$$

$\sigma \in \{-1,+1\}$

□.

多分类问题只取值最大的 prediction. ⇒ scaling 不影响结果 ⇒ 不影响泛化性质 $/\rho$

$$\Rightarrow R(f) \leq \frac{1}{m}\sum_{i=1}^{m} \log_{u_0}\left(\sum_{y \in \mathcal{Y}} e^{f(x_i,y) - f(x_i,y_i)}\right) + \frac{8c}{\rho}\mathcal{R}_m(\Pi_1(\mathcal{H})) + \sqrt{\frac{\log\log_2 \frac{4r}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

$$f \in \tilde{\mathcal{F}} = \left\{(x,y) \mapsto w \cdot \Phi(x,y) : \|w\|_1 \leq \frac{1}{\rho}\right\}$$

minimize RHS!

$\rho \to$ free parameter (cross-validation)

↳ select w to minimize $\frac{1}{m}\sum_{i=1}^{m} \log \sum_{y \in \mathcal{Y}} e^{f(x_i,y) - f(x_i,y_i)}$

s.t. $\|w\|_1 \leq \frac{1}{\rho}$ ← λ

$$\Rightarrow \min_{w} \lambda \cdot \|w\|_1 + \frac{1}{m}\sum_{i=1}^{m} \log\left(\sum_{y\in\mathcal{Y}} e^{f(x_i,y)-f(x_i,y_i)}\right)$$

Conditional MaxEnt Model !

Thm2

$$R(f) \le \frac{1}{m}\sum_{i=1}^{m}\log_{10}\left(\sum_{y\in\mathcal{Y}} e^{\frac{f(x_i,y)-f(x_i,y_i)}{\rho}}\right) + \frac{4Cr\cdot\sqrt{2\log(2CN)}}{\rho} + \sqrt{\frac{\log\log_2\frac{4r}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

N-dependent !

Proof: 只需证: $\hat{\mathcal{R}}_S(\Pi_1(\mathcal{F})) \le r\cdot\sqrt{2\log(2CN)}$

$$\hat{\mathcal{R}}_S(\Pi_1(\mathcal{F})) = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\substack{\|w\|_1\le 1\\ y\in\mathcal{Y}}} \sum_{i=1}^{m}\sigma_i\, w\cdot\Phi(x_i,y)\right]$$

$$= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\substack{\|w\|_1\le 1\\ y\in\mathcal{Y}}} w\cdot\sum_{i=1}^{m}\sigma_i\,\Phi(x_i,y)\right]$$

$$\le \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{y\in\mathcal{Y}} \left\|\sum_{i=1}^{m}\sigma_i\,\Phi(x_i,y)\right\|_\infty\right]$$

$$= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\substack{j\in[N]\\ y\in\mathcal{Y}\\ s\in\{-1,+1\}}} s\cdot\sum_{i=1}^{m}\sigma_i\,\Phi_j(x_i,y)\right] \le r\cdot\sqrt{2\log(2CN)} \quad \square$$

N C 2

(Maximal Inequality: $\mathbb{E}\left[\max_{j\in[N]} X_j\right] \le r\cdot\sqrt{2\log N}$.)