Lecture 17 - MaxEnt duality; exp. family

Tuesday, November 7, 2023 2:33 PM

today: • Max Ent & duality
• exponential family

MLE & KL minimization

$\{p_\theta\}_{\theta\in\Theta}$ parametric family for a discrete observation space

then ML for θ for i.i.d. data $\iff \min_{\theta\in\Theta} KL(\hat{p}_n \| p_\theta)$

$$\hat{p}_n(x) \triangleq \frac{1}{n}\sum_{i=1}^n \delta(x, x^{(i)})$$

"empirical distribution" kronecker-delta

proof:
$$KL(\hat{p}_n \| p_\theta) = \sum_x \hat{p}_n(x) \log \frac{\hat{p}_n(x)}{p_\theta(x)}$$

$$= -H(\hat{p}_n) - \sum_x \underbrace{\hat{p}_n(x)}_{\frac{1}{n}\sum_{i=1}^n \delta(x,x^{(i)})} \log p_\theta(x)$$

$$\frac{1}{n}\sum_{i=1}^n \underbrace{\sum_x \delta(x,x^{(i)}) \log p_\theta(x)}_{\log p_\theta(x^{(i)})}$$

$$= \underbrace{-H(\hat{p}_n)}_{\text{constant w.r.t. } \theta} - \frac{1}{n}\underbrace{\sum_{i=1}^n \log p_\theta(x^{(i)})}_{\log\left(\prod_{i=1}^n p_\theta(x^{(i)})\right)}$$

//

Maximum entropy principle:

idea: consider some subset of dist. over $\mathcal{X}$ according to some data-driven constraints

get a subset $\mathcal{M} \subseteq \Delta_{|\mathcal{X}|}$ ← prob. simplex over $|\mathcal{X}| = k$ elements

MAXENT principle: pick $\hat{p} \in \mathcal{M}$ which maximizes the entropy

i.e. $\boxed{\hat{p} = \underset{q\in\mathcal{M}}{\arg\max}\ H(q)}$

$$= \underset{q\in\mathcal{M}}{\arg\min}\ KL(q \,\|\, \underbrace{\text{uniform}}_{u})$$

$$KL(q\|u) = \sum_x q(x) \log \frac{q(x)}{1/|\mathcal{X}|} \quad = -H(q) + \text{constant}$$

$$KL(q \| u) = \sum_x q(x) \log \frac{q(x)}{u(x)} \Big] \frac{1}{k} = \text{constant} \quad = -H(q) + \text{constant}$$

"generalized max. entropy" $KL(q \| h_0)$ ← preferred dist. to bias towards

* <u>example from Wainwright</u>

$\hat{p}_L = \frac{3}{4}$ kangoroos are left-handed

$\hat{p}_B = \frac{2}{3}$ " drink Labatt beer

question: how many kang. are both L.H. & drink Lab. beer.

[here: max. entropy solution is that $p(B=1, LH.=1) = \hat{p}_B \cdot \hat{p}_L$ (indep.)]

* how do we get set $\mathcal{M}$

typically: through empirical "moments"

feature functions: $T_1(x), T_2(x), \dots, T_d(x)$ d features

Kangoroos: $T_1(x) = \mathbb{1}\{X \text{ drinks labatt}\}$, $T_2(x) = \mathbb{1}\{X \text{ is left handed}\}$

define $\mathcal{M} = \{ q : \underbrace{\mathbb{E}_q[T_j(X)]}_{\text{model expected feature "count"}} = \underbrace{\mathbb{E}_{\hat{p}_n}(T_j(X))}_{\text{empirical feature "count" / "moment constraints"}} \; j = 1, \dots, d \}$

then

$$\boxed{\begin{array}{ll} \text{MaxENT} & \\ \min_{q \in \mathbb{R}^{|X|}} & KL(q \| \text{unif}) \\ & q \in \mathcal{M} \\ & q \in \Delta_{|X|} \end{array}}$$

$q \in \mathcal{M}$ → $\sum_x q(x) T_j(x) = \frac{1}{n} \sum_{i=1}^n T_j(x^{(i)}) = \alpha_j$

i.e. $\langle \vec{q}, \vec{T_j} \rangle = \alpha_j$

↳ convex opt. problem over $q \in \Delta_{|X|} \subseteq \mathbb{R}^{|X|}$

quick presentation of Lagrangian duality

convex min. problem

$$\left.\begin{array}{ll} \min_{x \in \mathbb{R}^d} & f(x) \\ \text{s.t.} & f_j(x) \leq 0 \quad \forall j \in \{1, \dots, m\} \\ & g_k(x) = 0 \quad \forall k \in \{1, \dots, n\} \end{array}\right\} \text{"primal problem"}$$

convex opt. problem ⟸ { • f, f_j are convex fct. • g_k <u>affine fct.</u> }

Lagragian fct. $\mathcal{L}(x, \lambda, \nu) \triangleq f(x) + \sum_{j=1}^m \lambda_j f_j(x) + \sum_{k=1}^n \nu_k g_k(x)$

Lagrangian fct. $\mathcal{L}(x,\lambda,\nu) \triangleq f(x) + \sum_{j=1}^{m} \lambda_j f_j(x) + \sum_{i=1}^{n} \nu_k g_k(x)$

"Lagrange multipliers"

magic trick
(saddle pt. interpretation)

$$h(x) \triangleq \sup_{\lambda \geq 0, \nu} \mathcal{L}(x,\lambda,\nu) = \begin{cases} f(x) & \text{if } x \text{ is feasible} \\ +\infty & \text{if } x \text{ is not feasible} \end{cases}$$

$$h(x) = f(x) + \delta_M(x)$$ indicator fct.

$$\delta_M(x) = \begin{cases} +\infty & \text{if } x \notin M \\ 0 & \text{o.w.} \end{cases}$$

an equivalent problem to primal problem $\inf_x h(x)$ ← fancy non-smooth fct.

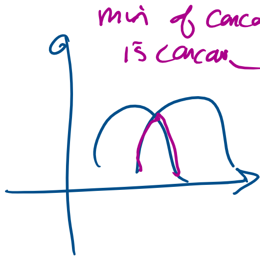
$$\inf_x \left(\sup_{\lambda \geq 0, \nu} \mathcal{L}(x,\lambda,\nu) \right)$$

duality trick is to swap inf & sup

$$\sup_{\lambda \geq 0, \nu} \underbrace{\inf_x \mathcal{L}(x,\lambda,\nu)}_{\triangleq g(\lambda,\nu)}$$ → this fct. is always concave

Lagrangian dual fct.

min of concave fct is concave



$$\boxed{\text{Lagrange dual problem} \quad \sup_{\underbrace{\lambda \geq 0, \nu}_{\text{"dual variables"}}} g(\lambda,\nu)}$$

"weak duality"

in general, $\sup_{\lambda,\nu} \underbrace{\inf_x \mathcal{L}(x,\lambda,\nu)}_{g(\lambda,\nu)} \leq \inf_x \sup \mathcal{L}(x,\lambda,\nu)$

strong duality when $\sup \inf \mathcal{L} = \inf \sup \mathcal{L}$

↳ sufficient conditions:
- when primal problem is convex
- + constraint qualification condition (e.g. Slater's condition)

(can get optimal primal variables $x^*(\lambda^*,\nu^*)$ using KKT conditions)

(→ see ch. 5 of Boyd's book)

see chapter 5 in Boyd's book for more info on duality: http://stanford.edu/~boyd/cvxbook/

15h33

dual problem for max. entropy

MaxENT in | min $KL(q\|u)$ | $u(x) \triangleq \frac{1}{|X|}$

dual problem for max. entropy

Max ENT in primal form

(P)

$$\min_{q \in \mathbb{R}^K} \underbrace{\sum_x q(x) \log \frac{q(x)}{u(x)}}_{KL(q\|u)}$$

$$\left.\begin{array}{l} q(x) \geq 0 \quad \forall x \\ c \rightarrow \sum_x q(x) = 1 \end{array}\right\} \Delta_{|X|}$$

eq. constraints $v \rightarrow \sum_x q(x) T_j(x) = \alpha_j \quad \forall j$ $\Big\}\ \mathcal{M}$

$u(x) = \frac{1}{|X|}$

absorb this constraint in domain of def. $KL(q\|u)$ i.e.

$$KL(q\|u) = \begin{cases} +\infty & \text{if } q(x)<0 \text{ for some } x \\ KL(q\|u) & \text{o.w.} \end{cases}$$

$$\mathcal{L}(q, v, c) = \sum_x q(x) \log \frac{q(x)}{u(x)} + \sum_{j=1}^{d} v_j \left(\alpha_j - \sum_x q(x) T_j(x)\right) + c\left(1 - \sum_x q(x)\right)$$

$q(x) = q_x$

$$\frac{\partial}{\partial q_x} = 1 + \log \frac{q(x)}{u(x)} - \underbrace{\sum_{j=1}^{d} v_j T_j(x)}_{v^T T(x)} - c \overset{\text{want}}{=} 0$$

$$\Rightarrow \boxed{q^*_{v,c}(x) = u(x) \exp\left(v^T T(x) + c - 1\right)}$$

exponential family!

dual fct:

plug back $q^*_{v,c}$ in $\mathcal{L}(\dots)$

$$g(v,c) = \mathcal{L}(q^*_{v,c}, v, c)$$

$$= \tilde{\mathbb{E}}_{q^*}\left[\cancel{v^T T(X)} + \cancel{c} - 1\right] + v^T\alpha - \tilde{\mathbb{E}}_{q^*}\left[\cancel{v^T T(X)}\right] + c - \tilde{\mathbb{E}}_{q^*}[\cancel{c}]$$

shorthand for $\sum_x q^*(x)$

$$= v^T\alpha + c - \underbrace{\sum_x u(x) \exp(v^T T(x))}_{\triangleq Z(v)} \exp(c-1)$$

max $g(v,c)$ with respect to c

$$\nabla_c = 0 \Rightarrow 1 - \exp(c-1) Z(v) \overset{\text{want}}{=} 0$$

$$\Rightarrow \exp(c^*-1) = \frac{1}{Z(v)} \Rightarrow c^*-1 = -\log Z(v)$$

plug back c^*

$$\max_c g(v,c) = v^T\alpha + \underbrace{c^* - \cancel{Z(v)} \frac{1}{\cancel{Z(v)}}}_{c^*-1 = -\log Z(v)}$$

dual problem $\max_v \tilde{g}(v)$ $\quad \boxed{\tilde{g}(v) \triangleq v^T\alpha - \log Z(v)}$

link with MLE!

if $\alpha = \frac{1}{n} \sum_{i} T(x^{(i)}) = \hat{\mathbb{E}}[T(x)]$

link with MLE:

if $\alpha = \frac{1}{n}\sum_{i=1}^{n} T(x^{(i)}) = \mathbb{E}_{\hat{p}_n}[T(x)]$

then $\tilde{g}(\nu) = \frac{1}{n}\sum_{i=1}^{n} [\underbrace{\nu^T T(x^{(i)}) - \log Z(\nu)}_{\log p(x^{(i)}|\nu) + cst.}]$

where $p(x|\nu) \triangleq u(x) \exp(\nu^T T(x) - \log Z(\nu))$

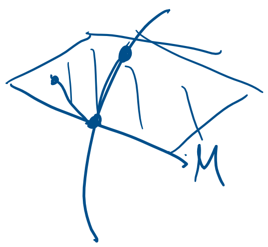
i.e. dual problem is $\max_\nu \tilde{g}(\nu) = \max_\nu \frac{1}{n} \log p(x_{1:n}|\nu) + cst.$

i.e. MLE //

to summarize: $\boxed{\text{ML in exp. family with } T(x) \text{ as sufficient statistics is } \underline{\text{equivalent}} \text{ to max ENT with empirical moment constraints on } T(x) \text{ where } \alpha = \mathbb{E}_{\hat{p}_n}[T(x)]}$

they are Lagrangian dual of each other?

$\boxed{\text{MLE in exp family} \Longleftrightarrow \text{moment matching in exp family}}$



KL pythagorean theorem

→ see lecture 16 in 2017

(see end of old lecture 16 2017 for "KL Pythagorean theorem" and I-projection vs. M-projection for KL + geometry)

note: $\nabla_\nu \log Z(\nu) = \frac{1}{Z(\nu)} \nabla_\nu \sum_x u(x) \exp(\nu^T T(x))$

$= \sum_x T(x) \underbrace{\frac{1}{Z(\nu)} u(x) \exp(\nu^T T(x))}_{p(x|\nu)}$

$\nabla_\nu \log Z(\nu) = \mathbb{E}_{p(x|\nu)}[T(x)] \triangleq \mu(\nu)$ "model moment"

$\nabla_\nu \tilde{g}(\nu) = \overbrace{\underbrace{\mathbb{E}_{\hat{p}_n}[T(x)]}_{\hat{\mu}_n \text{ "empirical moment"}}}^{\alpha} - \mu(\nu)$

$\nabla_\nu \tilde{g}(\nu) = 0 \Rightarrow \boxed{\mu(\nu^*) = \hat{\mu}_n}$ i.e. moment matching?