Lecture 19 - MaxEnt duality; exp. family

Tuesday, November 10, 2020 14:23

today : • max Ent & duality
• exponential family

MLE & KL minimization

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

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

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

δ ← Kronecker-delta

proof :

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

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

$$\frac{1}{n} \sum_i \sum_x \delta(x | x^{(i)}) \log p_\theta(x)$$

$$\log p_\theta(x^{(i)})$$

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

$$\log \prod_{i=1}^{n} p_\theta(x^{(i)})$$

//

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. $$\hat{p} = \operatorname{argmax}_{q \in \mathcal{M}} H(q)$$

$$= \operatorname{argmin}_{q \in \mathcal{M}} KL(q \| \underbrace{\text{uniform}}_{u})$$

$$KL(q \| u) = \sum_x q(x) \log \frac{q(x)}{u(x)} \quad = -H(q) + \text{cst.}$$

$u(x) = 1/k = \text{constant}$

"generalized max. entropy" $KL(q \| h_0)$

h_0 ← preferred dist. to be less towards

* example from Wainwright

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

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

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

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

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

typically: through empirical "moments"

kangaroos:
$T_1(x) = \mathbb{1}\{x \text{ kang. drinks Splatt}\}$
$T_2(x) = \mathbb{1}\{x \text{ is left handed}\}$

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

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

"moment constraints"

then MaxENT

$\min_{q \in \mathbb{R}^{|\mathcal{X}|}} KL(q \| \text{unif})$ s.t. $q \in \mathcal{M}$, $q \in \Delta_{|\mathcal{X}|}$

$\mathbb{E}_q[T_j(x)]$ want

$\sum_x q(x) T_j(x) = \frac{1}{n} \sum_{i=1}^n T_j(x^{(i)}) = a_j$

i.e. $\langle \vec{q}, \vec{T}_j \rangle = a_j$

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

quick presentation of Lagrangian duality

convex min. problem $\min_{x \in \mathbb{R}^d} f(x)$

convex problem ⇐ { • f, f_j are convex fct. s.t. $f_j(x) \leq 0 \; \forall j \in \{1, \dots, m\}$
• g_k affine fct. $g_k(x) = 0 \; \forall k \in \{1, \dots, n\}$ } "primal problem"

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

"Lagrange multipliers"

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

an equivalent problem to the primal problem

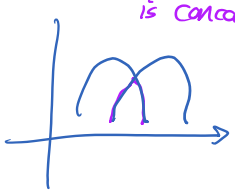
$\min_x \left(\sup_{\lambda \geq 0, \nu} f(x, \lambda, \nu) \right)$ — always non-smooth function

duality trick is to swap inf & sup

$\sup_{\lambda \geq 0, \nu} \inf_x f(x, \lambda, \nu)$ → this fct. is always concave

$\triangleq g(\lambda, \nu)$ Lagrange dual fct.

min of concave fct is concave



Lagrange dual problem $\sup_{\lambda \geq 0, \nu} g(\lambda, \nu)$

"dual variables"

weak duality

in general, $\sup_{\lambda,\nu} \inf_x f(x, \lambda, \nu) \leq \inf_x \sup_{\lambda,\nu} f(x, \lambda, \nu)$

strong duality when $\sup \inf f = \inf \sup f$

↳ sufficient conditions: • when primal problem is convex
• + constraint qualification condition (eg. Slater's condition)

(can get optimal primal variables x^* / (λ^*, ν^*) 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/

15h40

dual problem for max. entropy

MaxENT in primal form (P)

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

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

$q(x) \geq 0 \ \forall x$ } $\Delta_{|\mathcal{X}|}$

eq. constraints → $\sum_x q(x) = 1$

→ $\sum_x q(x) T_j(x) = \alpha_j \ \forall j$ } M

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

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

then eq. const. $\sum_x q(x) = 1$

$q(x) = q_x$

$$\mathcal{L}(q, \nu, c) = \sum_x q(x) \log \frac{q(x)}{u(x)} + \sum_j \nu_j \left(\alpha_j - \sum_x q(x) T_j(x) \right) + c \left(1 - \sum_x q(x) \right)$$

$$\frac{\partial}{\partial q(x)} = 1 + \log \frac{q(x)}{u(x)} - \sum_j \nu_j T_j(x) - c \overset{\text{want}}{=} 0$$

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

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

exponential family?

dual fct.:

plug back $q^*_{\nu,c}$ in $\mathcal{L}(\ldots)$ $\quad g(\nu,c) = \mathcal{L}(q^*_{\nu,c}, \nu, c)$

$= \mathbb{E}_{q^*}[\nu^T T(x) + c - 1] + \nu^T \alpha - \mathbb{E}_{q^*}[\nu^T T(x)]$

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

$+ c - \mathbb{E}_{q^*}[q]$

$= \nu^T \alpha + c - \sum_x u(x)\exp(\nu^T T(x))\exp(c-1)$

$\triangleq Z(\nu)$

$\max_{\text{with respect to } c} g(\nu,c) \qquad \nabla_c = 0 \Rightarrow 1 - Z(\nu)\exp(c-1) \overset{want}{=} 0$

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

plug back c^*: $\max_c g(\nu,c) = \nu^T\alpha + c^* - Z(\nu)\frac{1}{Z(\nu)}$

$c^* - 1 = -\log Z(\nu)$

dual problem $\max_\nu \tilde{g}(\nu) \qquad \boxed{\tilde{g}(\nu) \triangleq \nu^T\alpha - \log Z(\nu)}$

link with MLE:

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

then $g(\nu) = \frac{1}{n}\sum_{i=1}^n [\nu^T T(x^{(i)}) - \log Z(\nu)]$

$\log p(x^{(i)}|\nu) + \text{cst.}$

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

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

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

to summarize: ML in exp. family with $T(x)$ as sufficient statistics

is equivalent to MaxENT with moment constraints on $T(x)$

where $\alpha = \mathbb{E}_{\hat{p}_n}[T(x)]$

they are lagrangian dual of each other?

MLE in exp. family $\Leftrightarrow$ moment matching in exp. family

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

$= \sum_x T(x) \frac{u(x)\exp(\nu^T T(x))}{Z(\nu)}$

$p(x|\nu)$

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

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

$\hat{\mu}_n$ "empirical moment"

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

→ see lecture 16 in 2017 for "KL Pythagorean theorem"

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