Lecture 16 - scribbles

Friday, November 3, 2017
13:33

today: max. entropy
equivalence with ML
duality
exponential family

recall: last time $ML \iff \min_{\theta \in \Theta} KL(\hat{p}_n \| p_\theta)$

today: maximum entropy $\min_{q \in M} KL(q \| \text{uniform})$

Maximum entropy principle:

idea: consider some subset of dist. on X
according to some data-driven information

get a subset $M \subseteq \Delta_{|X|}$

principle pick $\hat{p} \in M$ by maximizing entropy

ie, $\hat{p} = \arg\max_{q \in M} H(q)$

$= \arg\min_{q \in M} KL(q \| \text{uniform})$ → "generalized max-entropy" which uses h_0 instead of uniform & favorite a priori dist.

$\sum_x q(x) \log \frac{q(x)}{\text{const.}} = -H(q) + \text{cst.}$

* example from Wainwright

$p_L = \frac{3}{4}$ kangaroos are left-handed

$p_B = \frac{2}{3}$ " drink Foster beer

question: how many " are both left-handed & drink F. beer

[here: max entropy solution is that $p(B, L) = p_B \cdot p_L$ (independence)]

* how do we get M?

typically: through empirical observation

feature functions $T_1(x), \ldots, T_d(x)$

define $M = \{ q : \underbrace{E_q[T_j(X)]}_{\text{model expected feature count}} = \underbrace{E_{\hat{p}_n}[T_j(X)]}_{\text{empirical feature count}} \quad j = 1, \ldots, d \}$

"moment constraints"

then MaxENT: $\min_{q \in \mathbb{R}^{|\mathcal{X}|}} KL(q \| \text{unif})$ (some scalar)

s.t. $q \in M$, $q \in \Delta_{|\mathcal{X}|}$ $\Big\}$ $\sum_x q(x) T_j(x) = a_j \quad \forall j$

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

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

Lagrange duality segway:

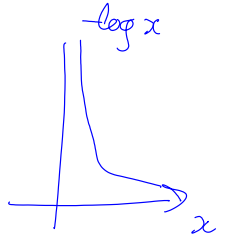
convex optimization problem $\min_x f(x)$ s.t. $f_i(x) \leq 0 \quad i = 1, \ldots, m$ $\Big\}$ "primal problem"

s.t. $f_j(x) \le 0 \quad j = 1, \dots, m$
$g_k(x) = 0 \quad k = 1, \dots, n$

- f_0, f_j are convex functions
- g_k is affine

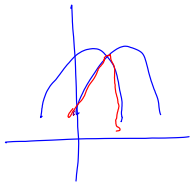
here, $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ "extended real valued function"

$\mathrm{dom}(f) \triangleq \{x : f(x) < \infty\}$ e.g. $f(x) = \begin{cases} -\log x & x > 0 \\ +\infty & \text{o.w.} \end{cases}$



Lagrangian fct. $\mathcal{L}(x, \lambda, \nu) \triangleq f_0(x) + \sum_{j=1}^{m} \lambda_j f_j(x) + \sum_{k=1}^{n} \nu_k g_k(x)$

"Lagrange multipliers"



magic trick (saddle point interpretation)

$$\sup_{\substack{\lambda \ge 0 \\ \nu}} \mathcal{L}(x, \lambda, \nu) = \begin{cases} f(x) & \text{if } x \text{ is feasible} \\ +\infty & \text{o.w.} \end{cases}$$

an equivalent problem to primal problem is

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

duality trick is swap inf & sup

$$\sup_{\substack{\lambda \ge 0 \\ \nu}} \inf_x \mathcal{L}(x, \lambda, \nu) \triangleq g(\lambda, \nu)$$ Lagrange dual fct.

dual problem

$$\sup_{\lambda, \nu} g(\lambda, \nu) \quad \lambda \ge 0$$

$\sup_{\lambda,\nu} g(\lambda,\nu)$ $\lambda \geq 0$ ← vector inequality

always true: Weak duality: $\sup_{\lambda \geq 0, \nu} \inf_x f(x,\lambda,\nu) \leq \inf_x \sup_{\lambda \geq 0, \nu} f(x,\lambda,\nu)$ ← always true

if $p^* = \inf_{x \text{ feasible}} f(x)$ ← global of primal

$g(\lambda,\nu) \leq p^* \quad \forall \lambda \geq 0, \nu$ { feasible dual variables

← always concave in λ, ν

Strong duality → when have equality i.e. $d^* = \sup_{\lambda \geq 0, \nu} g(\lambda,\nu) = p^*$

a sufficient condition for strong duality is Slater's condition: $\exists x \in \text{int}(\text{dom}(f))$

s.t. $f_j(x) < 0 \ \forall j$ where f_j is nonlinear

and x is feasible

KKT conditions:

when f etc. are differentiable

necessary conditions for strong duality

and x^* & (λ^*, ν^*) are respectively primal and dual optimal

- $g(\lambda^*, \nu^*) = f(x^*)$
  $= \inf_x f(x, \lambda^*, \nu^*)$

$$\nabla f(x^*) + \sum \lambda_j^* \nabla f_j(x^*) + \sum \nu_k^* \nabla g_k(x^*) = 0$$

necessary conditions are

KKT conditions

$\nabla_x \mathcal{L}(x^*, \lambda^*, \nu^*) = 0$

complementary slackness: $\lambda_j^* f_j(x^*) = 0$

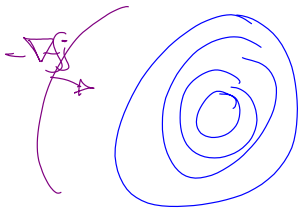
x^* is primal feasible

(λ^*, ν^*) is dual feasible (i.e. $\lambda^* \geq 0$)

see chapter 5 in Boyd's book: http://stanford.edu/~boyd/cvxbook/

if primal problem is convex + Slater's condition they are also sufficient

$f_j(x)$



dual problem for MaxENT

MaxENT primal form

(P)

$$\begin{cases} \min\limits_q \sum\limits_x q(x) \log \dfrac{q(x)}{u(x)} \\ q(x) \geq 0 \quad \Big\} \Delta_{|\mathcal{X}|} \\ \sum\limits_x q(x) = 1 \\ \sum\limits_x q(x) T_j(x) = q_j \quad \Big\} M \end{cases}$$

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

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

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

$$\Rightarrow q^*(x) = u(x) \exp\left(\nu^T T(x) + c - 1 \right)$$

exponential family?

(* we have strong duality by Slater
if $\exists q \in \mathcal{M}$ s.t. $q(x) > 0 \ \forall x$)

dual function: $g(\nu, c) = \mathcal{L}(q^*, \nu, c) = \mathbb{E}_{q^*}[\nu^T T(x) + c - 1] + \nu^T \alpha - \mathbb{E}_{q^*}[\nu^T T(x)] + c - \mathbb{E}_{q^*}[c]$

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

maximize with respect to c: $\nabla_c \rightarrow 1 - Z(\nu) e^{c-1} = 0$

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

plug back $\max_c g(\nu, c) = \nu^T \alpha - \underbrace{\log Z(\nu)}_{(c^*-1)} \triangleq \tilde{g}(\nu)$

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

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

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

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

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

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

note: $\nabla_\nu \ln Z(\nu) = \frac{1}{Z(\nu)} \nabla_\nu \sum_x u(x) \exp(\nu^T T(x)) = \sum_x \left(\underbrace{\frac{1}{Z(\nu)} u(x) \exp(\nu^T T(x))}_{p(x|\nu)} T(x) \right]$

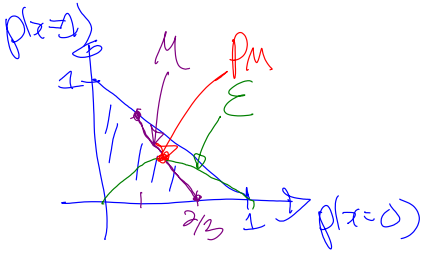
$\nabla_\nu \ln Z(\nu) = \mathbb{E}_{p(x|\nu)}[T(x)] \triangleq \mu(\nu)$ "model moment"

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

$\nabla_\nu \tilde{g}(\nu) = 0 \Rightarrow \mu(\nu) = \mathbb{E}_{\hat{p}_n}[T(x)]$ i.e. moment matching?

$\mathbb{E}_{p(x|\nu)}[T(x)]$

geometry: $X = \{0,1,2\}$ $T(x) = \begin{cases} 1 & \text{if } x=0 \\ 1/2 & \text{if } x=1 \\ 0 & \text{if } x=2 \end{cases}$



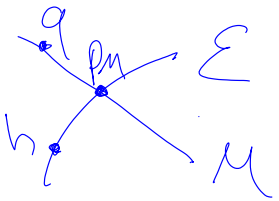
$\mathbb{E}_p[T(x)] = \frac{2}{3} = \alpha$

$\mathcal{E} = \{ p : p(x) = u(x) \exp(\nu^T T(x) - A(\nu)) \}$

"information Pythagorean thm!"

for any $q \in \mathcal{M}$ and $h \in \mathcal{E}$

$KL(q||h) = KL(q||p_M) + KL(p_M||h)$



restate our duality result : $p_M = \underset{q \in M}{\operatorname{argmin}}\ KL(q \| u)$ [Max ENT]

"I-projection" I → information

$p_M = \underset{q \in \mathcal{E}}{\operatorname{argmin}}\ KL(\hat{p}_n \| q)$ MLE in exp. family

"M-projection" M → moment