Lecture 17 - scribbles

Tuesday, November 12, 2019 14:30

- max Ent duality
- exponential family

quick presentation of Lagrangian duality

convex optimization problem : $\min_x f(x)$

- f, f_j are convex fct, s.t. $f_j(x) \leq 0 \quad \forall j = 1,\dots,m$
- g_k affine fct. $g_k(x) = 0 \quad \forall k = 1,\dots,n$

} "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 of Lagrangian duality)

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

an equivalent problem to primal problem

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

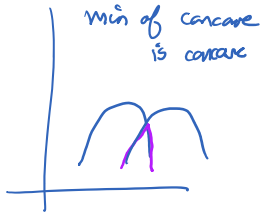
$h(x)$ fancy non-smooth fct.

duality trick is to swap inf & sup

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

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

this fct. is always concave



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

"dual variables"

weak duality (always true)

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

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

(↳ when primal is convex + constraint qualification condition (e.g. Slater's condition))

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

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

dual problem for max. entropy

MaxENT in primal form

(P)

$$\min_q \sum_x q(x) \log \frac{q(x)}{u(x)} \qquad u(x) \triangleq \frac{1}{|\mathcal{X}|}$$

$$q(x) \geq 0 \;\; \forall x \qquad \sum_x q(x) = 1 \qquad \Big\} \; \Delta_{|\mathcal{X}|}$$

$$\sum_x q(x) T_j(x) = \alpha_j \;\; \forall j$$

$T_j(x)$: feature fct.

$\Delta_{|\mathcal{X}|}$ → absorb this constraint in domain definition of $KL(q\|u)$ i.e.

$\mathcal{M}$

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

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

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

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

exponential family?

dual fct.:

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

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

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

$\triangleq Z(\nu)$

$\max g(\nu, c)$

with respect to c $\quad \nabla_c = 0 \Rightarrow \quad 1 - Z(\nu) \exp(c-1) = 0$

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

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

$\tilde{C}^{-1} = -\log(Z(\nu))$ ($Z(\nu)$)

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

link with MLE:

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

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

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

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

they are Lagrangian dual of each other

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

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

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

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

μ_n "empirical moment"

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

→ see lecture 16 in 2017 for "KL Pythagorean thm."

15h40

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

Exponential family

a (flat/canonical) exponential family on $\mathcal{X}$

is a parametric family of distributions defined by two quantities

is a parametric family of distributions defined by two quantities

I) $h(x)\, d\mu(x)$ → reference measure on $\mathcal{X}$

reference density, base measure — counting (discrete R.V.), Lebesgue (cts. R.V.)

II) $T: \mathcal{X} \to \mathbb{R}^p$ called "sufficient statistics" vector

aka. feature vector

members of family will have dist:

$$p(x;\eta)\, d\mu(x) = \exp\left(\eta^T T(x) - A(\eta)\right) h(x)\, d\mu(x)$$

defining pieces ($+\Omega_X$)

η: "canonical parameters"

$A(\eta)$: log normalizer or log partition fct. or cumulant generating fct.

if Ω_X is discrete, then $p(x;\eta)$ is a pmf

" " cts. " " " " pdf

$$\star \quad \text{want } 1 = \int_{\mathcal{X}} p(x;\eta)\, d\mu(x) = \int_{\mathcal{X}} \exp(\eta^T T(x))\, e^{-A(\eta)}\, h(x)\, d\mu(x)$$

$$A(\eta) \triangleq \log\left(\underbrace{\int_{\mathcal{X}} \exp(\eta^T T(x))\, h(x)\, d\mu(x)}_{Z(\eta)} \right)$$

Ω_X

domain $\Omega \triangleq \{\eta \in \mathbb{R}^p \mid A(\eta) < \infty\}$ (set of valid canonical parameters)

note: $A(\eta)$ is convex in η

$\Rightarrow \Omega$ is convex

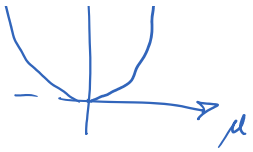
(*) more generally, consider a reparameterization of a subset of the family by defining the mapping $\eta: \Theta \to \Omega$

Θ: new set of parameters

consider $p(x;\theta) \triangleq p(x;\eta(\theta))$ for $\theta \in \Theta$

(get a "curved exponential family" if $\eta(\Theta)$ is a curved manifold in Ω)

e.g. could consider Gaussians where $N(\mu, \mu^2)$



* note: any single distribution can be put in an exponential family by using $h(x) \triangleq p(x)$ ($p(x)$)

* two examples of family not an exp. family:
  - mixture of Gaussians (latent variable models)
  - unif$(0, \theta)$

Example: (Multinoulli)

$X \sim \text{Mult}(\pi)$ $\quad \mathcal{X} = \{0,1\}^K$

$\Omega_X = \Delta_K \cap \mathcal{X}$ (one encoding)

parameter $\pi \in \Delta_K$; suppose $\pi_i > 0 \ \forall i$

$$p(x; \pi) = \prod_{j=1}^{K} \pi_j^{x_j} = \exp\left(\sum_j x_j \log \pi_j\right)$$

(π: think as "θ")

$$= \exp\left(\eta(\pi)^T x - 0\right)$$

we have $\eta_j(\pi) = \log \pi_j$

$T(x) = x$

$d\mu(x) =$ counting measure on $\mathcal{X}$

$h(x) = \mathbb{1}\{x \in \Omega_X\} = \mathbb{1}\{x$ has exactly one non-zero entry equal to $1\}$

$\mathcal{H} = \text{int}(\Delta_K)$ here, $A(\eta(\pi)) = 0 \ \forall \pi \in \mathcal{H}$

but here $\Omega = \mathbb{R}^K$

$\mathcal{H} \to$ dimension $K-1$

$\eta(\mathcal{H}) \to$ " $K-1$

$\Omega \to$ dimension K

we do not have a "minimal exponential family"

note: for any x s.t. $h(x) \neq 0$

here, $\sum_{j=1}^{K} T_j(x) x_j = 1$

here, $\sum_{j=1}^{K} T_j(x) \sum_{j} x_j = 1$

affine linear dep. between components of T

$\Rightarrow$ multiple η's give rise to same distribution; "overparametrization"

↳ not a "minimal" exp. family

⊛ for multinoulli, minimal exp. family

$$T(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{K-1} \end{pmatrix}$$

$$Z(\eta) = \sum_{x \in \Omega_X} \exp(\eta^T T(x)) = \sum_{j=1}^{K-1} \exp(\eta_j) + 1$$

$$p(x;\eta) = \exp\left(\sum_{j=1}^{K-1} \eta_j x_j - \underbrace{\log\left(\sum_{j=1}^{K-1} e^{\eta_j} + 1 \right)}_{A(\eta)} \right)$$

recall: $\nabla_\eta A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)]$ (valid for $\eta \in \text{int}(\Omega)$)

for multinoulli, $\frac{\partial A(\eta)}{\partial \eta_j} = \frac{1}{Z(\eta)} \exp(\eta_j) = p("x=j"|\eta)$

$= \mathbb{E}_{p(x;\eta)}[T_j(x)]$ as required