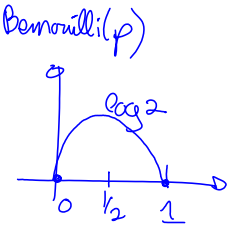
Lecture 15 - scribbles

Friday, October 21, 2016
13:35

today: KL divergence
exp family and duality

recall: Entropy $H(p) \triangleq -\sum_{x \in \mathcal{X}} p(x) \log p(x)$



in coding theory, use length of code $\propto -\log p(x)$
($\log_2 \to$ "bits", $\log_e \to$ "nats")

KL divergence: excess cost (in terms length of code) to use distribution q for coding instead of p

$$KL(p,q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)}$$

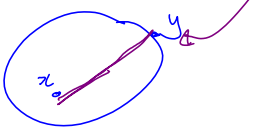
entropy of uniform on K states $\to -\sum_x \frac{1}{K} \log\left(\frac{1}{K}\right)$

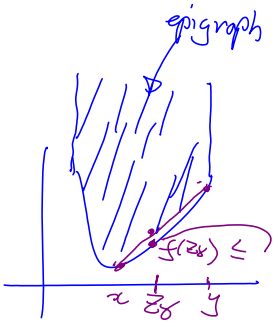
$\hookrightarrow H(p) = \log K$

properties of KL:

- $KL(p,q) \geq 0$
- is strictly convex in each of its argument
  - i.e. $KL(p, \cdot)$ is strictly convex as fct. of 2nd argument
  - $KL(\cdot, q)$ " "

convex set $A \iff \forall x, y \in A$, $(1-\gamma)x + \gamma y \in A$ for $\gamma \in [0,1]$





convex function $f \iff$ its epigraph is convex

$\hookrightarrow \{(x,t) : x \in \text{dom}(f), t \geq f(x)\}$

$\iff f((1-\gamma)x + \gamma y) \overset{\text{convex}}{\leq} (1-\gamma) f(x) + \gamma f(y)$

z_γ; $<$ strictly convex

if f is diff. $\to f(y) \geq f(x) + \langle \nabla f(x)$

If f is diff. $\Rightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$

ML and KL:

given $\{x_i\}_{i=1}^n$ observations — Kronecker-delta fct.

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

prop: $\{p_\theta\}_{\theta \in \Theta}$ parametric family on $\mathcal{X}$

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

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

$$= -H(\hat{p}_n) - \sum_x \hat{p}_n(x) \log p_\theta(x)$$

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

$$= -H(\hat{p}_n) - \frac{1}{n} \sum_{i=1}^n \log p_\theta(x_i)$$

$$\log \left(\prod_{i=1}^n p_\theta(x_i) \right)$$

//

also: can see ML as ERM for $L(P, p_\theta) = \mathbb{E}_{x \sim P}[-\log p_\theta(x)]$

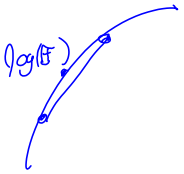
$$-KL(p, q) = \sum_x p(x) \log \frac{q(x)}{p(x)}$$

$$\mathbb{E}_p\left[\log \left(\frac{q(x)}{p(x)}\right)\right]$$

Jensen's

$$\leq \log \mathbb{E}_p\left[\frac{q}{p}\right] = \log 1$$

$$\sum_x p(x) \frac{q(x)}{p(x)}$$



$-KL(p, q) \leq 0$

Maximum entropy principle

idea: consider some subset of distributions on $\mathcal{X}$ according to some data-driven information

subset $M \subseteq \Delta_{|\mathcal{X}|}$

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

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

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

could use any h_0 as "prior" dist. in general ("generalized")

$q \in M$ $\quad \sum_x q(x) \log \frac{q(x)}{\text{cst.}}$ $\quad$ max. entropy

$= -H(q) + \text{cst.}$

example by Martin 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.B. . . .

[there Max-ent. solution is that $p(B, L) = p_B \cdot p_L$ (independance)]

* where do we get M?

typically: through empirical observations

feature functions $T_1(x) \dots T_d(x)$

define $M = \{ q : \underbrace{E_q[T_j(x)]}_{\text{model expected feature count}} = \underbrace{\hat{E}_{p_n}[T_j(x)]}_{\text{empirical feature count}}, j = 1, \dots, d \}$ ($\hat{q}_j$)

"moment constraints"

so Max-entropy principle: $\min_{q \in \Delta_{|\mathcal{X}|}} KL(q \,\|\, \text{unif.})$ $\quad q \in M$

$q \in M \Rightarrow \sum_x q(x) T_j(x) = \alpha_j$

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

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

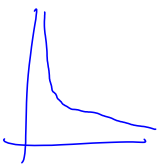
Lagrangian duality:

convex optimization problem $\min_x f(x)$ s.t. $f_j(x) \le 0$ $j = 1, \dots, m$, $g_k(x) = 0$ $k = 1, \dots, n$ } primal problem

- f, f_j are convex fct.
- g_k is affine

here, $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$

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



(extended real valued fct.)

$\sum_{i=1}^{m}$ $\quad \sum_{k=1}^{n}$

(extended real valued fct.)

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

λ_j, ν_k are "Lagrange" multipliers

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

an equivalent problem to primal is

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

duality trick is swap inf & sup

$$\sup_{\lambda \geq 0, \nu} \inf_{x} \mathcal{L}(x, \lambda, \nu) \triangleq g(\lambda, \nu)$$

Lagrange dual fct.

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

weak duality:

$$\sup_{\lambda \geq 0, \nu} \inf_{x} \mathcal{L}(x, \lambda, \nu) \leq \inf_{x} \sup_{\lambda \geq 0, \nu} \mathcal{L}(x, \lambda, \nu)$$

always true

moreover, $g(\lambda, \nu) \leq p^* \ \forall \lambda \geq 0, \nu$ where p^* is global min of primal

strong duality → have equality

a sufficient condition for strong duality is

Slater's condition: $\exists x \in \text{int}(\text{dom} f)$ s.t. $f_j(x) < 0 \ \forall j$ and x feasible

KKT conditions:

- necessary conditions for strong duality and $x^*, (\lambda^*, \nu^*)$ are respectively primal optimal and dual optimal:
  - $g(\lambda^*, \nu^*) = f(x^*)$
  - $= \inf_x \mathcal{L}(x, \lambda^*, \nu^*)$
  - complementary slackness: $\lambda_j^* f_j(x^*) = 0$

$\inf_x \rightarrow \nabla f(x) + \sum_j \lambda_j \nabla f_j(x) + \sum_k \nu_k \nabla g_k(x) = 0$

in other words, for differentiable fct.

necessary conditions for x^* be optimal and (λ^*, ν^*) be dual optimal is:

- x^* primal feasible
- (λ^*, ν^*) dual feasible ($\lambda^* \geq 0$)
- $\nabla f(x^*) + \sum_j \lambda_j^* \nabla f_j(x^*) + \sum_k \nu_k^* g_k(x^*) = 0$
- $\lambda_i^* f_i(x^*) = 0$

and if primal is convex, also sufficient

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

conditions → KKT conditions

dual problem for Max. Ent.:

$$(P)\quad \min_{q} \sum_x q(x)\log\frac{q(x)}{u(x)} \quad \left.\begin{array}{l} q(x)\geq 0 \ \}\ \forall x \\ \sum_x q(x)=1 \\ \sum_x q(x)T_j(x)=\alpha_j \end{array}\right\} \in M$$

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

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

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

exponential family

(we have strong duality by slater if $\exists q\in M$ s.t. $q(x)>0\ \forall x$)