Lecture 16 - scribbles

Friday, November 2, 2018 13:35

today: info theory & KL
max entropy
equivalence with ML in exp. family + duality

Information theory

KL divergence: for discrete dist. p & q

$$KL(p\|q) \triangleq \sum_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p\left[\log \frac{p(x)}{q(x)}\right]$$

$0 \cdot \log 0 = 0$

$\left(\lim_{x \to 0^+} x \log x = 0\right)$

motivation from density estimation

recall statistical decision theory

(statistical) loss $L(p_{\theta^*}, a)$ (θ^*: world)

↳ here, estimation of distribution, say $\hat{q}$

standard (ML) loss is log-loss $L(p_\theta, \hat{q}) = \mathbb{E}_{x \sim p_\theta}[-\log \hat{q}(x)]$

aside: "cross-entropy"

if use $\hat{q} = p_\theta$, then get $\sum_{x \in \Omega_X} -p(x) \log p(x) \triangleq H(p_\theta)$

entropy of θ

if use $q = p_\theta$, then get $\sum_{x \in \Omega_X} -p_\theta(x) \log p_\theta(x) = H(p_\theta)$ — entropy of p_θ

excess loss for action $a = \hat{q}$

$$L(p, \hat{q}) - \underbrace{\min_q L(p, q)}_{L(p,p)} = -\sum_x p(x) \log \frac{\hat{q}(x)}{p(x)} = KL(p \| \hat{q})$$

Coding theory:

$\lg \triangleq \log_2 \rightarrow$ "bits" $\quad \log_e \rightarrow$ "nats"

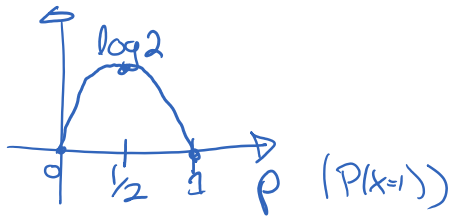
use length of code C_x $-\lg p(x)$

expected length of code: $\sum_x p(x)(-\lg p(x))$ entropy measured in bits

KL divergence → interpreted as the excess cost (in terms of length of code) to use dist q for coding vs. the optimal dist. (true p)

example:

entropy for a Bernoulli:



entropy of uniform dist. on K states: $-\sum_{x=1}^{K} \frac{1}{K} \log(\frac{1}{K}) = \log K$

(max entropy dist. over K states)

(max entropy dist. over K states)

properties of KL:

- $KL(p\|q) \geq 0$ ✓ to show this, use Jensen's inequality $f(\mathbb{E}x) \leq \mathbb{E}f(x)$ when f is convex
- is strictly convex in each argument
- not symmetric: $KL(p\|q) \neq KL(q\|p)$

i.e. $KL(p\|\cdot) : \mathbb{R}^K \to \mathbb{R}$, $KL(\cdot\|q)$ } strictly convex functions

$KL(\vec{p}\|\vec{p}) = 0 \quad \forall \vec{p} \in \Delta_K$

ML and KL minimization:

$\{p_\theta\}_{\theta \in \Theta}$ parametric family

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

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

proof:

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

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

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

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

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

14h29

Maximum entropy principle:

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

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

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

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

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

$KL(q\|u) = \sum_x q(x) \log \frac{q(x)}{u(x)} = -H(q) + \text{const.}$ $\quad u(x) = 1/K = \text{const.}$

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

h_0: preferred dist. to bias towards

* example from Wainwright:

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

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

question: how many kang. are both left-handed & drink S. beer?

[here: max entropy sol'n is that $p(B, L) = p_B \cdot p_L$ (independent)]

* how do we get M?

typically: through empirical "moments"

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

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

"moment constraints"

then MaxENT:

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

$$\rightsquigarrow \sum_x q(x) T_j(x) = \frac{1}{n} \sum_i T_j(x^{(i)}) = \alpha_j \quad (x^{(i)}: \text{data})$$

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

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

Lagrangian duality segway:

convex optimization problem:

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & f_j(x) \le 0 \quad \forall j = 1,\dots,m \\ & g_k(x) = 0 \quad \forall k = 1,\dots,n \end{aligned}$$

"primal problem"

- f, f_j are convex functions
- g_k affine fcts

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

$\text{dom}(f) \triangleq \{x : f(x) < \infty\}$

eg $f(x) = \begin{cases} -\log x & x > 0 \\ +\infty & \text{o.w.} \end{cases}$

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)

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

an equivalent problem to primal problem

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

duality trick is swap inf & sup

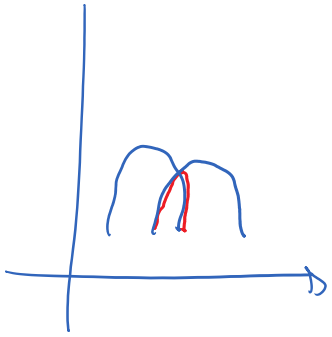
$h(x)$ fancy non-smooth fct.

sup of convex fct. $\Rightarrow$ convex
inf " convex " $\Rightarrow$ concave

duality trick is swap inf & sup $h(x)$ fancy non-smooth fct.

$$\sup_{\substack{\lambda \geq 0 \\ \nu}} \inf_{x} f(x, \lambda, \nu) \rightarrow \text{this fct. is always concave}$$

$$\triangleq g(\lambda, \nu) \quad \text{Lagrange dual fct.}$$



Lagrangian dual problem

$$\sup_{\substack{\lambda, \nu \\ \lambda \geq 0}} g(\lambda, \nu)$$

always true : weak duality : $\sup_{\nu} \inf_{x} f(x, \nu) \leq \inf_{x} \sup_{\nu} f(x, \nu)$

if $p^* = \inf_{\substack{x \\ x \text{ is feasible}}} f(x)$

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

g always concave in λ, ν

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

a sufficient condition for strong duality

is Slater's condition : $\exists x \in \text{int}(\text{dom}(f))$

Slater condition

+ s.t. $f_j(x) < 0$ ($\forall j$ where f_j is non-linear)

Slater condition
+
convex primal
$\Rightarrow$ strong duality

s.t. $f_j(x) < 0$ ($\forall j$ where f_j is non-linear)
and x is feasible

KKT conditions:

when f is differentiable

necessary conditions for x^* & ν^* to be primal & dual optimal

and strong duality

[when x^* is $\text{int}(\text{dom}(f))$]

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

$\Downarrow$

$$\nabla f(x^*) + \sum_k \nu_k^* \nabla g_k(x^*) = 0$$

+ x^* is primal feasible

KKT conditions

if primal is convex problem + Slater's condition $\Rightarrow$ sufficient conditions

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