Lecture 17 - MaxEnt

November 5, 2021 2:56 PM

today: information theory
Max Ent.
duality

Information theory

KL divergence: for discrete dist. p & q

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

convention: $0\cdot\log 0 = 0$ $\quad$ ($\lim_{x\to 0^+} x\log x = 0$)

[if $\exists x$ s.t. $q(x)=0$ but $p(x)\neq 0$]

$-p(x)\underbrace{\log q(x)}_{\to -\infty} = +\infty$

if support of p $\not\subseteq$ support of q $\Rightarrow KL(p\|q)=+\infty$

motivation from density estimation

recall statistical decision theory

(statistical) loss $L(p_\theta, a)$ — world; here, density estimation, say $\hat{q}$

Standard (MLE) loss is log-loss (cross-entropy loss) $L(p_\theta, \hat{q}) = \mathbb{E}_{X\sim p_\theta}[-\log \hat{q}(X)]$

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

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

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

$\min_q L(p,q) = L(p,p)$

Coding theory:

$\log \triangleq \log_2 \to$ "bits" $\quad \log_e \to$ "nats"

use length of code $\propto -\log p(x)$

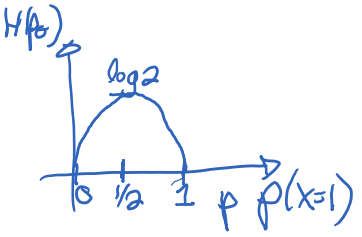
expected length of code: $\sum_x p(x)(-\log p(x))$ → entropy of p measure in bits

KL divergence → interpreted as excess cost (in terms of length of code) to use dist. q to design code vs

the optimal dist. (true p)

Example of H:

entropy for a Bernoulli: $-p\log p-(1-p)\log(1-p)$



entropy for a uniform dist. on k states: $-\sum_{x=1}^{k}\frac{1}{k}\log\frac{1}{k}=\log k$

(max entropy dist. over k states)

properties of KL divergence:

- $KL(p\|q)\geq 0$ ← to show this, use Jensen's ineq. $f(\mathbb{E}X)\leq \mathbb{E}f(X)$ when f convex
- KL is strictly convex in each of its arguments ie. $KL(p\|\cdot):\Delta_k\subseteq\mathbb{R}^k\to\mathbb{R}$ strictly convex fct. $KL(\cdot\|q)$
  - $KL(p\|p)=0\ \forall p\in\Delta_k$
- not symmetric $KL(p\|q)\neq KL(q\|p)$ in general

  Symmetrized version $\frac{1}{2}\left(KL(p\|q)+KL(q\|p)\right)$

MLE & KL minimization

$\{p_\theta\}_{\theta\in\Theta}$ parametric family of dist. for a discrete data space

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

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

proof:

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

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

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

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

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

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

//

cons. w.r. to θ $\quad \log \prod_i p_\theta(x^{(i)})$ //

Maximum entropy principle:

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

get a subset $M \subseteq \Delta_{|\mathcal{X}|}$ & prob simplex over $|\mathcal{X}| = k$ elements

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

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

$= \underset{q \in M}{\operatorname{argmin}}\ KL(q \,\|\, \text{uniform})$ $\quad u$

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

"generalized max. entropy" $KL(q \,\|\, h_\theta)$

prefered dist. to be biased towards

1607

* example from Wainwright

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

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

question: how many kang. are both L.H. & drink Sab. beer? (what proportion)

[here: max-entropy solution is that $p(B=1, L=1) = \hat{p}_B \cdot \hat{p}_L$ (indep.)]

* how do we get set M?

typically: through empirical "moments"

Kangaroo
$T_1(x) = \mathbb{1}\{x \text{ drinks Sabatt}\}$
$T_2(x) = \mathbb{1}\{x \text{ is L.H.}\}$

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

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 Max ENT

$\min_{q \in |\mathcal{X}|}$ $KL(q \,\|\, \text{unif.})$ $\rightarrow$ $\underbrace{\mathbb{E}_q[T_j(x)]}$ $\sum_x q(x) T_j(x) = \frac{1}{n} \sum_{i=1}^n T_j(x^{(i)}) = \hat{\alpha}_j$

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

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

ie. $\langle q, \vec{T}_j \rangle = \hat{q}_j$

↳ constraint.

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

Quick presentation of Lagrangian duality

convex min. problem

convex problem $\Leftarrow$ { f, f_j are convex fct. • g_k affine fct.

$\min_{x \in \mathbb{R}^d} f(x)$ s.t. $f_j(x) \leq 0$ $j = 1, \dots, m$; $g_k(x) = 0$ $k = 1, \dots, n$

} "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)

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

an equivalent form of the primal problem: $h(x)$ a fancy non-smooth fct.

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

duality trick is to swap inf & sup

min of concave fct. is concave

$\sup_{\lambda \geq 0, \nu} \left(\inf_x \mathcal{L}(x, \lambda, \nu) \right)$ → this is always concave (λ, ν)

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

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

$\sup \inf f \leq \inf \sup f$ "weak duality"

in general, $\sup_{\text{dual problem}} \inf f \leq \inf \sup f$

Strong duality when $\sup \inf f = \inf \sup f$ saddle pt.

↳ sufficient conditions: • when primal problem is convex + constraint qualif. cond. (e.g. Slater's condition)

(can get optimal primal variables $x^*(\lambda^*, \nu^*)$

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/