Density Estimation

ECE57000: Artificial Intelligence, Fall 2019

David I. Inouye

Announcements

- Resubmit HW2 since many formatting mistakes
- Quiz 3

<u>Density estimation</u> finds a density (PDF/PMF) that represents the data (or empirical distribution) well

- We <u>**always**</u> make an assumption about a <u>**density model class**</u> often parametrized by θ

- Assumption: Bernoulli density

$$\theta = [p], \qquad p \in [0,1]$$

- Assumption: Exponential density

$$\theta = [\lambda], \qquad \lambda \in \mathbb{R}_{++}$$

- Assumption: Gaussian density

$$\theta = [\mu, \sigma^2], \qquad \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{++}$$

- Assumption: DNN-based model

$$\theta = [\text{“}all\ neural\ network\ parameters\text{”}]$$

Informally, <u>Kullback-Leibler Divergence (KL)</u> measures the distance between distributions

- Formally, **<u>KL divergence</u>** for discrete variables

$$KL(P(x), Q(x)) = \mathbb{E}_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Formally, **<u>KL divergence</u>** for continuous variables

$$KL(p(x), q(x)) = \mathbb{E}_{X\sim p}\left[\log \frac{p(x)}{q(x)}\right] = \int_x p(x) \log \frac{p(x)}{q(x)} dx$$

- Note: **NO *negative sign*** compared to entropy
- Note: Not symmetric!
- Non-negative property: $KL\big(p(x), q(x)\big) \geq 0$
- Equal distribution property:

$$KL\big(p(x), q(x)\big) = 0 \Leftrightarrow p(x) = q(x)$$

One use of KL divergence is to estimate distribution parameters only from samples

- Let $p(x)$ denote the **real/true** distribution of the data
  - $p(x)$ is ***unknown***
  - We only have samples $\{x_i\}_{i=1}^{n}$ from $p(x)$
- Let $\hat{q}(x;\theta)$ denote an **estimate** of the true distribution
  - Parametrized by θ
- We want to find $\hat{q}(x;\theta)$ that is closest to $p(x)$

$$\theta^* = \arg\min_{\theta} \mathrm{KL}(p(x), \hat{q}(x;\theta))$$

One use of KL divergence is to estimate distribution parameters only from samples

- We want to find $\hat{q}(x;\theta)$ that is closest to $p(x)$

$$\theta^* = \arg\min_\theta \mathrm{KL}(p(x), \hat{q}(x;\theta))$$

- Wait, but we don't know $p(x)$, how do we do this?
  - (Simplify on board)
- Two main ideas for simplification
  - Constants with respect to (w.r.t.) θ can be ignored
  - Full expectation replaced by empirical expectation

<u>Maximum likelihood estimation (MLE)</u> is another way to estimate distribution parameters from samples

- **<u>Likelihood function</u>** how likely (or probable) a dataset $\mathcal{D} = \{x_i\}_{i=1}^n$ is under a distribution with parameters θ

$$\mathcal{L}(\theta; \mathcal{D}) = p(x_1, x_2, \dots, x_n; \theta)$$

- If we ***assume*** samples (or observations) of dataset are **<u>independent and identically distributed (iid)</u>**, then

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{n} p(x_i; \theta)$$

- Often simplified to the **<u>log-likelihood function</u>**

$$\ell(\theta; \mathcal{D}) = \log \mathcal{L}(\theta; \mathcal{D})$$

- Example: Coin flips with Bernoulli
- Non iid example: First flip Bernoulli, then alternating
- Example: Flight delays with exponential distribution

The likelihood function is a function of **parameters** θ as opposed to a density which is a function of x

- Sometimes written $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta})$

- Subtle but important difference with PDF/PMF
  - PDF/PMF are functions of x where θ is fixed
  - Likelihood is function of θ where x is fixed

- Additionally, likelihood function $\mathcal{L}$ is usually product of density functions (if **iid**)