Review of Probability

ECE57000: Artificial Intelligence

David I. Inouye

Sep. 18, 2019

Informally, **entropy** measures the "amount of randomness/disorder" of a distribution

- Formally, **entropy** for discrete variables

$$H\big(P(x)\big) = \mathbb{E}[-\log P(x)] = \sum_{x} -P(x) \log P(x)$$

- Formally, **differential entropy** for continuous variables

$$H\big(p(x)\big) = \mathbb{E}[-\log p(x)] = \int_{x} -p(x) \log p(x)\, dx$$

- Consider fair coin vs coin where both sides are heads

Maximum entropy probability distributions are the most “random” or “smooth” given expectation constraints

- Maximum entropy distribution problem

$$p^*(x) = \arg\max_{p(x)\in\mathcal{P}} H\big(p(x)\big)$$
$$s.t.\ \ \forall i \in \{1, \dots, k\}, \mathbb{E}_{x\sim p(x)}[f_i(x)] = \hat{\mu}_i$$

- “Maximal uncertainty while fitting data”

- Surprisingly simple solution:

$$p^*(x; \eta_1, \dots, \eta_k) \propto \exp\left(\sum_{i=1}^{k} \eta_i f_i(x)\right)$$

Maximum entropy principle is also similar to Occam's razor principle "Simple explanations better than complex ones."

- Suppose we only know that $X \in [0,1]$, what is the maximum entropy distribution $p(x)$?
- Uniform distribution $p(x) = 1$
- Suppose we know that $X \in [0,\infty)$ and that $\widehat{\mathbb{E}}[x] = \lambda$, what is the max entropy distribution?
- Exponential distribution
$$p(x) = \lambda \exp(-\lambda x)$$
  - (Check distribution properties on board, see if it matches form)

Gaussian distribution is the maximum entropy distribution given only mean and second moment/variance

- Suppose we know that $X \in \mathbb{R}$ and that $\mathbb{E}[x] = \eta_1$ and $\mathbb{E}[x^2] = \eta_2$, what is the maximum entropy distribution?
- Gaussian distribution:

$$p^*(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Wait, how does that have the same form as the solution? (derive on board)
- Check

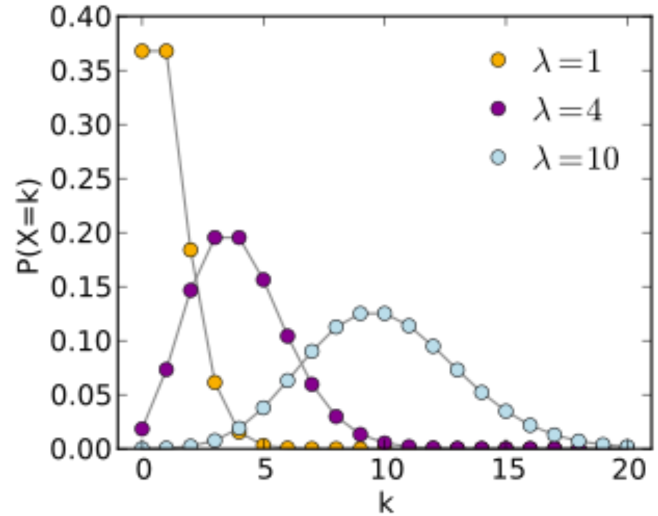
$$p^*(x) = \exp\left(\eta_1 x + \eta_2 x^2 + \frac{\eta_1}{4\eta_2} + \frac{1}{2}\log(-2\eta_2)\right)$$

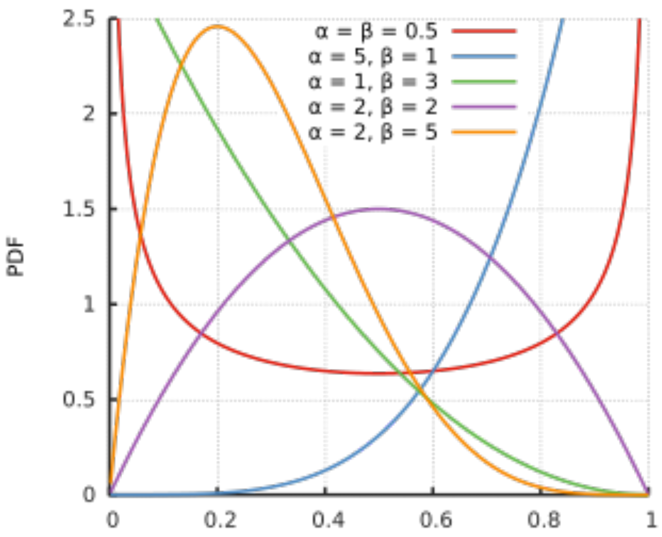
Many more common distributions are maximum entropy distributions

- Bernoulli (coin flip) distribution for $X \in \{0, 1\}$

- Poisson distribution for count data
$$X \in \mathbb{Z}_+$$

- Beta distribution for
$$X \in [0,1]$$





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