11:16

First Sec HOME INSERT DRAW VIEW

Stop

$x_1, x_2, \ldots, x_m$ $\quad p_{model}(x, \theta)$

$$\prod_{i=1}^{m} p(x_1, x_2, \ldots, x_m | \mu, \sigma^2) = \prod_{i=1}^{m} p(x_i | \mu, \sigma^2)$$

$$\arg\max_{\mu,\sigma^2} \prod_{i=1}^{m} p(x_i | \mu, \sigma^2) = \arg\max_{\mu,\sigma^2} \log\left(\prod_{i=1}^{m} p(x_i | \mu, \sigma^2)\right)$$

$$= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{m} \log\left(p(x_i | \mu, \sigma^2)\right)$$

$$= \arg\max_{\mu,\sigma^2} \frac{1}{m} \sum_{i=1}^{m} \log\left(p(x_i | \mu, \sigma^2)\right)$$

to make life easier

$X = x$ discrete

$$E(f(x)) = \sum_{i=1}^{m} f(x_i) P(X = x_i)$$

lognormal distribution $X \sim \hat{P}_{data}$

$$\frac{1}{m} \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$= \frac{1}{m} m \log\frac{1}{\sqrt{2\pi}\sigma} + \frac{1}{m} \sum_{i=1}^{m} \left((-1)\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{m} \sum_{i=1}^{m} \frac{(x-\mu)^2}{2\sigma^2}$$

$P_{data}(x)$ $\quad P_{model}(x_i;\theta)$ $\quad \hat{P}_{data}(x)$ (discrete)

actual distribution

$p_{model}(x;\Theta)$ $\qquad$ $\hat{p}_{data}(x)$
(discrete)

$D_{KL}(\hat{p}_{data} \| p_{model})$ shows how dissimilar these are.

$$\mathbb{E}_{x\sim\hat{p}_{data}}\left[\log \hat{p}_{data}(x) - \log p_{model}(x)\right]$$

$$= \mathbb{E}_{x\sim\hat{p}_{data}} \log\left(\hat{p}_{data}(x)\right) - \mathbb{E}_{x\sim\hat{p}_{data}}\left(\log p_{model}(x)\right)$$

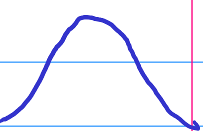
("from"/follows)

$$= \sum_{i=1}^{m} \frac{1}{m} \log\left(\frac{1}{m}\right) - \ldots$$

given (there are 'm' things & you can pick any)

$$= -\log m -$$

$$= \arg\min \mathbb{E}\left[\frac{-1}{\sqrt{2\pi}\,\sigma} + \frac{1}{m}\sum_{i=1}^{m}\left(\frac{x_i-\mu}{2\sigma^2}\right)\right]$$



KL divergence is called cross entropy (where the points taken are the same as the data or sth)

Cross entropy $(\hat{p}_{data}, \hat{p}_{model}(x;\mu,\sigma^2))$

$$= -\mathbb{E}_{x\sim\hat{p}_{data}}\left(\log p_{model}(x)\right)$$

Whenever you are finding the parameters by MLE, the error function ~~to~~ that is the result of the likelihood function, you call the error function as a "cross entropy" function

This applies to regression as well! if MSE ~~so~~ can be got from a likelihood function too, it's also a cross entropy error!

Deeplearning fixes the family of functions and uses that family!