18 mins gon

Apparently 15 mins spent by Sir in writing

If f is a function & p is a probability distribution for a r.v. X

$$E(f) = \sum f(x)\, p(x)$$
$$= \int f(x)\, p(x)\, dx$$

$$H(x) = E(-\log(p(x)))$$

$$H(y/x) = E(-\log(p(y/x)))$$
$$= \iint p(x,y) \log p(y/x)\, dy\, dx$$

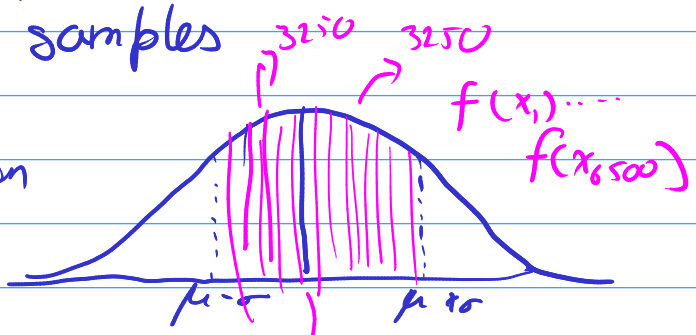
(just note this fn now)

$f(x_1)$ $f(x_2)$ ..... $f(x_n)$

$$E(f) \simeq \frac{1}{N} \sum_i f(x_i)$$

Let $x \sim N(\mu, \sigma^2)$

Collect 10,000 samples



When f is a function

$$E(f) = \int f(x) p(x) dx$$

if you don't know $p(x)$, how do we approximate $E(f)$ with $\frac{1}{N} \sum_{n=1}^{N} f(x_n)$

This is how: averaging out, we can get $f(\mu)$ LMAO

$$-\int p(x) \log(q(x))\, dx - \left(-\int p(x) \log(p(x))\, dx\right)$$

$p(x)$ → [A] → [B]

a b c d e f g h
000 001 010 011 100 101 110 111

Let all of them be in a basket. Let them all follow a distribution p.

if the 8 chars are uniform distribution

Avg code length $\frac{3(8)(10000)}{8000} = 3$

What's the encoding mechanism for least number of bits: Ans: $-\log(p(x))$

Shannon's: $\sum_x -p(x) \log(p(x))$ — Entropy

↳ min no. of bits

A → B

$-\log(p(x))$

$q(x) : -\log(q(x))$

Average information for $p(x)$ $= -\int p(x) \log(p(x))\,dx$

if $-\log(p(x)) = f(x)$

↓

Expectation of code length

then

$= -\int p(x) f(x)\,dx = E(f)$

if $p(x)$ dictates the universe
and we encode in $q(x)$
then avg code length

code length: $-\log(q(x))$
probability should be decided by universe : $p(x)$

$\Rightarrow -\int p(x) \log(q(x))\,dx$ → more

vs

$-\int p(x) \log(p(x))\,dx$ → lesser (by Shannon's encoding theorem)

increase of code length:

$\left(-\int p(x) \log(q(x))\,dx\right) - \left(-\int p(x) \log(p(x))\,dx\right)$

$$= -\int p(x) \log\left(\frac{q(x)}{p(x)}\right) dx$$ ⇒ This is the KL divergence of p and q

$KL(p||q)$

We dun no $p(x)$

$q(x|\theta)$, and we need to find the probability distributions for this!

So, $\min_{\theta} KL(p(x)||q(x|\theta))$

HOW TO FIND KL divergence without p?

(That's for another lecture) Tk it is @ bottom

$$E(f) \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

From here we can estimate

$$E(f) = -\int f(x)\, p(x) dx \simeq -\frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

$\min_{\theta}$(KL divergence)=

$$\min_{\theta} \; -\frac{1}{N} \sum_{n=1}^{N} \underbrace{f(x_n)}_{\log(q(x/\theta))} + \frac{1}{N} \sum_{n=1}^{N} \underbrace{g(x_n)}_{\log(p(x_n))}$$

$$= \min_{\theta} \; -\frac{1}{N} \sum_{n=1}^{N} \log q(x_n/\theta) + \frac{1}{N}(\log p(x_n))$$

$$= \min_{\theta} \; \frac{1}{N} \sum_{n=1}^{N} (-\log q(x_n/\theta) + \log(p(x_n)))$$

Now if we have θ_1 & θ_2, we can entirely disregard $\log(p(x_n))$

to find out which is better, since $p(x_n)$ is independent of θ, (But we need to sample points from p)

$$\Rightarrow \min_{\theta} \frac{1}{N} \sum_{n=1} (-\log(q(x_n/\theta)))$$

↳ likelihood (ol!

$$\min_{\theta} -\frac{1}{N} (\log q(x_1/\theta) + \log(q(x_2/\theta)) + \cdots + \log(q(x_n/\theta)))$$

$\min_{\theta}$ (negative log likelihood)! $\Rightarrow$ KL divergence is small

$\Rightarrow$ Average codelength is closest to minimum codelength

we see another face of log likelihood!

we however didn't actually use KL divergence from the START of the course, it was rather intuitive.

$p(x,y)$

(y/x)

given X, find average code length to specify y

will be a function $\int -p(y/x) \log(p(y/x))\, dy$ ← weighted average

now

$$H(y/x) = \int \int -p(y/x) \log(p(y/x))\, dy\, p(x)\, dx$$

$f(x)$

$\rightarrow$ expected value of average length

$$- \int\int p(y/x) p(x) \log(p(y/x))\, dy\, dx$$

$$H(y/x) = -\iint p(x,y)\log(p(y/x))\,dy\,dx$$

$$H(x/y) = -\iint p(x,y)\log(p(x/y))\,dx\,dy$$

$p(x,y) = p(x)p(y) \Rightarrow x, y$ are independent

$$H_x(x,y) = -\int p(x,y)\log(p(x,y))\,dy$$

$=$ ↳ for a given x

$$H(x,y) = -\iint p(x,y)\log(p(x,y))\,dy\,dx$$

$$= -\iint p(x,y)\log(p(y/x))\,dx\,dy$$

$$- \int\int p(x,y)\log(p(1))\,dx\,dy$$

$$= H(y/x) - \iint p(x,y)\log(p(x))\,dx\,dy$$

$$= H(y/x) - \int \log(p(x))\left(\int p(x,y)\,dy\right)dx$$

$$= H(y/x) - \int p(x)\log(p(x))\,dx$$

$$\boxed{H(x,y) = H(y/x) + H(x)} \rightarrow \text{Conditional Entropy}$$

Independence

(we can throw away a feature similar to another)

$p(x,y)$ $p(x)p(y)$ assumed [marginalized] on a feature that doesn't contribute

is the KL divergence less

To find the KL divergence in simpler terms as a measure of entropy

$$KL(p(x,y) \,||\, p(x)p(y))$$

$$= -\iint p(x,y) \log\left(\frac{p(x)p(y)}{p(x,y)}\right) dx\,dy$$

expand

$$= -\iint p(x,y)\left[\log(p(x)) + \log(p(y)) - \log(p(x)p(y/x))\right] dx\,dy$$

$$= -\iint p(x,y)\left[\log(p(x)) + \log(p(y)) - \log(p(x)) - \log(p(y/x))\right] dx\,dy$$

$$= -\iint p(x,y)\left[\log(p(y)) - \log(p(y/x))\right] dx\,dy$$

$$= -\iint H(x,y)\left[\log(p(y))\right] dx\,dy + \iint p(x,y) \log p(y/x)\, dx\,dy$$

$$= -\int \log p(y) \left(\int p(x,y)\,dx\right) dy - H(y/x)$$

$$= -\int \log p(y)\, p(y)\, dy - H(y/x)$$

$$= H(y) - H(y/x)$$

If we know the entropies of these we know the KL divergence

The Independent Mutual inference is hence given by

$$I(x,y) = H(y) - H(y/x)$$
$$= H(x) - H(x/y)$$

→ Similar elimination

if x & y are really independent

$H(y|x) = H(y)$ $\quad I(x,y) = 0$

$I(x,y)$ is the reduction of entropy of y, when x is involved

if y is dependent on x, then $I(x,y)$ can have many values, and this can give insights.

After this we look at a simple algorithm for classification & feature engineering, data preprocessing etc.