Cross Entropy and KL Divergence

Cross Entropy

Entropy: Expected(average) Surprise

Def: surprise score (ss) inverse of the probability

eg. P(A) = 0.01 if one observes A happens, one will be highly surprised.

```
SS(A) := log(1/P(A))
```

why log?

Initiation: 1/P(A) (Wrong)

Reason (one of): if P(A) = 1, see A happend, won't surprise (ss low) yet 1/P(A) = 1

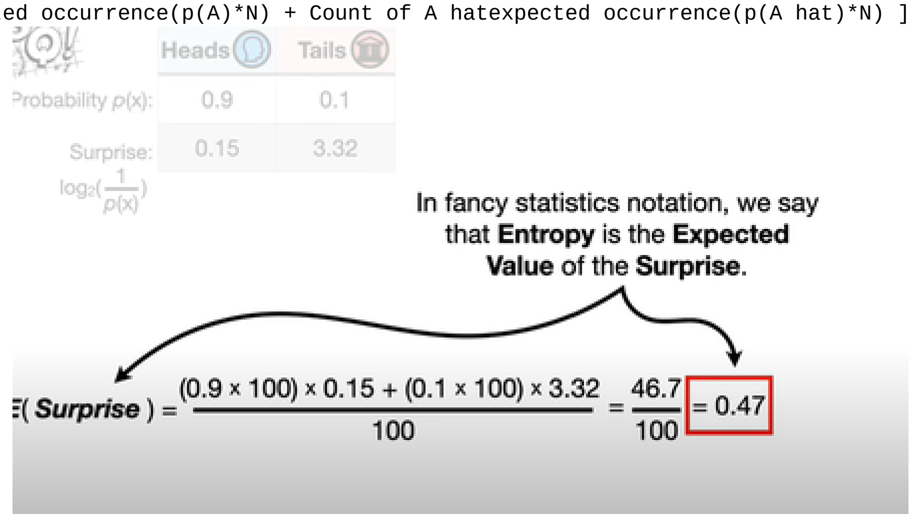
Alternative: log(1/P(A))

How to calculate expected (average) entropy?

Given p(A), P(A hat) ss(A) ss (A hat)

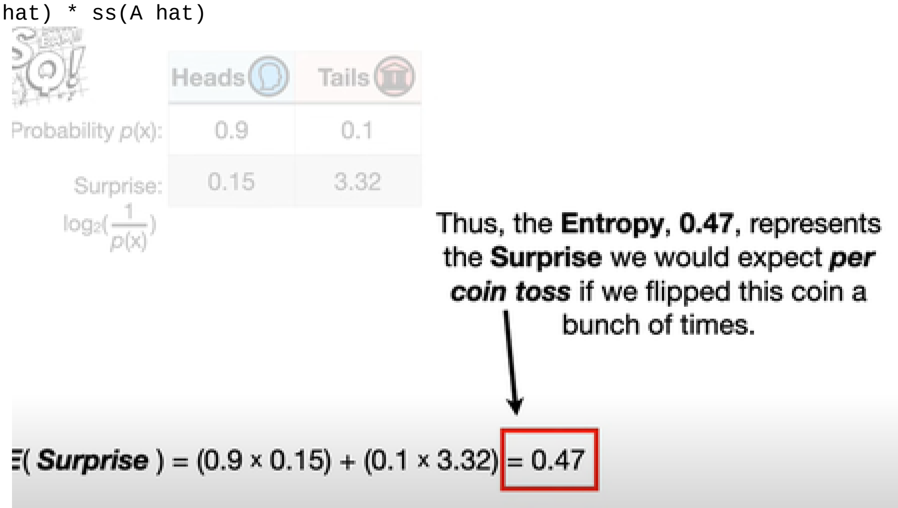
Averge interperation:

```
[Count of A expected occurrence(p(A)*N) + Count of A hatexpected occurrence(p(A hat)*N) ] / N
```



Expectation interpretation:

```
P(A)* ss(A) + P(A hat) * ss(A hat)
```



Derivation!

Def: surprise

$$E(ss(A)) = P(A)\cdot ss(A) + P(\bar{A})ss(\bar{A})$$
$$= P(A)\log\frac{1}{P(A)} + P(\bar{A})\log\frac{1}{P(\bar{A})}$$
$$= P(A)\cdot[\log(1) - \log(P(A))] + P(\bar{A})[\log(1) - \log P(\bar{A})]$$
$$= -P(A)\log(P(A)) + P(\bar{A})\log(P(\bar{A}))$$

discrete distribution X

$$E(ss(X)) = -\sum p(x)\log(p(x))$$

In information theory, the **entropy** of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes. Given a discrete random variable X, which takes values in the alphabet $\mathcal{X}$ and is distributed according to $p : \mathcal{X} \to [0,1]$:

$$\mathrm{H}(X) := -\sum_{x\in\mathcal{X}} p(x)\log p(x) = \mathbb{E}[-\log p(X)],$$

where Σ denotes the sum over the variable's possible values. The choice of base for $\log$, the logarithm, varies for different applications. Base 2 gives the unit of bits (or "shannons"), while base e gives "natural units" nat, and base 10 gives units of "dits", "bans", or "hartleys". An equivalent definition of entropy is the expected value of the self-information of a variable.[1]

KL Divergence

What is KL Divergence?

Kullback–Leibler divergence(**relative entropy**) is a type of statistical distance: a measure of how one probability distribution *P* is different from a second, reference probability distribution *Q*

How dose KL Divergence measure the different of two distribution?

Interperation: KL divergence of *P* from *Q* is the expected excess surprise from using *Q* as a model when the actual distribution is *P*

Relative entropy (difference) as element: `log(P(X) / Q(X))` using Q(X) beneath is because the Q(X) is model we want to measure against P (X) as the observed(real) model.

Definition [edit]

For discrete probability distributions P and Q defined on the same probability space, $\mathcal{X}$, the relative entropy from Q to P is defined[11] to be

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right).$$

which is equivalent to

$$D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x\in\mathcal{X}} P(x)\log\left(\frac{Q(x)}{P(x)}\right)$$

In other words, it is the expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the probabilities P.

Cross entropy

Same setting of surprise score: `SS(A) := log(1/Q(A))` normally Q(A) (the distrbution model which surprise score we wanna measure is the predicted model)

Yet when calculating expected surprise (entropy), the real distribution might be different (as P(A))

Thus cross-entropy

`P(A)* ss(Q(A)) + P(A hat) * ss(Q(A hat))`

H(p,q) = sum(p(x) * log(1/q(x))

$$H(p,q) = -\sum_{x\in\mathcal{X}} p(x)\,\log q(x)$$