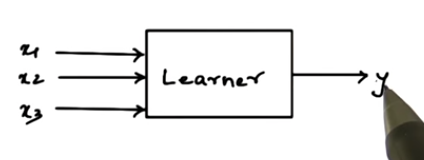
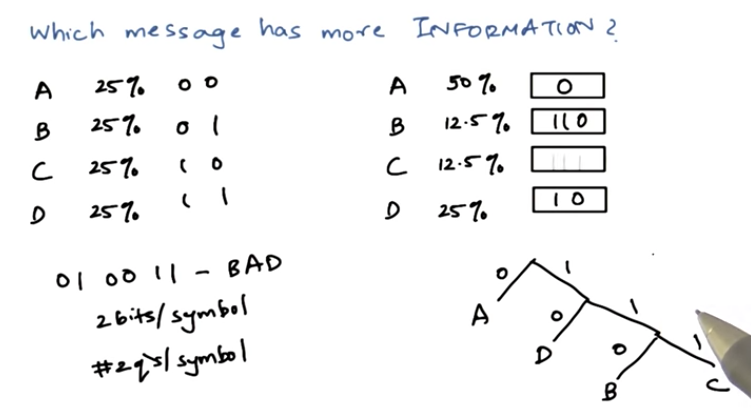
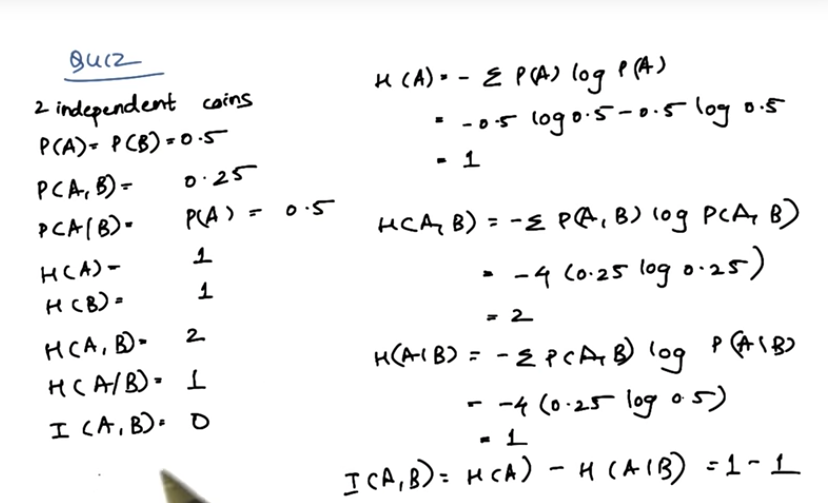
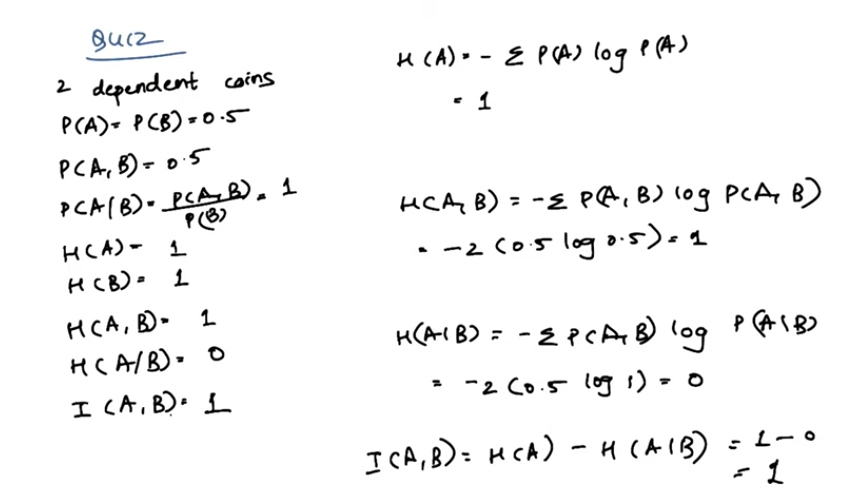
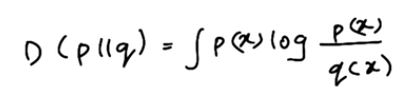
Information Theory



Which input gives you the most information about y?

Information theory: mathematical framework which allows us to compare the probability density functions of inputs

* Are the input vectors similar? “mutual information”
* Does this feature have any information? “entropy”
* Maxwell’s demon
  + Information is entropy
  + Erasing information increases the entropy of the environment
  + “demon” does not touch particles and can open and close a door such that more and more particles will gather on one side, which decreases entropy
  + If we have some idea of where a particle is, we need less information to describe where it is. If we know where it is, does this decrease the entropy of the universe? No, we must consider the demon’s knowledge entropy, which would offset any entropy decrease of having more knowledge of where the ball is.
  + It gets even more weird… As the demon tries to forget, she dissipates heat which then increases entropy again which offsets the loss of entropy from losing information.
* Coin flip example
  + If output is predictable, you don’t need to communicate anything
    - Less uncertainty = less information
    - Entropy
      * If you had to predict next example in sequence, what is the minimum number of yes/no questions needed (2 bits, 2 questions)
  + If the output is random, you need to communicate result of each and every flip
* ABCD example
  + 
  + What is the expected message size in this language?
    - P(symbol)\*#bits = 1.75 bits per message
    - Variable length encoding
* Information between two variables
  + Joint entropy - Is there some information in one variable that tells you something about another variable
  + conditional entropy – measure of randomness of one variable given other variable
* mutual information – measure of the deduction of randomness of a variable given knowledge of some other variable
* 
* 
* Kullback-Leibler divergence (KL Divergence)
  + Difference between any two distributions, it is a distance measure
  + 
  + Does not follow triangle law, so not “really” a distance measure
  + Can use well known dataset for p(x) and can sample our dataset for q(x)
  + Can use as substitute for least square method
* Summary
  + Information can be measured in terms of entropy
  + Joint and conditional entropy and mutual information to tell us information between two variables
  + KL divergence can be used as a distance measure between two distributions