Practice 8.1&8.2

**1.**Is the following statement true or false?

"The entropy of a random variable can be measured in bits."

* True
* False

**2.**Is the following statement true or false?

“The perplexity of a model is always higher than the entropy. Therefore we should choose the model with the smallest entropy, rather than the one that has the smallest perplexity.”

* True
* False

**3.** Is the following statement true or false?

“The reverse KL divergence DKL(q\bmx||p\bmx) exhibits mode-seeking behaviour when minimised under model misspecification, assuming that we are trying to approximate

p\bmx with q\bmx.”

* True
* False

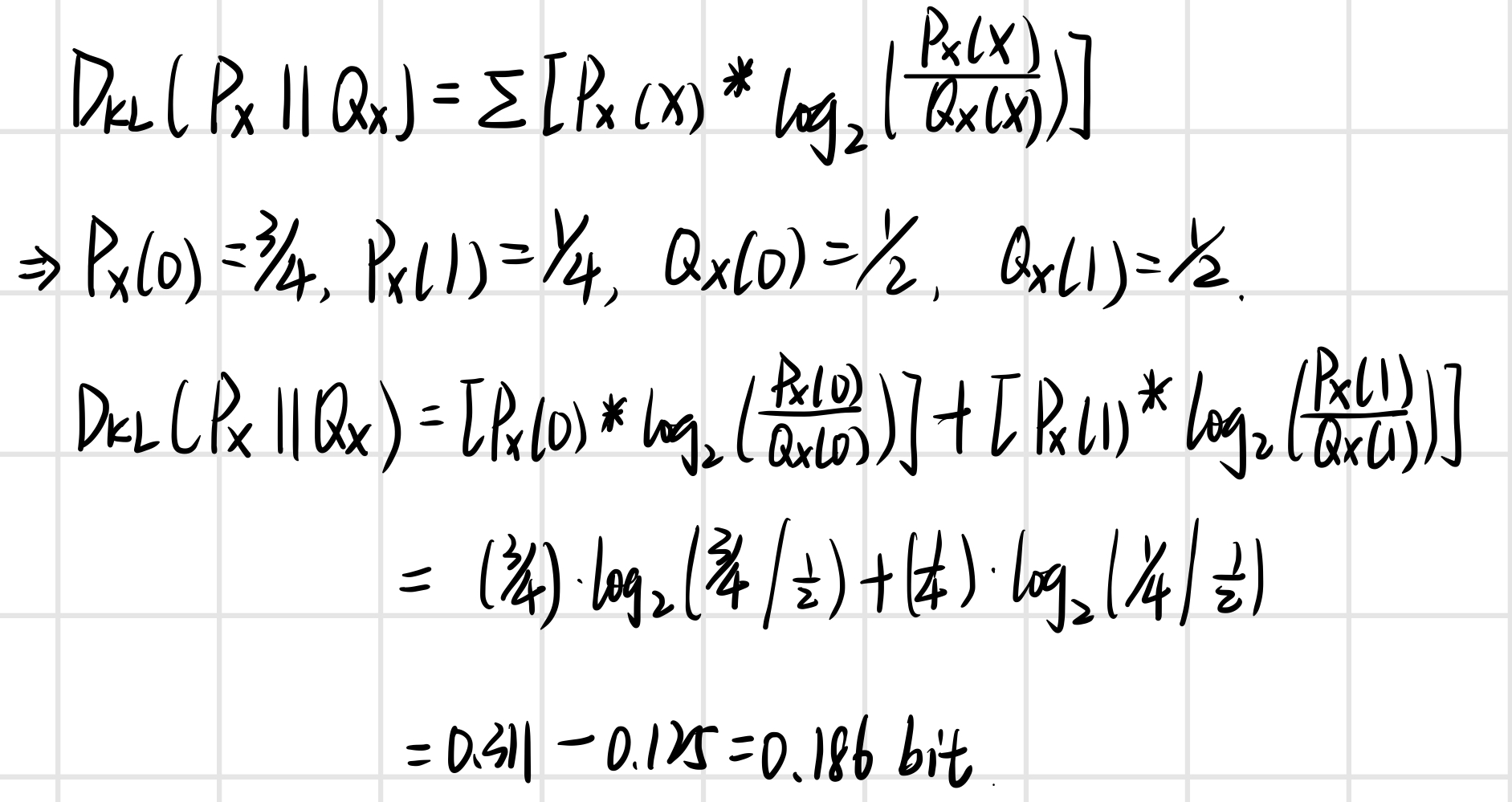
**4.**Is the following statement true or false?  
"The KL divergence is an excellent distance function for densities, since it behaves analogously to the Euclidean distance between two points."

* True
* False

5. Let Px and Qx be two distributions on {0,1}. Px(0)=3/4 and Px(1)=1/4, while Qx(0)=Qx(1)=1/2.

Compute the Kullback-Leibler divergence DKL(Px||Qx). Give your answer in bits. A margin of error of ±0.01 is allowed.

Answer: 0.19



**6.**Choose the correct answer.

* The Shannon information content of an outcome is **more than** the probability of the outcome being observed.
* The Shannon information content of an outcome **has a monotonic relationship** with the probability of the outcome being observed: **as the probability decreases, the information content decreases as well.**
* The Shannon information content of an outcome is **less than** to the probability of the outcome being observed.
* The Shannon information content of an outcome **has a monotonic relationship** with the probability of the outcome being observed: **as the probability decreases, the information content increases.**
* The Shannon information content of an outcome is **equal to** the probability of the outcome being observed.

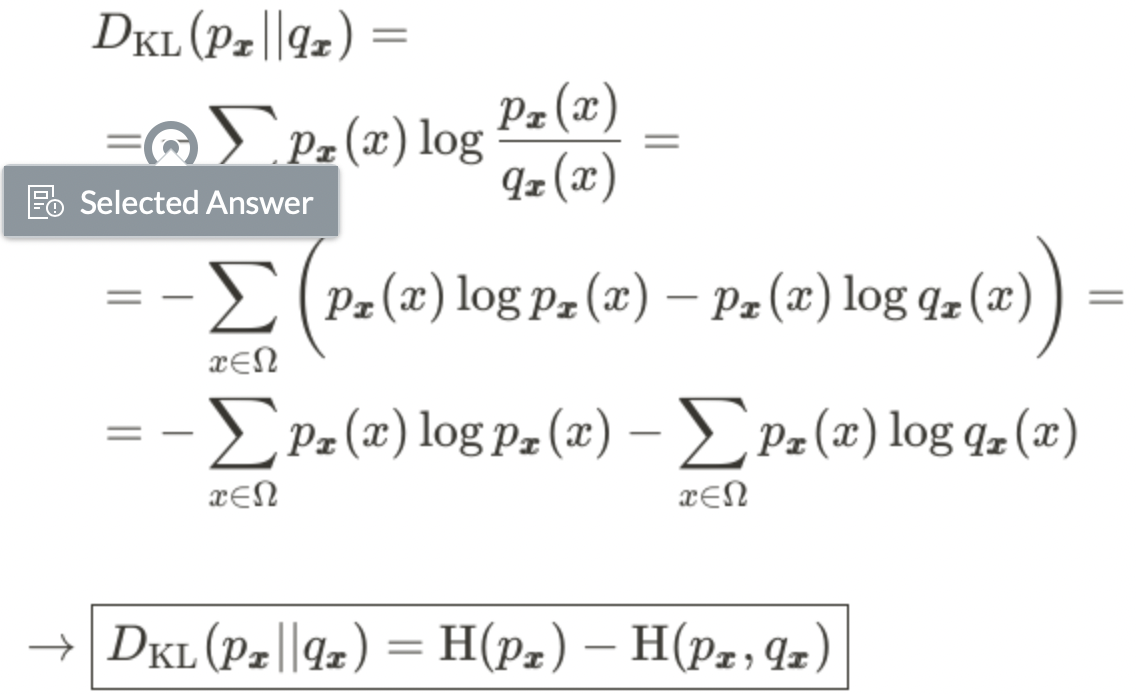
7.Is the following statement true or false?

“The information content of the joint outcome (x,y) of any two random variables \bmx and \bmy

is equal to the sum of the information content of the two individual outcomes.”

* True
* False

**8.**Place a mark somewhere on the first incorrect line in the derivation below.

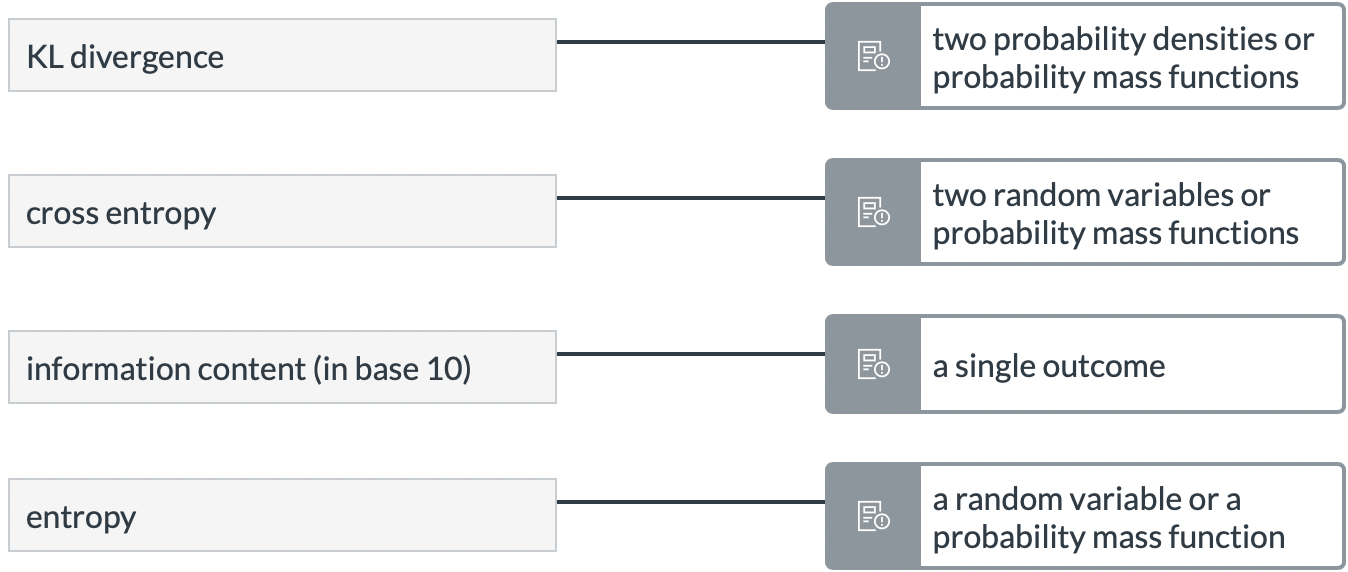


9.Let X be a dataset of n i.i.d. random samples from the density p\bmx(x). Is the following statement true or false?

“In the limit n→∞, the maximum-likelihood estimated distributionq^\bmx(x)=q\bmx(x;θ^ML(X)) converges on the distribution that minimises the Kullback-Leibler divergence DKL(p\bmx||q^\bmx).”

* True
* False

**10**.Match the following information-theoretic functions with the options that best describe their inputs.

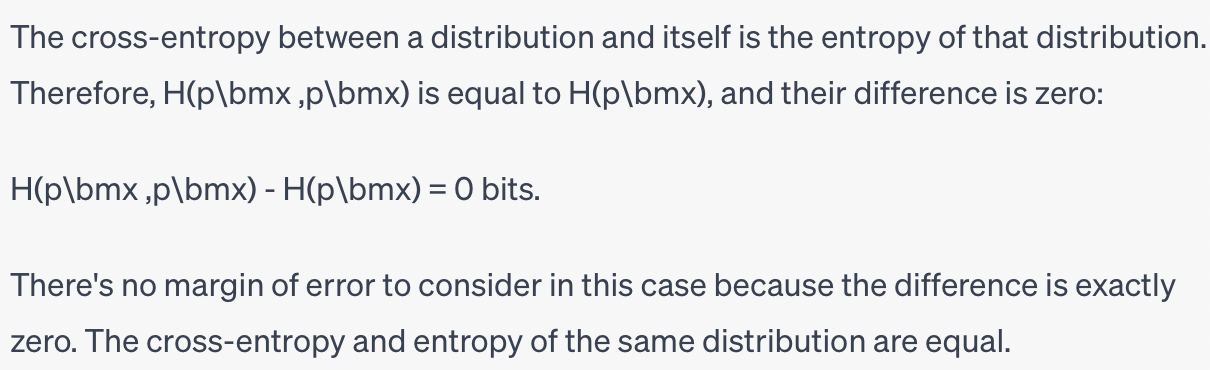


**11**.Let p\bmx be a Poisson distribution with mean 1. Let further H(p\bmx) be the Shannon entropy of the distribution and let H(p\bmx,p\bmx)

be the cross entropy between the distribution and itself.

What is H(p\bmx,p\bmx)−H(p\bmx), measured in bits? A margin of error of ±0.1 is accepted.

Answer: 0

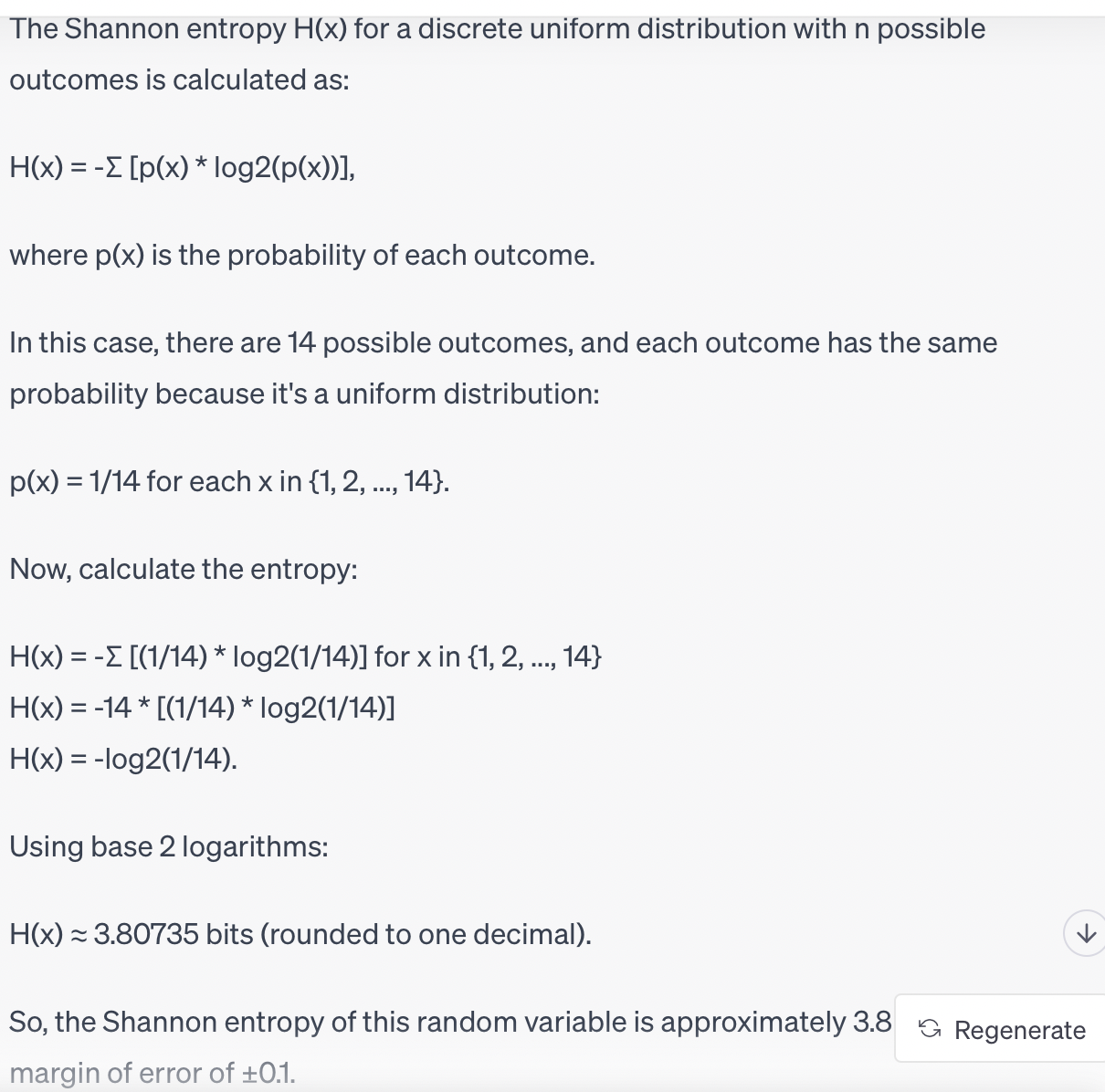


12.Let x be a random variable with a uniform distribution on the random integers {

1, 2, ..., 14}. What is the Shannon entropy H(x) of this random variable?

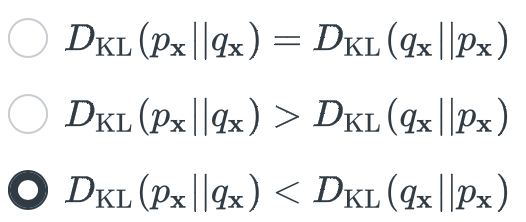
Give your answer in bits, rounded to one decimal. A margin of ±0.1 suffices.

Answer: 3.8

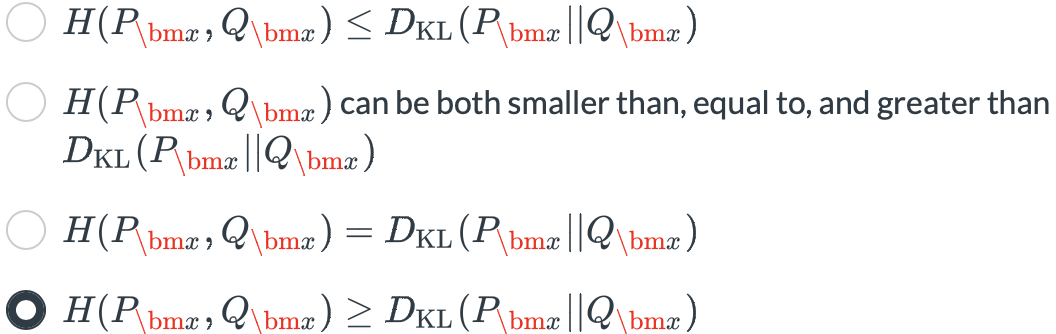


13. Let px and qx be uniform distributions on the real line that both have the same mean μ. More specifically, we let px be uniform on [μ−1,μ+1] while qx is uniform on [μ−10,μ+10].

Both distributions have the same mean, but qx has 100 times the variance of px. Which statement is correct?



14.Let P\bmx and Q\bmx be the probability mass functions functions of two probability distributions on the same space. Let further H(P\bmx,Q\bmx) be the cross entropy and DKL(P\bmx||Q\bmx) be the Kullback-Leibler divergence between the two distributions.

Mark the correct statement below.

Practice 8.3

**1.**Please indicate whether the following statement is true or false:

“Because we are using a greedy algorithm, decision-tree learning will find the shallowest tree that is globally optimal.”

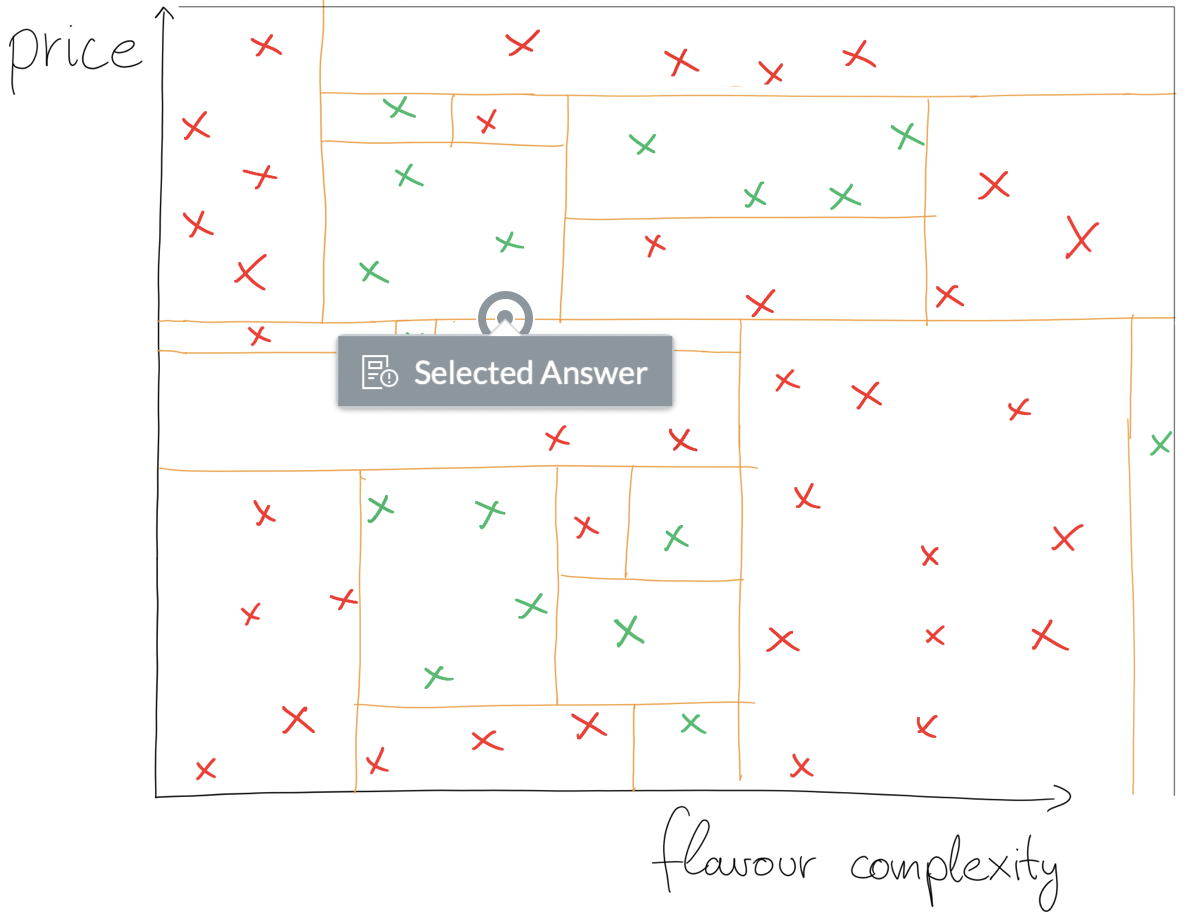
* True
* False

**2.** Please indicate whether the following statement is true or false:

“The leaf nodes in a decision tree always contain a class label.”

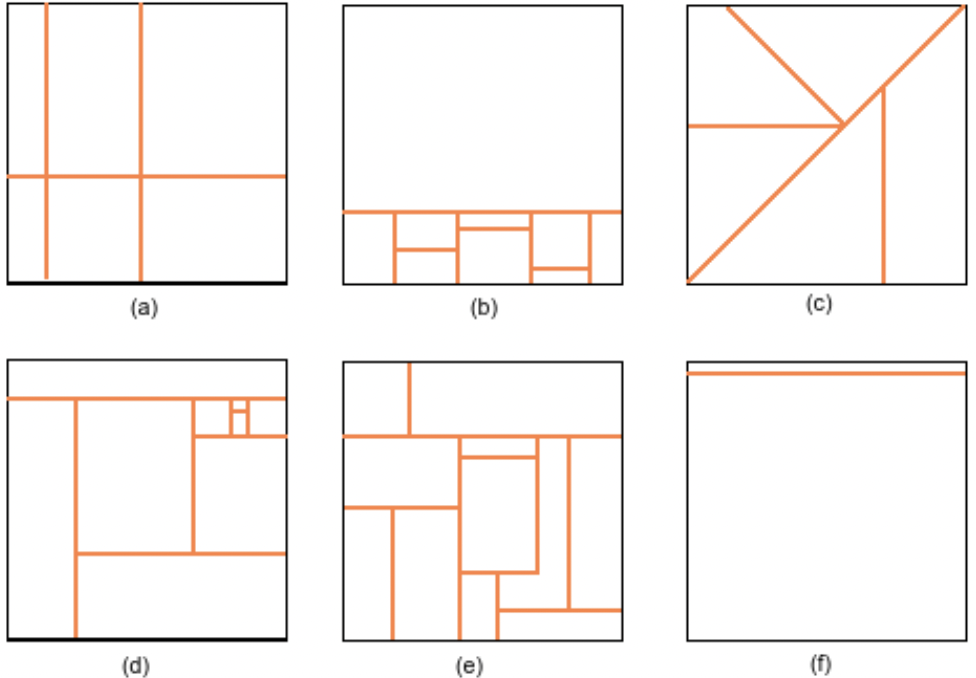
* True
* False

**3.**The below picture illustrates the axis-parallel splits learned by a decision tree model. Mark the line that corresponds to the very first split.



**4.** Mark all true statements below.

* Decision trees can always reach 100% accuracy on the training dataset, even when paired with regularisation.
* Unless decision tree learning is paired with regularisation, it will not stop until every datapoint is separated into a different region. This ensures that the model achieves 100% accuracy on the training dataset.
* Decision trees are high-variance models. Even a few new datapoints could radically change the predictions.
* In practice, decision trees should be paired with some form of regularisation. For example, setting a maximum depth limit is a good default choice, as it is guaranteed to yield good results.

**5**. The above picture shows 6 possible partitionings of a 2D input space. Mark all choices that correspond to a valid decision tree.

* (a)
* (b)
* (c)
* (d)
* (e)
* (f)

**6.** Is the following statement true or false?

“Classic decision trees only allow for binary questions. This design choice reduces the time it takes to fit the tree, but it reduces the expressiveness of the model. For example, binary decision trees cannot handle categorical input features.”

* True
* False

**7.** Decision trees are very popular models, especially in applied machine learning. They have numerous advantages: they **can be more interpretable** than neural networks, and they are

**easier to implement** . On the other hand, decision trees are considered **high-variance models**, and they **are inefficient for certain problems** .

**8.** Is the following statement true or false?

“When making predictions, decision trees always ask the same sequence of binary questions about the input x, regardless of the value of x.”

* True
* False

**9.** Decision tree learning is considered **a greedy, recursive** algorithm.