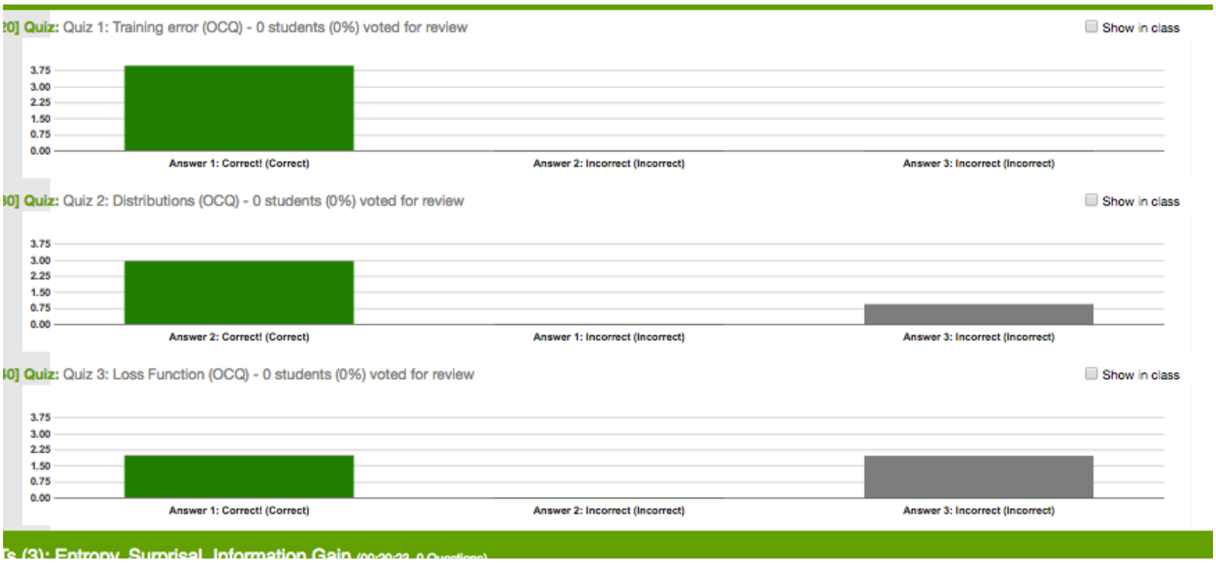
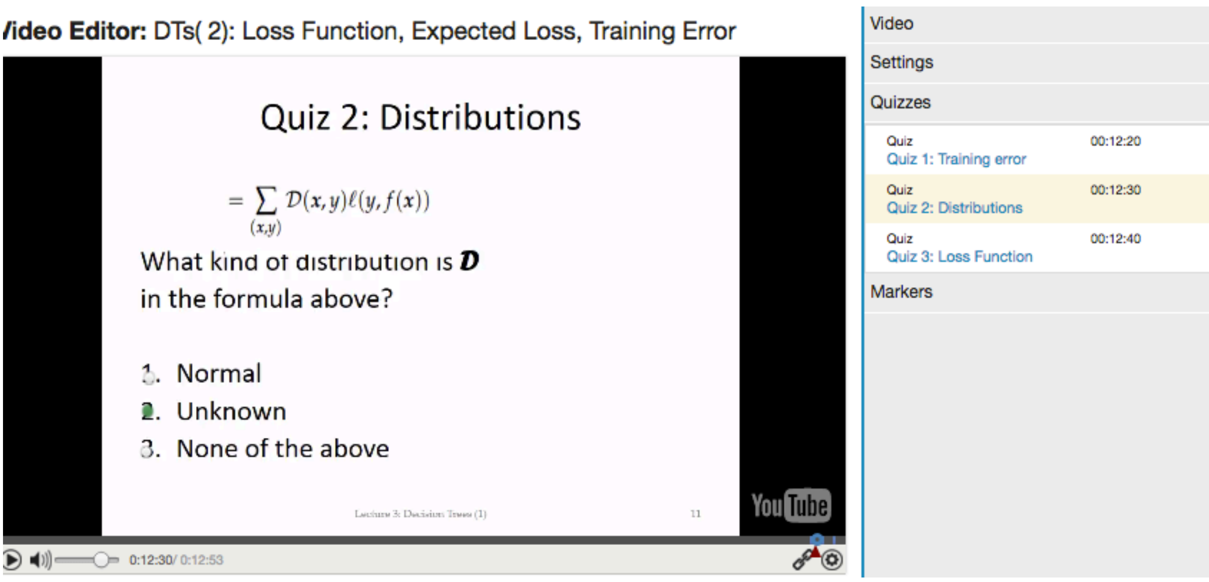
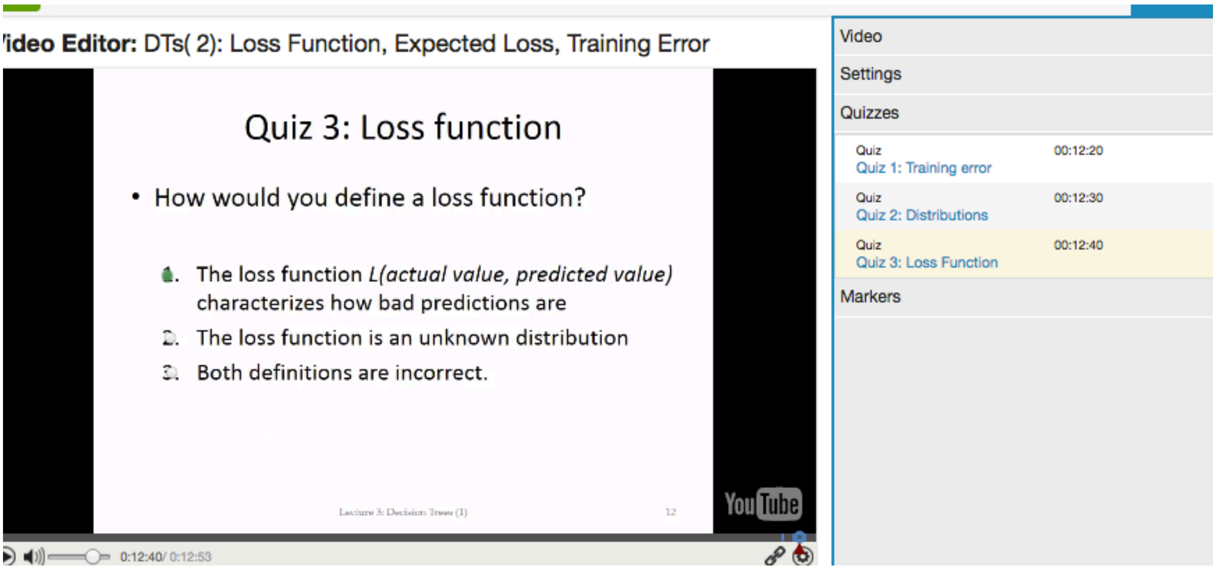
Feedback – DTs –Quizzes – 14 Nov 2106

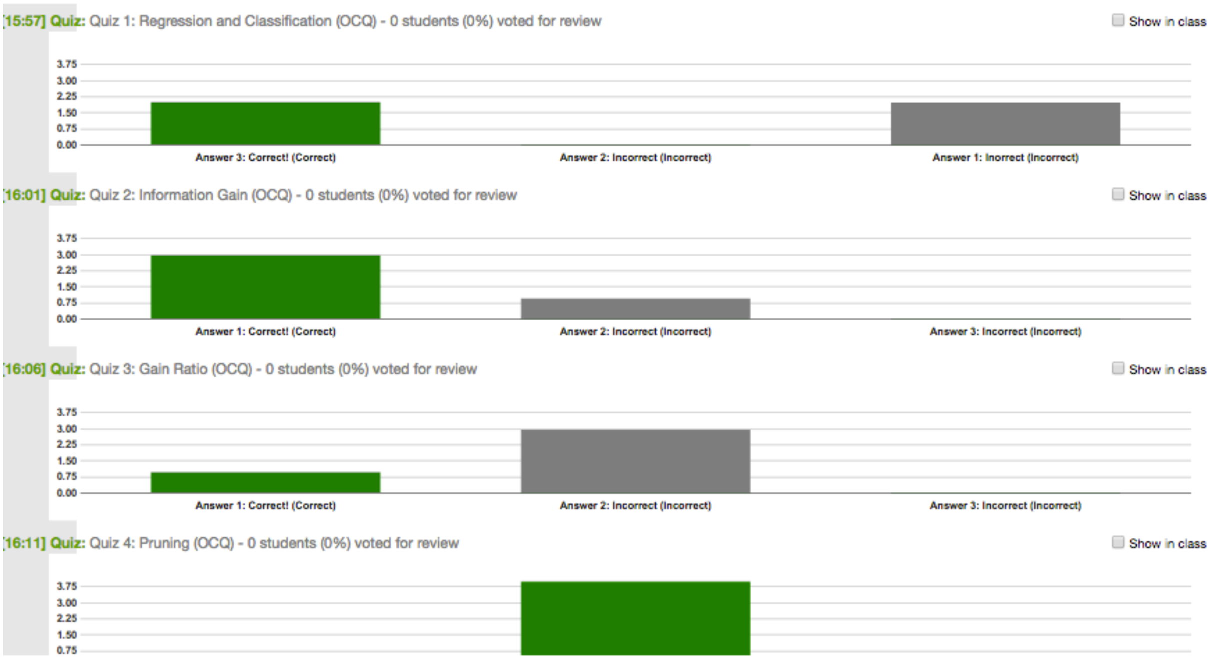


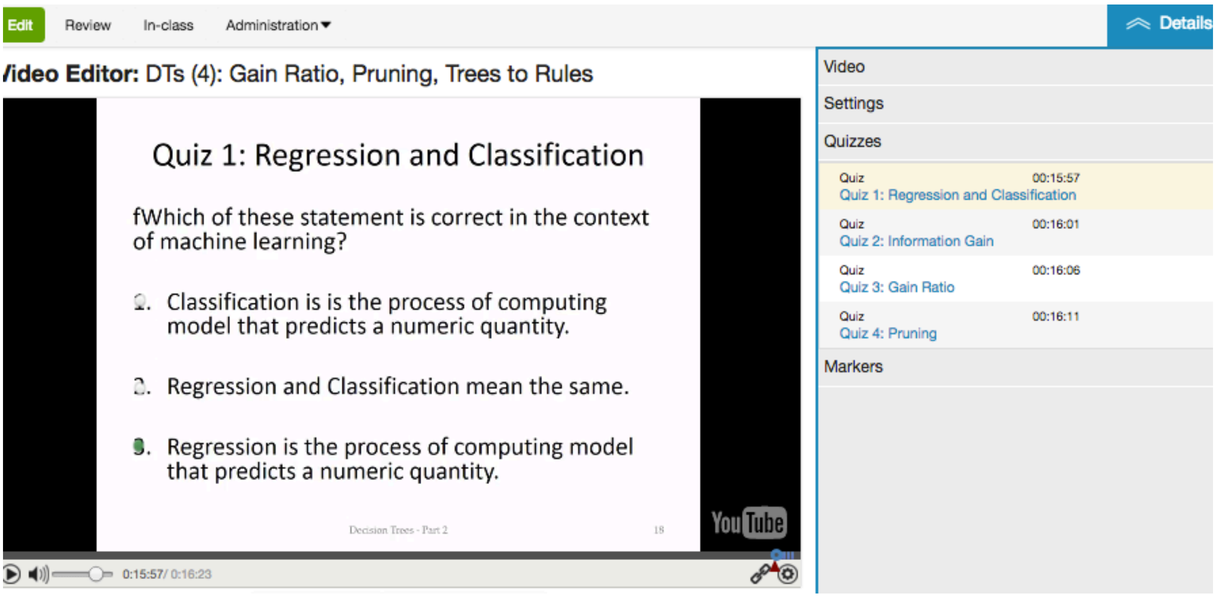


This formula is extracted from formula 1.1. (Daume': 15). Daume' says that (in this case) we are not making any assumptions about the distribution D ("For instance, we are not assuming it looks like a Gaussian or some other, common distribution). Read also pag 16 "a sample D from some unknown distribution D".



a very simple words (maybe too simple?), you could say that the loss function quantifies the amount of wrong predictions. There different types of loss functions (see next lectures). The one explained by Daume' is 0-1 loss: you count 1 for every error the classifier makes. There are several types of loss function and they will be described in Lect 04. Several loss functions are described in Witten et al (2011), Chap 5.





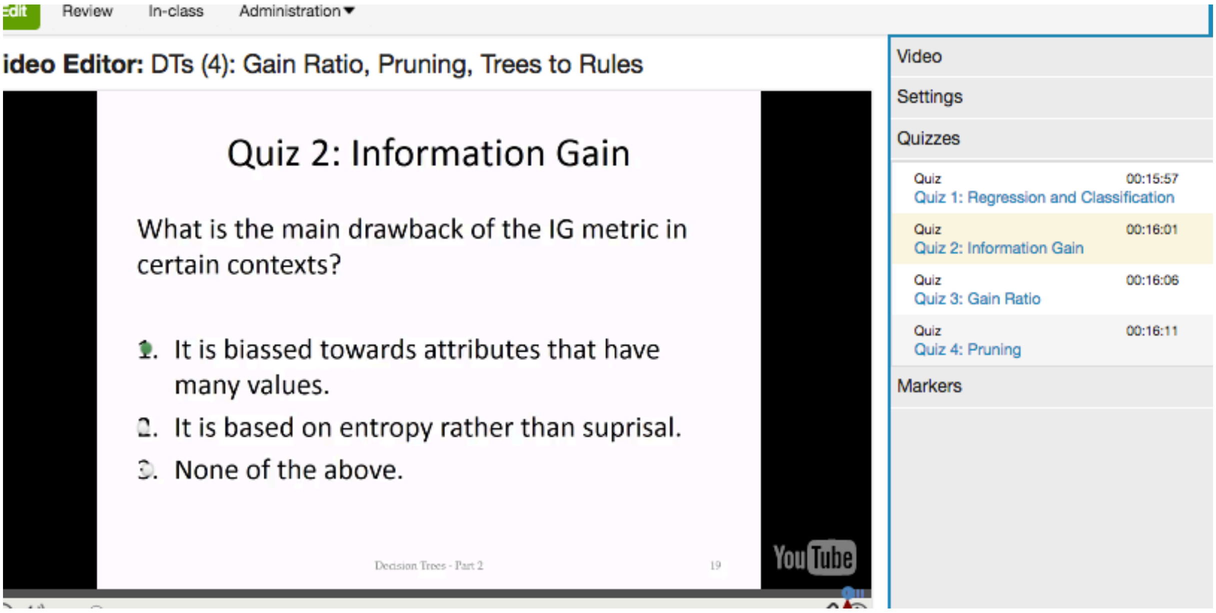
The key difference is that the output.

Regression is used to predict continuous values. For regression the output is a continuous number (e.g. "what's the expected amount the stock market will go up or down today?" or "how carcinogenic is this compound?").

Classification is used to predict which class a data point is part of (discrete value). For classification is one of a discrete set (e.g. "what type of fruit is this?" or "what medical condition is the person suffering from?")

Att! We are going to work with the spam dataset shortly: are the class labels discrete or continous? Discrete data can be numeric -- like numbers of apples -- but it can also be categorical -- like red or blue, or male or female, or good or bad. Continuous data are not restricted to defined separate values, but can occupy any value over a continuous range.

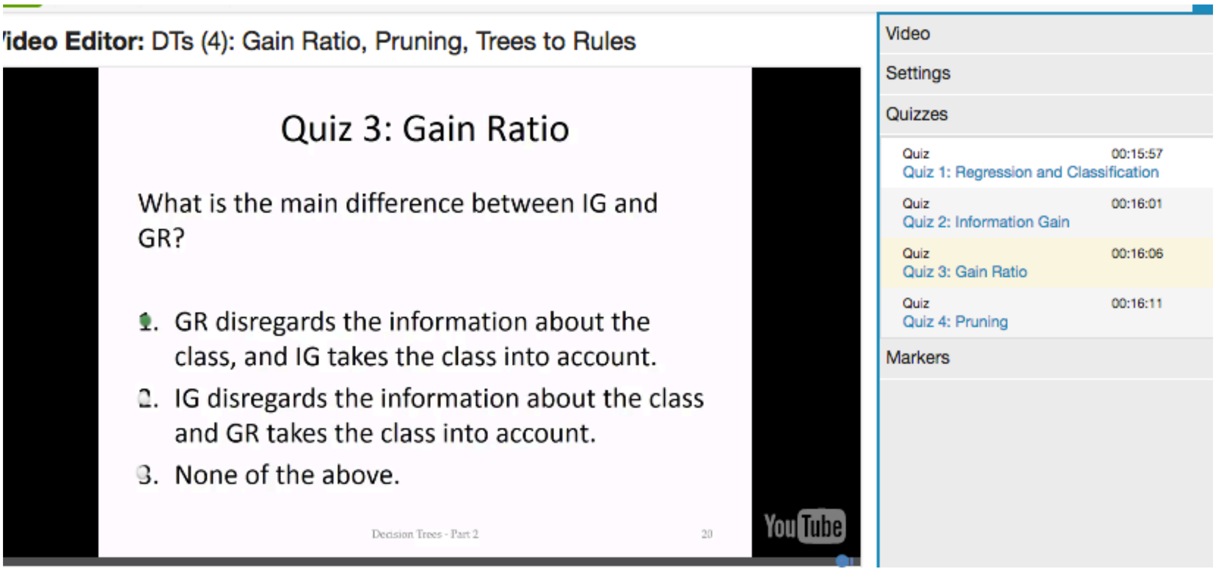
Some learning algorithms are targeted at one, others at the other. Many of them have the same core algorithm, but with minor modifications -- e.g. decision trees can be easily modified to support regression instead.



First of all, watch out terminological uses: sometimes gain ratio is called ”normalized information gain”. Sometimes you can read that entropy is the same as information gain. Sometimes the word ”information” is used indistinctly for entropy-based measure. Some other times, everything is unified under the concept of “information entropy”. Make sure you understand the basic concepts beyond the labeling terms, and always specify what you mean by the term you use. Apply your critical thinking when you browse the web for additional information.

To simplify, let’s use the following terminology here:

- Entropy H(S) is a measure of the amount of uncertainty in a set of example S. Entropy only computes the quality of a single set of examples. How can we compute the quality of the entire split, ie of the entire attribute?
- With Information Gain we compute the weighted average over all sets resulting from the split. Information gain **IG(A)** is the measure of the difference in entropy from before to after the set S is split on an attribute **A**. In other words, how much uncertainty in S was reduced after splitting set S on attribute **A**. The weka book explains with a very comprehensive example why IG is biased towards multi values. Another useful example on IG: https://courses.cs.washington.edu/courses/cse455/10au/notes/InfoGain.pdf
- Gain Ratio is modification of the information gain that (1) reduces its bias towards multi-valued attributes, (2) takes number and size of branches into account when choosing an attribute (3) corrects the information gain by taking the **intrinsic information** of a split into account.



Check the concept of "intrinsic information" in the weka book.

"intrinsic information" = the information about the class is disregarded. *If we neglect the class, we get the intrinsic information of a node*. We divide the information gain by the intrinsic information and we get the GR.