Feature Engineering: 2nd lesson – Mutual Information

First encountering a new dataset can sometimes feel overwhelming. You might be presented with hundreds or thousands of features without even a description to go by. Where do you even begin?

A great first step is to construct a ranking with a **feature utility metric**, a function measuring associations between a feature and the target. Then you can choose a smaller set of the most useful features to develop initially and have more confidence that your time will be well spent.

The metric we'll use is called "mutual information". Mutual information is a lot like correlation in that it measures a relationship between two quantities. The advantage of mutual information is that it can detect any kind of relationship, while correlation only detects *linear* relationships.

Mutual information is a great general-purpose metric and especially useful at the start of feature development when you might not know what model you'd like to use yet. It is easy to use and interpret, computationally efficient, theoretically well-founded, resistant to overfitting, and able to detect any kind of relationship.

Mutual information:

Mutual information describes relationships in terms of uncertainty. The **mutual information** (MI) between two quantities is a measure of the extent to which knowledge of one quantity reduces uncertainty about the other. If you knew the value of a feature, how much more confident would you be about the target? Here's an example from the *Ames Housing* data. The figure shows the relationship between the exterior quality of a house and the price it sold for. Each point represents a house.



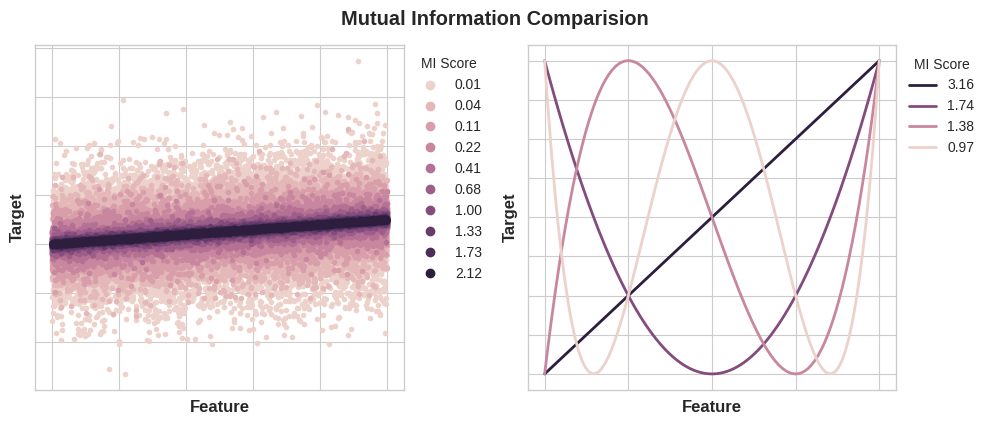
From the figure, we can see that knowing the value of ExterQual should make you more certain about the corresponding SalePrice -- each category of ExterQual tends to concentrate SalePrice to within a certain range. The mutual information that ExterQual has with SalePrice is the average reduction of uncertainty in SalePrice taken over the four values of ExterQual. Since Fair occurs less often than Typical, for instance, Fair gets less weight in the MI score.

Technical note:

What we're calling uncertainty is measured using a quantity from information theory known as "entropy". The entropy of a variable means roughly: "how many yes-or-no questions you would need to describe an occurance of that variable, on average." The more questions you have to ask, the more uncertain you must be about the variable. Mutual information is how many questions you expect the feature to answer about the target.

Interpreting mutual information scores:

The least possible mutual information between quantities is 0.0. When MI is zero, the quantities are independent: neither can tell you anything about the other. Conversely, in theory there's no upper bound to what MI can be. In practice though values above 2.0 or so are uncommon (mutual information is a logarithmic quantity, so it increases very slowly). The next figure will give you an idea of how MI values correspond to the kind and degree of association a feature has with the target.



Here are some things to remember when applying mutual information:

* Mutual information can aid understanding the *relative potential* of a feature as a predictor of a target, considered by itself.
* It is possible for a feature to be very informative when interacting with other features, but not so informative all alone. Mutual information *cannot detect interactions* between features. It is an unvariate metric.
* The *actual* usefulness of a feature *depends on the model* used it with. A feature is only useful to the extent that its relationship with the target is one that model can learn. Just because a feature has a high MI score doesn't mean that model will be able to do anything with that information. Programmers may need to transform the feature first to expose the association.