Feature Importance

Feature Importance

After coming up with a bunch of features to describe your data, you might be tempted to investigate which of them deliver the most bang for their buck. Try not to fall into this trap by making too many assumptions about which features are truly relevant! There will be plenty of time and better tools for doing that later in the data analysis process. While out collecting, time spent deliberating whether you should move forward with a particular feature is time not spent adding more samples into your dataset! What more, if you were already aware of a single *golden feature* that completely resolved your challenge and answered the question you had in mind, rather than approaching the problem through data analysis and machine learning, you could probably directly engineer a solution around that single attribute.

Data-driven problems can reach a level of complexity where even with your expertise in the problem's domain, you still are only vaguely able to derive a few noisy features that do a bad job of describing the relationship you hope to model. Such sub-par features might only partially help answer your question by providing marginal information, but do not throw them away. Instead, use them as your feature set.

One of the beauties of machine learning is its ability to discover relationships within your data you might be oblivious to. Two or more seemingly weak features, when combined, might end up being exactly that golden feature you've been searching for. Unless you've collected as much data as possible before leaving the feature discovery stage of machine learning, you will never have the opportunity to test for that.

Let's say you want to get your pet iguana Joey a special treat for his birthday. Joey is super picky in his eating habits. You know he generally likes dark, green, leafy vegetables like kale and mustard greens; but occasionally he *violently* rejects them. Joey can't communicate to you *why* he sometimes like's them or other times doesn't. But being a data driven person, you've long since theorized there must be a method to the madness, and have been keeping some stats on all the food you've ever given him. Of particular interest are two features: fiber, and antioxidant content.

Individually, these features seem like poor indicators of Joey's preference of food. It looks like he likes some low-fibrous greens, but he also likes greens that are packed with them. What gives? By combining these two features, you realize there truly is a succinct way of correctly identifying veggies Joey likes, each time every time (the blue plots)!