CIS 7000-005: Debugging Data & Models Fall 2022

Explainabilty — September 29

Prof. Eric Wong

To find problems in our models, we need ways to inspect and debug them. How can we do this for complex models often seen as opaque? This problem is called explainability in machine learning.

- What is an ideal explainability method?
- What do explainability methods currently provide?

What is an explanation?

1 Desiderate for explainabilty

- Faithfulness - explanations should reflect an actual change in the model
- Usability - explanation should be meaningful to the end user
- Necessity - model needs the identified explanation to do a prediction
- Sufficient - explanation covers the entire model's prediction

1.1 Interpretabile by design

Some models are "interpretable by design". These

Linear models

$$y = \beta^T x + \beta_0$$

What is the claim of interpretability here? it depends on the feature and the R^2.

$$R^2 = 1 - SSE/SST$$

where $SSE = \sum_i (y^i - \hat{y}^i)^2$ and $SST = \sum_i (y^i - \bar{y})^2$. If R^2 is low, then little variance is explained by the linear model.

How to actually interpret the weights? From Christoph Molnar's textbook:

- Numerical feature: Increasing the numerical feature by one unit changes the estimated outcome by its weight. An example of a numerical feature is the size of a house.

- Binary feature: A feature that takes one of two possible values for each instance. An example is the feature "House comes with a garden" One of the values counts as the reference category (in some programming languages encoded with 0), such as "No garden". Changing the feature from the reference category to the other category changes the estimated outcome by the feature's weight.
- Categorical feature with multiple categories: A feature with a fixed number of possible values. An example is the feature "floor type", with possible categories "carpet", "laminate" and "parquet". A solution to deal with many categories is the one-hot-encoding, meaning that each category has its own binary column. For a categorical feature with L categories, you only need L-1 columns, because the L-th column would have redundant information (e.g. when columns 1 to L-1 all have value 0 for one instance, we know that the categorical feature of this instance takes on category L). The interpretation for each category is then the same as the interpretation for binary features. Some languages, such as R, allow you to encode categorical features in various ways, as described later in this chapter.
- Intercept : The intercept is the feature weight for the "constant feature", which is always 1 for all instances. Most software packages automatically add this "1"-feature to estimate the intercept. The interpretation is: For an instance with all numerical feature values at zero and the categorical feature values at the reference categories, the model prediction is the intercept weight. The interpretation of the intercept is usually not relevant because instances with all features values at zero often make no sense. The interpretation is only meaningful when the features have been standardised (mean of zero, standard deviation of one). Then the intercept reflects the predicted outcome of an instance where all features are at their mean value.

Logistic regression

$$P(y=1) = \frac{1}{1+\exp(-(\beta^T x))}$$

How to interpret the weights now? Take a log transofrm of the odds:

$$\ln \frac{P(y=1)}{1-P(y=1)} = w^T x \tag{1}$$

Lets call this the log odds. What happens if we increase the unit of β_i by one?

$$\frac{Odds_{x_j+1}}{Odds_{x_j}} = \exp(\beta_j) \tag{2}$$

So one unit of change corresponds to an increase in the "log odds ratio" by β_j.

1.2 LIME

Suppose we are happy with linear models. We can try to fit a local surrogate linear model to explain local decision boundaries. For a specific example x, model f, and linear model class G, we can solve

$$\min_g \ell(f, g; x) + \Omega(g)$$

- Select x to explain
- Perturb x locally to get a dataset of $(x_i, f(x_i))$
- Weight each x_i according to its distance to x with w_i
- Train a g on the weighted dataset
- Interpret g

For example, we can use LASSO:

$$\min_w \|w^T X - y\| + \|w\|_1$$

Under the assumption that the local linear model is accurate, this tells us how the model behaves locally. But how many weights to interpret? Is it faithful? Or usable?

2 References

Notes largely sourced from `https://christophm.github.io/interpretable-ml-book/`