Progetto Categorical Data

Handling categorical data is useful with models that only accepts numerical data. While some models can accept categorical data, like tree-based models, a lot of models don't accept categorical data, and in this case this type of data needs to be transformed. We also must consider the order of the model

Dummy variables

The first approach in this case can consist in creating dummy variables, or indicator variables, for categorical data. These are artificial numerica values that capture some aspect of one of the categorical values. For example in the OkCupid dataset we can transform the categorical data represented by the day of week column into a binary dummy variable. The mathematical function that make this translation is referred to contrast or parametrization function. One example is the "reference cell" function that leaves one value out to binary encode the others, this prevents the X'X matrix to be uninvertible, while "one hot" encodings does not consider this thing and can be used in methods that does not have the problem of matrix inversion

Encoding predictors with many categories

How to handle categorical predictors with any categories, like the ZIP Code?
With the previous methods we will achieve an overabundance ov dummy variables. This can cause the matrix to be overdetermined and prevent the use od certain models. Resampling can exlude some of the rarer categories. When a predictor contains only a single value, this is a zero variance predictor. The first way to handle this case is to create a full set of dummy variables and simply remove the zero-variance predictors, this is a simple approach. A proble of this method is that with resampling, every time the model will have different predictors because of the change of data. We can also determine id any of the variables are near zero variance or have the potential to become near-zero variance during resampling. These predictors usually have a few value(like a binary predictor) and occur infrequently in the data. We can calculate for training set as the ratio between the most commongly occuring value and the second-most occuring value. We can choose a cutoff of 19 to declare a variable near zero variance.

Avoiding filtering near zero variance.

We can redefine the predictor's categories before creating the dummy variables. For example an "other" category can be created to group the rarer categories and reduce the overall number of categories. This should occur during the resampling.

Hashing functions

Hashes are used to map one set of values to another set of values. The original values are valled keys. In our context the keys are the categories values and we want an hash function to represent them in a smaller set of values. The number od possible hashes is specified by the user and is a power of 2. In our case the process is called feature hashing or "hash trick", After transforming the hash integer to one of the specified feature number, we can transform the number into a dummy variable. We can used signed dummy representation to separate eventual aliases

Unseen data

In case of unseen categories during testing it will be useful to have the mentioned "Other" category in order tu put unseen data inside this category. Otherwise the unseen category can be a zero-variance predictor of the dummy variable and not considered during preprocessing.

Supervised encoding methods

There are methods of encoding unseen data using the outcome as a guide. This techniques are useful when the predictor has many possible values or when new levels appear after model training.

Effect or Likelihood encoding

It is a simple translation. In essence the effect of the factor level on the outcome is measured and this effect is used as numerical encoding. For example we can calcualte the mean or median sale price for each neighborhood and use this statistic to represent the factor level in the model.

Classification problems

For classification problems a logistic regression model can be used to measure the effect betweenthe categorical outcome and the predictor. For each predictor the model calculate the log-odds and this value can be used as a representation. If the predictor have a single value the log odds should be infinite, but numerically is capped at a large and inaccurate value. A way around is to use some type of shrinkage method, in this case the effect of a factor level can be biased towards an overall estimate that disregards the levels of the predictor. These methods can move extreme estimates towards the middle of the distribution

Word embeddings

It reduces the dimension of a text by prepresenting words with dense vectors, and there is the possibility that similar words will have similar embedding vectors. This techniques is not limited to text data but can be used to encode any type of qualitative data. Once the number of features are specified, the model takes the traditional indicators variables and randomly assign them to one of the new features. The model the tries to optimize the allocation of indicators to features and the coefficients for the features themselves. The outcome of the model can be the same as the predictive model. usually root mean squared error is used as loss function for numeric values and categorical crossentropy for categorical outcomes. Once the model is trained the values of the embeddings are saved for each bserved values of the quantitative factor. These serve as lookup table that is used for the predction. An extra level can be allocated to the ordiginal predictor to serve as place-holder for new values. We can also use a traditional neural network structure in order to permit the model to generate more complex representations

Ordered data

If we transform ordered categorical data("low, medium high") with simple dummy variables, we loose the information about the order. Ordered categories may have linear relationship with the response. A way to encode this type of data is called *polynomial contrast*. A contrast has the charateritich that is a single comprison(one degree of freedom) and its coefficients sum to zero. For the previous example the contrast to uncover a linear trend would be (-0.71, 0, 0,71). *Ploynomial contrast* can be used also for non linear relations. These contrasts can be generated for predictors with any number of ordered factors, but the complexity of the contrast is constrained to one less than the number of the categories in the original predictor, for example we cannot generate a cubic relationship for a predictor with only 3 categories. Using this contrast we can investigate multiple relationships(linear, quadratic, etc) simultaneously. Patterns described by polynomial contrast may not effectively relate a predictor to the response. Another downside ìoccurs when there are a moderate high number of categories: if a predictor has C levels, the encoding uses polynomial up to degree C - 1. It is very unlikely thhat these higher-level polynomials are modeling important trends and it might make sense to place a limit on the polynomial degree. In practice, we rerely explore the effectiveness of anything more than a quadratic polynomial.

alternatives

As alternatives on could:

- Treat the predictors as unordered factors.

- Translate the ordered categories into a single set of numeric scores based on context-specific information: for example, discussing failure modes of a piece of computer hardware, experts are able to rank the severity of a failure on an integer scale. Simple visualizations of context-specific expertise can be used to undestand whether either of these approaches are good ideas.

Creating features from text data

In open text infromation there could be important info that should be used in the model. The type of words used can describe the type of person. This data is qualitative and requires more effort to put in a form that models can consume.
For example in the OkCupid dataset if a profile description contains an link to a website it could be related to the person's profession. For example in OkCupid dataset 21% of STEM profiles have an hyperlink, while 12.4 % of non-STEM profiles have an hyperlink. A way to evaluate a difference in two proportions is the *Odds-ratio*. The odds of an event that occurs with rate p is defined as $\frac{p}{1-p}$. For STEM the odds of containing an hyperlink are relatively small(0.27). For non-STEM is even lower(0.142). The ratio between these two quantities can be used to understand the effect of having a hyperlink would be betweem the two professions . With these data the odds of a STEM profile is 1.9 times higher when the profile contains a link.If we use statistics to assign a lower 95% confidence to this intervall, the lower bound is 1.7, which indicates that the increase in the odds is unlikely to be due to random noise since it doesn not include the value . Given these results the indicator of the presence of an hyperlink should be included

Words

We also need to understand of there are words that would make a good predictor of the outcome. We need to clean the text. We can compute features on the cleaned texts, like the number of commas, hashtags, mentions, exclamation points, etc.
We can also filter the words based on non sense words, punctuation, word frequency, etc.
We can compute for each singular word the odds-ratio and its associated p-value. The p-value tests the hypothesis that the odds of the keyord occurring in either professional groups are equal(i.e 1.0). p-value can provide misleading results:

- In isolation the p-value does not measure the magnitude of the differences.
  As a rough criteria of importance, keywords with odds-ratios od at least 2 and a FDR values less than 10^{-5} will be conidered for modeling.

Other features.

Other features can be computed related to the sentiment and language. Words can be assigned with qualitative assessments, or numeric scores . We can also consider a measure of point of view(first, second third person text)

Factors Vs Dummy variables in Tree-Based Models

A tree-based model can handle both categoriacal and dummy variables encoded data. Does it matter if the data is encoded or not?