Martinez Assignment 1 – Part 3

Supervised learning is useful when the dependent variable is known. We have a set of example vectors and independent variable matrices that can be used to train the model. This model can then be used to extrapolate or generalize to other, similar data.

Supervised learning consists of two types of algorithms: Classification and Regression.

Classification is used to "classify" or determine an outcome of finite choices. Whether a person has diabetes or not, a particular handwritten digit, some prognosis based on criteria. Regression on the other hand, is for a continuous dependent variable. This is the ideal method for house prices, projected income or cost, and other things that require a numeric value that could be any possible number.

The article, Using Machine Learning to Translate Applicant Work History Into Predictors of Performance and Turnover is a clear example of a classification algorithm at work. The writers are focused on categorizing text data into four clear outcome categories that they then use to predict the quality of the teacher hire. The article infers that leveraging different job turnover hazards that they classified using their model strongly depict the quality and duration of a particular hire.

There is a two step approach used here. The main focus of the article is the classification algorithm used to parse free text to bucket applicant turnover into four outcomes: Involuntary, Avoiding a bad job, approaching a better job, and other. The model builds vectors of the text words and then the model assigns probabilities to each of these categories and then buckets the observations based on the weights. They use a 3% training and 97% testing split on the data to train the model. This learning task is common among supervised models.

Page 1215 clearly describes the model that was developed and the several other that were tried. Each one is a particular variety of supervised classification. Among them were Naive Bayes, Logistic Regression, Decision Tree, Random Forest and K-Nearest Neighbor.

The second step was to build another supervised classification model that maps the job descriptions to the O\*NET standardized occupation codes. The results are then tested against a manually classified group of individuals by an HR representative. The type of model they use is known as a Naive Bayes Classifier. This is a simple model that has surprisingly good results. It’s known as naive because it assumes independence between the features, which may not always be the case. In spite of this, it still performs quite well. The equation for the Naive Bayes is:



The models were used in several tests, including voluntary versus involuntary turnover, the impact of things such as student evaluation, expert observations and others. What was particularly interesting is that the researchers controlled for the fact that the data only had hired applicants.

The model was then used to recommend a series of applicants for each job type (e.g. special education) and those applicants were compared to the existing list that was built based on heuristics. They found that their selection quality was actually higher than the manual efforts of the HR employees.

Coefficients from the model indicated strong evidence that work experience relevance, tenure history and history of leaving previous jobs (built by the prior classification effort) indicate statistical significance. In line with these conclusions, causal inference is evident here. Involuntary and voluntary reasons for leaving a particular job provide excellent indicators on the performance and competence of future hires. Intercorrelations between the variables show strong negative relationships between involuntary turnover and student evaluation and performance composite. While whether correlation indicates causation is a matter of philosophical reasoning, it could be inferred here that teachers who were removed forcibly from a position had generally poor performance and their turnover was a product of that. The researchers controlled well for the verbiage used by applicants and created clever ways to distinguish between someone who left a job involuntarily but were looking for a new one (exhibited personal agency) and a person who left and did not provide any intrinsic reasoning.

Support vector machines (SVM) are a type of model that was not leveraged in the study. It can be used for both classification and regression problems. SVMs map points into space and then categories placed around them. In other words, it tries to define an algorithm to find a hyperplane that most effectively classifies the data points. It does this by maximizing the margins of the hyperplane between the data points.

The following article discusses use of an SVM to improve image classification using a bag of words scheme. This is similar to the applicant history article that was the focus of this write-up. It may be possible to leverage some of the content within this article to improve the classification algorithm built by Sjjadiani et. al. since they both use a similar approach.

References

Eduardo Fidalgo Fernández. (2018). Selection of relevant information to improve Image Classification using Bag of Visual Words. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, *16*(2), 5–8. https://doi-org.lopes.idm.oclc.org/10.5565/rev/elcvia.1102

Brownlee, J. (2019). 14 Different Types of Learning in Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/types-of-learning-in-machine-learning/

Evolutionary algorithm. (2020). In Wikipedia. https://en.wikipedia.org/wiki/Evolutionary\_algorithm

Sajjadiani, S., Sojourner, A., Kammeyer-Mueller, J., Mykerezi, E. (2019). Using Machine Learning to Translate Applicant Work History Into Predictors of Performance and Turnover. Journal of Applied Psychology. http://ds.doi.org/10.1037/apl0000405