Emerging Tools for a "Driverless" Legal System

Comment

by

Elliott Ash*

1 Introduction

Professor Talley (Talley, 2018) has provided an exciting introduction to how machine learning has changed and could change legal practice and research. He asks the provocative question of whether the legal system could become automated in the same way that cars are becoming driverless. Due to its simplicity, the central example in Professor Talley's piece (COW waivers) is useful as an introduction, but also somewhat limited in scope. The goal of this comment is to discuss more recent developments in machine learning that might be well suited for more ambitious legal tasks.

2 Featurization

This section describes some recent developments in featurization – the steps taken to transform raw text data into numerical data. Denny and Spirling (2017) show that featurization choices can make big differences in results, especially in unsupervised learning tasks. More generally, the researcher has to trade off predictiveness, computational tractability, and interpretability.

As mentioned in Talley (2018), the workhorse for text-based feature sets is a frequency distribution over N-grams. N-grams extract information (predictiveness) and familiarity (interpretability) through word order. But N-grams can be taxing on computational resources, with feature sets easily blowing up into the hundred-thousands or millions of columns. Even sparse matrices of this size can be troublesome to work with.

Fortunately there are recent developments in excluding the less informative N-grams. The first obvious thing to do is impose minimum frequency thresholds. Talley mentions this step in his example.

* Assistant Professor of Economics, The University of Warwick, Coventry, United Kingdom; Visiting Scholar, Center for Study of Democratic Politics, Princeton University, Princeton (NJ), U.S.A.

Journal of Institutional and Theoretical Economics 174, 1–8 – ISSN 0932-4569
DOI: 10.1628/093245618X15135819919262 – © 2018 Mohr Siebeck

Second, one can filter out N-grams by their relative collocation frequencies. A useful metric for this purpose is pointwise mutual information,

$$\begin{aligned}\text{PMI}(w_1, w_2) &= \frac{F(w_1\_w_2)}{F(w_1)F(w_2)} \\ &= \frac{\text{Prob. of collocation, actual}}{\text{Prob. of collocation, if independent}},\end{aligned}$$

where w_1 and w_2 are words in the vocabulary, w_1, w_2 is the N-gram $w_1\_w_2$, and $F(\cdot)$ returns the term frequency. This metric does a nice job ranking idiomatic phrases (e.g., "income tax") and can be generalized to an arbitrary number of words (e.g., "corporate income tax").

Another technique for filtering N-grams, introduced by Denny, O'Connor, and Wallach (2015), is to filter the N-gram set based on the sequences of parts of speech in the N-gram. This arises from the idea that informative, recognizable concepts are represented in language as noun and verb phrases. Removing other types of phrases greatly reduces the dimensionality of the feature space while keeping the most informative content. While Denny, O'Connor, and Wallach (2015) provide sequences up to length 3, Ash (2016) extends the list to length 4, while adding sequences that often turn up as key phrases in legal documents. For example, "beyond a reasonable doubt" is "preposition article adjective noun," while "earned income tax credit" is "adjective noun noun noun."

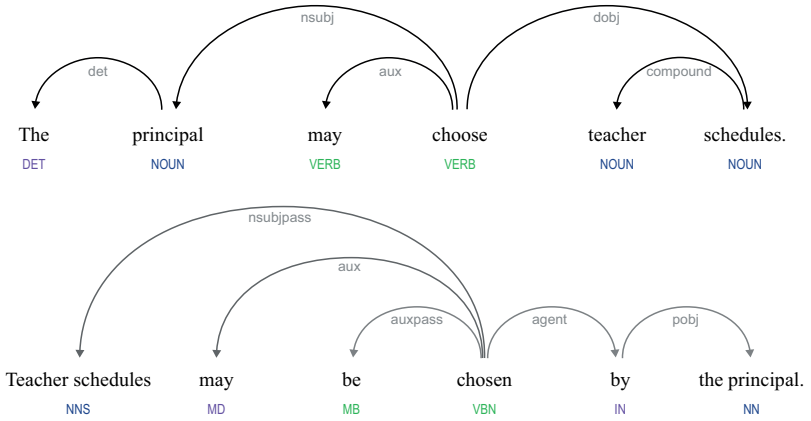
As mentioned, N-grams work by extracting information and familiarity from word ordering. Recent work has innovated by using the syntactic relations between words, rather than just order. Figure 1 shows two parsed sentences from teacher's union contracts. These sentences give the same authority to the school principal, but the word ordering is different. By recovering the longer-range relations between words, syntactic dependencies allow the researcher to extract a richer set of features.

3 Dimension Reduction

Professor Talley's (2018) paper provides good background on the importance of dimension reduction when using text data. Even after aggressive filtering of the feature space, there will be too many features. Normal regression methods like OLS will fail due to computational constraints, and due to bias from multicollinearity. Talley uses principal-component analysis, which is a popular and successful method, especially for prediction/classification. More recently, there have emerged dimension-reduction techniques that are specifically suited for text, either because they have better performance in prediction tasks, or because they have more interpretable components.

In the case of unsupervised learning, the researcher would like the algorithm to provide summaries of the corpus in a data-driven but intepretable way. The classic model for this task is latent Dirichlet allocation, or LDA (e.g., Blei, 2012). LDA

Figure 1
Example of Syntactic Dependencies



represents documents as vectors – distributions over topics. Unlike PCA components, LDA's topics tend to be interpretable.

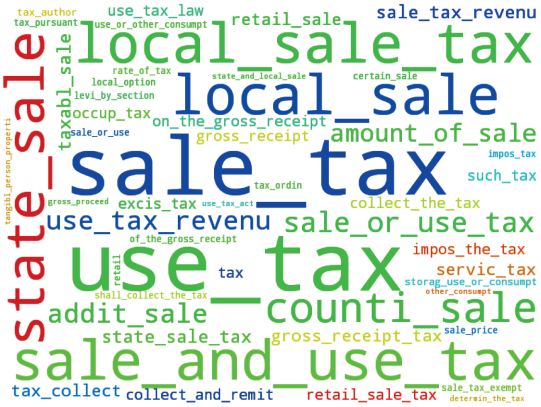
Another way to understand a corpus with unsupervised learning is to use word embeddings (e.g., Mikolov et al., 2013). Just as topic models represent documents as vectors, word embeddings represent words as vectors. These word vectors can then be used for description or prediction tasks. Figure 2 shows an example of the words and phrases from Ash (2016) that are closest to "sales tax" in the word embedding space. In that paper, this word distance was used to classify statutes by their relation to different tax revenue sources.

When prediction or classification is needed, there are useful learning models that both reduce dimensionality and improve predictive power of data. Data scientists have gotten good performance from regularized linear regression. With the addition of penalty parameters, Lasso, Ridge, and Elastic Net shrink the coefficients toward zeros and exclude weak predictors altogether. This results in a more parsimonious model that is more computationally tractable, avoids multicollinearity problems, and tends not to overfit the data.

An alternative supervised dimension reduction method, similar to principal components, is partial least squares (PLS) regression, which finds the optimal linear combination of predictors for predicting an outcome variable. Figure 3 shows an example of PLS prediction using text data; in Ash (2016), PLS was used to predict tax revenue changes using the text of enacted tax legislation.

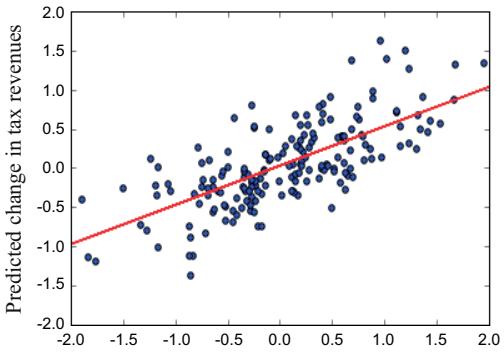
Oftentimes machine learning models are needed for classification, rather than regression, where there are multiple categories without a linear ordering. For these types of tasks, regression methods will not work. Data scientists have obtained good performance with regularized multinomial logit, random forests, and boosted trees.

Figure 2
Example of Word Embeddings



Notes: Words and phrases that are most similar to "sales tax" in a Word2Vec model trained on the corpus of state session laws (Ash, 2016).

Figure 3
Partial Least-Squares Example
(out-of-sample prediction: income tax (OLS))



Notes: Predicted change in state income tax revenue based on enacted statutes (vertical axis), plotted against true change in tax revenue (horizontal axis) (correlation between truth and prediction: 0.89).

4 Causal Inference

When machine learning techniques are used for measurement or data coding, selection bias is not a problem. But many prediction tasks are less straightforward. For example, are tax code features actually driving revenue effects, or are they merely correlated with nontext factors (e.g., strict enforcement by the tax authority) that are driving the effect? New developments in high-dimensional econometrics are allowing us to answer these types of questions. For example, Ash (2016) adapted the sparse instrumental-variables methods from Belloni et al. (2012) and Lin, Feng, and Li (2015) to recover "the effective tax code," the set of tax-code text features that have a causal effect on revenues.

This is an active research area. An important recent development is double or debiased machine learning (Chernozhukov et al., 2017). That paper provides (relatively broad) conditions for producing consistent estimates, and doing statistical inference, for a low-dimensional causal parameter using a high-dimensional set of exogenous instruments. The method involves a residualization step in which the instruments are used to form predictions for the treatment and outcome variables. Both the auxiliary and main prediction problems can be solved using a broad set of machine learning methods including random forests, penalized regression, boosted trees, and deep neural nets. The use of K-fold sample-splitting avoids overfit and results in an unbiased, normally distributed estimator for which valid confidence intervals may be constructed.

These results pave the way for development and application of new methods to causal inference using text data. For example, Hartford et al. (2016) use deep neural networks in a two-stage least-squares setup. These DNNs can form flexible predictions that harness interactions between the predictors, allowing for more flexible estimates of heterogeneous treatment effects. These methods will play an important role in legal automation.

5 Interpretability

Professor Talley (2018) makes the key point that PCA cannot be interpreted easily. Even if it predicts well, it is difficult to assess and visualize which features are most important. One option is to regress the components on the features – but this does not work well in practice using text data.

Lack of interpretability is a major barrier to a driverless legal system. In civil law, clients and judges want to understand the model before making decisions based on it. In the context of criminal justice, understanding the workings of a model is important, given that preexisting institutional or social biases could easily be perpetuated in an algorithm (Kleinberg et al., 2017; Caliskan, Bryson, and Narayanan, 2017).

With regard to interpretability, the featurization steps and modeling choices can both help. As mentioned in section 2, using idiomatic phrases, rather than just word

counts, can help for representation of legal language. One can then rank features by their predictiveness of an outcome, using univariate regression. For larger models, one can use the feature importance rankings given by random forests and boosted trees.

One of the best NLP technologies for interpretability is the structural topic model, developed by Roberts et al. (2013). The STM is an extension of LDA where the underlying model is allowed to depend on some observable covariates. While this model is not as effective as random forests or deep neural nets for predictive accuracy, it does a superior job for interpretability. STM provides word clouds with indications of which topics are more predominant with respect to your covariate (e.g., political affiliation of the writing judge), and even the words within each topic that are most associated with your covariate.

The broader problem of model interpretability is an active area of research in machine learning. For example, Ribeiro, Singh, and Guestrin (2016) provide a method for approximating any (black-box) predictive model with an interpretable (white-box) model, such as linear regression or decision trees. That paper provides an illustration where the method could observe the predictions of a (black-box) support vector machine classifier using text, and then highlight the words and phrases in documents that are most important.

6 Algorithms and Experimentation

I conclude with the question: "Can algorithms experiment?" Professor Talley (2018) answers this in the negative. But I think there is an emerging recognition that algorithms can and do experiment. For example, search engines use stochastic multi-armed bandit models to design optimal policy experiments for maximizing advertising revenue (see e.g., Burtini, Loeppky, and Lawrence, 2015). Could these algorithms be applied to the law?

There are many legal decisions, such as settlement in litigation, which could be subjected to algorithm-driven experimentation. An algorithm could design an experiment for making bail decisions that would update dynamically in response to follow-up information about its previous decisions. One could even imagine an "optimal" stare decisis rule, where appellate judges experiment with different legal rulings across jurisdictions and collect information on the socioeconomic impacts of those rulings.

Is this rendering fanciful? It is not so different, after all, from the well-known efficiency-of-common-law hypothesis, whose proponents argue that common-law rules tend toward efficiency via the evolutionary process of adjudication and the gradual accretion of precedent (e.g., Parisi, 2004). Supporting this hypothesis is a large literature, reviewed in La Porta, Lopez-de-Silanes, and Shleifer (2008), documenting higher growth in common-law countries. An influential argument is that judge gap-filling results in better outcomes than legislation because there is no need for a complex network of rules.

But, one should ask, could robot judges reverse this preference? Ash and Morelli (2017) provide a model to motivate this idea. One can think of the laws as a network of clauses, with an additional layer of interpreters. An increase in complexity makes law more precise but increases search costs for the interpreters. With human interpreters, search costs are high enough to put civil law at a disadvantage. But eventually, robot interpreters will reduce search costs enough to make civil law the superior institution.

References

Ash, Elliott (2016), "The Political Economy of Tax Laws in the U.S. States," unpublished Manuscript, University of Warwick, Coventry.

— and Massimo Morelli (2017), "Robot Judges and the Efficiency-of-Common-Law Hypothesis," unpublished Manuscript, University of Warwick, Coventry.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80(6), 2369–2429.

Blei, David M. (2012), "Probabilistic Topic Models," *Communications of the Association for Computing Machinery*, 55(4), 77–84.

Burtini, Giuseppe, Jason Loeppky, and Ramon Lawrence (2015), "A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit," electronic Preprint, arXiv:1510.00757v4, Cornell University, Ithaca (NY).

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (2017), "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases," *Science*, 356(6334), 183–186.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, et al. (2017), "Double Machine Learning for Treatment and Causal Parameters," electronic Preprint, arXiv:1608.00060v5, Cornell University, Ithaca (NY).

Denny, Matthew J., Brendan O'Connor, and Hanna Wallach (2015), "A Little Bit of NLP Goes a Long Way: Finding Meaning in Legislative Texts with Phrase Extraction," Paper presented at the Midwest Political Science Association (MPSA) 73rd Annual Conference, Chicago (IL), April 16–19, 2015.

— and Arthur Spirling (2017), "Text Preprocessing for Unsupervised Learning: Why it Matters, When it Misleads, and What to Do about it," http://www.nyu.edu/projects/spirling/documents/preprocessing.pdf, accessed November 11, 2017.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy (2016), "Counterfactual Prediction with Deep Instrumental Variables Networks," electronic Preprint, arXiv:1612.09596v1, Cornell University, Ithaca (NY).

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2017), "Human Decisions and Machine Predictions," Working Paper 23180, National Bureau of Economic Research, Cambridge (MA).

La Porta, Rafael, Florencio Lopez-de-Silanes, and Andrei Shleifer (2008), "The Economic Consequences of Legal Origins," *Journal of Economics Literature*, 46(2), 285–332.

Lin, Wei, Rui Feng, and Hongzhe Li (2015), "Regularization Methods for High-Dimensional Instrumental Variables Regression with an Application to Genetical Genomics," *Journal of the American Statistical Association*, 110(509), 270–288.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean (2013), "Distributed Representations of Words and Phrases and their Compositionality," in: Christopher J. C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.),

Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, Nevada, USA, December 5–10, 2013, Curran Associates, Red Hook (NY), pp. 3136–3144.

Parisi, Francesco (2004), "The Efficiency of the Common Law Hypothesis," in: Charles K. Rowley and Friedrich Schneider (eds.), *The Encyclopedia of Public Choice*, Springer, New York, pp. 519–522.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016), " 'Why Should I Trust you?" Explaining the Predictions of Any Classifier," in: Balaji Krishnapuram and Mohak Shah (eds.), *KDD '16: Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, August 13–17, 2016*, Association for Computing Machinery (ACM), New York, pp. 1135–1144.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi (2013), "The Structural Topic Model and Applied Social Science," Paper presented at the Neural Information Processing Systems (NIPS) 2013 Workshop on Topic Models: Computation, Application, and Evaluation, December 10, Lake Tahoe, Nevada, USA, https://sites.google.com/site/nips2013topicmodels/papers, accessed November 10, 2017.

Talley, Eric L. (2018), "Is the Future of Law a Driverless Car? Assessing How the Data-Analytics Revolution will Transform Legal Practice," *Journal of Institutional and Theoretical Economics (JITE)*, 174(1), forthcoming.

Elliott Ash
Department of Economics
The University of Warwick
Coventry
CV4 7AL
United Kingdom
e.ash@warwick.ac.uk