Text Mining for Economics and Finance
Supervised Learning

Stephen Hansen

Introduction

Supervised learning is the problem of predicting a response variable (i.e. dependent variable) associated with documents using features of the text data.

For example, we may want to predict political party affiliation using state-of-the-union addresses.

Supervised learning has many commercial applications: predict which links a visitor to a website will click; which songs a listener will like; which emails are spam; and so on.

Supervised learning and econometrics share many of the same tools, but the emphasis is different.

Limitations of Prediction

The ability to predict response variables with a high degree of accuracy is not necessarily useful for many of the goals of economic research:

1. In many cases, no specific behavioral model is being tested.
2. Often economists care about counter-factual analysis and policy experiments, which the output of supervised learning algorithms does not often facilitate. E.g. relationship between price and quantity.
3. No guidance as to why predictive relationships exist, which is problematic for building forecasting models. E.g. Google Trends and flu outbreak.
4. Lucas critique. E.g. when consumers know firms set prices based on their observed behavior, consumers can strategically respond.

Application of Supervised Learning I

In some cases, we have some observations with a response variable and others without.

We can use a predictive model to associate observation features with responses, and then apply the model to associate out-of-sample observations with labels.

Useful when we have a theory that requires measuring observations in terms of labels, but which does not provide guidance about the relationship between features and labels.

One advantage of supervised over unsupervised learning is that the output is more interpretable since it is linked explicitly to labels.

Application of Supervised Learning II

Many empirical papers boil down to testing the significance of a single covariate's relationship with a dependent variable while controlling for other covariates.

Theory often motivates the relationship of interest, but not necessarily the set of variables to control for.

A typical approach is to run many different specifications to test how robust the predicted relationship is.

Supervised learning can allow us to be more honest and simply include all covariates and let the data decide which are the important ones.

Supervised Learning Problems

Supervised learning problems can be divided into two:

1. In classification problems, we place each document into one of a finite number of categories.
2. In regression problems, we use features of the document-term matrix to predict a continuous variable.

There is a strong connection between classification problems and latent variable models. Essentially, what are unobserved in the latter are observed in the former.

Rocchio Classification

Let each document d belong to one of C finite classes, and let D_c be the set of documents in class $c \in \{1, \ldots, C\}$.

The *centroid* of the documents in class c is $\vec{u}_c = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{x}_d$.

We then assign out-of-sample documents to whichever class has the closest class centroid.

This defines a *linear classifier* in which the decision boundary between two classes is a hyperplane.

Supervised analogue of k-means.

K Nearest Neighbor Classification

In k nearest neighbor model (kNN), we have a set of labeled documents, and classify new documents by looking at the labels of the k closest documents.

We assign the new document the label corresponding to the most frequent label among the neighbors.

This yields a locally linear, but globally non-linear classifier. More complex decision rule than Rocchio, but potentially more accurate for inherently non-linear classification.

Discriminative Classification

We now discuss parametric models of the relationship between covariates $\mathbf{x}_d$ and the dependent variable y_d.

A *discriminative* classifier estimates the conditional distribution $\Pr[\, y_d \mid \mathbf{x}_d \,]$.

Discriminative classifiers (logistic regression, support vector machines, etc.) can be applied to text data without modification since we do not have to model the generation of documents.

Since we usually do not have many documents relative to vocabulary terms, fitting such models requires regularization.

Feature Selection

To apply a discriminative classifier, we need to choose how to represent the features of our text.

Can use raw term counts (unigrams, bigrams, etc.); topic share representations; or even both together.

Can also use non-labeled texts along with labeled texts in topic modeling, since LDA uses no information from labels in estimation of topic shares.

Blei et. al. (2003) show that topic share representation is competitive with raw counts in discriminative classification exercise.

Linear Regression

We are all familiar with one discriminative classifier: the linear regression model in which $\Pr[y \mid \boldsymbol{\beta}, \mathbf{x}] = \mathcal{N}(y \mid \boldsymbol{\beta} \cdot \mathbf{x}, \sigma^2)$.

The log-likelihood function is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \Pr[y_i \mid \boldsymbol{\beta}, \mathbf{x}_i] = \sum_{i=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp\left(-\frac{1}{2\sigma^2} (y_i - \boldsymbol{\beta} \cdot \mathbf{x}_i)^2 \right) \right] =$$
$$-\frac{1}{2\sigma^2} \mathrm{RSS}(\boldsymbol{\beta}) - \frac{N}{2} \log(2\pi\sigma^2).$$

where $\mathrm{RSS}(\boldsymbol{\beta})$ is the residual sum of squares.

So the OLS coefficients minimize $\mathrm{RSS}(\boldsymbol{\beta})$, which yields $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

OLS and Overfitting

Especially with many covariates, or high-degree polynomials, OLS is at risk of overfitting.

One popular solution is to punish model complexity through introducing penalization terms into the objective function.

These can be statistically founded by introducing priors on the regression coefficients that encourage them to be smaller in magnitude.

Ridge Regression

Suppose we draw each regression coefficient β_i from a normal prior $\mathcal{N}(0, \tau^2)$. Lower values of τ are associated with smaller values of the coefficients.

The posterior distribution over $\boldsymbol{\beta}$ is then proportional to $\Pr[y \mid \beta, \mathbf{x}] \Pr[\beta]$.

Finding the $\boldsymbol{\beta}$ at which the posterior is highest is called *maximum a posteriori* (MAP) estimation.

The objection function for MAP estimation can be written

$$\mathrm{RSS}(\boldsymbol{\beta}) + \lambda \sum_j \beta_j^2$$

which is called the *ridge regression* model.

Solution is $\widehat{\boldsymbol{\beta}}_R = (\lambda \mathbf{I}_D + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The higher is λ, the smaller are the coefficients. Shrinkage but not sparsity.

LASSO

Now suppose we draw each regression coefficient β_j from a Laplace prior so that $\Pr[\beta_i \mid \lambda] \propto \exp(-\lambda|\beta_j|)$.

Unlike the normal distribution, the Laplace distribution has a spike at 0 which promotes sparsity.

The objection function for MAP estimation can be written

$$\mathrm{RSS}(\boldsymbol{\beta}) + \lambda \sum_j |\beta_j|$$

which is called the *LASSO*.

One reason for the popularity of LASSO for variable selection is that it is a convex optimization problem and can be solved with extremely efficient algorithms.

LASSO and Causal Inference

A typical goal in econometrics is to identify the effect of a (low-dimensional) treatment variable on an outcome of interest.

There are typically also a high-dimensional set of controls whose dimensionality is reduced through *ad hoc* decisions.

These controls may include text data, e.g. effect of identity of author on reader opinions requires controlling for content of text.

Increasing interest in using variable selection techniques to aid in causal inference.

Belloni et. al. (2014) provide an overview of current research on this topic in econometrics.

Instrumental Variables Model

$$y_i = \alpha d_i + \varepsilon_i$$
$$d_i = z_i' \Pi + r_i + \nu_i$$

where $\mathbb{E}[\,\varepsilon_i \mid z_i\,] = \mathbb{E}[\,\nu_i \mid z_i, r_i\,] = 0$, but $\mathbb{E}[\,\varepsilon_i \nu_i\,] = 0$.

This is the standard instrumental variables model in econometrics, but what about when there are many instruments?

Under an assumption of approximate sparsity in the relationship between d_i and the instruments, performing LASSO as part of a two-stage procedure allows for valid inference on α.

Intuitively, selection mistakes in the first stage have only mild effects since the true coefficient of any omitted instruments will be small.

Model with Treatment and Controls

$$y_i = \alpha d_i + x_i' \theta_y + r_{yi} + \psi_i$$

where $\mathbb{E}[\, \psi_i \mid d_i, x_i, r_{yi} \,] = 0$.

Suppose that d_i is low-dimensional; that x_i is high-dimensional; and that the model satisfies approximate sparsity.

A naive procedure for selection of controls is to estimate a LASSO that penalizes the controls but not the treatments.

The problem is that the procedure will omit controls that are highly correlated with d_i since they are penalized while d_i is not.

This creates an omitted-variable bias whenever the relationship between the omitted variables and the outcome is moderately sized.

Double Selection

The proposed solution is to run LASSO twice:

1. Regress outcome variable on controls, select set P_1.
2. Regression treatment variable on controls, select set P_2.

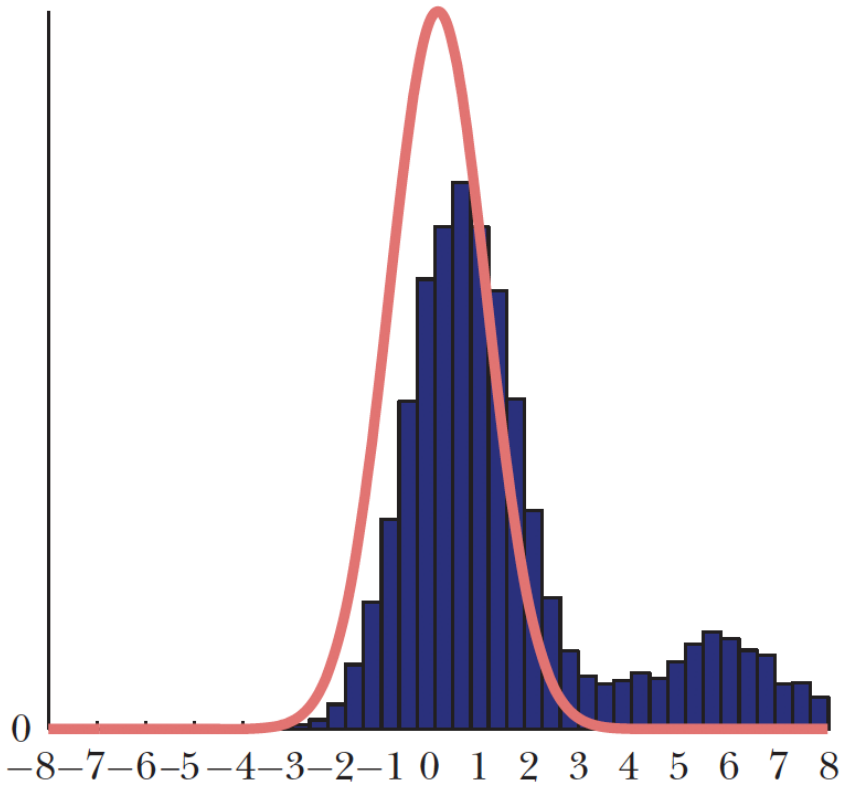
Perform OLS with outcome as dependent variable, and treatment + P_1 + P_2 as independent variables.

This guards against excluding controls that are either highly correlated with the treatment and moderately correlated with the outcome, or moderately correlated with the treatment and highly correlated with the outcome.

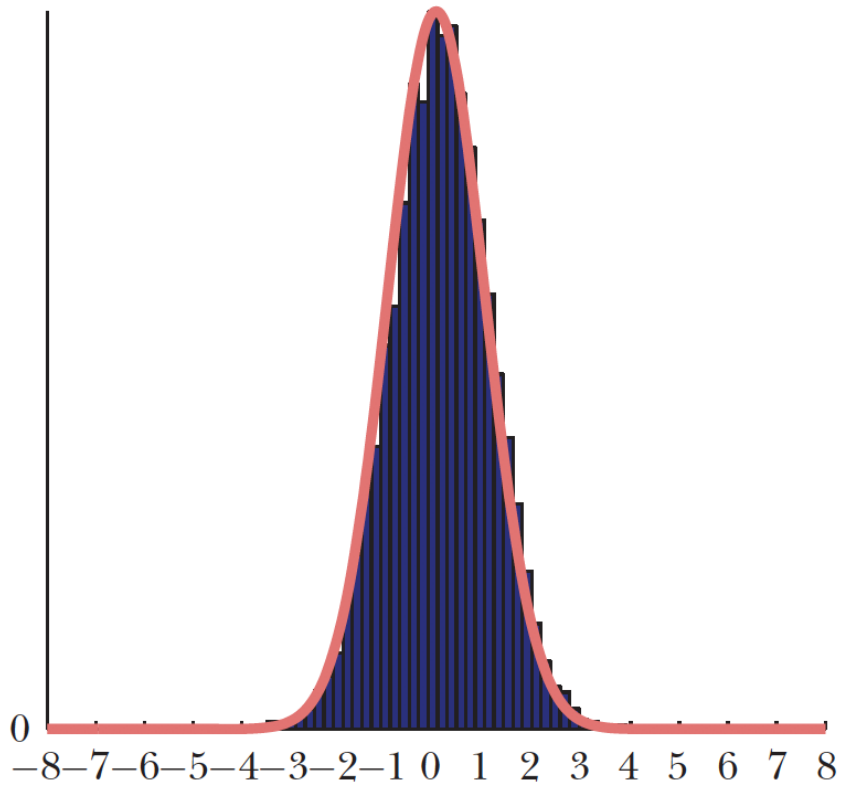
Inference on α using double selection is valid even if we allow selection mistakes.

Simulation Results

A: A Naive Post-Model Selection Estimator



B: A Post-Double-Selection Estimator



Post-Selection Inference on Selected Variables

The previous model emphasized inference about a non-selected treatment effect by tuning the selection of controls.

However at times we wish to perform inference on the selected variables themselves.

This is problematic because even if random relationships exist in the data, LASSO will find some covariates that correlate with the dependent variable.

Cross-validation can mitigate the problem, but few guarantees exist about the estimated coefficients—one can prove (Meinshausen and Bühlmann 2006) that cross validation selects too many variables.

Bottom line: LASSO was designed for prediction problems, and research into parameter inference is an ongoing and unresovled issue.

Solutions

More robust approach is to explicitly compute model inclusion probabilties in a Bayesian framework.

A simpler solution (Bach 2008, Meinshausen and Bühlmann 2010) is based on bootstrapping.

We can repeatedly resample the observations, fit LASSO for each draw, and record whether every variable was selected or not.

The fraction of times that each variable appears across the samples then acts as an approximation of posterior inclusion probabilities in a Bayesian model.

Selecting variables that appear more than a fixed fraction of times produces a sparser estimator than cross validation.

Example

In recent work with Michael McMahon and Matthew Tong, we study the impact of the release of the Bank of England's Inflation Report on bond price changes at different maturities.

IR contains forecast variables we use as controls: (i) mode, variance, and skewness of inflation and GDP forecasts; (ii) their difference from the previous forecast.

To represent text, we estimate a 30-topic model and represent each IR in terms of (i) topic shares and (ii) evolution of topic shares from previous IR.

First step in the analysis is to partial out the forecast variables from bond price moves and topic shares by constructing residuals.

We are then left with 69 bond price moves (number of IRs in the data) and 60 text features.

Results of Hold-One-Out CV

	Gilt1yspot	Gilt3y	Gilt5y
λ_{min}	.0000204	.0000306	.0000244
Features selected	54	57	55

No guidance on what the key dimensions of variation in the text are.

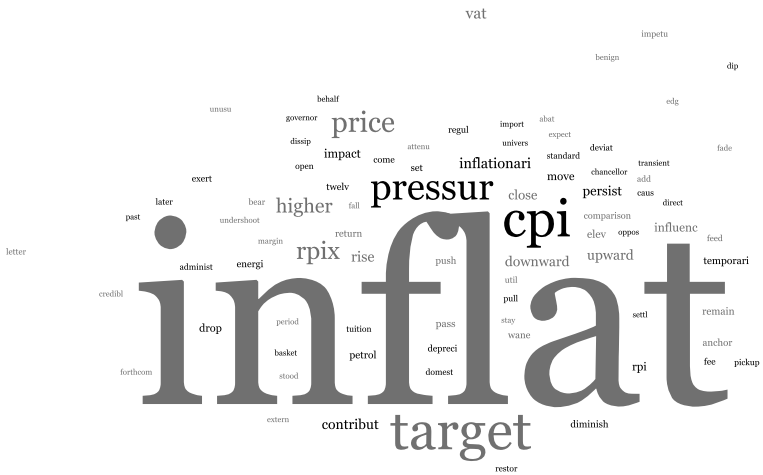
Results of Boostrapping

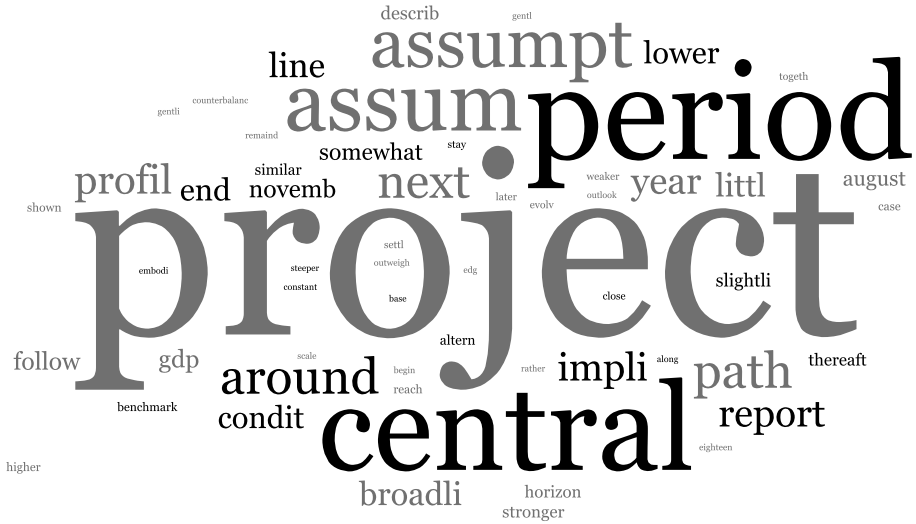
For each maturity, we constructed 100 boostrap draws and kept text features present in two-thirds or more.

	Gilt1yspot	Gilt3y	Gilt5y
Features selected	5	4	5

Gilt1yspot: T1 and T5

Gilt3y and Gilt5y: D6 and D15





Generative Modeling

A generative classifier estimates the full joint distribution $\Pr[y, \mathbf{x}_d]$.

Discriminative models have lower asymptotic error, but generative models:

1. May approach their asymptotic error faster (Ng and Jordan 2001).
2. Allow one to generate new data.

In practice, we estimate models of the form $\Pr[\mathbf{x_d} \mid y]$, a relationship we then need to invert for classifying out-of-sample documents.

Naive Bayes Classifier

A simple generative model is the *Naive Bayes classifier*.

The "naive" assumption is that the elements of $\mathbf{x}_d$ are independent within a class. This is equivalent to the unigram model we discussed earlier.

Let $x_{c,v}$ be the count of term v among all documents in class c, and $|D_c|$ the number of documents in class c. Then the joint log-likelihood is

$$\sum_c |D_c| \log(\rho_c) + \sum_c \sum_v x_{c,v} \log(\beta_{c,v})$$

with MLE estimates

$$\widehat{\rho}_c = \frac{|D_c|}{D} \text{ and } \widehat{\beta}_{c,v} = \frac{x_{c,v}}{\sum_v x_{c,v}} \left(= \frac{x_{c,v}+1}{\sum_v x_{c,v} + V} \quad \text{with smoothing} \right).$$

This is like the multinomial mixture model but with observed rather than latent class labels.

Classification

We can obtain $\Pr[\, c_d \mid \mathbf{x}_d \,] \propto \Pr[\, \mathbf{x}_d \mid c_d \,] \Pr[\, c_d \,]$ from Bayes' rule, where the probabilities on the RHS are already estimated.

To assign a class-label c_d to an out-of sample document we can use MAP estimation:

$$c_d = \underset{c}{\operatorname{argmax}} \; \log(\widehat{\rho}_c) + \sum_v x_{d,v} \log(\widehat{\beta}_{c,v}).$$

While the probabilities themselves are not generally accurate, classification decisions can be surprisingly so.

Generative Classification with LDA

To build a generative classifier with LDA, one can estimate separate models for each class labels, and thereby obtain $\boldsymbol{\alpha}_c$ and $\boldsymbol{\beta}_{1,c}, \ldots, \boldsymbol{\beta}_{K,c}$ for each class label.

For an out-of-sample document d, one can then obtain an estimate of $\widehat{\theta}_{d,c}$ given the class-label-specific parameters, for example by querying according to the procedure in the previous lecture slides.

Finally, one can assign the document to whichever class has a highest probability, which is easily computed—the probability of observing term v in class c is $\sum_k \widehat{\theta}_{d,c,k} \widehat{\beta}_{k,c,v}$.

Inverse Regression

Modeling and inverting the relationship $\Pr[\mathbf{x_d} \mid y]$ is more difficult when y is continuous and/or multidimensional.

Well-known example of this inverse regression problem is Gentzkow and Shapiro (2010).

Drawing on this paper as motivation, Taddy (2013) and Taddy (2015) have proposed fully generative models for inverse regression.

Measuring Media Slant

Gentzkow and Shapiro (2010) explore the determinants of newspapers' ideological slant.

The key measurement problem is that we observe the text of newspaper articles, but not their location on a political ideology scale.

Their solution is to determine the relationship between bigram and trigram frequencies used in US Congressional speeches and political party affiliation, and then to use these estimates to predict the ideology of newspaper.

The theory relies on observing newspapers' ideologies, but the relationship between words and ideology is left completely open ex ante.

Text Data

2005 *Congressional Record*, which contains all speeches made by any member of US Congress during official deliberations. (Full text).

After stopword removal and stemming, compute all bigrams and trigrams in the data. Millions in total.

Text Data

2005 *Congressional Record*, which contains all speeches made by any member of US Congress during official deliberations. (Full text).

After stopword removal and stemming, compute all bigrams and trigrams in the data. Millions in total.

Consider all English language daily newspapers available in either ProQuest or NewsLibrary for a total sample of 433 newspapers. (Access to phrase searches).

Consider only bigrams and trigrams that appear in not too few and not too many headlines.

Identifying Partisan Phrases

Let x_{vD} and x_{vR} denote the total counts of term v among Democratic and Republican speeches, respectively.

Let x_{vD}^{-} and x_{vR}^{-} denote the total counts of all terms besides term v.

One can then compute Pearson's χ^2 statistic for each term v as

$$\chi_v^2 = \frac{(x_{vR}x_{vD}^{-} - x_{vR}^{-}x_{vD})}{(x_{vR} + x_{vD})(x_{vR} + x_{vD}^{-})(x_{vR}^{-} + x_{vD})(x_{vR}^{-} + x_{vD}^{-})}.$$

Identify the 500 bigrams and 500 trigrams with the highest test statistic.

Democratic Phrases

Most Partisan Phrases From the 2005 *Congressional Record*[a]

Panel A: Phrases Used More Often by Democrats

Two-Word Phrases		
private accounts	Rosa Parks	workers rights
trade agreement	President budget	poor people
American people	Republican party	Republican leader
tax breaks	change the rules	Arctic refuge
trade deficit	minimum wage	cut funding
oil companies	budget deficit	American workers
credit card	Republican senators	living in poverty
nuclear option	privatization plan	Senate Republicans
war in Iraq	wildlife refuge	fuel efficiency
middle class	card companies	national wildlife
Three-Word Phrases		
veterans health care	corporation for public broadcasting	cut health care
congressional black caucus	additional tax cuts	civil rights movement
VA health care	pay for tax cuts	cuts to child support
billion in tax cuts	tax cuts for people	drilling in the Arctic National
credit card companies	oil and gas companies	victims of gun violence
security trust fund	prescription drug bill	solvency of social security
social security trust	caliber sniper rifles	Voting Rights Act
privatize social security	increase in the minimum wage	war in Iraq and Afghanistan
American free trade	system of checks and balances	civil rights protections
central American free	middle class families	credit card debt

Republican Phrases

TABLE I—*Continued*

Panel B: Phrases Used More Often by Republicans		
Two-Word Phrases		
stem cell	personal accounts	retirement accounts
natural gas	Saddam Hussein	government spending
death tax	pass the bill	national forest
illegal aliens	private property	minority leader
class action	border security	urge support
war on terror	President announces	cell lines
embryonic stem	human life	cord blood
tax relief	Chief Justice	action lawsuits
illegal immigration	human embryos	economic growth
date the time	increase taxes	food program
Three-Word Phrases		
embryonic stem cell	Circuit Court of Appeals	Tongass national forest
hate crimes legislation	death tax repeal	pluripotent stem cells
adult stem cells	housing and urban affairs	Supreme Court of Texas
oil for food program	million jobs created	Justice Priscilla Owen
personal retirement accounts	national flood insurance	Justice Janice Rogers
energy and natural resources	oil for food scandal	American Bar Association
global war on terror	private property rights	growth and job creation
hate crimes law	temporary worker program	natural gas natural
change hearts and minds	class action reform	Grand Ole Opry
global war on terrorism	Chief Justice Rehnquist	reform social security

Constructing Newspaper Ideology

For each member of Congress i compute relative term frequencies $f_{iv} = x_{iv} / \sum_v x_{iv}$; for each newspaper n compute similar measure f_{nv}.

1. For each term v regress f_{iv} on the share of votes won by George W Bush in i's constituency in the 2004 Presidential election $\rightarrow$ slope and intercept parameters a_v and b_v. Provides mapping from ideology to language.
2. For each newspaper n, regress $f_{nv} - a_v$ on b_v, yielding slope estimate $\widehat{y}_n = \sum_v b_v (f_{nv} - a_v) / \sum_v b_v^2$. Measures how the partisanship of term v affects language of newspaper n.

If $f_{nv} = a_v + b_v y_n + \varepsilon_{nv}$ with $\mathbb{E}[\, \varepsilon_{nv} \mid b_v \,] = 0$, then $\mathbb{E}[\, \widehat{y}_n \,] = y_n$.

Use $\widehat{y}_n$ as a measure of n's ideology in econometric work.

Taddy (2013)

"Multinomial Inverse Regression for Text Analysis" proposes a more formal statistical model in the spirit of Gentzkow and Shapiro.

Let $\mathbf{x}_y = \sum_{d:y_d=y} \mathbf{x}_d$ and $N_y = \sum_{d:y_d=y} N_d$.

Then we can model

$$\mathbf{x}_y \sim \mathrm{MN}(\mathbf{q}_y, N_y) \text{ where } q_{y,v} = \frac{\exp(a_v + b_v y)}{\sum_v \exp(a_v + b_v y)}.$$

This is a generalized linear model with a (multinomial) logistic link function.

Gamma-Lasso

The prior distribution for the b_v coefficients is Laplace with a term-specific Gamma hyperprior:

$$p(b_v, \lambda_v) = \frac{\lambda_v}{2} \exp(-\lambda_v |b_v|) \frac{r^s}{\Gamma(s)} \lambda_v^{s-1} \exp(-r\lambda_v).$$

This is a departure from the typical lasso model in which all coefficients share the same λ_v. This allows for heterogeneous coefficient penalization, which increases robustness in the presence of many spurious regressors.

Taddy proposes a simple inference procedure that maximizes penalized likelihood (implemented in 'textir' package in R).

Sufficient Reduction Projection

There remains the issues of how to use the estimated model for classification.

Let $z_d = \mathbf{b} \cdot \mathbf{f}_d$ be the *sufficient reduction projection* for document d, where $\mathbf{f}_d = \mathbf{x}_d / N_d$ and $\mathbf{b}$ is the vector of estimated coefficients.

The sufficient reduction projection is sufficient for y_d in the sense that $y_d \perp \mathbf{x}_d, N_d \mid z_d$.

This can be seen as an alternative dimensionality reduction technique (specific to the label of interest): all the information contained in the high-dimensional frequency counts relevant for predicting y_d can be summarized in the SR projection.

Classification

For classification, one can use the SR projections to build a forward regression that regresses y_d on some function of the z_d: OLS; logistic; with or without non-linear terms in z_d, etc.

To classify a document d in the test data:

1. Form z_d given the estimated **b** coefficients in the training data.
2. Use the estimated forward regression to generate a predicted value for y_d.

Taddy shows that MNIR outperforms other classifiers (LASSO, LDA, sLDA, etc.).

Taddy (2015)

Taddy (2015) constructs an algorithm for fitting MNIR with y_d itself having multiple dimensions, i.e. $y_d = (y_{d,1}, \ldots, y_{d,M})$.

The SR projection idea extends to this environment in the sense that $y_{d,m} \perp \mathbf{x}_d, N_d \mid z_{d,m}$ where $z_{d,m} = \mathbf{f}_d \cdot \mathbf{b}_m$.

Prediction application. Suppose some $y_{d,m}$ is only observed for a subset of documents while $\mathbf{y}_{d,-m}$ is observed for all documents. We can build a forward regression in the training data that relates $y_{d.m}$ to $\mathbf{y}_{d,-m}$ and $z_{d,m}$ which can be applied to test data.

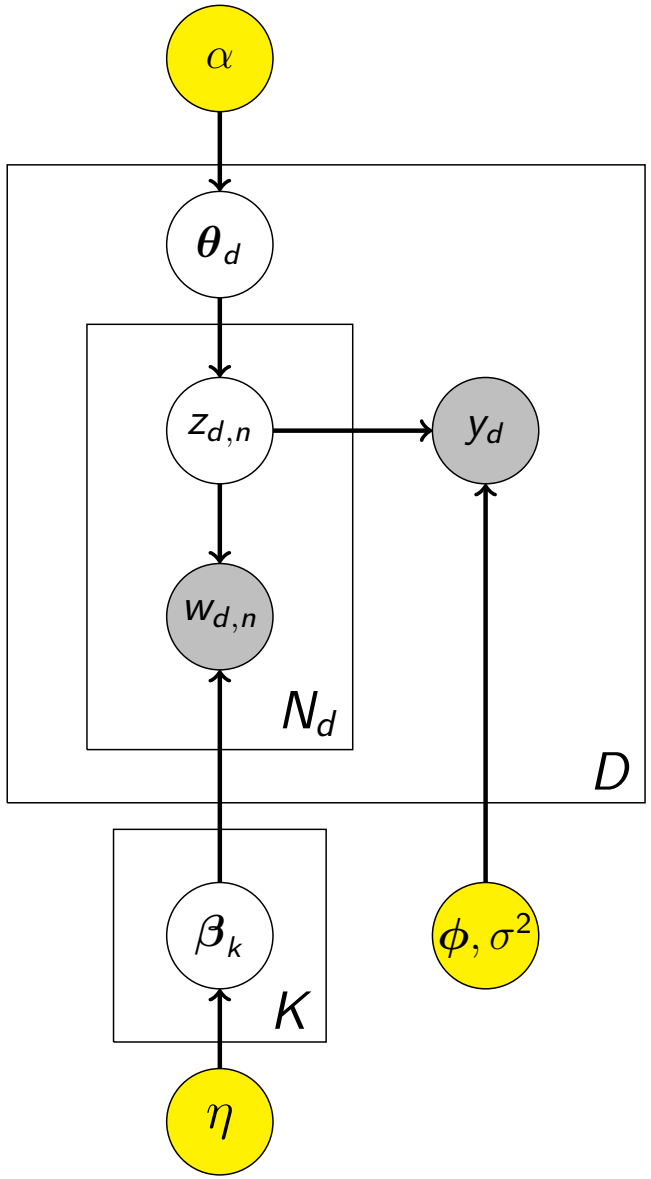
Treatment effect application. Suppose we want to estimate the treatment effect of $y_{d,m}$ on $y_{d,1}$ but want to also control for $\mathbf{x}_d$, which is high dimensional. (Similar problem to Belloni et. al.) The SR result implies that $y_{d,m}, y_{d,1} \perp \mathbf{x}_d$ given $z_{d,m}$, $z_{d,1}$, N_d, and controls other than m. We can then perform a forward regression with just the SR projections.

Supervised LDA (Blei and McAuliffe)

1. Draw $\boldsymbol{\theta}_d$ independently for $d = 1, \ldots, D$ from Dirichlet($\boldsymbol{\alpha}$).
2. Each word $w_{d,n}$ in document d is generated from a two-step process:
   - 2.1 Draw topic assignment $z_{d,n}$ from $\boldsymbol{\theta}_d$.
   - 2.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{z_{d,n}}$.
3. Draw y_d from $\mathcal{N}(\bar{\mathbf{z}}_d \cdot \boldsymbol{\phi}, \sigma^2)$ where $\bar{\mathbf{z}}_d = (n_{1,d}/N_d, \ldots, n_{K,d}/N_d)$.

Essentially plain LDA with a linear regression linking topic allocations with observed variables.

sLDA Plate Diagram



Joint Likelihood

Applying the factorization formula for Bayesian networks to sLDA yields

$$\left(\prod_d \Pr\left[\boldsymbol{\theta}_d \mid \alpha\right]\right)\left(\prod_k \Pr\left[\boldsymbol{\beta}_k \mid \eta\right]\right) \times$$
$$\left(\prod_d \prod_n \Pr\left[z_{d,n} \mid \boldsymbol{\theta}_d\right]\right) \times$$
$$\left(\prod_d \prod_n \Pr\left[w_{d,n} \mid z_{d,n}, \mathbf{B}\right]\right) \times$$
$$\left(\prod_d \Pr\left[y_d \mid \mathbf{z}_d, \boldsymbol{\phi}, \sigma^2\right]\right)$$

Inference

One can apply a stochastic EM algorithm. The sampling equation for the topic allocations becomes

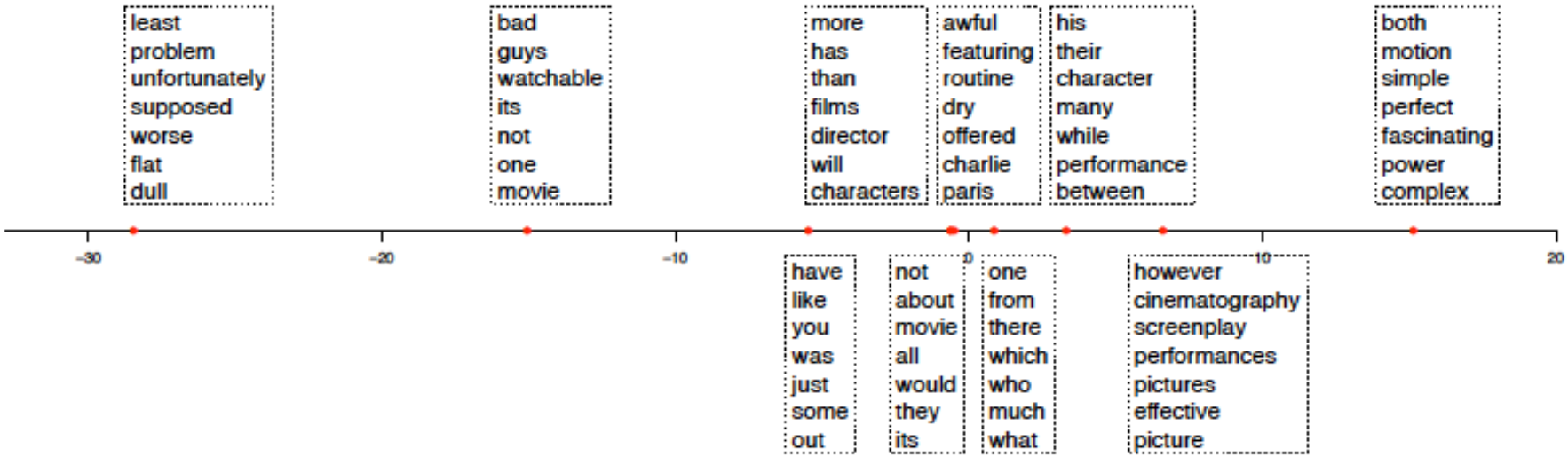
$$\Pr\left[z_{d,n} = k \mid \mathbf{z}_{-(d,n)}, \mathbf{w}, \alpha, \eta \right] \propto$$

$$\frac{m^{-}_{k,v_{d,n}} + \eta}{\sum_v m^{-}_{k,v} + \eta V} \left(n^{-}_{d,k} + \alpha \right) \exp[-(y_d - \boldsymbol{\phi} \cdot \bar{z}_d)^2] \propto$$

$$\frac{m^{-}_{k,v_{d,n}} + \eta}{\sum_v m^{-}_{k,v} + \eta V} \left(n^{-}_{d,k} + \alpha \right) \exp[2\phi_k / N_d (y_d - \boldsymbol{\phi} \cdot \bar{z}^{-}_{d}) - (\phi_k / N_d)^2].$$

Alternate between drawing samples for topic allocations (E-step), and updating the estimated coefficients $\boldsymbol{\phi}$ through standard OLS (M-step).

Movie Review Example



Conclusion

We have seen numerous discriminative and generative models for supervised learning.

Text regression is essentially a particular instance of a more general model in machine learning, which brings along the same qualifications and caveats.

Generative models take seriously the count nature of text in building a likelihood function, and are recommended with there are relatively few documents.

A new possibility which you should explore is word embeddings, a neural network approach for representing vocabulary terms and documents in a latent space.