Machine Learning Methods for Economists
Generative Models for Supervised Learning

Stephen Hansen
University of Oxford

Introduction

In lecture 1 we discussed discriminative classifiers that estimated models of the form $p(y_i \mid \mathbf{x}_i)$, which can be applied directly to text data.

Recall that a generative classifier estimates the full joint distribution $p(y_i, \mathbf{x}_i)$.

Efron (1975)[1] shows that discriminative classifiers obtain a lower asymptotic error than generative ones.

Why then study generative classifiers?

1. Ng and Jordan (2001)[2] show that generative classifiers can approach their (higher) asymptotic error faster.
2. They can reveal interesting structure, e.g. $p(\mathbf{x}_i \mid y_i)$.

Applying a generative classifier requires a probability model for $\mathbf{x}_i$, which have developed in previous lectures.

[1] JASA, "The efficiency of logistic regression compared to Normal Discriminant Analysis".

[2] NIPS, "On Discriminative vs Generative Classifiers: A comparison of logistic regression and naive Bayes".

Feature Selection for Discriminative Classifier

One question is how to represent text: can use unigram, bigram, trigrams counts.

One can also apply a dimensionality-reduction algorithm to map $\mathbf{x}_d$ into a K-dimensional latent space, and use these as the features.

This technique is related to principal components regression, and is particularly appropriate when terms are highly correlated.

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 classification.

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

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

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:

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

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 $\boldsymbol{\alpha}_y$ and $\boldsymbol{\beta}_{1,y}, \ldots, \boldsymbol{\beta}_{K,y}$ for each unique class label y.

For an out-of-sample document d, one can then obtain an estimate of $\boldsymbol{\theta}_{d,y}$ given the estimated hyperparameters and topics for class label y, 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 y is $\sum_k \widehat{\theta}_{d,y,k} \widehat{\beta}_{k,y,v}$.

Inverse Regression

Modeling and inverting the relationship $p(\mathbf{x_i} \mid y_i)$ is more difficult when y_i 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)[3] and Taddy (2015)[4] have proposed fully generative models for inverse regression.

3 JASA, "Multinomial Inverse Regression for Text Analysis".

4 Annals of Applied Statistics, "Distributed Multinomial 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 (2015)

Taddy (2015) formulates a model that is also relevant for treatment effect estimation in the presence of high-dimensional (discrete) controls.

We can extend MNIR to allow the response variable y_d to have multiple dimensions, i.e. $y_d = (y_{d,1}, \ldots, y_{d,M})$.

One of these can be a policy and a response variable associated with some document d.

Taddy (2015)

Taddy (2015) formulates a model that is also relevant for treatment effect estimation in the presence of high-dimensional (discrete) controls.

We can extend MNIR to allow the response variable y_d to have multiple dimensions, i.e. $y_d = (y_{d,1}, \dots, y_{d,M})$.

One of these can be a policy and a response variable associated with some document d.

For example, the Bank of England publishes its Inflation Report (IR) and forecasts on the same day.

Suppose we are interested in the impact of the low-dimensional forecast variables on the daily change in bond prices.

Two strategies:

1. Double LASSO.
2. Model the distribution of terms in IR as a function of bond price changes and forecast variables.

Treatment Effect Estimation

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 $z_{d,m} = \mathbf{f}_d \cdot \mathbf{b}_m$.

Suppose $y_{d,1}$ is an outcome variable and $y_{d,2}$ is a treatment.

The SR result implies that $y_{d,1}, y_{d,2} \perp \mathbf{x}_d$ given $z_{d,1}$, $z_{d,2}$, N_d, and other controls $m > 2$.

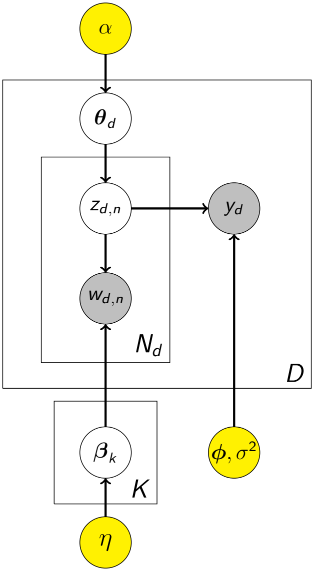
We can then perform a forward regression with just the SR projections to estimate the treatment effect.

Supervised LDA (Blei and McAuliffe)

1. Draw $\boldsymbol{\pi}_d$ 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 $\boldsymbol{\pi}_d$.
   2.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{z_{d,n}}$.
3. Draw y_d from $\mathcal{N}(\bar{\mathbf{z}}_d \cdot \phi, \sigma^2)$ 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

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

Inference

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

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

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

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

Alternate between drawing samples for topic allocations (E-step), and updating the estimated coefficients ϕ through standard OLS (M-step).

Movie Review Example