Topics in Empirical Economics, Part III
Supervised Learning

Stephen Hansen
Universitat Pompeu Fabra

Introduction

Supervised learning is the problem of predicting a response variable (i.e. dependent variable) associated with observations with one or more features (i.e. covariates).

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

Supervised learning has many commercial applications: predict which products a shopper will buy; 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.

We will explore supervised learning algorithms in the context of text data, and their role in economics research.

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.

The 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.

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.

Classification Problem

More generally speaking, the above problem is one of binary classification.

Suppose all documents belong to one of two classes A and B.

We want to use the data in the document-term matrix to classify documents into one of the two classes.

There are many approaches in the machine-learning literature for this problem.

Assessing Classifier Performance

A typical problem in predictive modeling is over-fitting, meaning we tune the model to idiosyncratic features of the data.

One popular way of addressing this problem is through cross-validation:

1. Divide the data into a training portion and a test portion.
2. Estimate a predictive model using the training data.
3. Use the estimated model to predict the value of the response variable in the test data.
4. Compare the predicted and actual values for the response variable in the test data.

Cross-validation helps establish how well models generalize to unseen data.

Note that simple mis-classification rates may obscure classifier performance. More informative to look at performance for each different category of the response variable.

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.

Close connection to 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.

Can select k using cross-validation exercise.

Parametric Classification

There are two kinds of parametric model for modeling the relationship between covariates $\mathbf{x}_i$ and dependent variable y_i.

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

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.

Linear Regression

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

The log-likelihood function is

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

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

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

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

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

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.

Lasso Regression

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 is even more likely to promote sparsity.

The objection function for MAP estimation can be written

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

which is called the *lasso regression* model.

Increasingly popular in economics, see work for example of Victor Chernozhukov.

Naive Bayes Classifier

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

Consider again the problem of C distinct classes. We want to model $\Pr[c \mid \mathbf{x}_d] \propto \Pr[\mathbf{x}_d \mid c] \Pr[c]$, and use this for classification.

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

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

Classification

To assign a class-label c_d to an out-of sample document we can use MAP estimation given the estimated relationship $\Pr[\, c \mid \mathbf{x}_d \,]$.

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

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

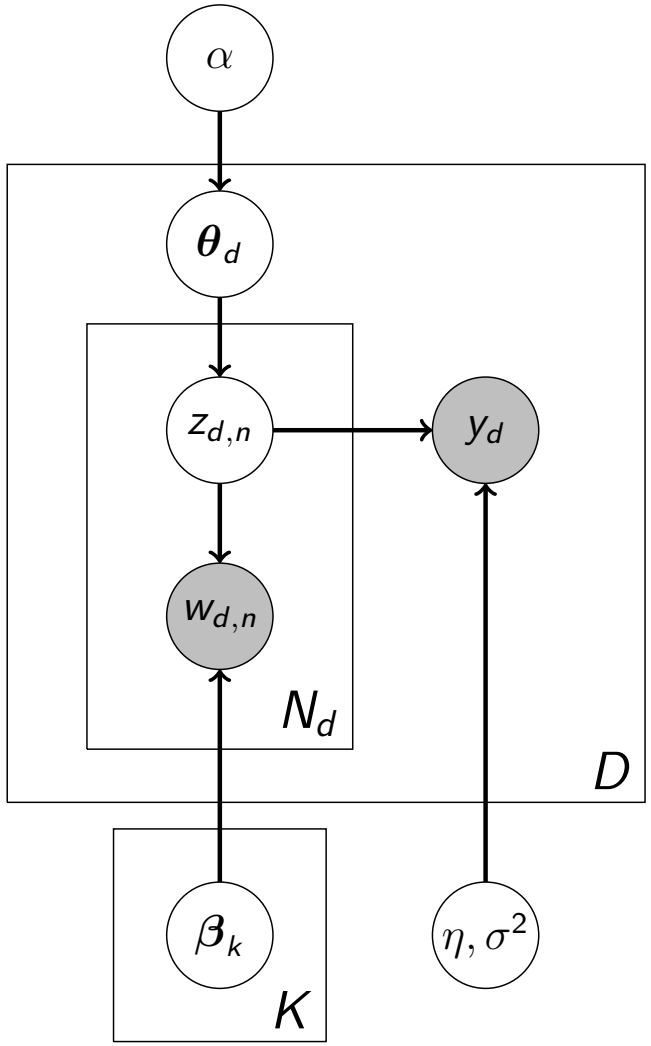
Close relationship to logistic regression (discriminative classifier).

Supervised LDA (Blei and McAuliffe)

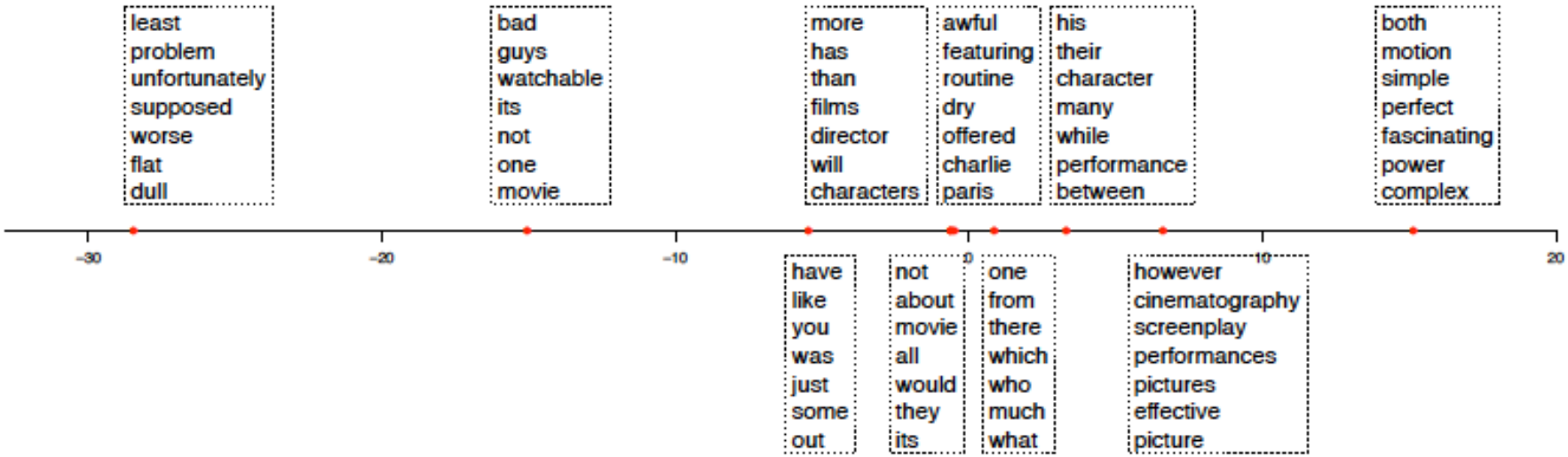
1. Draw $\boldsymbol{\theta}_d$ independently for $d = 1, \ldots, D$ from Dirichlet($\boldsymbol{\alpha}$).
2. Each word $w_{d,n}$ in document d is generated from a two-step process:
   - 2.1 Draw topic assignment $z_{d,n}$ from $\boldsymbol{\theta}_d$.
   - 2.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{z_{d,n}}$.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

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

sLDA Plate Diagram



Movie Review Example



Conclusion

Key ideas from this lecture:

1. Supervised learning is more widely applied than unsupervised learning.
2. The machine learning literature essentially evaluates the performance of supervised learning algorithms in terms of a sole criterion: prediction.
3. The utility of these for economists is still a somewhat open question given our emphasis on causal modeling.
4. Various useful techniques for correcting over-fitting.