Lecture 9: Bayesian and Machine Learning Methods

Chris Walters

University of California, Berkeley and NBER

Introduction

- This lecture gives an introduction to **Bayesian** and **machine learning** (ML) methods
- Useful for "regularizing" or "shrinking" estimates to reduce the influence of statistical noise ("overfitting")
- Often motivated as methods for dealing with "big data"
  - But primarily useful for dealing with finite-sample problems in settings with many predictor variables
  - Perhaps more accurate to think of these as "rich data" methods
- ML methods focus on prediction rather than causal inference
  - Open question (Angrist and Frandsen, 2020): Is ML useful for estimating causal effects?

Empirical Bayes

- **Empirical Bayes** (EB) methods are used in settings with large numbers of parameters
- Examples:
  - Teacher/school effects
  - Firm effects
  - Neighborhood effects
- We are often interested both in individual parameters (how effective is a specific teacher?) and in the *distribution* of parameters (how much does effectiveness vary across teachers?)
- Useful for analyzing **hierarchical** data: students within classrooms, workers within firms, etc.
- See Robbins (1964) and Morris (1983) for background

EB Framework: Level I

- We'll start with an abstract description of EB, then illustrate with an example
- Let $j \in \{1...J\}$ index groups (e.g. classes), and let $i \in \{1...N\}$ index individuals within groups (e.g. students)
- θ_j is an unknown parameter for group j (e.g. the causal effect of teacher j)
- Y_{ij} is an observed outcome for individual i in group j, assumed to follow the distribution

$$Y_{ij}|\theta_j \sim f(y;\theta_j)$$

EB Framework: Level II

- At the next level of the hierarchy, we posit a second distribution that governs the cross-group distribution of parameters:

$$\theta_j \sim g(\theta; \Omega)$$

- In the Bayesian framework, $g(\cdot)$ is a **prior distribution**, and Ω is a **hyperparameter** describing the prior
- Equivalently, think of this as a random coefficients model with $g(\cdot)$ the distribution of random coefficients
- We hope to estimate
  - The individual θ_j's, which tell us about specific groups
  - The hyperparameter Ω, which tells us about cross-group heterogeneity

EB Framework: Estimating Hyperparameters

- To estimate Ω, construct an integrated likelihood function that expresses the distribution of the data for group j, $Y_j = (Y_{1j} ... Y_{Nj})$, as a function of the hyperparameters:

$$\mathcal{L}(Y_j|\Omega) = \int \prod_{i=1}^{N} f(Y_{ij}; \theta) \, g(\theta; \Omega) \, d\theta$$

- EB maximum likelihood estimator:

$$\hat{\Omega}_{EB} = \arg\max_{\Omega} \sum_{j=1}^{J} \log \mathcal{L}(Y_j|\Omega)$$

- Alternatively, we could estimate Ω by method of moments

EB Posteriors

- Using Bayes' rule, the **posterior density** for the group-specific parameter θ_j conditional on the observed data is:

$$h(\theta_j|Y_j;\Omega) = \frac{\prod_i f(Y_{ij};\theta_j)g(\theta_j;\Omega)}{\mathcal{L}(Y_j|\Omega)}$$

- Often we are interested in a particular feature of the posterior distribution, such as the posterior mean:

$$\theta_j^* = \int \theta h(\theta|Y_j;\Omega)d\theta$$

- Putting the "E" in "EB": an empirical Bayes posterior mean plugs $\hat{\Omega}_{EB}$ into this formula

EB Example: Teacher Effects

- Suppose students are randomly assigned to classrooms, and student test scores are given by

$$Y_{ij} = \theta_j + \epsilon_{ij},$$

$$\epsilon_{ij}|\theta_j \sim N\left(0, \sigma_\epsilon^2\right)$$

- θ_j is the mean potential score for teacher j, which measures teacher j's effectiveness
- The distribution of teacher effects is

$$\theta_j \sim N(\mu_\theta, \sigma_\theta^2)$$

- Here μ_θ and σ_θ^2 are hyperparameters

EB Example: Teacher Effects

- With equal class sizes of N, estimates of hyperparameters in the teacher effects model are

$$\hat{\mu}_\theta = \frac{1}{J}\sum_j \bar{Y}_j,$$

$$\hat{\sigma}^2_\theta = \frac{1}{J}\sum_j \left(\bar{Y}_j - \hat{\mu}_\theta\right)^2 - \hat{\sigma}^2_\epsilon/N$$

$$\hat{\sigma}^2_\epsilon = \frac{1}{NJ}\sum_i \sum_j \left(Y_{ij} - \bar{Y}_j\right)^2$$

- Here $\bar{Y}_j = \frac{1}{N}\sum_i Y_{ij}$ is the mean for class j
- The variance of teacher effects, σ^2_θ, is inferred from **overdispersion** in class averages beyond what we'd expect from random chance

EB Example: Teacher Effects

- In the normal/normal model, the posterior mean takes a simple form:

$$\theta^*_j = \tau\bar{Y}_j + (1-\tau)\mu_\theta,$$

$$\tau = \frac{\sigma^2_\theta}{\sigma^2_\theta + \sigma^2_\epsilon/N}$$

- The EB posterior mean plugs $\hat{\mu}_\theta$, $\hat{\sigma}^2_\theta$ and $\hat{\sigma}^2_\epsilon$ into this formula
- The posterior mean shrinks the unbiased estimate $\bar{Y}_j$ towards the grand mean μ_θ in proportion to one minus the signal-to-noise ratio
- With large enough classes, $\tau \to 1$ and shrinkage disappears

To Shrink or Not to Shrink?

- Note that the sample mean $\bar{Y}_j$ is an unbiased estimate of θ_j. Why should we prefer the shrunken posterior mean θ_j^*?
- It's a matter of perspective. Consider the mean squared error (MSE) of $\bar{Y}_j$ and θ_j^*, treating teacher j's effect θ_j as a fixed parameter:

$$E\left[\left(\bar{Y}_j - \theta_j\right)^2 | \theta_j\right] = \sigma_\epsilon^2 / N$$

$$E\left[\left(\theta_j^* - \theta_j\right)^2 | \theta_j\right] = \tau^2 \sigma_\epsilon^2 / N + (1-\tau)^2 \left(\theta_j - \mu_\theta\right)^2$$

- For any particular teacher, it is not clear which is better – it depends on the (unknown) true effect θ_j
- θ_j^* is also biased, in the sense that $E\left[\theta_j^* | \theta_j\right] \neq \theta_j$
- If we only cared about one specific teacher's effect, we might prefer the sample mean

To Shrink or Not to Shrink?

- But θ_j^* has lower mean squared error averaged over all teachers:

$$E\left[\left(\bar{Y}_j - \theta_j\right)^2\right] = \sigma_\epsilon^2 / N$$

$$E\left[\left(\theta_j^* - \theta_j\right)^2\right] = \tau \sigma_\epsilon^2 / N$$

- If we care about getting the answer right for every teacher, θ_j^* does better (in an MSE sense) than $\bar{Y}_j$
- In this sense θ_j^* is the minimum MSE predictor of θ_j
- We assumed a normal distribution, but θ_j^* is the Best Linear Predictor (BLP) of θ_j even if the normality assumption is wrong

EB Application: Teacher Value-Added

- Chetty, Friedman and Rockoff (CFR; 2014) look at the effects of teachers on short-run test scores and long-run outcomes
- CFR study **value-added models**: OLS estimates of teacher effects, controlling for previous test scores and demographics
- Motivated by a selection-on-observables assumption: this year's teacher is as good as random conditional on last year's score
- Re-interpret $\bar{Y}_j$ from our simple example as a class average residual after adjusting for past test scores
- CFR introduce some subtlety by allowing effects for a given teacher to "drift" over time, but the core EB approach is the same as in our example

EB Application: Teacher Value-Added

- CFR are interested in the regression:

$$Income_i = \alpha + \beta\theta_{j(i)} + X_i'\gamma + \epsilon_i$$

- θ_j is the test score value-added of teacher j, $j(i)$ is i's teacher, $Income_i$ is income in adulthood, and X_i is a vector of control variables
- If teachers are as good as random conditional on X_i, the OLS coefficient β answers the question: Do teachers who improve short-run test scores improve long-run outcomes?
- Problem: We don't observe θ_j
- Substituting in the unbiased but noisy estimate $\bar{Y}_j$ would cause attenuation bias, pulling the OLS estimate of β towards 0

EB Application: Teacher Value-Added

- A useful fact: putting the posterior mean θ_j^* on the right-hand side of a regression fixes attenuation bias
- If the model is right, EB shrinkage reduces variance to exactly offset measurement error
- A regression of $Income_i$ on $\theta_{j(i)}^*$ therefore produces the same coefficient we'd get if we knew the true $\theta_{j(i)}$ (though the estimate is less precise)
- Note: Shrinkage fixes bias when applied to a right-hand side variable, but causes bias when applied to a left-hand side variable (recall that classical measurement error on the left does not generate bias)

TABLE 3—IMPACTS OF TEACHER VALUE-ADDED ON EARNINGS

	Earnings at age 28 ($) (1)	Earnings at age 28 ($) (2)	Earnings at age 28 ($) (3)	Working at age 28 (%) (4)	Total income at age 28 ($) (5)	Wage growth ages 22–28 ($) (6)
Teacher VA	349.84 (91.92)	285.55 (87.64)	308.98 (110.17)	0.38 (0.16)	353.83 (88.62)	286.20 (81.86)
Mean of dep. var.	21,256	21,256	21,468	68.09	22,108	11,454
Baseline controls	X	X	X	X	X	X
Parent chars. controls		X				
Lagged score controls			X			
Observations	650,965	650,965	510,309	650,965	650,965	650,943

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample (as described in the notes to Table 1). There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure described in Section IIIA. The dependent variable in columns 1–3 is the individual's wage earnings reported on W-2 forms at age 28. The dependent variable in column 4 is an indicator for having positive wage earnings at age 28. The dependent variable in column 5 is total income (wage earnings plus self-employment income). The dependent variable in column 6 is wage growth between ages 22 and 28. All columns control for the baseline class-level control vector; column 2 additionally controls for parent characteristics, while column 3 additionally controls for twice-lagged test scores (see notes to Table 2 for details). We use within-teacher variation to identify the coefficients on all controls as described in Section IA; the estimates reported are from regressions of outcome residuals on teacher VA with school by subject level fixed effects.

Combining Estimators

- Bayesian methods are also useful in settings where multiple estimates of the effect of interest are available
- Suppose we observe experimental or quasi-experimental estimates $\hat{\theta}_j^E$, as well as (possibly biased) non-experimental estimates, $\hat{\theta}_j^{NE}$:

$$\hat{\theta}_j^E = \theta_j + u_j^E$$

$$\hat{\theta}_j^{NE} = \theta_j + b_j + u_j^{NE}$$

- u_j^E and u_j^{NE} are mean-zero estimation errors
- b_j is bias in the non-experimental estimate for group j
- The experimental estimates are unbiased, but also less precise: $Var(u_j^E) >> Var(u_j^{NE})$
- This creates a bias/variance tradeoff. What is our best estimate of θ_j using all available information?

Combining Estimators: Examples

$$\hat{\theta}_j^E = \theta_j + u_j^E$$

$$\hat{\theta}_j^{NE} = \theta_j + b_j + u_j^{NE}$$

- Chetty and Hendren (2018) study causal effects of neighborhoods on child income
  - $\hat{\theta}_j^E$ are quasi-experimental estimates based on families that move between neighborhoods
  - $\hat{\theta}_j^{NE}$ are OLS estimates based on permanent residents
- Angrist, Hull, Pathak and Walters (2017) study causal effects of schools on test scores
  - $\hat{\theta}_j^E$ are quasi-experimental estimates based on randomized admission lotteries
  - $\hat{\theta}_j^{NE}$ are OLS value-added estimates that control for past test scores

Combining Estimators

- Suppose causal effects, bias, and estimation errors are normally distributed. Then the posterior mean is

$$\theta_j^* = \tau^E \hat{\theta}_j^E + \tau^{NE}\left(\hat{\theta}_j^{NE} - \mu_b\right) + \left(1 - \tau^E - \tau^{NE}\right)\mu_\theta$$

- The posterior mean is a weighted average of the unbiased experimental estimate, the biased non-experimental estimate (net of mean bias), and the prior mean
- When $Var(u_j^{NE}) \approx 0$, the shrinkage factors are

$$\tau^E = \frac{\sigma_\theta^2(1-R^2)}{\sigma_\theta^2(1-R^2)+Var(u_j^E)}\ ,\ \tau^{NE} = \psi\left(1-\tau^E\right)$$

- Here $\psi = Cov(\theta_j, \theta_j + b_j)/Var(\theta_j + b_j)$ is the reliability ratio from a regression of the true effect on the non-experimental estimate, and R^2 is the R-squared from this regression
- This "hybrid" approach trades off the bias in non-experimental estimates against variance in experimental estimates to minimize MSE
- EB version plugs in estimates of prior means and shrinkage factors

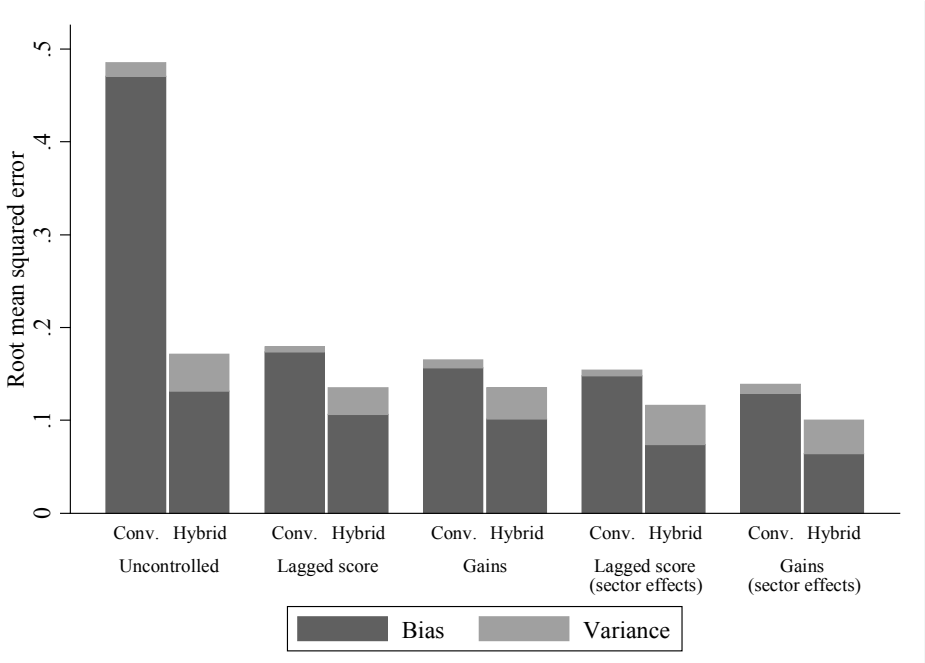


Figure VI. Root Mean Squared Error for Value-Added Posterior Predictions

Notes: This figure plots root mean squared error (RMSE) for posterior predictions of sixth grade math value-added. Conventional predictions are posterior means constructed from OLS value-added estimates. Hybrid predictions are posterior modes constructed from OLS and lottery estimates. The total height of each bar indicates RMSE. Dark bars display shares of mean squared error due to bias, and light bars display shares due to variance. RMSE is calculated from 500 simulated samples drawn from the data generating processes implied by the estimates in Table VI. The random coefficients model is re-estimated in each simulated sample.

EB vs. Full B

- A fully Bayesian analysis of teacher effects would add a third level to the hierarchy: a **hyperprior distribution** over $(\mu_\theta, \sigma_\theta)$, with parameters chosen by the researcher rather than estimated
- We would then compute posterior distributions for μ_θ, σ_θ, and θ_j
- Why should we opt for empirical Bayes rather than fully Bayesian methods?

EB vs. Full B

- Pros of full Bayes:
  - If you're a committed Bayesian, none! But if you're a frequentist...
  - EB posteriors do not account for estimation error in hyperparameters, so can overstate precision (though we can adjust for this; see Morris, 1983)
  - In some cases hyperparameters are difficult to estimate, and smoothing via a hyperprior can help
- Cons of full Bayes:
  - Fully Bayesian estimation requires simulation methods (Markov Chain Monte Carlo, MCMC), which are harder to implement and less transparent
  - Where do the parameters of the hyperprior come from? To the extent that these affect the estimates, why should we believe the results?
- EB estimates have desirable frequentist properties and are easier to understand – arguably less "harmful"
- If hyperparameters are estimated precisely, there won't be much difference

Machine Learning

- **Machine learning** (ML) refers to a class of data-driven statistical methods for selecting and estimating predictive models
- The challenge is avoiding overfitting: with N observations and N predictors, we can fit the data perfectly in-sample, and (probably) fail miserably out of sample
- ML techniques penalize model complexity (regularize) to improve out-of-sample fit
- In econometrics we are typically interested in causal inference, not prediction
  - One explanatory treatment variable D_i is often privileged at the expense of other controls X_i
  - We don't usually care about correctly specifying $E[Y_i|X_i]$ – controls are only included to increase precision
  - After introducing ML, we will discuss prospects for improving causal inference

ML Prediction Problem

- Suppose we are interested in predicting the value of some variable Y_i using a vector of P predictors $X_i = (X_{i1}...X_{iP})$, with X_{ik}'s normalized to mean zero and variance 1
- We observe (Y_i, X_i) for a sample of size N
- Goal is out of sample prediction: use X_i to predict Y_i in a new sample drawn from the same population
- Minimum MSE predictor is $E[Y_i|X_i]$. The challenge is getting a good estimate of this CEF
- Obvious first impulse: Run an OLS regression of Y_i on X_i
- Problem: What if $P > N$? We can't run OLS
- Even if $P < N$, OLS will likely perform poorly out of sample if P is close to N

Regularized Regression

- Regularized least squares regression:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \left[Y_i - X_i' \beta \right]^2 + \lambda p(\beta)$$

- $p(\cdot)$ is a penalty function that depends on the complexity of the model through β
- λ is a tuning parameter that determines how severely to penalize
- Several common ML procedures have this structure

LASSO

- Common approach: Least Absolute Shrinkage and Selection Operator (LASSO)
- LASSO uses the penalty function

$$p(\beta) = \sum_{p=1}^{P} |\beta_p|$$

- LASSO penalizes the absolute value of each coefficient
  - Penalty function is kinked at zero, so will result in lots of coefficients set exactly to zero
  - This makes LASSO useful for variable selection
- Motivated by the idea of **sparsity**: the set of variables with non-zero coefficients in the true model is small (equivalently, a lot of totally irrelevant variables are included in X_i)
- LASSO may not work as well if the true model is not sparse

Variants of LASSO

- Post-LASSO
  - First, run LASSO. Then, take the set of predictors selected to have non-zero coefficients, and run OLS without regularization
  - LASSO penalizes coefficients for both irrelevant variables and variables that belong in the model, pushing all coefficients towards zero
  - Post-LASSO gets rid of "shrinkage bias" by eliminating the penalty for variables selected for inclusion
- $\sqrt{LASSO}$
  - Objective function: $\sqrt{\sum_i [Y_i - X_i'\beta]^2} + \lambda \sum_p |\beta_p|$
  - Unlike LASSO, the optimal λ for $\sqrt{LASSO}$ does not depend on unknown parameters (more on this later)

Ridge Regression

- Ridge regression uses the penalty function:

$$p(\beta) = \sum_{p=1}^{P} \beta_p^2$$

- Quadratic penalty rather than linear
- Unlike LASSO, Ridge will not set coefficients to exactly zero – Ridge does not perform variable selection
- But Ridge's convex penalty punishes large coefficients more harshly
- Ridge tends to perform better than LASSO when the true model is **dense** rather than sparse (few exact zero coefficients)
- See Abadie and Kasy (forthcoming) for more on the relative performance of LASSO and Ridge

Choosing the Tuning Parameter

- Choice of the tuning parameter λ is crucial for good performance of regularized models
  - $\lambda = 0$ recovers OLS, while $\lambda \to \infty$ shrinks all coefficients to zero
  - How should we choose λ?
- There are some theoretical results on optimal (MSE-minimizing) penalization, but these often involve unknown parameters
  - Optimal λ's for LASSO and Ridge depend on σ, the unknown standard deviation of errors in the true model (Bickel et al., 2009)
  - Optimal λ for $\sqrt{LASSO}$ is independent of σ, a virtue of this approach (Belloni et al., 2011)
- Researchers commonly select λ with a data-driven approach such as **k-fold cross validation**

Cross Validation

- K-fold cross validation algorithm:
  - Randomly partition the sample into k separate subsets (folds)
  - For a given value of λ, leave out one fold, and estimate the model on the remaining $k-1$ folds
  - Compute the fit (MSE) of the estimated model on the left-out fold
  - Do this k times, once for each fold, and combine results to compute overall goodness of fit
  - Repeat for many λ's to find the tuning parameter that generates best fit
- Common choices for k: 5, 10, and N (leave one out)
- Cross validation is useful more generally for assessing out-of-sample fit

ML and Bayesian Methods

- Let's return to our teacher effects example to explore the connection between ML and Bayesian methods
- Since we assumed a normal prior and normal data, the posterior distribution for θ_j is normal, so the posterior mean θ_j^* is also the posterior mode
- This implies that θ_j^* maximizes the posterior density:

$$\theta_j^* = \arg\max_{\theta} \log h(\theta | Y_j; \Omega)$$

$$= \arg\max_{\theta} \log \left(\prod_i f(Y_{ij}; \theta) \right) + \log g(\theta; \Omega)$$

- The posterior mode is also known as a **maximum a posteriori** (MAP) estimate

ML and Bayesian Methods

$$\theta_j^* = \arg\max_{\theta} \log \left(\prod_i f(Y_{ij}; \theta) \right) + \log g(\theta; \Omega)$$

- Plugging in normal densities gives

$$\theta_j^* = \arg\max_{\theta} -\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{N} [Y_{ij} - \theta]^2 - \frac{1}{2\sigma_\theta^2} (\theta - \mu_\theta)^2$$

$$= \arg\min_{\theta} \sum_{i=1}^{N} [Y_{ij} - \theta]^2 + \underbrace{\sigma_\epsilon^2}_{\lambda} \times \underbrace{\left(\frac{\theta - \mu_\theta}{\sigma_\theta} \right)^2}_{p(\theta)}$$

- The Bayesian posterior mean can be rewritten as a regularized least squares estimate with a Ridge-style quadratic penalty

ML and Bayesian Methods

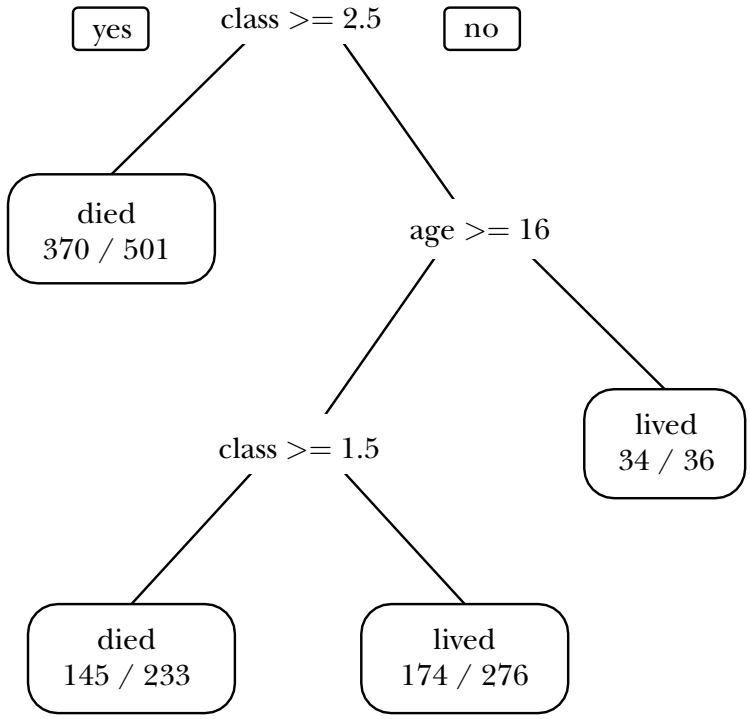
- ML procedures often have a Bayesian interpretation
  - LASSO estimates can be interpreted as posterior modes from a model with double exponential (Laplace) priors on the elements of β
  - Ridge estimates can be interpreted as posterior means from a model with normal priors on the elements of β
- The Bayesian posterior mode solves a regularized maximum likelihood problem, with the prior density serving as the regularization term
- No clear distinction between ML and Bayesian approaches

Other Useful ML Methods

- Decision trees
  - Classify observations via a sequence of nodes defined by values of predictor variables. Useful method for capturing complex nonlinearities and interactions
  - "Prune" trees by introducing a cost for the number of nodes
  - Bagging (bootstrap aggregating): Construct many trees on random samples of size N, drawn with replacement. Final classification for an observation is based on a "vote" of these trees
  - Random forest: Modify bagging by using a random subset of the predictor variables at each node
- K-means clustering
  - Assign observations to k groups to minimize sum of squared errors relative to a group-specific mean
  - Example of an **unsupervised** learning algorithm: infer latent grouping structure in the data, rather than predicting one variable as a function of others (**supervised** learning)
- There are many variations on these themes, and many other ML techniques – see Varian (2014) for an overview from an economist's perspective

Figure 1

A Classification Tree for Survivors of the *Titanic*



ML and Causation

- Sophisticated prediction algorithms give more accurate measures of correlation, but they do not distinguish causation from selection bias
- ML is not a substitute for a research design
- In some cases, however, accurate predictive models are important for estimating causal relationships. Examples:
  - Selection on observables with a high-dimensional set of control variables
  - Instrumental variables estimation with many/weak instruments
- Angrist and Frandsen (2020): In labor applications, ML seems to do better at selecting control variables than picking instruments

ML Example: Dale and Krueger Redux

- Angrist and Frandsen (AF; 2020) consider the use of ML methods in a reanalysis of Dale and Krueger (DK; 2002, 2014)
- DK seek to estimate the causal returns to college selectivity
  - For future earnings, is it better to attend UPenn, or Penn State?
  - Research design = selection on observables: compare students who applied/were admitted to the same schools but made different attendance choices, assuming CIA
  - Basic finding: matching on application/admission sets eliminates the apparent returns to selectivity
- DK control for a large set of application/attendance dummies, reducing sample size and precision
- AF ask whether ML can find a sufficient but more parsimonious control set, maintaining CIA
  - Post double selection (Belloni et al., 2014): LASSO regressions of Y_i and D_i on X_i; retain union of variables selected in either model

Table 1: OLS Estimates of Elite College Effects

	Basic Controls		DK02 Selection controls			
					Self-revelation	
	None (1)	Personal charac-teristics (2)	Barron's matches only (3)	matches w/pers. char. (4)	Barron's sample (5)	Full sample (6)
A. Private School Effects						
Estimated Effect	0.212	0.139	0.007	0.013	0.036	0.037
	(0.060)	(0.043)	(0.038)	(0.025)	(0.029)	(0.039)
R**2	0.019	0.107	0.058	0.138	0.111	0.114
No. of controls	0	10	150	160	13	14
N	14238			5583		14238
B. Effects of School Average SAT/100						
Estimated Effect	0.109	0.076	0.008	0.004	0.004	0.000
	(0.026)	(0.016)	(0.029)	(0.016)	(0.017)	(0.018)
R**2	0.019	0.107	0.066	0.140	0.107	0.113
No. of controls	0	10	334	344	13	14
N	14238			9166		14238
C. Effects of Attending Schools Rated Highly Competitive +						
Estimated Effect	0.225	0.153	0.018	0.022	0.031	0.068
	(0.046)	(0.030)	(0.047)	(0.035)	(0.032)	(0.029)
R**2	0.020	0.108	0.048	0.129	0.106	0.114
No. of controls	0	10	128	138	13	14
N	14238			4945		14238

Table 2: Post-Lasso Estimates of Elite College Effects

	Double-selection (PDS)			Outcome selection			All controls
	plugin (16)	C.V. λ	cvlasso	plugin (16)	C.V. λ	cvlasso	OLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	A. Private School Effects						
Estimated Effect	0.038	0.020	0.040	0.046	0.043	0.042	0.017
	(0.040)	(0.039)	(0.041)	(0.041)	(0.043)	(0.043)	(0.039)
No. of controls	18	100	112	10	35	50	303
	B. Effects of School Average SAT/100						
Estimated Effect	-0.009	-0.013	-0.009	-0.008	-0.009	-0.008	-0.012
	(0.020)	(0.018)	(0.019)	(0.020)	(0.019)	(0.019)	(0.018)
No. of controls	24	151	58	10	34	43	303
	C. Effects of Attending Schools Rated Highly Competitive +						
Estimated Effect	0.068	0.051	0.073	0.076	0.080	0.082	0.053
	(0.033)	(0.033)	(0.033)	(0.031)	(0.032)	(0.032)	(0.033)
No. of controls	17	185	106	10	34	43	303

References

- Abadie, A., and Kasy, M. (forthcoming). "The risk of machine learning." *Review of Economics and Statistics*.
- Angrist, J., and Frandsen, B. (2020). "Machine labor." Working paper.
- Angrist, J., Hull, P., Pathak, P., and Walters, C. (2017). "Leveraging lotteries for school value-added: testing and estimation." *Quarterly Journal of Economics* 132(2).
- Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica* 80(6).
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). "Inference on treatment effects after selection among high-dimensional controls." *Review of Economic Studies* 81(2).
- Belloni, A., Chernozhukov, V., and Wang, L. (2011). "Square-root LASSO: pivotal recovery of sparse signals via conic programming." *Biometrika* 98(4).
- Bickel, T., Ritov., Y., and Tsybakov, A. (2009). "Simultaneous analysis of LASSO and the Dantzig selector." *Annals of Statistics* 37(4).
- Chetty, R., Friedman, J., and Rockoff, J. (2014). "Measuring the impacts of teachers II: teacher value-added and outcomes in adulthood." *American Economic Review* 104(9).

References

- Chetty, R., and Hendren, N. (2018). "The imact of neighborhoods on intergenerational mobility II: county-level estimates." *Quarterly Journal of Economics* 133(3).
- Dale, S., and Kruger, A. (2002). "Estimating the payoff to attending a more selective college: an application of selection on observables and unobservables." *Quarterly Journal of Economics* 117(4).
- Dale, S., and Kruger, A. (2014). "Estimating the return to college selectivity over the career using administrative earnings data." *Journal of Human Resources* 49(2).
- Gilchrist, D., and Sands, E. (2016). "Something to talk about: social spillovers in movie consumption." *Journal of Political Economy* 124(5).
- Morris, C. (1983). "Parametic empirical Bayes inference: theory and applications." *Journal of the American Statistical Association* 78(381).
- Robbins, H. (1964). "An empirical Bayes approach to statistical decision problems." *Annals of Mathematical Statistics* 35(1).
- Varian, H. (2014). "Big data: new tricks for econometrics." *Journal of Economic Perspectives* 28(2).