Introduction and Background

Robert McCulloch and Rodney Sparapani

2023-05-29

Modern Machine/Statistical Learning

Basic Bayesian Ideas

Regularization Priors

Modern Machine/Statistical Learning

What is modern statistics/machine learning/data science?

Lots of things.

But, in particular, we have truly amazing tools for finding high dimensional patterns in data.

The Stars of the show?

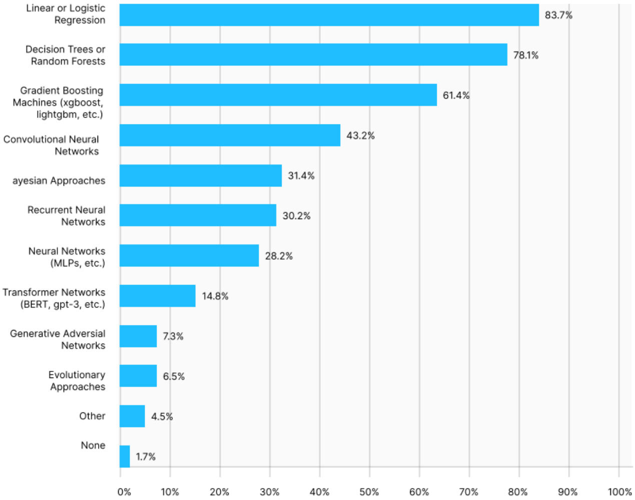
ensembles of trees, *neural networks*

kaggle™

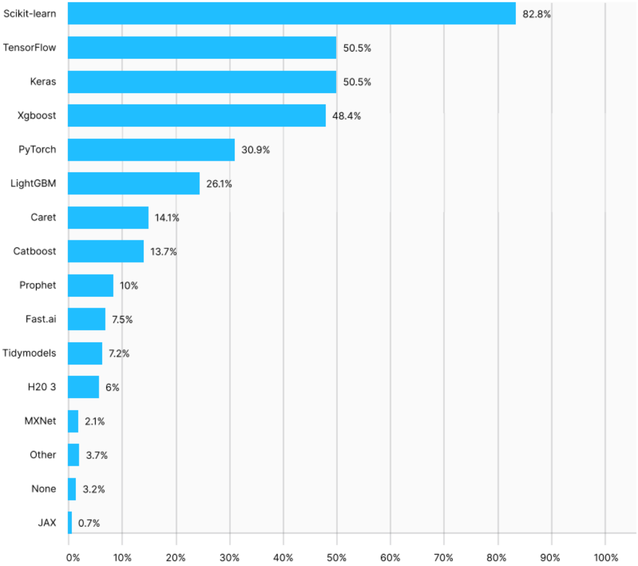
State of Machine Learning and Data Science 2020

Enterprise Executive Summary Report

METHODS AND ALGORITHMS USAGE



MACHINE LEARNING FRAMEWORK USAGE



Clearly:

- linear methods are still important
- after that, tree based methods are huge, *Random Forests*, *boosting*
- python and R are king

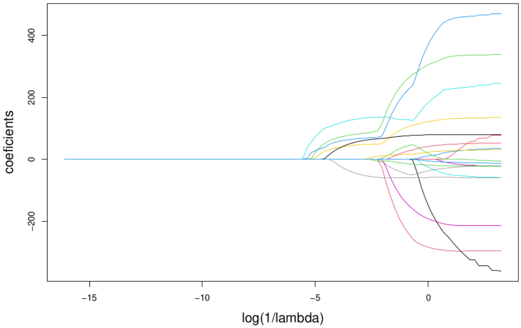
Bayesian approaches are also important !!!

Note that linear methods usually means *regularized regression*

Lasso:

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

We can have a complicated model (lots of x and corresponding β s) if we *shrink* the coefficients.

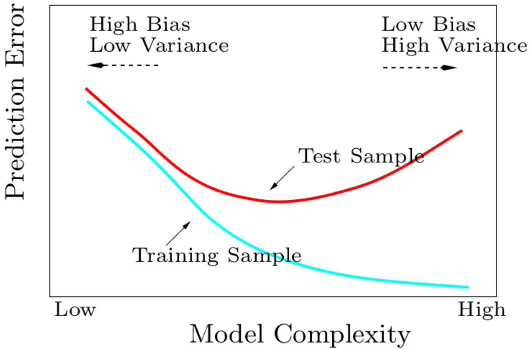


As λ get bigger, the coefficients are shrunk to zero.

With the Lasso L1 penalty, the solutions sets some of the coefficients to zero.

A big λ gives you a simple model, a small λ gives you a complicated model !!

The bias variance tradeoff



If your model is too simple (e.g shrink too much) you will miss important patterns in the data and get poor predictions.

If your model is too complex (e.g don't shrink enough) will will overfit the data, find false patterns, and get poor predictions.

What are the big ideas ???!!!

What want to have complex models with *many* parameters which can capture high dimensional nonlinear relationships.

We don't want to overfit, so we need to be able to *shrink towards simple models!!*

We can use the prior, in a Bayesian approach, to shrink towards simple models !!!

BART: Bayesian Additive Regression Trees

We build a high dimensional nonlinear model inspired by boosting and tree ensembles.

We use a *sophisticatedly simple* prior to shink towards a simple model.

We use MCMC to get draws from the high-dimensional posterior so that we can stochastically search the model space and gauge our uncertainty.

Often works well with little tuning because the prior does the regularization/shrinkage work.

Chapter 8 of the wonderful *An Introduction to Statistical Learning* is on tree based methods.

In the “lab’’, they illustrate the methods.

https://hastie.su.domains/ISLR2/Labs/Rmarkdown_Notebooks/Ch8-baggboost-lab.Rmd

On this data set, the test error of BART
is lower than the test error of random forests and boosting.

BART is a notably effective ensemble of trees method with some Bayesian advantages.

Basic Bayesian Ideas

The fundamental beauty of a Bayesian approach is its simplicity.

We have one big probability model !!.

In our most basic version we start with a model as giving the distribution for potentially observable y given parameters θ.

$$p(y|\theta)$$

A basic example is the Bernoulli model.

$y = (y_1, y_2, \ldots, y_n)$, with

$$Y_i \sim \text{Bernoulli}(p),\ \text{IID}$$

So here "θ is p".

Bayesian Inference for parametric models

Given a model for observables, $p(y|\theta)$, we think about our probabalistic beliefs about θ and capture them with a prior distribution $p(\theta)$.

We then have a joint model for y and θ

$$p(y, \theta) = p(\theta)\, p(y|\theta)$$

Note that we are reusing the symbol p and its meaning is determined by its arguments.

Given we observe y, our beliefs about θ are updated in light of this information by conditioning on y:

$$p(\theta|y) \propto p(y,\theta) = p(\theta)\, p(y|\theta)$$

The object $p(y|\theta)$ viewed as a function of θ for a fixed y is called the *likelihood function*.

$$L(\theta) \propto p(y|\theta)$$

Giving us the famous equation,

$$p(\theta|y) \propto p(\theta)\, L(\theta).$$

Bayesian Inference for the Bernoulli parameter

Let's suppose the Leafs have score on 26 of their last 100 powerplays and we believe that it is IID Bernoulli whether or not they score.

Given the data and our model, what are our beliefs about θ where $Y_i \sim$ Bernoulli(θ) (that is, θ is p).

To specify a prior on $\theta \in (0,1)$ we use the Beta distribution

$$\theta \sim Beta(\alpha, \beta)$$

$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \;\; \theta \in (0,1).$$

What is the likelihood?

$y = (y_1, y_2, \ldots, y_n)$, $y_i \in \{0, 1\}$.

$$p(y|\theta) = \Pi_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} = \theta^{S} (1-\theta)^{n-S}$$

where $S = \sum y_i$ the number of ones.

So,

$$p(\theta|y) \propto [\theta^{\alpha-1}(1-\theta)^{\beta-1}] [\theta^{S} (1-\theta)^{n-S}] = \theta^{\alpha+S-1} (1-\theta)^{\beta+n-S-1}$$

So,

$$\theta|y \sim \text{Beta}(\alpha + S, \beta + (n - S))$$

For the Leafs data lets use $\alpha = 2$, $\beta = 8$.
$S = 26$, $n - S = 74$

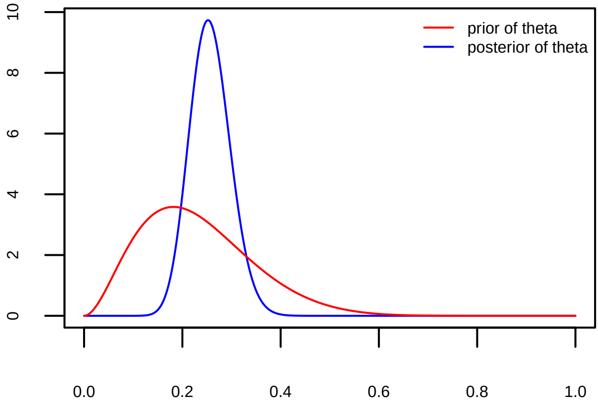
Let's plot the prior and posterior.

Choose the prior and compute the posterior.

```
# beta parameters for our prior
apri = 3; bpri = 10
# beta parameters for our posterior
apost = apri + 26
bpost = bpri + 74
```

In this case our prior in *conjugate* so computation of the posterior is very simple!!!

```
# plot prior and posterior at theta values in tvec
tvec = seq(from=.0001,to=.99999,length.out=1000)
plot(tvec,dbeta(tvec,apost,bpost),type='l',col='blue',xlab='',ylab="",
     cex.lab=.5,cex.axis=.5,cex=.5)
lines(tvec,dbeta(tvec,apri,bpri),col='red')
legend('topright',legend=c("prior of theta","posterior of theta"),
       col=c('red','blue'),lty=c(1,1),bty="n",cex=.5)
```



The difference between the prior and posterior is a great way to understand the information in the data.

Monte Carlo and Bayes

Suppose we have $p(\theta \mid y)$ and we want the marginal distribution of $\gamma = f(\theta)$.

- draw θ_j , $j = 1, 2, \ldots, J, \ \sim \theta|y$
- compute $\gamma_j = f(\theta_j)$
- look at distribution of the γ_j.

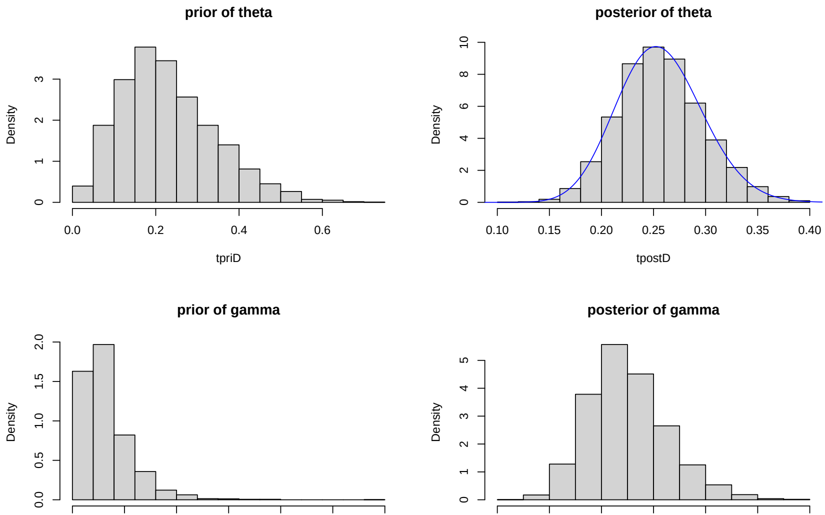
In our Bernoulli example suppose I want to know what the data tells me about the odd ratio.

$$\gamma = \frac{\theta}{1-\theta}$$

```
nd=5000
set.seed(99)
## prior of gamma
tpriD = rbeta(nd,apri,bpri) ## iid draws from the prior of theta = p
gpriD = tpriD/(1-tpriD) ## iid draws from the prior of gamma = odds ratio
## post of gamma
tpostD = rbeta(nd,apost,bpost) ## draws from the posterior of theta = p
gpostD = tpostD/(1-tpostD) ## iid draws from the posterior of gamma = odds ratio
```

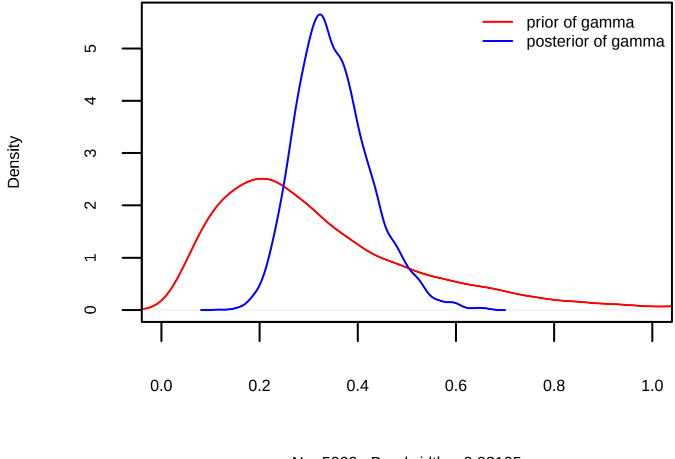
Plot prior and posterior of $\theta = p$ and γ.

```
par(mfrow=c(2,2))
hist(tpriD,freq=F,main='prior of theta')
hist(tpostD,freq=F,main='posterior of theta')
lines(tvec,dbeta(tvec,apost,bpost),type='l',col='blue')
hist(gpriD,freq=F,main='prior of gamma')
hist(gpostD,freq=F,main='posterior of gamma')
```



Let's plot the prior and posterior of γ using a density smooth of the draws.

```
dspri = density(gpriD)
dspost = density(gpostD)
plot(dspri,xlim=c(0,1),ylim=range(c(dspri$y,dspost$y)),main='',
     col='red',cex.lab=.5,cex.axis=.5,cex=.5)
lines(dspost,col='blue')
legend('topright',legend=c("prior of gamma","posterior of gamma"),
       col=c('red','blue'),lty=c(1,1),bty="n",cex=.5)
```



Prediction and Bayes

Suppose we have $p(\theta \mid y)$ and we want the predictive distribution of *future* y.

- draw θ_j , $j = 1, 2, \dots, J, \ \sim \theta|y$
- draw $y_j \sim f(y \mid \theta_j)$
- look at distribution of the y_j.

Higher Dimension Bayesian Modeling

The Bayesian approach lends itself to building higher dimensional model with a hierarchical approach.

That is, we build a complex model by building it up from model pieces just like a car.

For example, suppose we want to consider mixtures of normals

$$Y \mid p, \mu, \sigma \sim \sum_{j=1}^{J} p_j \, n(y \mid \mu_j, \sigma_j)$$

We can build a probability model which includes latent indicators I assigning a Y to one of the normal components.

$$p = (p_1, p_2, \cdots p_J), \quad \mu = (\mu_1, \mu_2, \cdots \mu_J)$$

$$\sigma = (\sigma_1, \sigma_2, \cdots \sigma_J)$$

$$p(I \mid p) \sim p(I = j) = p_j$$

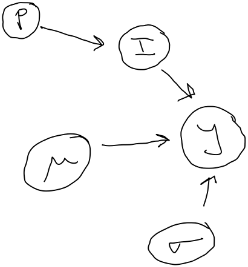
$$p(\mu) = \prod p(\mu_j)$$

$$p(\sigma) = \prod p(\sigma_j)$$

$$y = (y_1, y_2, \cdots y_n), \quad I = (I_1, I_2, \cdots I_n)$$

$$y_i \mid \mu, \sigma, I_i = j \sim N(y \mid \mu_j, \sigma_j)$$

We can understand the model with a DAG.



$$p(p, I, \mu, \sigma, y) = p(p)\, p(I \mid p)\, p(\mu)\, p(\sigma)\, p(y \mid \mu, \sigma, I)$$

This can be done quite effectively for very complex models.

Build a complex model up out of simple pieces !!!!!.

Draw From the Posterior Using Markov Chain Monte Carlo

We can define Markov chains in the conditional space whose stationary distribution is the posterior.

We then iterate the Markov chain and treat each iteration as a draw from the posterior much as we do with simple IID Monte Carlo

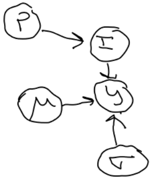
The *Gibbs Sampler* just iterates through conditionals.

For example in our mixture model we want

$$p(p, I, \mu, \sigma \mid y)$$

We can iterate through the conditionals:

$$p(p, I, \mu, \sigma \mid y)$$



$$p \mid I, \mu, \sigma, y = p \mid I$$

$$I \mid p, \mu, \sigma, y$$

$$\mu \mid p, I, \sigma, y = \mu \mid I, \sigma, y$$

$$\sigma \mid p, I, \mu, y = \sigma \mid I, \mu, y$$

Given current values of (p, I, μ, σ), we get the next set of values by drawing from the above conditionals sequentially.

Bayesian Nonparametrics

Bayesian non-parametrics is just Bayesian analysis with *lots* of parameters so you can get flexibility.

But, we can *let the dimension of the parameter space vary*.

For example DPM models fit the mixture model *with an unknown number of mixture components* !!!

BART works with an ensemble of trees where *each tree has an unknown size* !!!

Bayesian Foundations

Bayesian modeling is popular because:

- appeal of probability modeling.
- build up complex models out of simple pieces.

However, Bayesian thinking also as some interesting theoretical foundations:

- Complete class theorem: any admissable rule is a Bayes rule and vice versa.
- "coherent" decision making corresponds to Bayesian decision making.
- de Finetti's theorem states that exchangeable observations are conditionally independent relative to some latent variable.
- The likelihood principle: If two problems (model/data) give the same likelihood, they should give the same inference (e.g. p-values out!!)

Of course, Bayesian thinking also is intimately tied with the view the probability is fundamentally subjective
(obviously correct!!).

Bayesian thinking also avoids the absurdities associated with such frequentist concepts as confidence intervals and p-values.

Why isn't everyone a Bayesian ??

It can be hard to build a full probability model, even with the hierachical approach.

For example, choosing priors in high dimensions can be tricky.

With big data and big models MCMC may not be feasible.

A huge advantage of BART is that even though the model is high and random dimensional, the prior and MCMC are "easy".

Regularization Priors

Our fundamental Bayes equation is

$$p(\theta|y) \propto p(y,\theta) = p(\theta)\, p(y|\theta)$$

If we simply take the log we get

$$\log(p(\theta|y)) = \log(p(\theta)) \, + \, \log(p(y|\theta)) \, + \, C$$

The posterior mode will optimize the fit through the log likelihood but the prior will provide shrinkage and can be designed to “prefer” simpler models even thought θ is complicated enough to capture quite a complicated model.

This is how BART work.
We will call our prior a *regularization prior*.