Prediction: Motivations, Problems and Methods

Katherine Evans

Saberseminar, 2017

Prediction: Motivations, Problems and Methods

- What do I mean by "prediction?"
  - Any time a model is fit: $P(Event \mid Covariates)$ or $E[Outcome \mid Covariates]$
  - Looking at effect sizes
  - Could also call it "estimation"
- Infinite applications
  - Probability a given pitch is a strike
  - How effective is the infield shift
  - How will a draft prospect develop

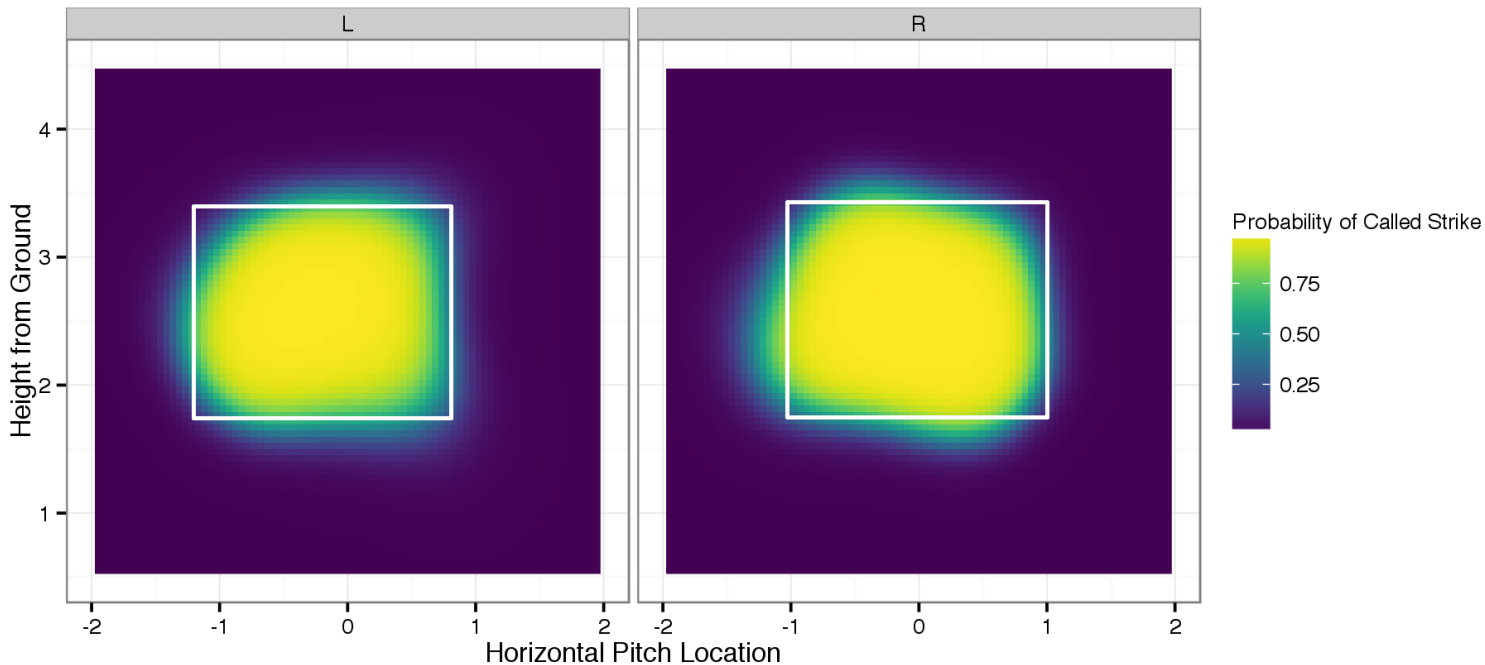
Strike Prediction

- What is the probability that a pitch is a called strike?
- Data:
  - PITCHf/x
  - I'm using Carson Sievert's pitchRx package in R and the **pitches** data frame

    Four-seam and cut fastballs thrown by Mariano Rivera and Phil Hughes during the 2011 season
- General method:
  - Look at x and z coordinates for pitches that were not swung at
  - Split out by other variables - stance, umpire
  - Don't really care about coefficients.

Strike Prediction

- The package happens to fit a Generalized Additive Model (GAMs)
- Probably don't have to get more complicated
- Could also include umpire - further stratify



Infield Shift

- What is the effect of the infield shift?
- Now we are thinking about a much more specific question - need to be careful with the definition because that will determine what we target with our estimation
- To me this is a very specific treatment - shift vs no shift (though I acknowledge there are degrees of shifting)
- Many potential outcomes, e.g.
  - Individual level: batting average
  - Team level: runs saved
- Average treatment effect vs effect of treatment on the treated?
  - Effect of shift on those who were shifted

Infield Shift

- Formula for the effect of the shift on those who were shifted

$$E\left[Avg|Shift\right] \quad - \quad \frac{1}{P(Shift)}E\left[(1-Shift)\frac{P(Shift|Covariates)}{P(No\ Shift|Covariates)}Avg\right]$$

- Target the shift rather than having it as one of many variables in a model
- Need to be smart in how we model $P(Shift|Covariates)$
  - ▷ What variables go into deciding whether or not to shift a player on a given at bat?
  - ▷ What model best describes the process? Likely more complicated than a simple logistic regression (more later)
- Expected values can be evaluated empirically

Prospect Success

- How will a draft prospect perform?
- Now we are getting complicated
- Still need to define a clear outcome - how to define success?
- No "treatment" variable
- Many, many variables that may predict future performance
  - Not necessarily interested in the predictors, just the prediction

Ensemble Learners

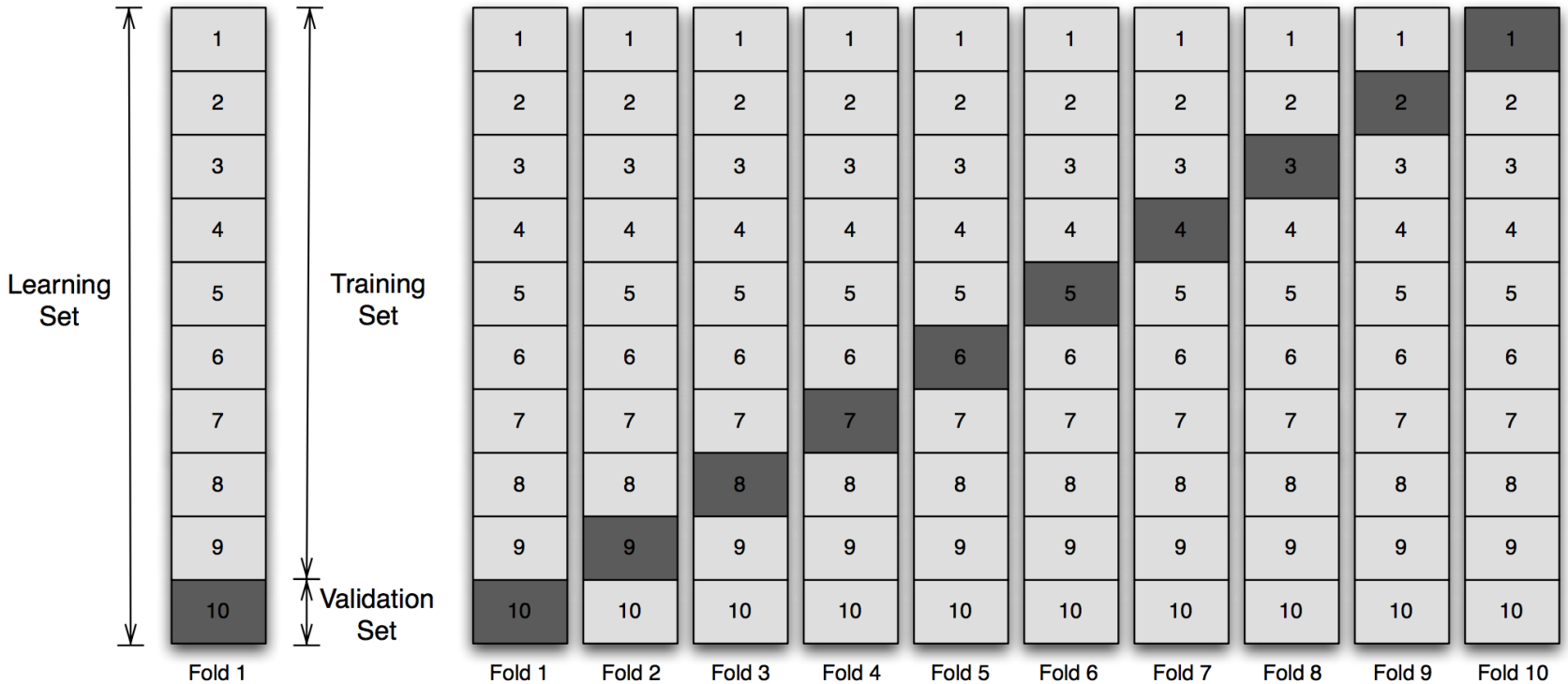
- "In statistics and machine learning, **ensemble methods** use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone." - Wikipedia
- Uses an ensemble/group of weak learners/methods (e.g. Random Forest, Lasso, KNN)
  - ▷ Any mapping from the data into a predictor.
- Stacked generalization to combine the predictions from the multiple models
- No model will ever be perfect or 100% true
  - ▷ Ensemble learners can give a good approximation of the true prediction function

Super Learner (van der Laan, Polley, Hubbard; 2007)

- The **Super Learner algorithm** is a loss-based supervised learning method that finds the optimal combination of an ensemble of prediction algorithms/models/methods
- Super Learner performs asymptotically as well as best possible weighted combination of the base learners.

Performance Evaluation: Cross Validation

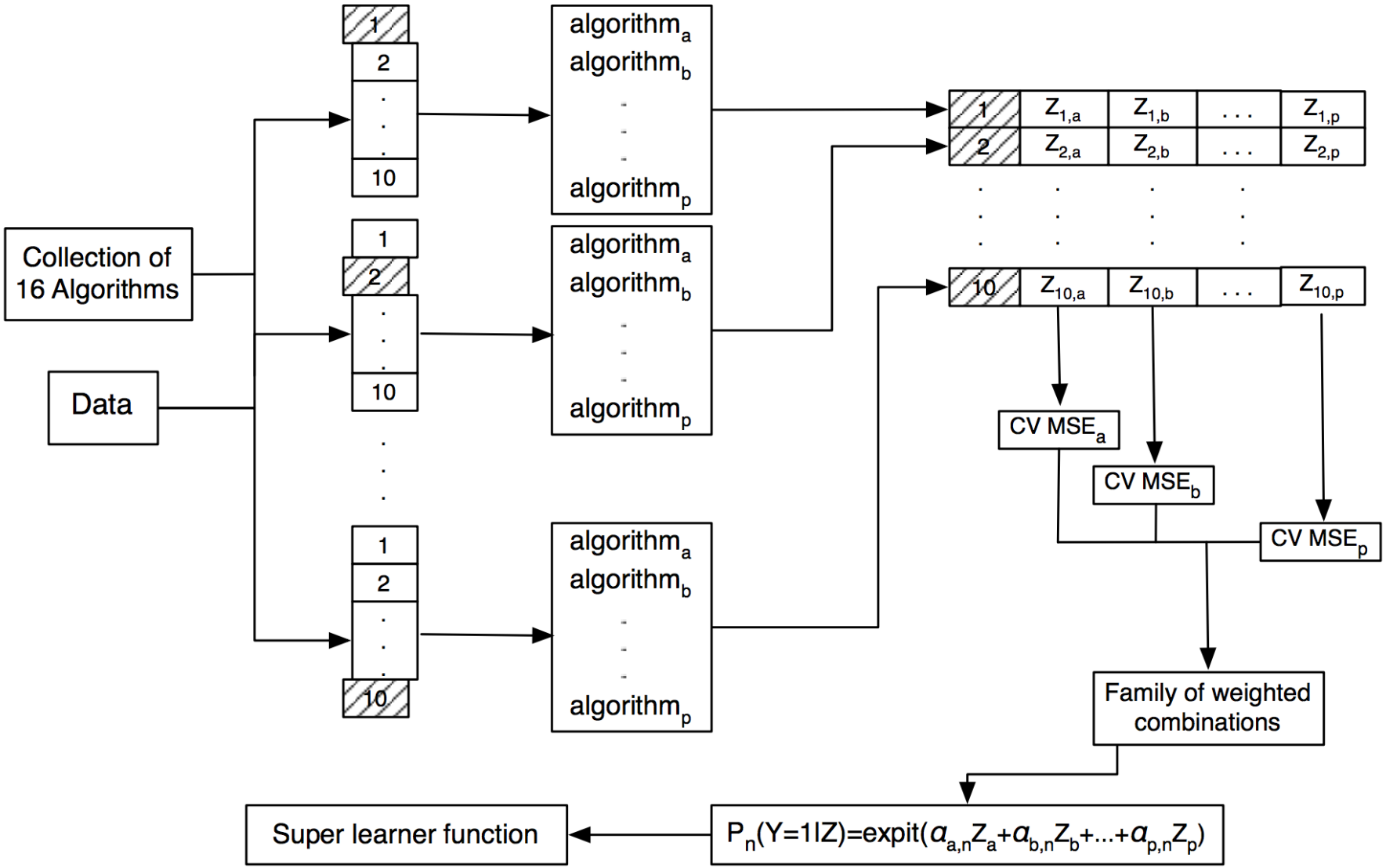
- Ensemble methods, such as Super Learner, allow us to use many methods
- We do not have to decide ahead of time which single technique to use
  - ▷ We can use several by incorporating cross validation.



Super Learner

- Build a library of algorithms consisting of all weighted averages of the pre specified algorithms.
- One of these weighted averages might perform better than one of the algorithms alone.
- It is this principle that allows us to map a collection of algorithms into a library of weighted averages of these algorithms.
- The effects of the individual variables are obscured since weights are assigned to algorithms, not variables.

Super Learner



Conclusion

- Prediction can be relatively straightforward - stratified (x, z) coordinates.
- It is important to clearly define the question as this will help determine the proper method.
- At times we don't care much about the inputs, just the final prediction.
  - ▷ Ensemble learners can improve prediction significantly.
- Cross validation is great - use it!

Thank You!

Questions?

Contact Information:

CausalKathy@gmail.com

Twitter: @CausalKathy

CausalKathy.com