Imputing the epigenome

http://sahirbhatnagar.com/talks/

March 12, 2015

(Ernst and Kellis 2015)

Main Idea

Matrix of Observed and Imputed Data

1. Leverage other marks in same sample

2. Leverage same mark in different sample

Types of data used to impute

Advantages of Imputation

- Beneficial even if observed data is available

Advantages of Imputation

- Beneficial even if observed data is available
- Combining information –> robust to experimental noise, confounders

Advantages of Imputation

- Beneficial even if observed data is available
- Combining information –> robust to experimental noise, confounders
- Achieve a higher sequencing depth –> higher signal to noise ratio

Advantages of Imputation

- Beneficial even if observed data is available
- Combining information –> robust to experimental noise, confounders
- Achieve a higher sequencing depth –> higher signal to noise ratio
- Improve GWAS enrichments –> epigenomic maps as an unbiased approach for discovering disease-relevant tissues and cell types

Advantages of Imputation

- Beneficial even if observed data is available
- Combining information –> robust to experimental noise, confounders
- Achieve a higher sequencing depth –> higher signal to noise ratio
- Improve GWAS enrichments –> epigenomic maps as an unbiased approach for discovering disease-relevant tissues and cell types
- Quality Control –> Are there discrepancies between imputed and observed datasets

Advantages of Imputation

- Beneficial even if observed data is available
- Combining information –> robust to experimental noise, confounders
- Achieve a higher sequencing depth –> higher signal to noise ratio
- Improve GWAS enrichments –> epigenomic maps as an unbiased approach for discovering disease-relevant tissues and cell types
- Quality Control –> Are there discrepancies between imputed and observed datasets
- Feature importance

Advantages of Imputation

- Beneficial even if observed data is available
- Combining information –> robust to experimental noise, confounders
- Achieve a higher sequencing depth –> higher signal to noise ratio
- Improve GWAS enrichments –> epigenomic maps as an unbiased approach for discovering disease-relevant tissues and cell types
- Quality Control –> Are there discrepancies between imputed and observed datasets
- Feature importance
- Chromatin state annotation

Limitations

- If the presence of mark signal is highly specific to one or a few samples, and it does not correlate with other marks mapped in the sample or has a different correlation structure than in samples used for training, then it would not be possible to accurately impute the mark at those locations

Limitations

- If the presence of mark signal is highly specific to one or a few samples, and it does not correlate with other marks mapped in the sample or has a different correlation structure than in samples used for training, then it would not be possible to accurately impute the mark at those locations
- When the target mark has been mapped in only a few samples, the features pertaining to the same mark in other samples may be less informative or more biased e.g. TFBS

Limitations

- If the presence of mark signal is highly specific to one or a few samples, and it does not correlate with other marks mapped in the sample or has a different correlation structure than in samples used for training, then it would not be possible to accurately impute the mark at those locations
- When the target mark has been mapped in only a few samples, the features pertaining to the same mark in other samples may be less informative or more biased e.g. TFBS
- For tissue samples that reflect mixtures of multiple cell types, our imputed maps will most likely reflect the same mixture as the observed data, though deconvolution of mixed samples is a potentially important direction for future work

ChromImpute Software

- Command line tool written in `JAVA`
- http://www.biolchem.ucla.edu/labs/ernst/ChromImpute/

Not a new idea

Leo Breiman (1928-2005)

(Breiman 2001)

`randomForest` package in R

MissForest

(Stekhoven and Bühlmann 2012)

`missForest` package in `R`

Introduction to Regression Trees

Some intuition behind the imputation approach

```
total sales = 7.1 + 0.0475 x # of TV's sold
```

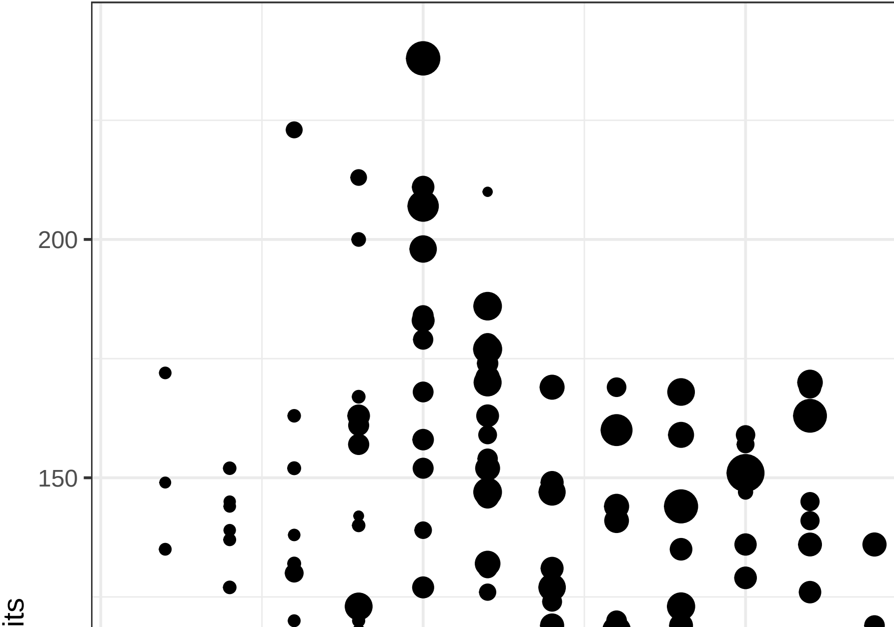
Tree-based Methods

- Involves *splitting* the predictor space into simple regions
- Since the set of splitting rules used to segment the predictor space can be summarized in a tree, these types of approaches are known as decision-tree methods (James et al. 2013)

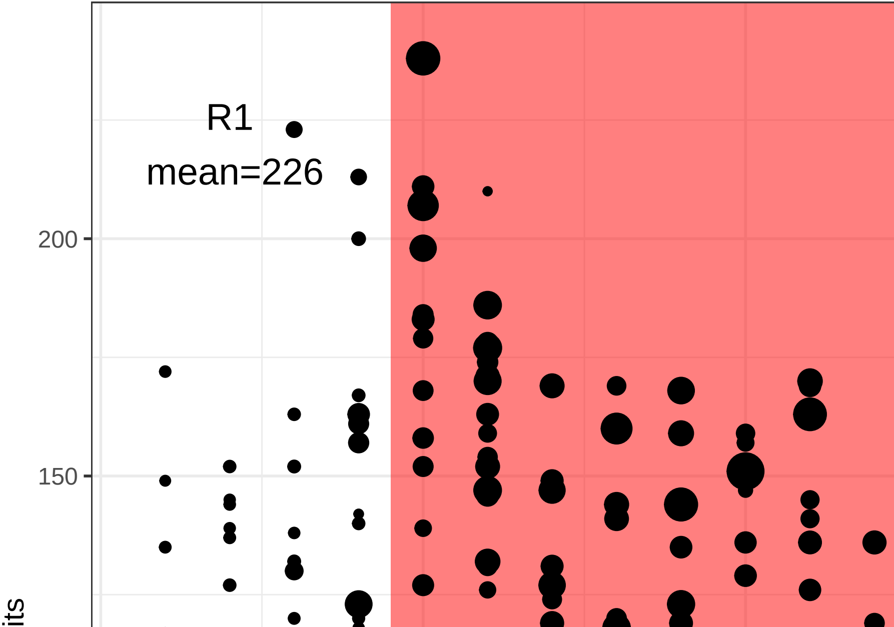
Baseball Data

```
## PhantomJS not found. You can install it with webshot::ir
```

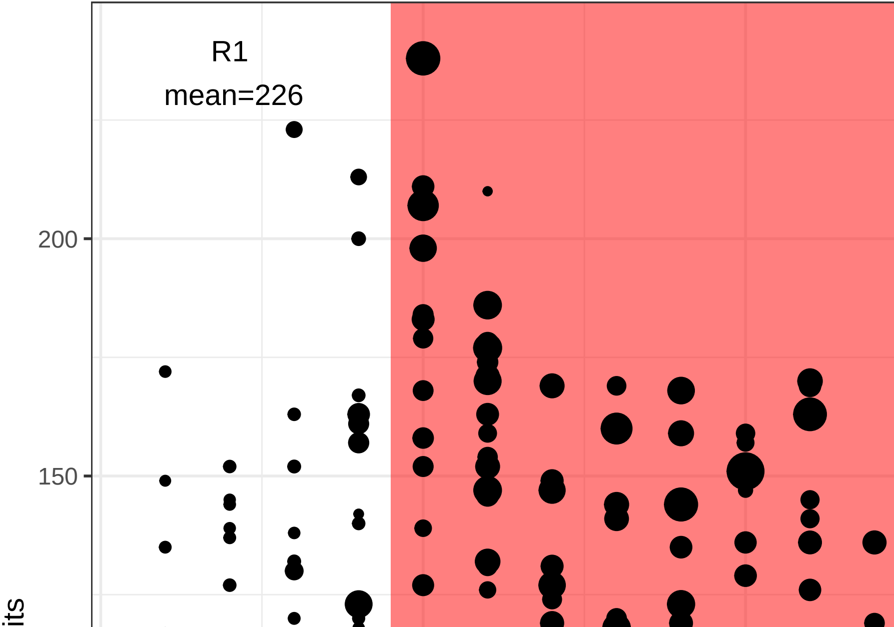
Predict salary based on Hits and Years Played



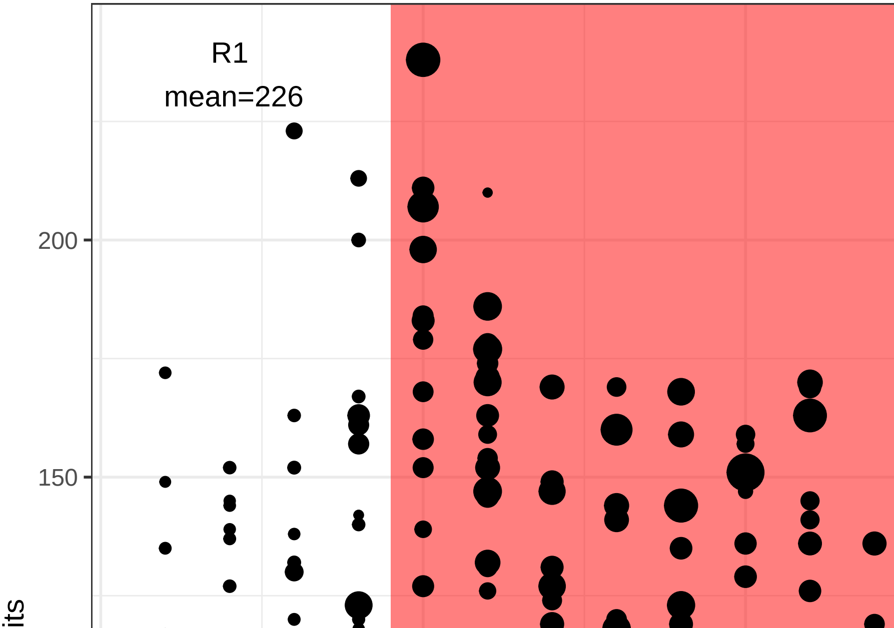
How to split the data



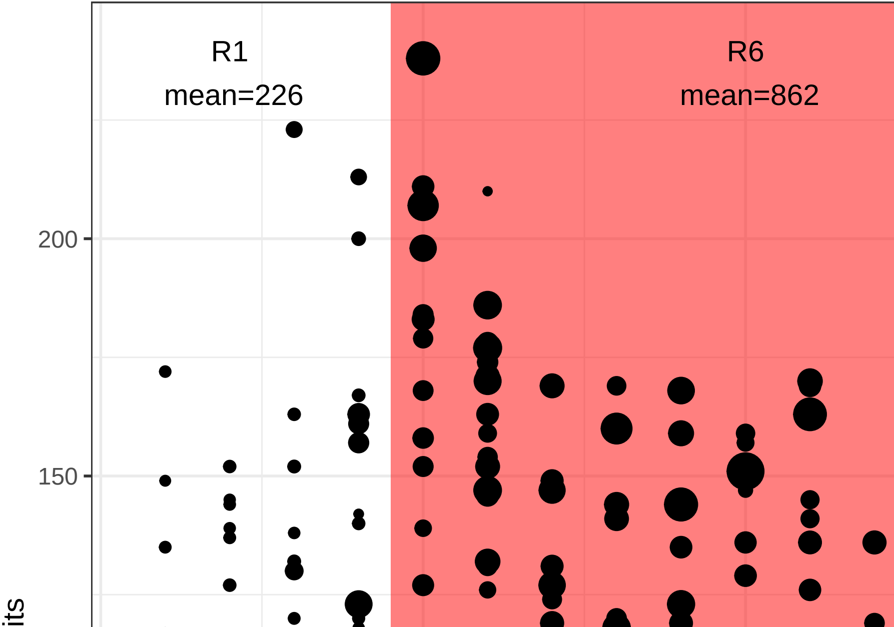
How to split the data



How to split the data

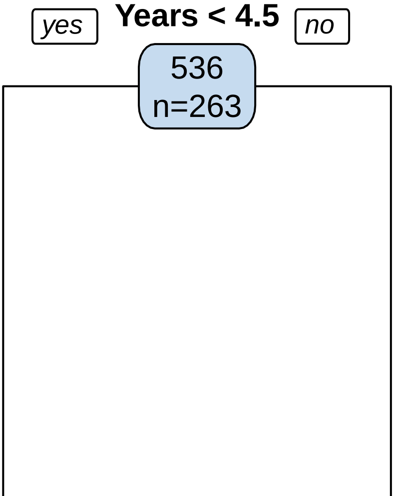


How to split the data



Regression Tree for Baseball data

```
## Warning: Bad 'data' field in model 'call' (expected a da
## To silence this warning:
##     Call rpart.plot with roundint=FALSE,
##     or rebuild the rpart model with model=TRUE.
```



Decision Tree

More Details of Tree Building

- The goal is to find boxes $R_1, \ldots, R_J$ that minimize the residual sum of squares give by

More Details of Tree Building

- The goal is to find boxes $R_1, \ldots, R_J$ that minimize the residual sum of squares give by
- $$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})$$

More Details of Tree Building

- The goal is to find boxes $R_1, \ldots, R_J$ that minimize the residual sum of squares give by
- $$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})$$
- y_i is the subjects response, $\hat{y}_{Rj}$ is the mean in box j

More Details of Tree Building

- The goal is to find boxes $R_1, \ldots, R_J$ that minimize the residual sum of squares give by
- $$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})$$
- y_i is the subjects response, $\hat{y}_{Rj}$ is the mean in box j
- Computationally infeasible to consider every single partition of the feature space into J boxes

More Details of Tree Building

- The goal is to find boxes $R_1, \ldots, R_J$ that minimize the residual sum of squares give by
- $$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})$$
- y_i is the subjects response, $\hat{y}_{Rj}$ is the mean in box j
- Computationally infeasible to consider every single partition of the feature space into J boxes
- Solution: take a top-down, greedy approach

More Details of Tree Building

- The goal is to find boxes $R_1, \ldots, R_J$ that minimize the residual sum of squares give by
-
$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})$$
- y_i is the subjects response, $\hat{y}_{Rj}$ is the mean in box j
- Computationally infeasible to consider every single partition of the feature space into J boxes
- Solution: take a top-down, greedy approach
- Begins at the top, and never looks back

Pros and Cons

- Tree-based methods are simple and useful for interpretation

Pros and Cons

- Tree-based methods are simple and useful for interpretation
- Highly sensity to the first split

Pros and Cons

- Tree-based methods are simple and useful for interpretation
- Highly sensity to the first split
- Solution: Combining a large number of trees can often result in dramatic improvements in prediction accuracy, at the expense of some loss interpretation.

Bagging

The Bootstrap

(James et al. 2013)

Pull yourself up by your bootstraps

Random Forests

ETH Zurich Slides

Acknowledgements

Regression tree slides are based on

Free PDF book

Leo Breiman (1928-2005)

References

```
## No encoding supplied: defaulting to UTF-8.
```


Breiman, Leo. 2001. "Random Forests." *Mach. Learn.* 45 (1). Hingham, MA, USA: Kluwer Academic Publishers: 5–32. https://doi.org/10.1023/A:1010933404324.

Ernst, Jason, and Manolis Kellis. 2015. "Large-Scale Imputation of Epigenomic Datasets for Systematic Annotation of Diverse Human Tissues." *Nature Biotechnology*. Nature Publishing Group.

James, G, D Witten, T Hastie, and R Tibshirani. 2013. *An Introduction to Statistical Learning*.

Stekhoven, Daniel J, and Peter Bühlmann. 2012. "MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data." *Bioinformatics* 28 (1). Oxford Univ Press: 112–18.