

Kinds of Ensemble

Tested on apple quality dataset



David Nardi

June 11, 2024

MSc in AI, University of Florence

Apple quality dataset

Variables

- Size
- Weight
- Sweetness
- Crunchiness
- Juiciness
- Ripeness
- Acidity

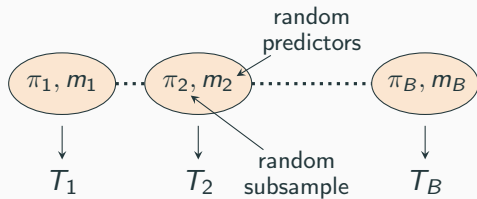
Binary classification task

Class distribution: 0.49 – 0.51

Methods

- *k*NN, Decision tree, Logistic regression
- Random forest
- AdaBoost
- Super learner

Random forest



```
randomForest(x, y,  
             importance=TRUE,  
             ntree=500,  
             ntree=B.oob ←  $B^*$ )
```

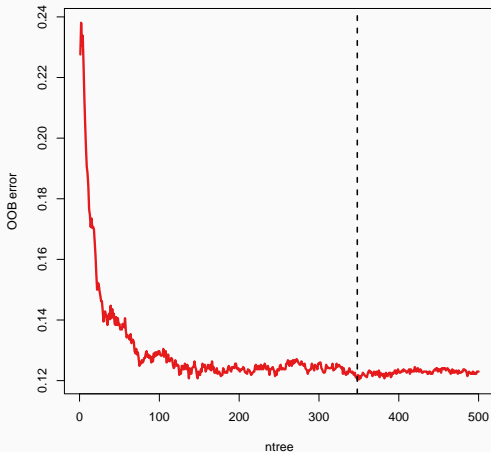
Loss-based variable
importance

$$G(x) = \arg \max_k \sum_{b=1}^B \mathbb{I}(T_b(x) = k)$$

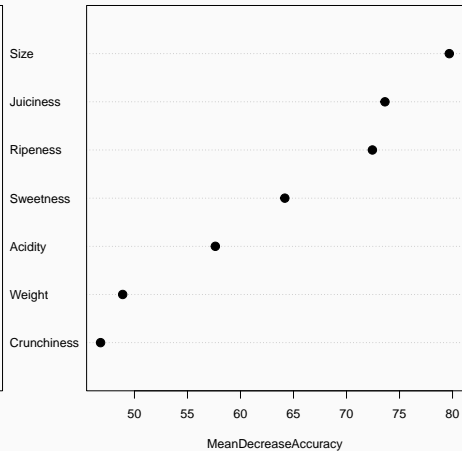
$$mtry = \lfloor \sqrt{p} \rfloor$$

Random forest tuning

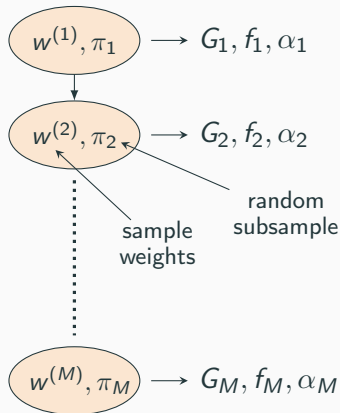
Bias-Variance tradeoff



Variable importance



AdaBoost algorithm



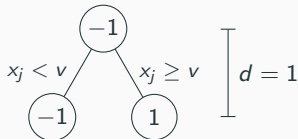
$$f_m(x) = f_{m-1}(x) + \lambda \alpha_m G_m(x)$$

$$G(x) = \text{sign}(f_M(x))$$

$$L(y, f(x)) = \exp(-yf(x))$$

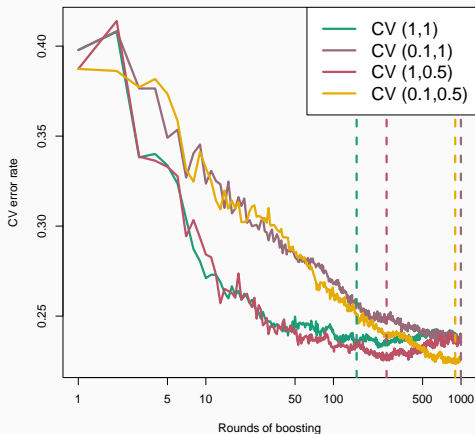
Encoding $\mathcal{Y} \in \{-1, 1\}$

```
ada :: ada(x, y,  
  loss="exponential",  
  type="discrete",  
  iter <- M*, nu <- λ*,  
  bag.frac <- π*,  
  control=base.learner)
```

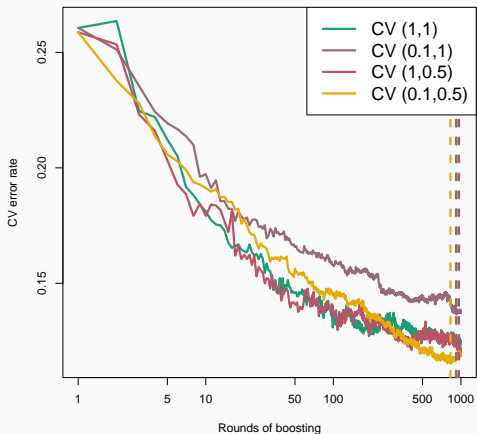


AdaBoost tuning

Bias-Variance tradeoff $d=1$

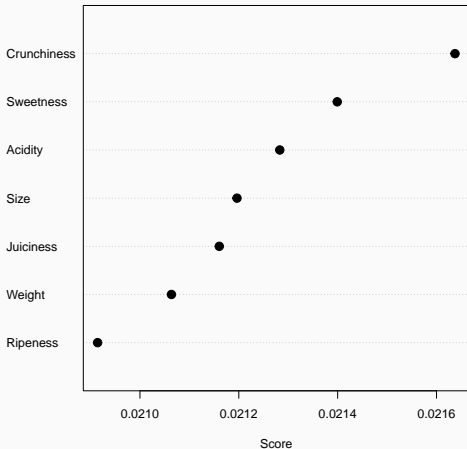


Bias-Variance tradeoff $d=4$

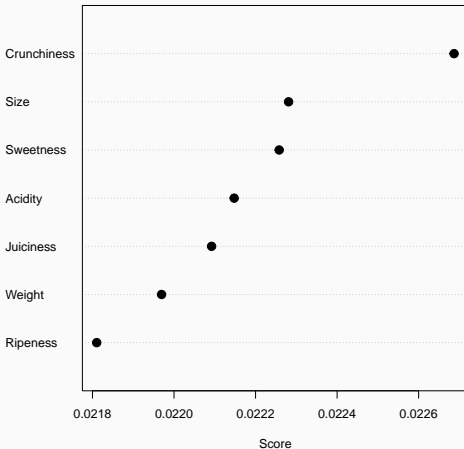


AdaBoost variable importance

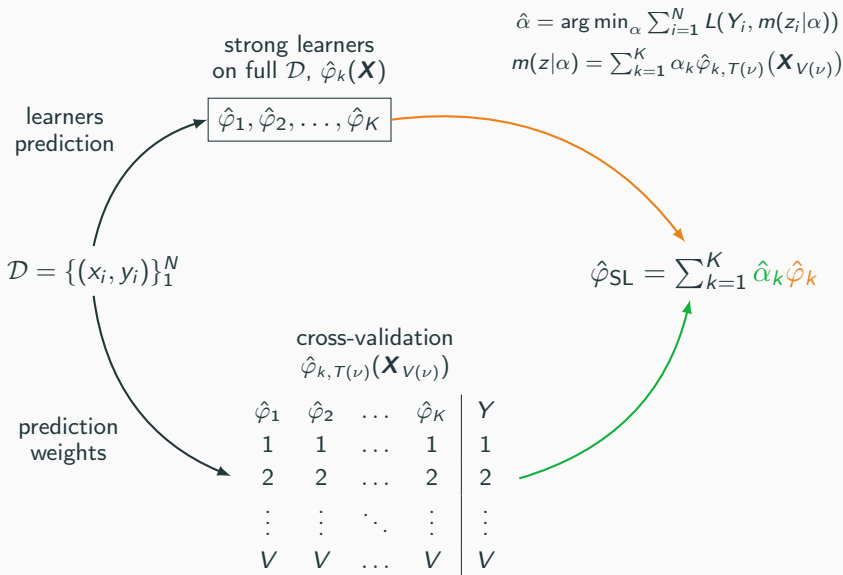
Average variable importance d=1



Average variable importance d=4



Super Learner flow diagram



Super Learner in practice

```
SuperLearner(Y, X,  
  cluster,  
  SL.library ←  $\{\varphi_k\}$ ,  
  cvControl=list(  
    V=10, shuffle=FALSE))
```

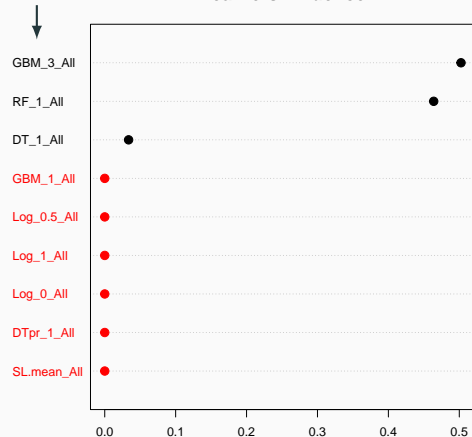
What's in the ensemble?

- Response variable mean \bar{y}
- Logistic Regression with $\alpha = 0, 1$ and 0.5
- Grown and pruned Decision Tree
- Random Forest
- Gradient Boosting Machine with $d = 1$ and $d = 4$

Super Learner library CV error

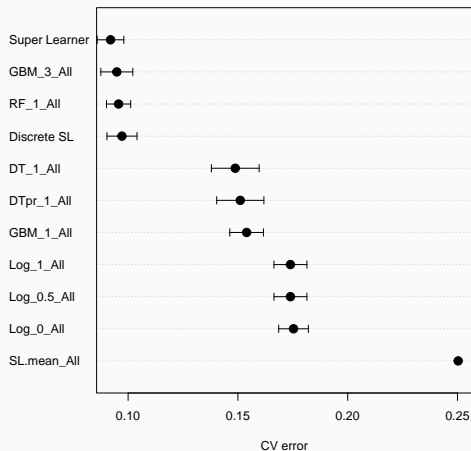
$$\{\hat{\varphi}_k\}$$

Learners influence



$\hat{\alpha}_k$ → Coefficient

Learners 5-fold CV error



Model	Train score	Test score
k NN	0.9029	0.8950
CART	0.8841	0.8290
Random forest	0.8808	0.8890
AdaBoost _{$d=1$}	0.7983	0.7922
AdaBoost _{$d=4$}	0.9996	0.8845
Super learner	0.9861	0.8777



T. Hastie, R. Tibshirani, and J. H. Friedman

The Elements of Statistical Learning

Springer, 2009.



E. C. Polley, and M. J. van der Laan

Super Learner in Prediction

U.C. Berkeley Division of Biostatistics Working Paper Series.

Working Paper 266, 2010

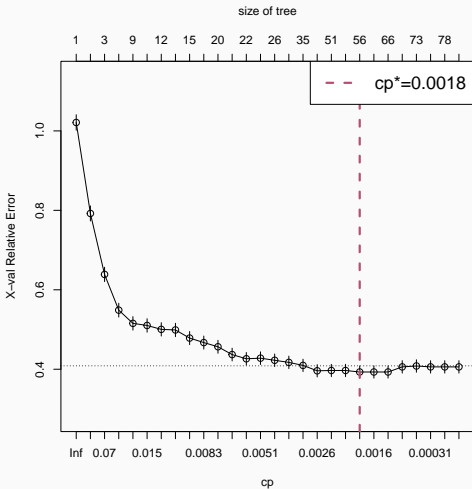
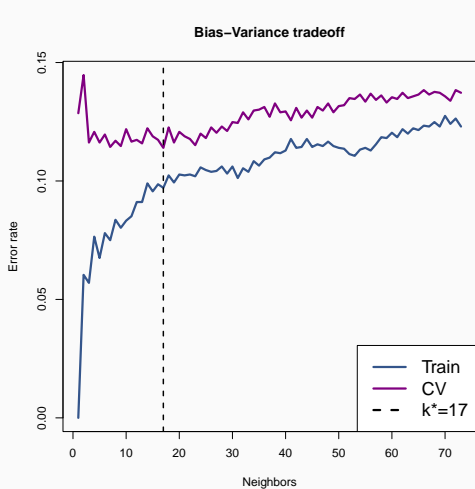


M. Culp, K. Johnson and G. Michailidis

ada: The R Package Ada for Stochastic Boosting

Journal of Statistical Software, 17(2), 1–27, 2006

kNN and CART tuning



Discrete AdaBoost algorithm

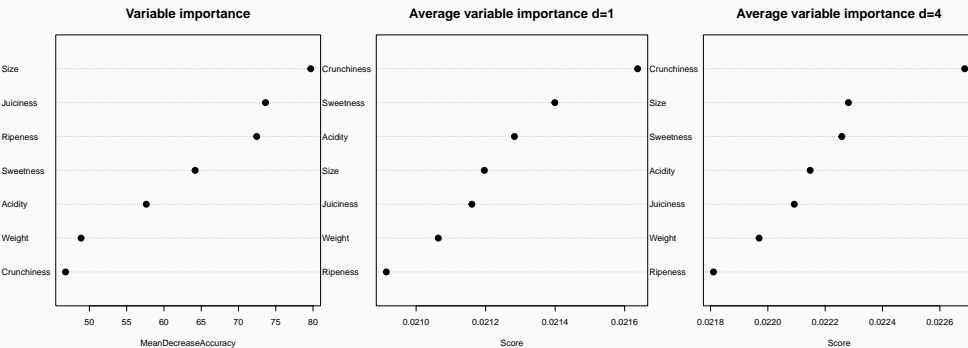
Discrete AdaBoost with shrinkage and out-of-bag, as an additive model with prediction function $f_m(x)$

Input: $M, \{(x_i, y_i)\}_1^N, x_i \in \mathbb{R}^p$

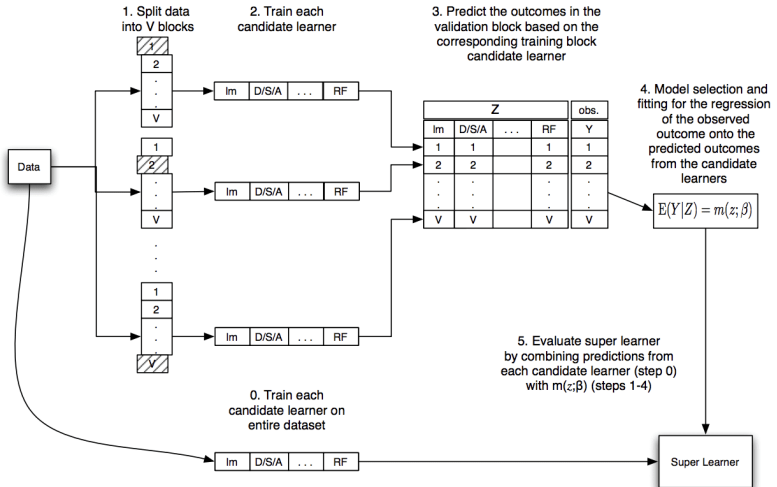
- 1 Initialize $f_0(x) = 0$;
- 2 **for** $m = 1$ **to** M **do**
- 3 Set $w_i^{(m)} = -\frac{\partial L(y, g)}{\partial g} \Big|_{g=f_m(x)}$ s.t. $\sum_{i=1}^N w_i^{(m)} = 1$;
- 4 Fit classifier $G_m(x)$ using $w_i^{(m)}$ with samples from π_m ;
- 5 Weighted error rate $\text{err}_m = \sum_{i=1}^N w_i^{(m)} \mathbb{I}(y_i \neq G(x_i))$;
- 6 Set $\alpha_m = \frac{1}{2} \log\left(\frac{1-\text{err}_m}{\text{err}_m}\right)$;
- 7 Update $f_m(x) \leftarrow f_{m-1}(x) + \lambda \alpha_m G_m(x)$;
- 8 **end**

Output: $G(x) = \text{sign}(f_M(x))$

Variable importance comparison



Super Learner algorithm flow diagram



Input: $\mathcal{D} = \{(x_i, y_i)\}_1^N$, $\mathcal{L} = \{\varphi_k(X)\}_{k=1}^K$

1 **foreach** *strong learner* in \mathcal{L} **do**

2 Fit φ_k on $\mathcal{D} \Rightarrow \hat{\varphi}_k(\mathbf{X}) \rightarrow \hat{\mathcal{L}} = \{\hat{\varphi}_k\}_{k=1}^K$;

3 **end**

4 **for** $\nu = 1, 2, \dots, V$ **do**

5 **foreach** *strong learner* in \mathcal{L} **do**

6 Fit φ_k on $T(\nu)$, predict $\hat{\varphi}_{k, T(\nu)}(X_i \in V(\nu))$;

7 **end**

8 **end**

9 Stack output in an $N \times K$ matrix $Z = \{\hat{\varphi}_{k, T(\nu)}(X_{V(\nu)})\}$;

10 Propose a family of weighted combinations

$$m(z|\alpha) = \sum_{k=1}^K \alpha_k \hat{\varphi}_{k, T(\nu)}(\mathbf{X}_{V(\nu)}) \rightarrow \hat{\alpha} = \arg \min_{\alpha} \sum_{i=1}^N L(Y_i, m(z_i|\alpha))$$

of size N s.t. $\alpha_k \geq 0$, $\sum_k \alpha_k = 1$ and minimizes $\sum_k \alpha_k \hat{\varphi}_k$;

11 Combine $\hat{\alpha}$ with the library $\hat{\mathcal{L}} \rightarrow \hat{\varphi}_{\text{SL}}(\mathbf{X}) = \sum_{k=1}^K \hat{\alpha}_k \hat{\varphi}_k(\mathbf{X})$;

Output: $\hat{\varphi}_{\text{SL}}$
