Miscellaneous and recap

Aaditya Ramdas

Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

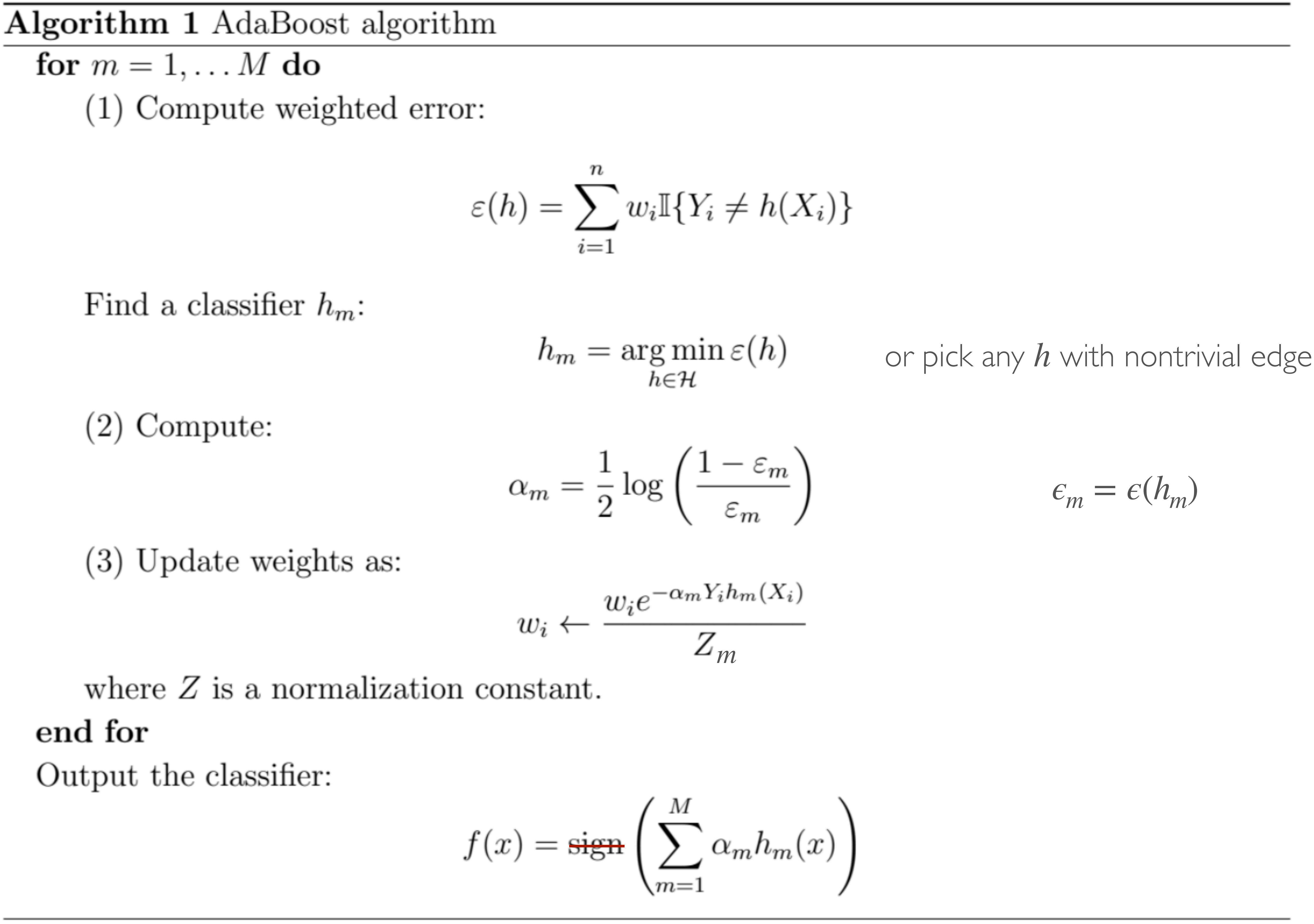
Outline

1. *Coordinate descent view of adaboost (1/4 class)*

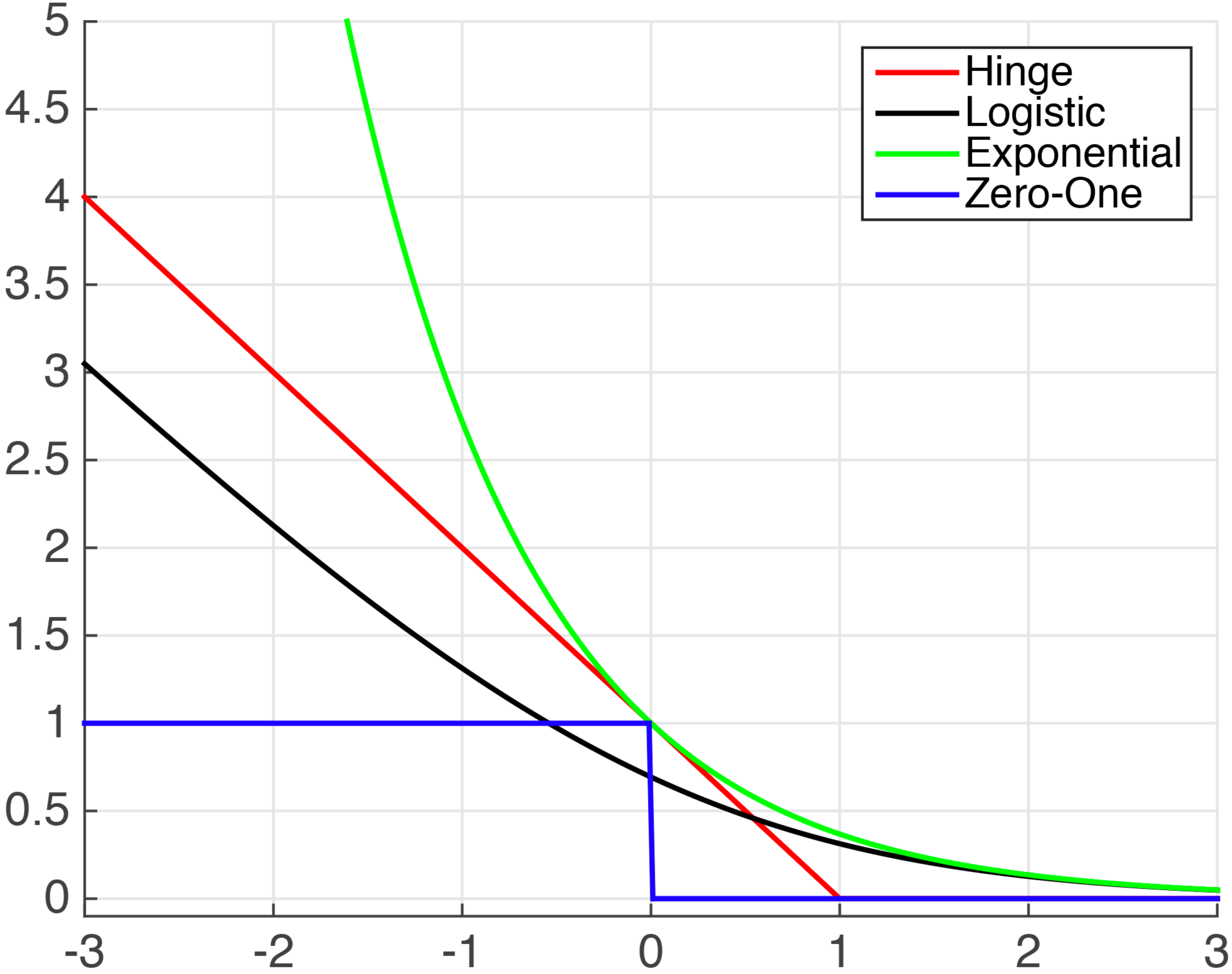
2. *Variable importance measures for trees: OOB (1/4 class)*

3. *Calibration reliability diagrams (1/4 class)*

4. *Recap and looking ahead to second half*



```
Algorithm 1 AdaBoost algorithm
for $m = 1, \ldots M$ do
    (1) Compute weighted error:
        $\varepsilon(h) = \sum_{i=1}^{n} w_i \mathbb{I}\{Y_i \neq h(X_i)\}$
    Find a classifier $h_m$:
        $h_m = \arg\min_{h \in \mathcal{H}} \varepsilon(h)$        or pick any $h$ with nontrivial edge
    (2) Compute:
        $\alpha_m = \frac{1}{2} \log \left( \frac{1 - \varepsilon_m}{\varepsilon_m} \right)$        $\epsilon_m = \epsilon(h_m)$
    (3) Update weights as:
        $w_i \leftarrow \frac{w_i e^{-\alpha_m Y_i h_m(X_i)}}{Z_m}$
    where $Z$ is a normalization constant.
end for
Output the classifier:
    $f(x) = \cancel{\text{sign}} \left( \sum_{m=1}^{M} \alpha_m h_m(x) \right)$
```


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
-3
-2
-1
0
1
2
3
Hinge
Logistic
Exponential
Zero-One

Convex surrogate loss minimization by coordinate descent

$$\hat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i \sum_{j=1}^{J} \beta_j h_j(x_i)\right)$$

Adaboost solves $\min_{\beta \in \mathbb{R}_+^J} \hat{R}(\beta) = \min_{f \in \text{span}(\mathscr{H})} \hat{R}(f)$ by "coordinate descent".

1. Begin at $\beta^{(0)} = [0,0,\ldots,0]$
2. At step t, pick direction $e_t \in \{e_j\}_{j \in J}$ and stepsize $\alpha_t \geq 0$ to minimize $\hat{R}(\beta^{(t-1)} + \alpha_t e_t)$
3. Gradient with respect to coordinate j is $\hat{R}'(\beta^{t-1})_j \propto (2\epsilon_{t,j} - 1)\prod_{s=1}^{t-1} Z_s$, where $\epsilon_{t,j}$ is weighted error of h_j
4. h_t is chosen to minimize the weighted error, optimal stepsize happens to equal $\log(\frac{1-\epsilon_t}{\epsilon_t})$

Outline

1. *Coordinate descent view of adaboost (1/4 class)*
2. *LOO and OOB for bagging or RFs (1/4 class)*
3. *Calibration reliability diagrams (1/4 class)*
4. *Recap and looking ahead to second half*

LOOCV via OOB for bagging and forests

Usually, for each X_i , we calculate $Y_i - f^{-i}(X_i)$, where f^{-i} is the model trained on all except the i-th point.
(N model fits)

For forests, we can calculate $Y_i - f^{OOB_i}(X_i)$, where f^{OOB_i} is the prediction of all the trees that do not use the i-th point.
(Single model fit)

OOB as a variable importance measure for RFs

*Previous slide left out *points**

Instead, suppose we wish to calculate the "feature importance" of feature j.

We could retrain the whole procedure without feature j, and see how the out-of-sample (holdout) error changes.

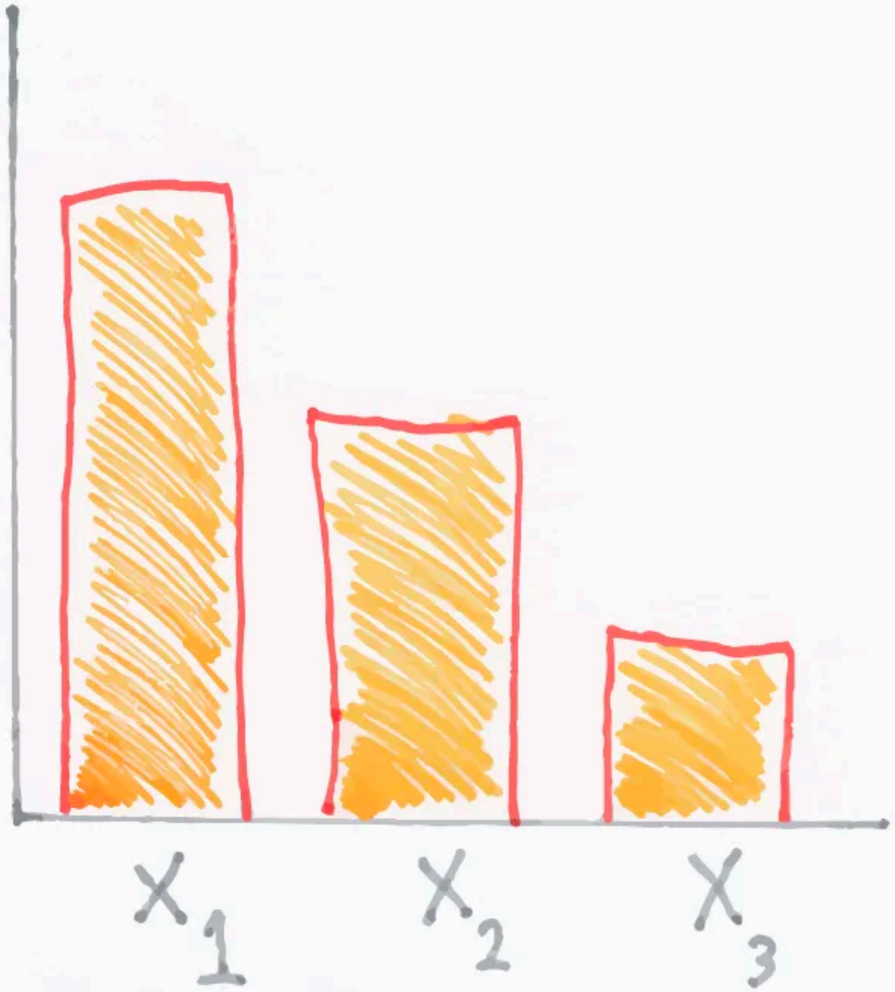
Eg: $\frac{1}{n'}\sum_{i=1}^{n'} (Y_i - f(X_i))^2 - (Y_i - f_{-j}(X_i))^2$ *on* n' *held out points.*

Instead of retraining from scratch,
one can simply consider those (OOB) trees which did not use feature j.

FEATURE IMPORTANCE

Decision trees make splits that maximize the decrease in impurity.

By calculating the mean decrease in impurity for each feature across all trees we can know that feature's importance.



ChrisAlbon

Outline

1. *Coordinate descent view of adaboost (1/4 class)*
2. *Variable importance measures for trees: OOB (1/4 class)*
3. *Calibration reliability diagrams (1/4 class)*
4. *Recap and looking ahead to second half*

Distribution shift

Distribution of training data P_{XY}
is not the same as distribution of test data.

Covariate shift (special type of distribution shift)

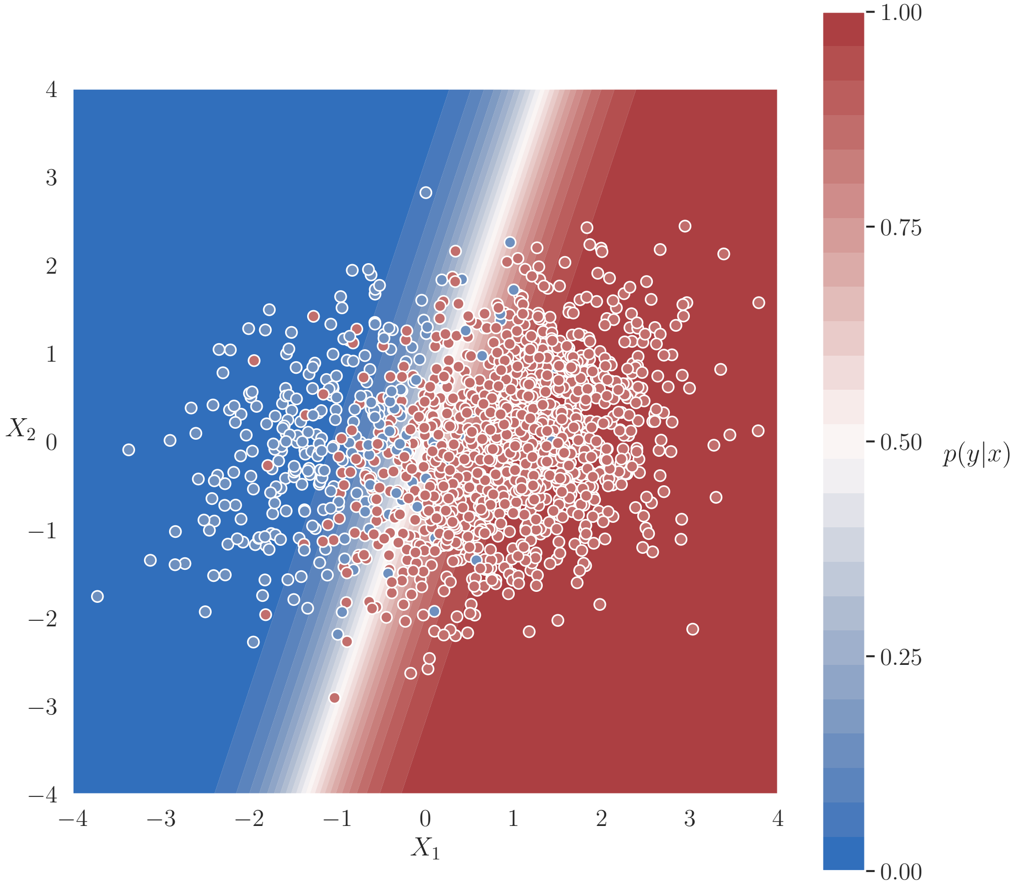
P_X *changes,* $P_{Y|X}$ *remains same*

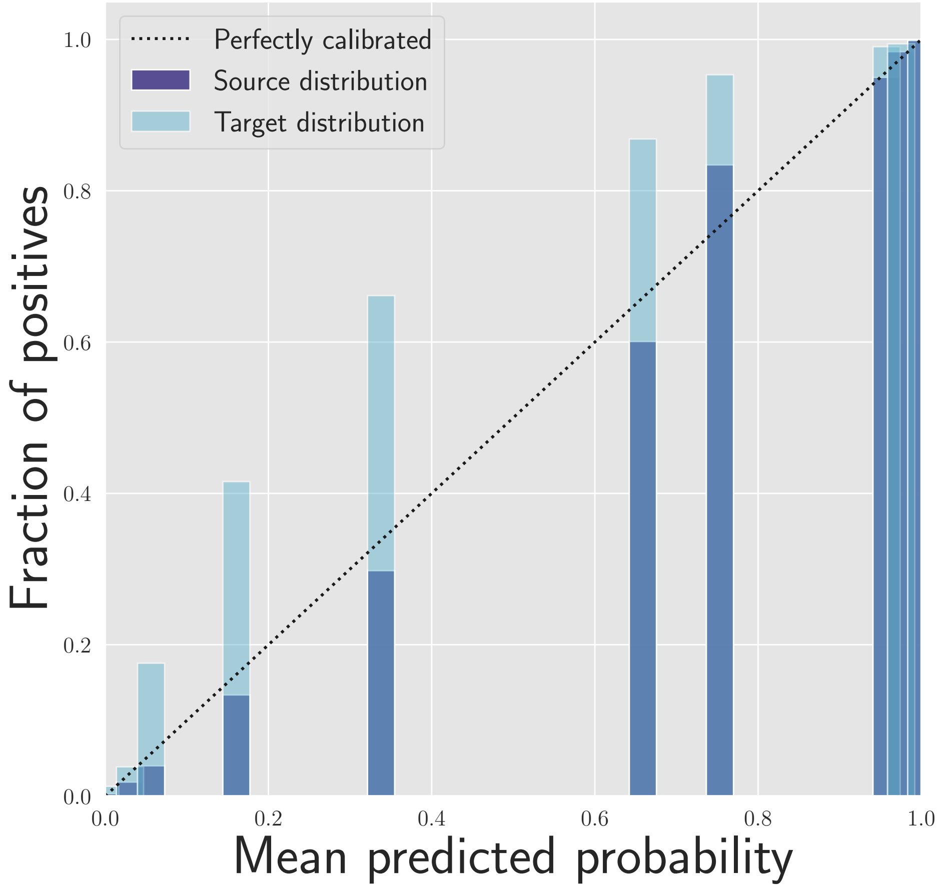
Label shift (special type of distribution shift)

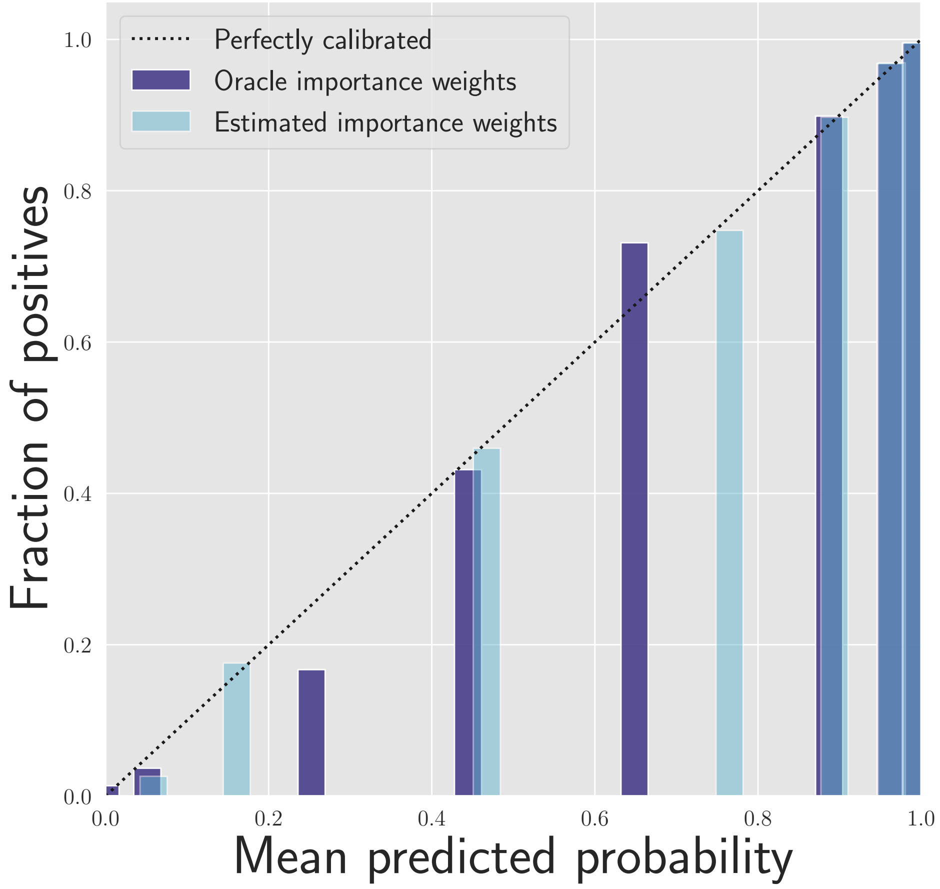
P_Y *changes,* $P_{X|Y}$ *remains same*

“Label shift”

P_Y changes, $P_{X|Y}$ remains same (for the test distribution)







Outline

1. *Coordinate descent view of adaboost (1/4 class)*
2. *Variable importance measures for trees: OOB (1/4 class)*
3. *Calibration reliability diagrams (1/4 class)*
4. *Recap and looking ahead to second half (1/4 class)*

	K-nn	CART	Bagging	RF	Boosting	Stacking
Algorithmic idea						
Bias/variance						
Calibration/Conformal						
Data Analysis						
Explainability						