Lecture 9: Ensemble Learning

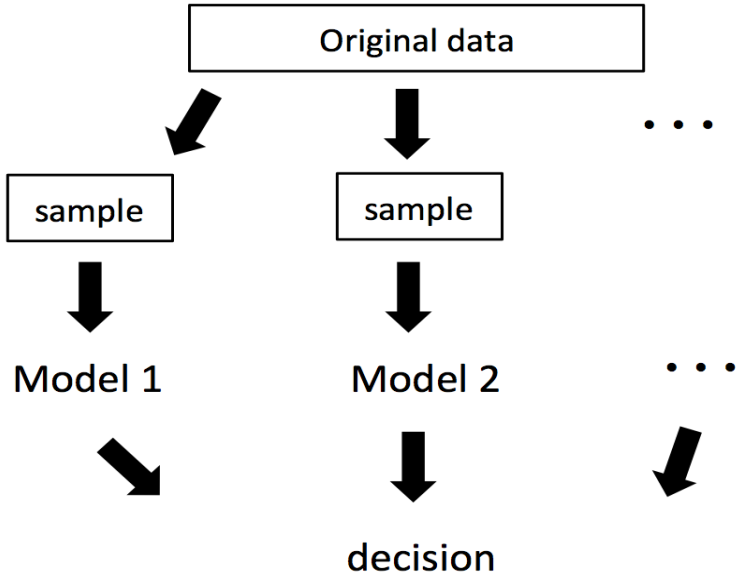
Instructor: Prof. Shuai Huang

Industrial and Systems Engineering

University of Washington

Random forest is one example

- Ensemble learning is very flexible. It is a hodgepodge of models, combined in different ways, trained in different ways …
- It is the cocktail treatment in machine learning to handle complex datasets



The AdaBoost algorithm

AdaBoost algorithm for two-class classification problems

Given N data points (x_1,y_1), (x_2,y_2), …, (x_N,y_N), build T trees

Initialize equal weights for all data points $w_1 = (\frac{1}{N}, \dots, \frac{1}{N})$

For t in 1 to T

Build base learner h_t on the data points with weights w_t

Calculate weighted error $\epsilon_t = \sum_{n=1}^{N} w_{t,n}\{h_t(x_n) \neq y_n\}$

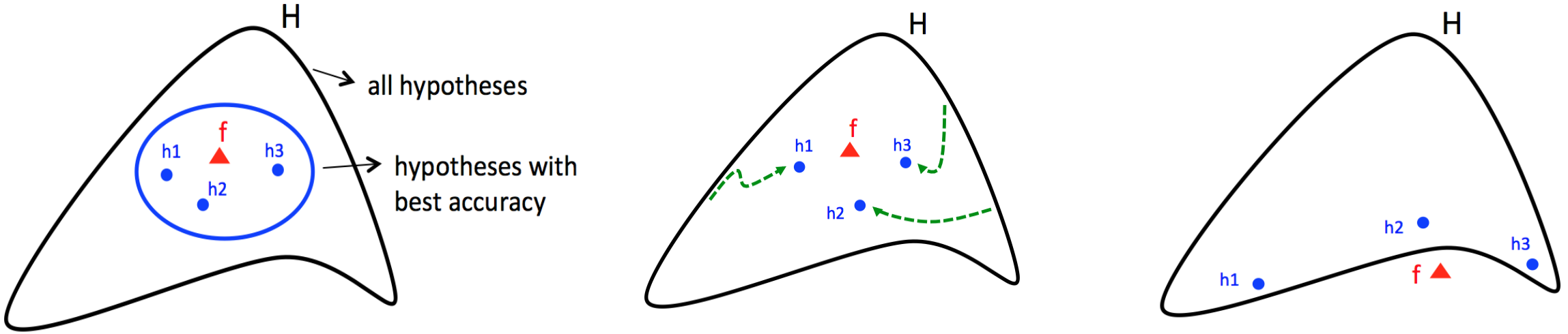
Calculate coefficient for h_t: $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$

Updated weights: $w_{t+1,i} = \frac{w_{t,i}}{Z_t} * \begin{cases} e^{-\alpha_t} \; if \; h_t(x_n) = y_n \\ e^{\alpha_t} \; if \; h_t(x_n) \neq y_n \end{cases}$

where Z_t is a normalization factor so that $\sum_{n=1}^{N} w_{t+1,n} = 1$

Calculate final decision: $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$

Why ensemble learning makes sense?



R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets