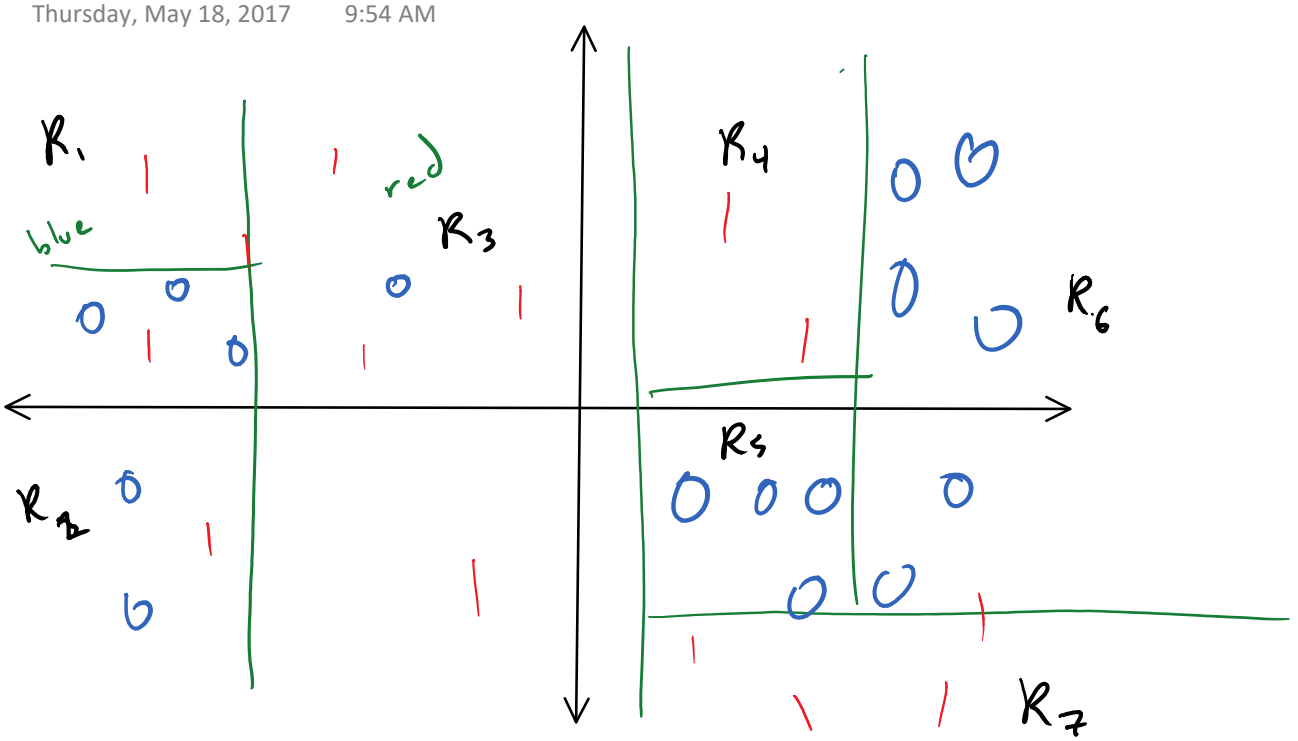
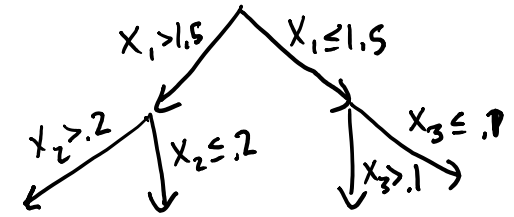
Decision Trees

Thursday, May 18, 2017 9:54 AM



ex covariates X_1, X_2, X_3



$$\hat{f}(x) = \sum_{m=1}^{r} \hat{y}_{R_m} 1\{x \in R_m\}$$

(1) Split into halfplanes $R_1(j,s) = \{X : X_j \le s\}$ $R_2(j,s) = \{X : X_j > s\}$

(2) $$\min_{j,s} \left[\min_{\hat{y}_1} \sum_{x_i \in R_1(j,s)} \ell(y_i, \hat{y}_1) + \min_{\hat{y}_2} \sum_{x_i \in R_2(j,s)} \ell(y_i, \hat{y}_2) \right] \quad (\star)$$

(3) $y_i \leftarrow y_i - \hat{y}_{1/2}$ ← depend on region

(4) repeat (1)-(3) w/ R_1, R_2

Regression: $\ell(y_i, \hat{y}) = (y_i - \hat{y})^2$ $\quad \hat{y}_i = \frac{1}{|R_1(j,s)|} \sum_{x_i \in R_1(j,s)} y_i$

Classification: classes $k \in \{1, \dots, K\}$

- $\ell(y_i, \hat{y}) = 1\{y_i \neq \hat{y}\}$ (not use (3))

  w/ rectangle pred. R_m $\quad \hat{p}_{mk} = \frac{1}{|R_m|} \sum_{x_i \in R_m} 1\{y_i = k\}$

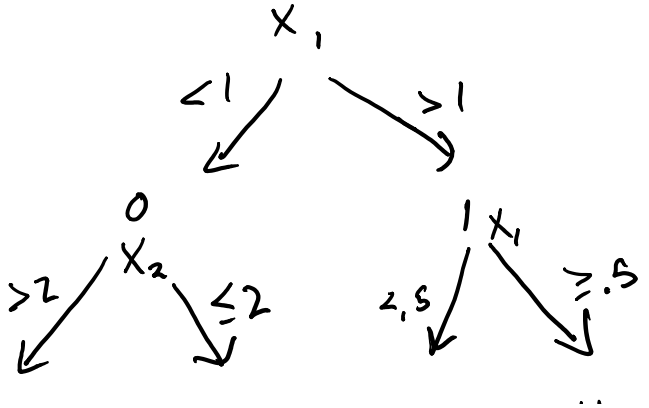
  $\hat{y}_i = \underset{k}{\arg\max}\{\hat{p}_{mk}, x_i \in R_m\}$ (ERM)

- Gini Index $\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_k \hat{p}_{mk}(1 - \hat{p}_{mk})$

- Cross-Entropy $-\sum_k \hat{p}_{mk} \log \hat{p}_{mk}$

- Gini & CE are differentiable, prefer "balanced" trees, outputs "soft classifier"

Code: Binary string - digits represent splits



Length of codeword is depth of tree

MDL principle use L as a tuning parameter

00 01 10 11 AIC: $-\text{loglik} + L$