Decision Trees

Gautam Kamath

Example: Should I wear a jacket?

- (Draw example on board in tree format)
  - Features: temperature, is cloudy?
  - Label: Jacket or no jacket
- (Draw example on board in 2D picture)

Decision Trees

- Simple and intuitive
- “Interpretable”
- Classification or Regression
- Can tend to overfit
- Can handle non-linear functions, but may fail on linear ones
  - (Draw example of diagonal separator, not axis-aligned)
  - (Draw example and tree with –’s on left, +’s on right, but +’s on top left and –’s on top right)
    - (Do recursive splitting, informally)
- Prediction: Walk down tree, use leaf’s prediction

Building a tree

- Start with one node
- Recursively split node
- Split(leaf):
  - Choose variable and threshold
  - Create two leaves and partition of points
- (Draw 1D example with dense +’s on left, dense –’s on right)
- (Show a few example choices of threshold)
- Intuition: Pick split that makes leaves “pure”

How to split?

- Choose a loss function for a node which is small for pure nodes and large for mixed ones
- Split at threshold t which minimizes sum of new node costs
$$t^* = \arg\min_t \ell(\{(x_i, y_i) : x_i \leq t\}) + \ell(\{(x_i, y_i) : x_i > t\})$$
- Only have to try at most n different values of t
- What loss function to use?

Loss functions? (Draw them for binary)

- Computing $\ell(S)$, where $S = \{(x_i, y_i)\}$ is some subset of examples
- Let $\hat{p}_c =$ fraction of S with label $c = \frac{1}{|S|} \sum_{i \in S} \mathbf{1}\{y_i = c\}$
- (Give example with 4 0's and 2 1's)
- $\hat{y} = \arg\max_c \hat{p}_c$
- Misclassification loss: $\ell(S) = 1 - \hat{p}_{\hat{y}}$
  - If all $0's$, $\hat{p}_0 = 1, \ell(S) = 1 - 1 = 0$. If 50-50, $\hat{p}_0 = \hat{p}_1 = 1/2, \ell(S) = 1/2$
- Entropy: $\ell(S) = -\sum_{classes\ c} \hat{p}_c \log \hat{p}_c$
- Gini index: $\ell(S) = \sum_{classes\ c} \hat{p}_c (1 - \hat{p}_c)$
- In regression setting: $\ell(S) = \min_p \sum_{i \in S} (y_i - p)^2 = \sum_{i \in S} (y_i - \bar{y}_S)^2$

Which variable to split on? Try all, pick best

$$S_L = \{(x_i, y_i) : x_{ij} \le t\}, S_R = \{(x_i, y_i) : x_{ij} > t\}$$
$$(j^*, t^*) = \arg\min_{j,t} |S_L|\ell(S_L) + |S_R|\ell(S_R)$$

Gini index: $\sum_{classes\ c} \hat{p}_c(1 - \hat{p}_c)$

Split on smokes?

No: $\hat{p}_0 = \frac{1}{4}, \hat{p}_1 = \frac{3}{4}$. Yes: $\hat{p}_0 = \frac{2}{3}, \hat{p}_1 = \frac{1}{3}$.

Cost: $4 \cdot \left(\left(\frac{3}{4}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\right) + 6 \cdot \left(\left(\frac{2}{3}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right) = 4.16$

Split on age? (Cheat to save time: use 35 as split)

≤ 35: $\hat{p}_0 = 1, \hat{p}_1 = 0$. > 35: $\hat{p}_0 = \frac{1}{6}, \hat{p}_1 = \frac{5}{6}$.

Cost: $4 \cdot ((0)(1) + (1)(0)) + 6 \cdot \left(\left(\frac{1}{6}\right)\left(\frac{5}{6}\right) + \left(\frac{5}{6}\right)\left(\frac{1}{6}\right)\right) = 1.66$

Age	Smokes	Cancer?
10	No	0
18	Yes	0
25	No	0
35	Yes	0
50	No	1
55	Yes	1
70	Yes	1
80	No	0
85	Yes	1
90	Yes	1

Stopping?

- Depth
- Running time
- Few examples at each leaf
- Leaves are all homogeneous
- Small improvements via a split
  - $\Delta = \ell(S) - (\ell(S_L) + \ell(S_R))$

Pruning

- Grow tree “fully” without stopping early, regularize over subtrees
- $\min \sum_{\text{leafs } v} \text{"error" in leaf } v + \alpha(\text{\# of leafs})$
- (Draw different trees at stages, indicate $\alpha = 0$ vs ∞ cases)

Decision Stump

- A three node decision tree (draw)
- Fast, but not very good