Introduction to Machine Learning II

Dihui Lai

dlai@wustl.edu

July 13, 2019

CART: Impurity Measurement

At node t, the fraction of class i is denoted as $p(i|t)$

Entropy: $H(t) = -\sum_{i=1}^{C} p(i|t) log(p(i|t))$

Gini-Index: $Gini(t) = \sum_{i=1}^{C} [p(i|t)]^2$

Classification Error: $Error(t) = 1 - \max_i p(i|t)$

Entropy

Considering the entropy of a data set that contains two types of outcomes: 0, 1. Running 10 experiments, we get the following out comes (0, 0, 1, 0, 1, 0, 1, 1, 0, 1). The entropy is

$$H(t) = -\sum_{i=1}^{2} \frac{1}{2} log_2(\frac{1}{2}) = 1$$

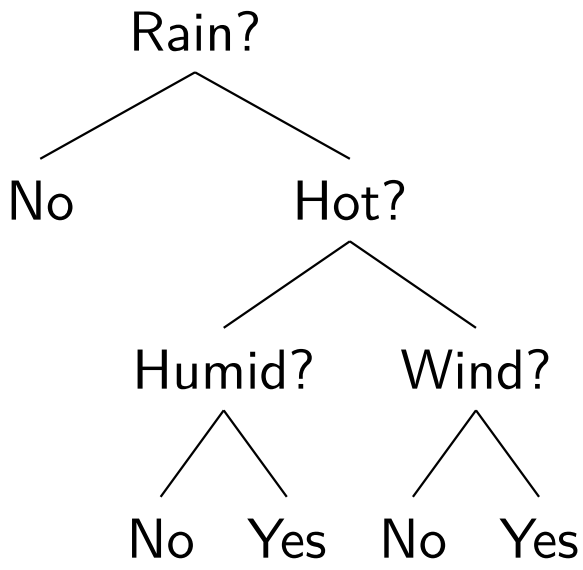
What if we have out come of (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)? The entropy is

$$H(t) = 1log_2(1) + 0log_2(0) = 0$$

The less variance in the data, the smaller the entropy is. For binary outcome, a probability of 1/2 leads to the largest uncertainty/entropy.

CART Model

John likes playing tennis on Saturday, The liklihood of John playing tennis depends on the weather. John has never played on a rainy. He may or may not play on a windy day. Given the weather, can you predict if John will play tennis this weekend?



How to Grow a Tree?

- At each node, the tree algorithm will check every variable for a possible split.
- A variable is selected for the split if it maximally reduces the impurity in the child nodes (e.g. the largest reduction of entroy, the largest reduction of variance etc.)
- Pruning: a tree can grow till each leave only contains one data point. Pruning the tree is needed to avoid overfitting (e.g. level of depth, minimum sample number req)

Information Gain at a Split Node

Assuming we collect John's tennis activities for 20 weeks. Out of the 20 weekends, we have 8 sunny day and John plays tennis on all of them. Out of the rest 12 days, John played tennis on 7 days. Before any split, we have the entropy calculated as

$$H(root) = -\frac{15}{20} log(\frac{15}{20}) - \frac{5}{20} log(\frac{5}{20}) = 0.562$$

Split based on the weather, we have on sunny days $H(sunny) = 0$ and on other days $H(!sunny) = -\frac{7}{12} log(\frac{7}{12}) - \frac{5}{12} log(\frac{5}{12}) = 0.679$ Therefore the child nodes have average entropy of

$$H(childs) = \frac{8}{20} \cdot 0 + \frac{12}{20} \cdot 0.679 = 0.407$$

We have entropy reduced by $\Delta H = 0.562 - 0.390 = 0.154$

Information Gain at a Split Node

Instead of splitting by forcast, we look at the wind speed. Out of the 20 weekends, 10 days are windy and 10 days are not windy. Out of the windy day, John played 6 times and 9 times for the non-windy days.
Split based on the wind condition, we have on windy days
$H(!windy) = -\frac{6}{10} log(\frac{6}{10}) - \frac{4}{10} log(\frac{4}{10}) = 0.673$ and on other days
$H(windy) = \frac{9}{10} \cdot log(\frac{9}{10}) + \frac{1}{10} \cdot log(\frac{1}{10}) = 0.325$
Therefore the child nodes have average entropy of

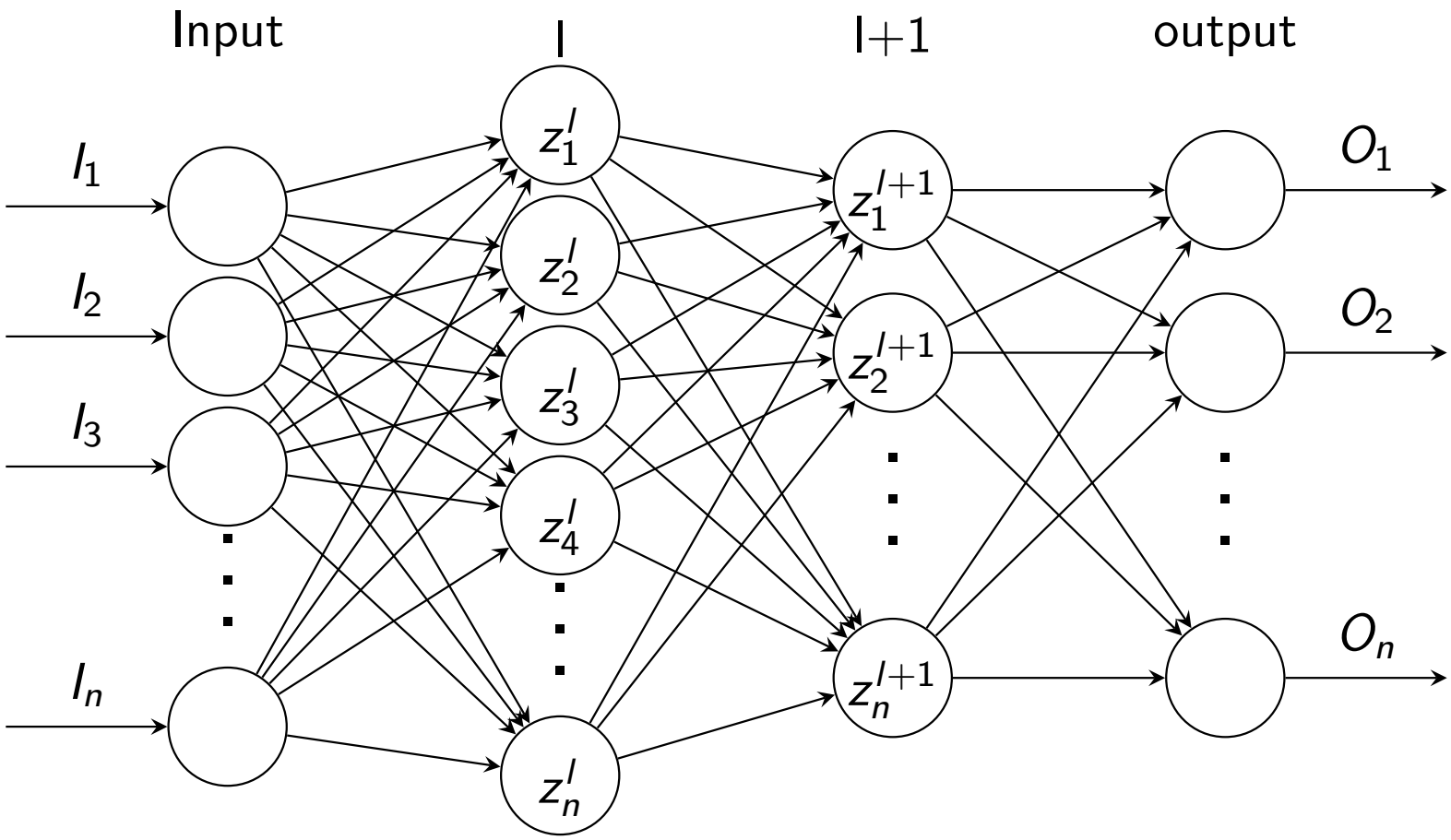
$$H(childs) = \frac{1}{2} \cdot 0.673 + \frac{1}{2} \cdot 0.325 = 0.499$$

We have entropy reduced by $\Delta H = 0.562 - 0.5 = 0.063$

At the first split, what condition shall we use to do the split?

Introduction to Neural Network

Neural Network: Topology



Neural Network: Forward

Each neuron at layer l recieves inputs from all neuron from the previous layer $l-1$

$$z_k^l = \sum_j w_{kj}^{l-1} a_j^{l-1}$$

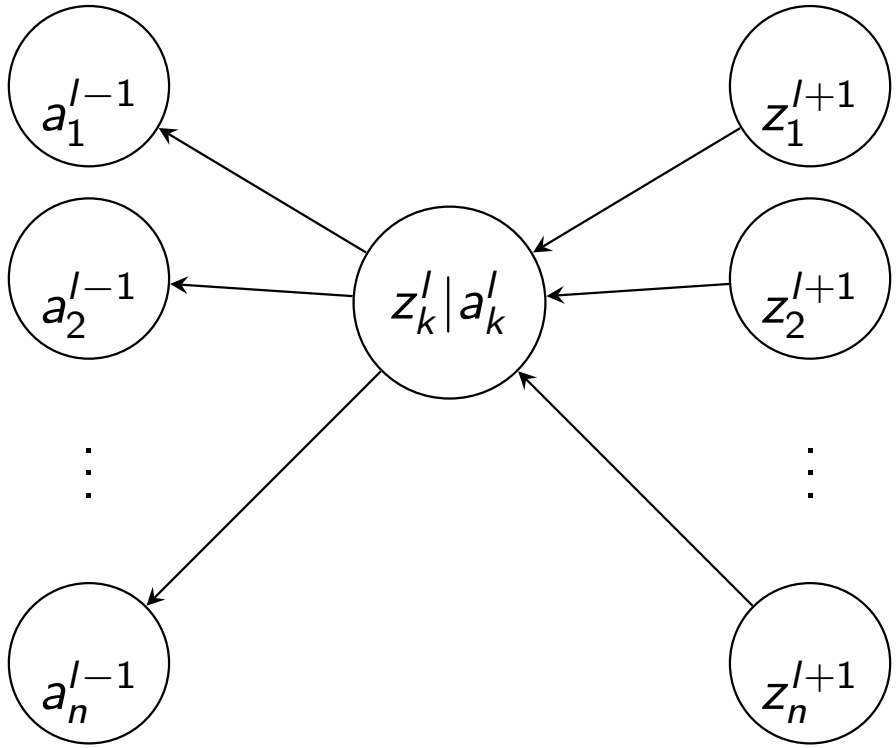
The neuron tansfer the input signal z_k^l via a transfer function σ and send as input to to the next layer

$$a_k^l = \sigma(z_k^l)$$

The cost function of the neural network is dependeng on all the zs of neurons in all layers

$$C\left(z_1^l, z_2^l, ...z_k^l(z_1^{l-1}, z_2^{l-1}, z_3^{l-1}, ...), ...\right)$$

Neural Network: Backpropogation



Neural Network: Backpropogation

The contribution to the cost function from a neuron in layer l can be cauculated iteratively as

$$\begin{aligned}\delta_k^l = \frac{\partial C}{\partial z_k^l} &= \sum_m \frac{\partial C}{\partial z_m^{l+1}} \frac{\partial z_m^{l+1}}{\partial z_k^l} \\ &= \left(\sum_m \frac{\partial C}{\partial z_m^{l+1}} \frac{\partial z_m^{l+1}}{\partial a_k^l} \right) \frac{\partial a_k^l}{\partial z_k^l} \\ &= \sum_m \delta_m^{l+1} w_{mk}^l \sigma'(z_k^l)\end{aligned}$$

The partial derivative of a cost function w.r.t the weight w_{kj}^{l-1} is

$$\frac{\partial C}{\partial w_{kj}^{l-1}} = \frac{\partial C}{\partial z_k^l} \frac{\partial z_k^l}{\partial w_{kj}^{l-1}} = \delta_k^l a_j^{l-1}$$