Decision trees

Victor Kitov

Skoltech
Skolkovo Institute of Science and Technology

November-December 2015.

Definition of decision tree

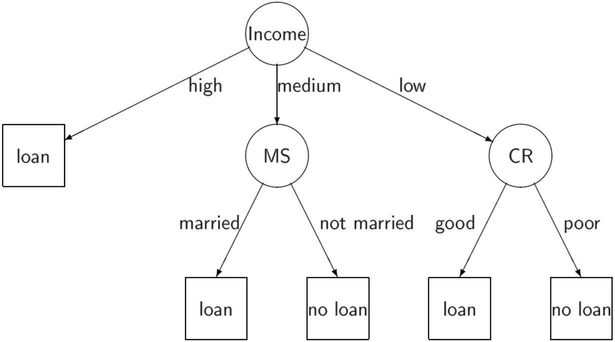
- Prediction is performed by tree T:
    - directed graph
    - without loops
    - with single root node

Definition of decision tree

- for each internal node t a check-function $Q_t(x^1, x^2, ...x^D)$ is associated
  - most commonly single feature value is considered: $Q_t(x^1, x^2, ...x^D) = x^{i(t)}$
- for each edge $r_1(t), ...r_{K(t)}(t)$ a set of values of check-function $Q_t(x^1, ...x^D)$ is associated: $S_1(t), ...S_{K(t)}(t)$ such that:
  - $S_1(t), ...S_{K(t)}$ cover the whole range of values of Q_t and $S_i \cap S_j = \emptyset \; \forall i \neq j$, $i, j \in \{r_1(t), ...r_{K(t)}(t)\}$.
  - most commonly $K(t) = 2$, $S_1 = \{x^{i(t)} \leq threshold(t)\}$, $S_2 = \{x^{i(t)} > threshold(t)\}$
  - variants: $S_i = \{l_i < x \leq h_i\}$, or $S = \{v_k\}$, where $\{v_1, v_2, ...\}$-is a set of individual values of $Q_t(x^1, x^2, ...x^D)$.

Definition of decision tree

- a set of nodes is divided into:
  - internal nodes *int*(T), each having ≥ 2 child nodes
  - terminal nodes *terminal*(T), which do not have child nodes but have associated prediction values.



Prediction process

- Each leaf (terminal) node performs prediction with a constant:
    - classification: class number
    - regression: real value
- Prediction process for tree T:
    - $t = root(T)$
    - while $t$is not leaf node:
        - calculate $Q_t(x)$
        - determine S_j out of $S_1(t), ...S_{K(t)}(t)$, where $Q_t(x)$ belongs: $Q_t(x) \in S_j(t)$
        - follow edge $r_j(t)$ to child node $\tilde{t}_j : t = \tilde{t}_j$
    - return prediction, associated with leaf t.

Specification of decision tree

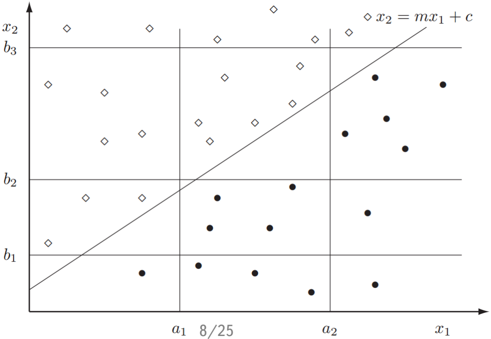
- To define a decision tree one needs to specify:
  - the check-function: $Q_t(x)$
  - the splitting criterion: $K(t)$ and $S_1(t), ...S_{K(t)}(t)$
  - the termination criteria (when node is defined as a terminal node)
  - the predicted value for each leaf node.

Most commonly used check-function

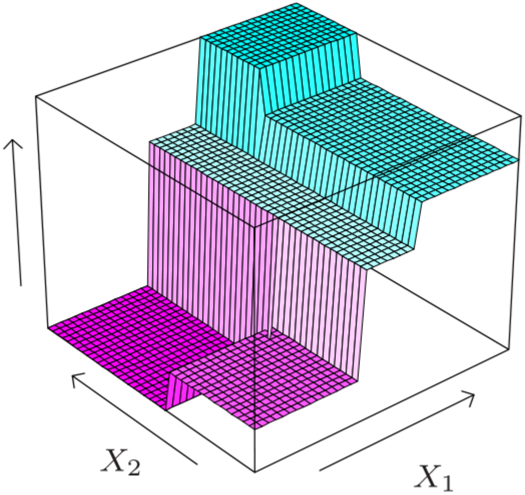
- Most commonly:
  - Q_t is defined as $Q_t(x) = x^{i(t)}$
  - $K(t) = 2\,\forall t \in \mathit{int}(T)$, where $\mathit{int}(T)$ - is a set of internal nodes.
  - $S_1(t) = \{x^{i(t)} \leq \mathit{threshold}(t)\}$, $S_2(t) = \{x^{i(t)} > \mathit{threshold}(t)\}$
    - $\mathit{threshold}(t) \in \{x_1^{i(t)}, x_2^{i(t)}, ...x_N^{i(t)}\}$
    - applicable only for real, ordinal and binary features
    - nominal features should be transformed, for example, using one-hot encoding

Analysis of single feature check-function

- Advantages:
  - simplicity
  - interpretability
- Drawbacks:
  - many nodes may be needed to describe boundaries not parallel to axes:

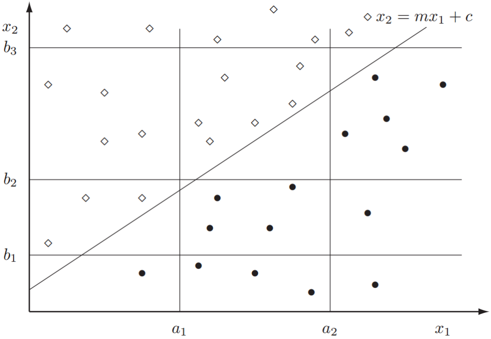


Piecewise constant solution of decision trees



More general check-functions

- Instead of considering value of individual feature, $Q(x)$ may be more general: $Q_t(x) =< a_t, x >$ or even non-linear.
    - also gives piecewise constant solution
    - less interpretable
    - may need much fewer nodes



Termination criterion

- Bias-variance tradeoff:
  - very large complex trees -> overfitting
  - very short simple trees -> underfitting
- Approaches to stopping:
  - rule-based
  - based on pruning

Rule-base termination criteria

- Rule-based: a criterion is compared with a threshold.
- Variants of criterion:
    - depth of tree
    - number of objects in a node
    - minimal number of objects in one of the child nodes
    - impurity of classes
    - change of impurity of classes after the split

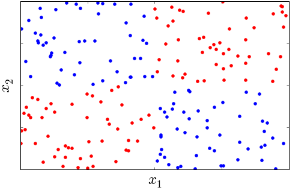
Analysis of rule-based termination

Advantages:

- simplicity
- interpretability

Disadvantages:

- specification of threshold is needed
- impurity change is suboptimal: further splits may become better than current one
    - example:



Impurity function

- Let t be any node and $u(t)$ - associated objects with node t,
- $N(t)$ - total number of objects and $N_j(t)$ - number of objects of class j in t
- Probabilities of classes within node t:

$$p(\omega_j|x \in u(t)) = p(\omega_j|t) \approx \frac{N_j(t)}{N(t)}$$

- Impurity function $I(t) = \phi(p(\omega_1|t), ...p(\omega_C|t))$ has the following properties:
  - $\phi(q_1, q_2, ...q_C)$ is defined for $q_j \geq 0$ and $\sum_j q_j = 1$.
  - ϕ attains maximum for $q_j = 1/C$, $k = 1, 2, ...C$.
  - ϕ attains minimum when $\exists j : \; q_j = 1$, $q_i = 0 \; \forall i \neq j$.
  - ϕ is symmetric function of $q_1, q_2, ...q_C$.

Typical impurity functions

- **Gini criterion**
  - interpretation: probability to make mistake when classifying object randomly with class probabilities $[p(\omega_1|t), ...p(\omega_C|t)]$:

$$I(t) = \sum_i p(\omega_i|t)(1 - p(\omega_i|t)) = 1 - \sum_i [p(\omega_i|t)]^2$$

- **Entropy**
  - interpretation: measure of uncertainty of random variable

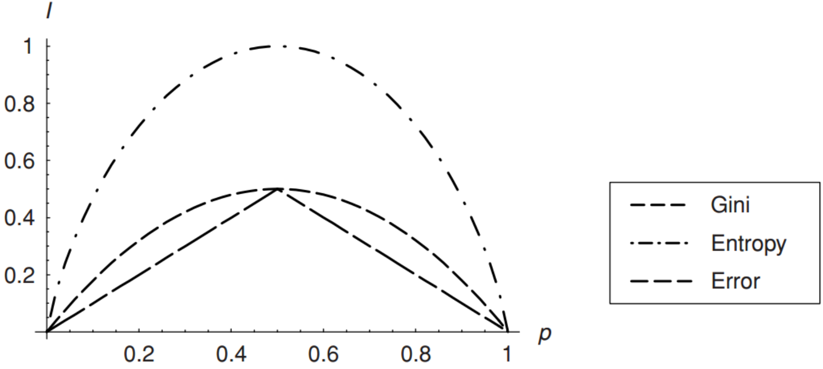
$$I(t) = -\sum_i p(\omega_i|t) \ln p(\omega_i|t)$$

- **Classification error**
  - interpretation: frequency of errors when classifying with the most common class

$$I(t) = 1 - \max_i p(\omega_i|t)$$

Typical impurity functions

Impurity functions for binary classification with class probabilities $p = p(\omega_1|t)$ and $1 - p = p(\omega_2|t)$.



Splitting criterion selection

- Select splitting criterion, maximizing:

$$\Delta I(t) = I(t) - \sum_{i=1}^{S} I(t_j) \frac{N_j(t)}{N(t)}$$

  where $t_1, ...t_S$ are the child nodes of node t.
- If $I(t)$ is entropy, then $\Delta I(t)$ is called *information gain*.

Regression: prediction assignment for leaf nodes

- Define $I_t = \{i : x_i \in u(t)\}$, N_t - number of elements in I_t.
- For quadratic loss $(\widehat{y} - y)^2$:

$$\widehat{y} = \arg\min_{\mu} \sum_{i \in I} (y_i - \mu)^2 = \frac{1}{N_t} \sum_{i \in I} y_i,$$

- For abs. deviation loss $|\widehat{y} - y|$:

$$\widehat{y} = \arg\min_{\mu} \sum_{i \in I} |y - \mu| = median\{y_i : \ i \in I\}.$$

Classification: prediction assignment for leaf nodes

- Define $\lambda(\omega_i, \omega_j)$ - the cost of predicting object of class ω_i as belonging to class ω_j
    - Minimum loss class assignment:

$$c = \arg\min_{\omega} \sum_{i:\, x_i \in u(t)} \lambda(c_i, \omega)$$

    - For $\lambda(\omega_i, \omega_j) = \mathbb{I}[\omega_i \neq \omega_j]$ most common class will be associated with the leaf node:

$$c = \arg\max_{\omega} \left|\{i : \, x_i \in u(t), \, y_i = \omega\}\right|$$

CART

- Let T be some subtree of out tree, $\tilde{T}$ be a set of leaf nodes of tree T.
- Define $R(t) = \frac{M(t)}{N}$ the error-rate loss for leaf node $t \in \tilde{T}$, where $M(t)$ - is the number of mistakes by the tree on the **validation set** and N is the validation set size.
- Also define

$$\begin{array}{ll} \text{error-rate loss :} & R(T) = \sum_{t \in \tilde{T}} R(t) \\ \text{complexity+error-rate loss:} & R_\alpha(T) = \sum_{t \in \tilde{T}} R_\alpha(t) = R(T) + \alpha|\tilde{T}| \end{array}$$

- Condition when $R_{\alpha_t}(T_t) = R_{\alpha_t}(t)$:

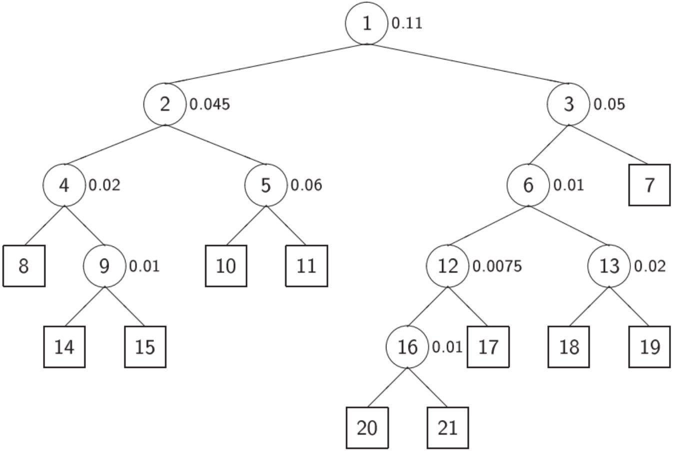
$$\alpha_t = \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}$$

Pruning algorithm

1. Build tree until each node contains representatives of only single class - obtain tree T.
2. Build a sequence of nested trees $T = T_0 \supset T_1 \supset ... \supset T_{|T|}$ containing $|T|$, $|T| - 1$,...1 nodes, repeating the procedure:
   - replace the tree T_t with smallest α_t with its root t
   - recalculate α_t for all ancestors of t.
3. For trees $T_0, T_1, ...T_{|T|}$ calculate their validation set error-rates $R(T_0), R(T_1), ...R(T_{|T|})$.
4. Select T_i, giving minimum error-rate:

$$i = \arg\min_i R(T_i)$$

Example


1 0.11
2 0.045
3 0.05
4 0.02
5 0.06
6 0.01
7
8
9 0.01
10
11
12 0.0075
13 0.02
14
15
16 0.01
17
18
19
20
21

Example

Logs of the performance metrics of the pruning process:

step num.	α_k	$\lvert\tilde{T}^k\rvert$	$R(T^k)$
1	0	11	0.185
2	0.0075	9	0.2
3	0.01	6	0.22
4	0.02	5	0.25
5	0.045	3	0.34
6	005	2	0.39
7	0.11	1	0.5

Handling missing values

If checked feature is missing:

- fill missing values:
    - with feature mean
    - with new categorical value “missing” (for categorical values)
    - predict them using other known features
- CART uses prediction of unknown feature using another feature that best predicts the missing one: “surrogate split” - technique
- ID3 and C4.5 decision trees use averaging of predictions made by each child node with weights $N(t_1)/N(t), N(t_2)/N(t), ... N(t_S)/N(t)$.

Analysis of decision trees

- Advantages:
  - simplicity
  - interpretability
  - implicit feature selection
  - naturally handles both discrete and real features
  - prediction is invariant to monotone transformations of features for $Q_t(x) = x^{i(t)}$
    - in particular, to normalization of features
- Disadvantages:
  - non-parallel to axes class separating boundary may lead to many nodes in the tree for $Q_t(x) = x^{i(t)}$
  - one step ahead lookup strategy for split selection may be insufficient (XOR example)
  - not online - slight modification of the training set will require full tree reconstruction.