Decision trees & Random forests

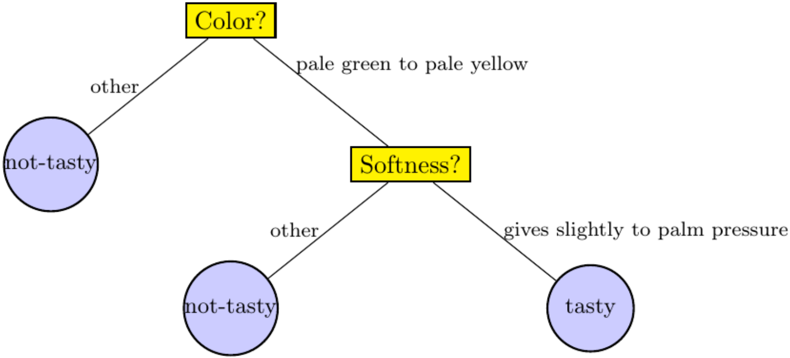
T. Romary

Mines ParisTech, PSL Research University

05/20/19

Outline

Example: Papayas



Notations

- The training set is $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$
- $X_i \in \mathbb{R}^p$
- classification: $Y_i \in \{0, 1\}$
- regression: $Y_i \in \mathbb{R}$

Decision trees

Principle

Divide recursively the observations through splitting rules based on classification/regression features. The recursion is completed when the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions

Operations:

- Decide if a node is terminal
- Select a segmentation rule
- Affect a class/value to a leaf

CART algorithm (Breiman et al. 1984)

Classification

A node is terminal if the associated value of the criterion is less than a threshold or if it contains less than a predefined number of observations

Splitting criterion (based on the Gini impurity index)

cut in cell A (j, z), $j \in [p]$, $z \in [0, 1]$ such that

$$\begin{aligned} L_{\text{class, } n}(j, z) =& p_{0,n}(A) p_{1,n}(A) - \frac{N_n(A_L)}{N_n(A)} p_{0,n}(A_L) p_{1,n}(A_L) \\ & - \frac{N_n(A_R)}{N_n(A)} p_{0,n}(A_R) p_{1,n}(A_R) \end{aligned}$$

where $p_{0,n}(A)$ is the proportion of 0 in node A
the class affected to A_L (resp. A_R) is obtained by majority vote
Then the tree is pruned to avoid overfitting

CART algorithm (Breiman et al. 1984)

Regression

A node is terminal if the splitting criterion cannot be improved or if it contains less than a predefined number of observations

Splitting criterion

cut in cell A (j, z), $j \in [p]$, $z \in [0, 1]$ such that

$$
\begin{aligned}
L_{\text{reg},\, n}(j, z) = & \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_A)^2 \mathbb{1}_{X_i \in A} \\
& - \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_L} \mathbb{1}_{X_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{X_i^{(j)} \geq z})^2 \mathbb{1}_{X_i \in A}
\end{aligned}
$$

the value affected to A_L (resp. A_R) is $\bar{Y}_{A_L}$ (resp. $\bar{Y}_{A_R}$)

Outline

Random forests

J. Howard (Kaggle) and M. Bowles (Biomatica) in Howard and Bowles (2012) claim that

Ensembles of decision trees – often known as "random forests" – have been the most successful general-purpose algorithm in modern times

Random forests (Breiman 2001)

Based on "bagging", i.e. bootstrap-aggregating

- bootstrap: resampling
- aggregating
  - regression: mean over M trees
  - classification: majority vote over M trees

Random forests

4 important parameters

- M, number of trees
- $a_n \in [n]$, number of sampled data points in each tree
- `mtry` $\in [p]$ number of possible splitting directions at each node
- `nodesize` $\in [a_n]$ number of examples in each cell below which the cell is not split

Some results in the regression framework

The j-th tree estimate takes the form

$$m_n(x;\Theta_j,\mathcal{D}_n)=\sum_{i\in\mathcal{D}_n^\star(\Theta_j)}\frac{\mathbb{1}_{X_i\in A_n(x;\Theta_j,\mathcal{D}_n)}Y_i}{N_n(x;\Theta_j,\mathcal{D}_n)}$$

then the finite forest estimates writes

$$m_{M,n}(x;\Theta_1,\ldots,\Theta_M,\mathcal{D}_n)=\frac{1}{M}\sum_{j=1}^{M}m_n(x;\Theta_j,\mathcal{D}_n)$$

and there is a law of large numbers result

$$\lim_{M\to\infty}m_{M,n}(x;\Theta_1,\ldots,\Theta_M,\mathcal{D}_n)=m_{\infty,n}(x;\mathcal{D}_n)=\mathbb{E}_\Theta(m_n(x;\Theta,\mathcal{D}_n))$$

The analysis of original random forests is difficult and people work on simplified versions (e.g. "pure" random forests")

Some results in the regression framework

The j-th tree estimate takes the form

$$m_n(x;\Theta_j,\mathcal{D}_n) = \sum_{i\in\mathcal{D}_n^\star(\Theta_j)} \frac{\mathbb{1}_{X_i\in A_n(x;\Theta_j,\mathcal{D}_n)} Y_i}{N_n(x;\Theta_j,\mathcal{D}_n)}$$

then the finite forest estimates writes

$$m_{M,n}(x;\Theta_1,\ldots,\Theta_M,\mathcal{D}_n) = \frac{1}{M}\sum_{j=1}^{M} m_n(x;\Theta_j,\mathcal{D}_n)$$

and there is a law of large numbers result

$$\lim_{M\to\infty} m_{M,n}(x;\Theta_1,\ldots,\Theta_M,\mathcal{D}_n) = m_{\infty,n}(x;\mathcal{D}_n) = \mathbb{E}_\Theta(m_n(x;\Theta,\mathcal{D}_n))$$

The analysis of original random forests is difficult and people work on simplified versions (e.g. "pure" random forests")

Variable importance measures

- Mean decrease impurity (MDI)

$$\widehat{MDI}(X^{(j)}) = \frac{1}{M} \sum_{l=1}^{M} \sum_{\substack{t \in \mathcal{T}_l \\ j^\star_{n,t} = j}} p_{n,t} L_{\text{reg},\, n}(j^\star_{n,t}, z^\star_{n,t})$$

- Mean decrease accuracy (MDA)

$$\widehat{MDA}(X^{(j)}) = \frac{1}{M} \sum_{l=1}^{M} \Big[R_n[m_n(\cdot; \Theta_l), \mathcal{D}^j_{l,n}] - R_n[m_n(\cdot; \Theta_l), \mathcal{D}_{l,n}] \Big]$$

where $\mathcal{D}_{l,n}$ is the out of the bag sample, $\mathcal{D}^j_{l,n}$, the same where the values of variable j have been randomly permuted

$$R_n[m_n(\cdot; \Theta_l), \mathcal{D}] = \frac{1}{|\mathcal{D}|} \sum_{i:(X_i, Y_i) \in \mathcal{D}} (Y_i - m_n(X_i; \Theta_l))^2$$

Extensions

- Weighted forests
- Ranking forests
- Quantile forests
- etc.

References

- Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.
- Breiman, L. (1984). Classification and regression trees. Routledge.
- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.