Assignment 5
Decision Trees

Kuang Yu Li, st169971@stud.uni-stuttgart.de, 3440829
Ya Jen Hsu, st169013@stud.uni-stuttgart.de, 3449448
Gabriella Ilena, st169935@stud.uni-stuttgart.de, 3440942

1 Inductive Construction

1. Entropy root node
   - ➔ 240 instances, 125 − and 115 +
   - ➔ Entropy $H(P) = -(\frac{125}{240} \times log_2(\frac{125}{240}) + \frac{115}{240} \times log_2(\frac{115}{240})) = 0.9987$
     - ◆ Attribute F1

Value	Instance −	Instance +	Total	Entropy
0	70	50	120	$-(\frac{70}{120} \times log_2(\frac{70}{120}) + \frac{50}{120} \times log_2(\frac{50}{120})) = 0.9799$
1	55	65	120	$-(\frac{55}{120} \times log_2(\frac{55}{120}) + \frac{65}{120} \times log_2(\frac{65}{120})) = 0.995$

     - Expected entropy: $\frac{120}{240} \times 0.9799 + \frac{120}{240} \times 0.995 = 0.9875$
     - Information Gain: $0.9987 - 0.9875 = 0.0112$

     - ◆ Attribute F2

Value	Instance −	Instance +	Total	Entropy
0	50	75	125	$-(\frac{50}{125} \times log_2(\frac{50}{125}) + \frac{75}{125} \times log_2(\frac{75}{125})) = 0.971$
1	70	45	115	$-(\frac{70}{115} \times log_2(\frac{70}{115}) + \frac{45}{115} \times log_2(\frac{45}{115})) = 0.9656$

     - Expected entropy: $\frac{125}{240} \times 0.971 + \frac{115}{240} \times 0.9656 = 0.9684$
     - Information Gain: $0.9987 - 0.9684 = 0.0303$

     - ◆ Attribute F3

Value	Instance −	Instance +	Total	Entropy
0	65	15	80	$-(\frac{65}{80} \times log_2(\frac{65}{80}) + \frac{15}{80} \times log_2(\frac{15}{80})) = 0.6962$
1	50	30	80	$-(\frac{50}{80} \times log_2(\frac{50}{80}) + \frac{30}{80} \times log_2(\frac{30}{80})) = 0.9544$
2	10	70	80	$-(\frac{10}{80} \times log_2(\frac{10}{80}) + \frac{70}{80} \times log_2(\frac{70}{80})) = 0.5436$

- Expected entropy:
  $\frac{80}{240} \times 0.6962 + \frac{80}{240} \times 0.9544 + \frac{80}{240} \times 0.5436 = 0.7314$
- Information Gain: $0.9987 - 0.7314 = 0.2673$

◆ Attribute F4

Value	Instance −	Instance +	Total	Entropy
0	70	55	125	$-(\frac{70}{125} \times log_2(\frac{70}{125}) + \frac{55}{125} \times log_2(\frac{55}{125})) = 0.9896$
1	50	65	115	$-(\frac{50}{115} \times log_2(\frac{50}{115}) + \frac{65}{115} \times log_2(\frac{65}{115})) = 0.9877$

- Expected entropy: $\frac{125}{240} \times 0.9896 + \frac{115}{240} \times 0.9877 = 0.9887$
- Information Gain: $0.9987 - 0.9887 = 0.01$

➔ We choose F3 to be root

2. For value=0 in F3

➔ 80 instances, 65 − and 15 +

➔ Entropy $H(P) = -(\frac{65}{80} \times log_2(\frac{65}{80}) + \frac{15}{80} \times log_2(\frac{15}{80})) = 0.6962$

◆ Attribute F1

Value	Instance −	Instance +	Total	Entropy
0	35	5	40	$-(\frac{35}{40} \times log_2(\frac{35}{40}) + \frac{5}{40} \times log_2(\frac{5}{40})) = 0.5436$
1	30	10	40	$-(\frac{30}{40} \times log_2(\frac{30}{40}) + \frac{10}{40} \times log_2(\frac{10}{40})) = 0.8113$

- Expected entropy: $\frac{40}{80} \times 0.5436 + \frac{40}{80} \times 0.8113 = 0.6775$
- Information Gain: $0.6962 - 0.6775 = 0.0187$

◆ Attribute F2

Value	Instance −	Instance +	Total	Entropy
0	25	15	40	$-(\frac{25}{40} \times log_2(\frac{25}{40}) + \frac{15}{40} \times log_2(\frac{15}{40})) = 0.9544$
1	40	0	40	0

- Expected entropy: $\frac{40}{80} \times 0.9544 + 0 = 0.4772$
- Information Gain: $0.6962 - 0.4772 = 0.219$

◆ Attribute F4

Value	Instance −	Instance +	Total	Entropy

0	30	10	40	$-(\frac{30}{40} \times log_2(\frac{30}{40}) + \frac{10}{40} \times log_2(\frac{10}{40})) = 0.8113$
1	35	5	40	$-(\frac{35}{40} \times log_2(\frac{35}{40}) + \frac{5}{40} \times log_2(\frac{5}{40})) = 0.5436$

- Expected entropy: $\frac{40}{80} \times 0.8113 + \frac{40}{80} \times 0.5436 = 0.6775$
- Information Gain: $0.6962 - 0.6775 = 0.0187$

➔ We choose F2 to be the attribute

3. For value=1 in F3
   ➔ 80 instances, 50 − and 30 +
   ➔ Entropy $H(P) = -(\frac{50}{80} \times log_2(\frac{50}{80}) + \frac{30}{80} \times log_2(\frac{30}{80})) = 0.9544$
      ◆ Attribute F1

Value	Instance −	Instance +	Total	Entropy
0	25	15	40	$-(\frac{25}{40} \times log_2(\frac{25}{40}) + \frac{15}{40} \times log_2(\frac{15}{40})) = 0.9544$
1	25	15	40	$-(\frac{25}{40} \times log_2(\frac{25}{40}) + \frac{15}{40} \times log_2(\frac{15}{40})) = 0.9544$

- Expected entropy: $\frac{40}{80} \times 0.9544 + \frac{40}{80} \times 0.9544 = 0.4772$
- Information Gain: $0.9544 - 0.4772 = 0.4772$

◆ Attribute F2

Value	Instance −	Instance +	Total	Entropy
0	25	15	40	$-(\frac{25}{40} \times log_2(\frac{25}{40}) + \frac{15}{40} \times log_2(\frac{15}{40})) = 0.9544$
1	25	15	40	$-(\frac{25}{40} \times log_2(\frac{25}{40}) + \frac{15}{40} \times log_2(\frac{15}{40})) = 0.9544$

- Expected entropy: $\frac{40}{80} \times 0.9544 + \frac{40}{80} \times 0.9544 = 0.4772$
- Information Gain: $0.9544 - 0.4772 = 0.4772$

◆ Attribute F4

Value	Instance −	Instance +	Total	Entropy
0	30	10	40	$-(\frac{30}{40} \times log_2(\frac{30}{40}) + \frac{10}{40} \times log_2(\frac{10}{40})) = 0.8113$
1	20	20	40	$-(\frac{20}{40} \times log_2(\frac{20}{40}) + \frac{20}{40} \times log_2(\frac{20}{40})) = 1$

- Expected entropy: $\frac{40}{80} \times 0.8113 + \frac{40}{80} \times 1 = 0.9057$
- Information Gain: $0.9544 - 0.9057 = 0.0487$

➔ We can choose F1

4. For value=2 in F3
   - ➔ 80 instances, 10 − and 70 +
   - ➔ Entropy $H(P) = -(\frac{10}{80} \times log_2(\frac{10}{80}) + \frac{70}{80} \times log_2(\frac{70}{80})) = 0.5436$
     - ◆ Attribute F1

Value	Instance −	Instance +	Total	Entropy
0	10	30	40	$-(\frac{10}{40} \times log_2(\frac{10}{40}) + \frac{30}{40} \times log_2(\frac{30}{40})) = 0.8113$
1	0	40	40	0

       - Expected entropy: $\frac{40}{80} \times 0.8113 + 0 = 0.4057$
       - Information Gain: $0.5436 - 0.4057 = 0.1379$

     - ◆ Attribute F2

Value	Instance −	Instance +	Total	Entropy
0	0	40	10	0
1	10	30	40	$-(\frac{10}{40} \times log_2(\frac{10}{40}) + \frac{30}{40} \times log_2(\frac{30}{40})) = 0.8113$

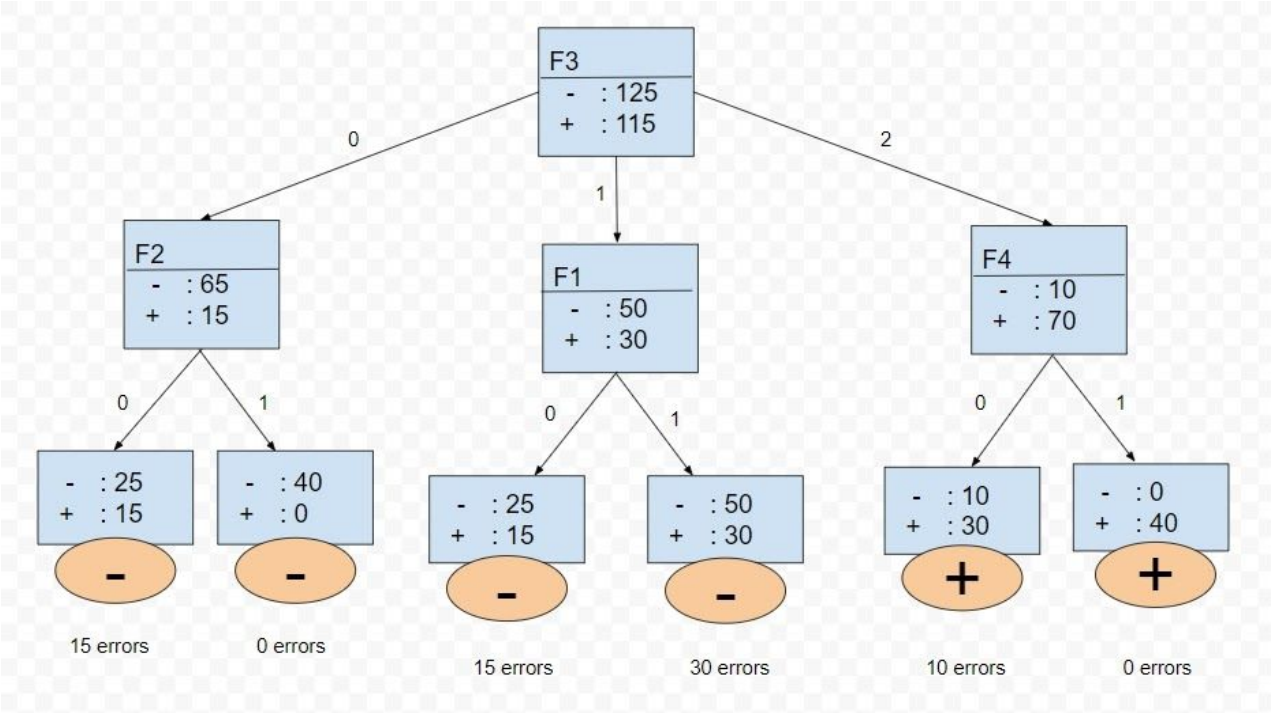
       - Expected entropy: $0 + \frac{40}{80} \times 0.8113 = 0.4057$
       - Information Gain: $0.5436 - 0.4057 = 0.1379$

     - ◆ Attribute F4

Value	Instance −	Instance +	Total	Entropy
0	10	30	40	$-(\frac{10}{40} \times log_2(\frac{10}{40}) + \frac{30}{40} \times log_2(\frac{30}{40})) = 0.8113$
1	0	40	40	0

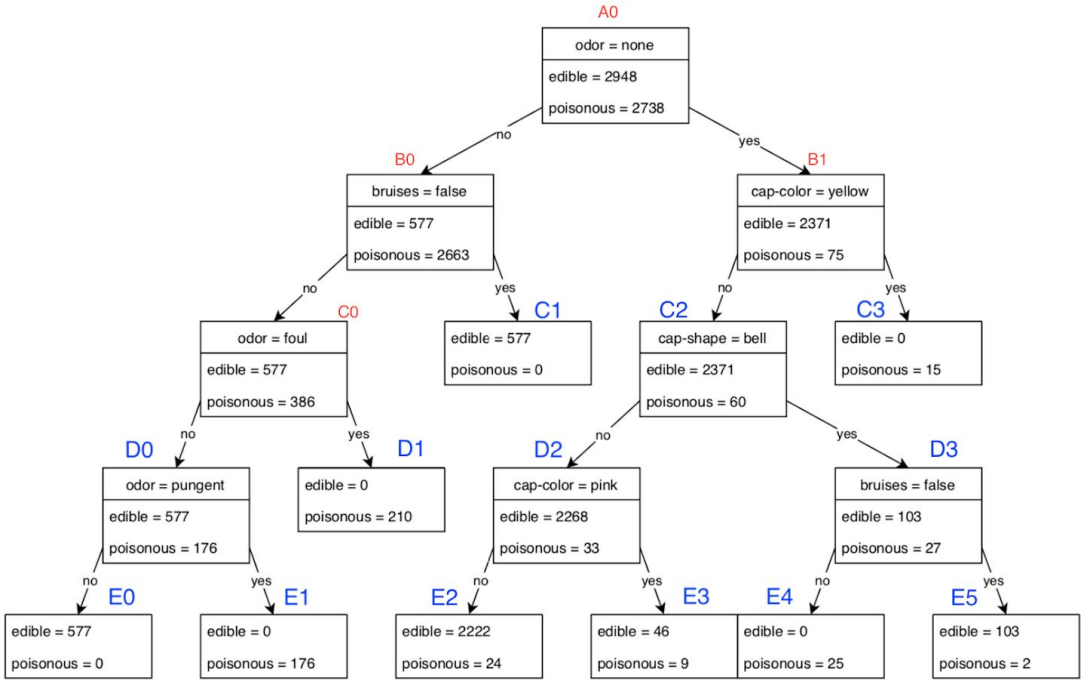
       - Expected entropy: $\frac{40}{80} \times 0.8113 + 0 = 0.4057$
       - Information Gain: $0.5436 - 0.4057 = 0.1379$

   - ➔ We can choose F4



=> Error rate: 70 errors / 240 training data = 0.292

2 Minimal Error Pruning

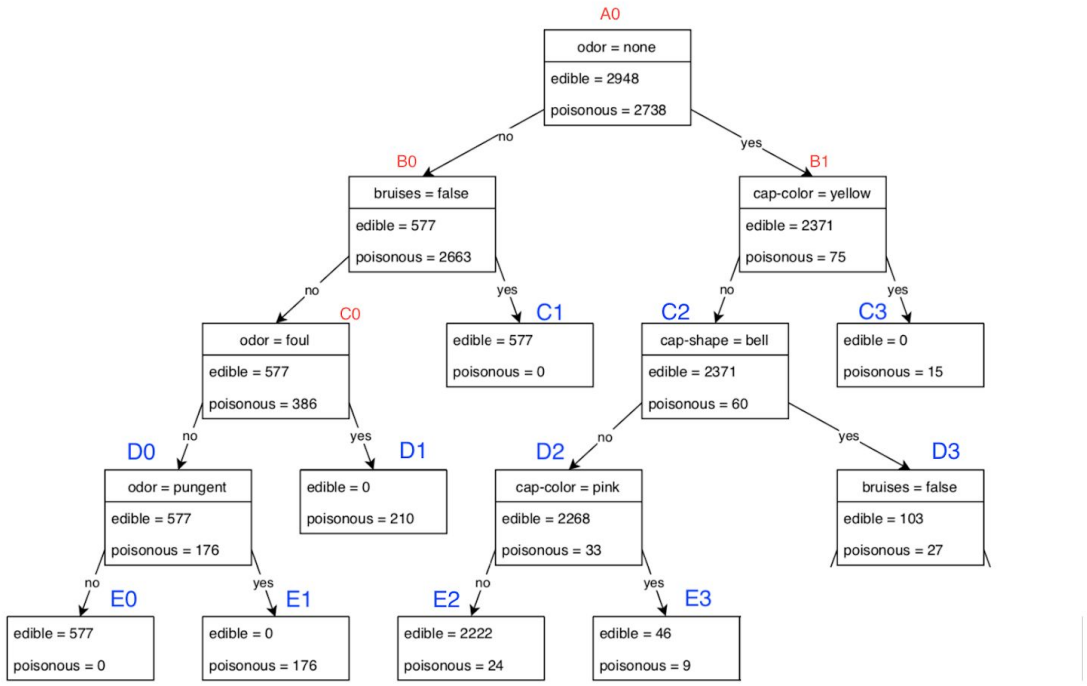


First, we calculated the total error from original decision tree, which is $\frac{35}{5686} = 0.6155\%$

No Pruning	Overall rate	Overall error	C1	C3	D1	E0	E1	E2	E3	E4	E5
	0.6155%	35	0	0	0	0	0	24	9	0	2

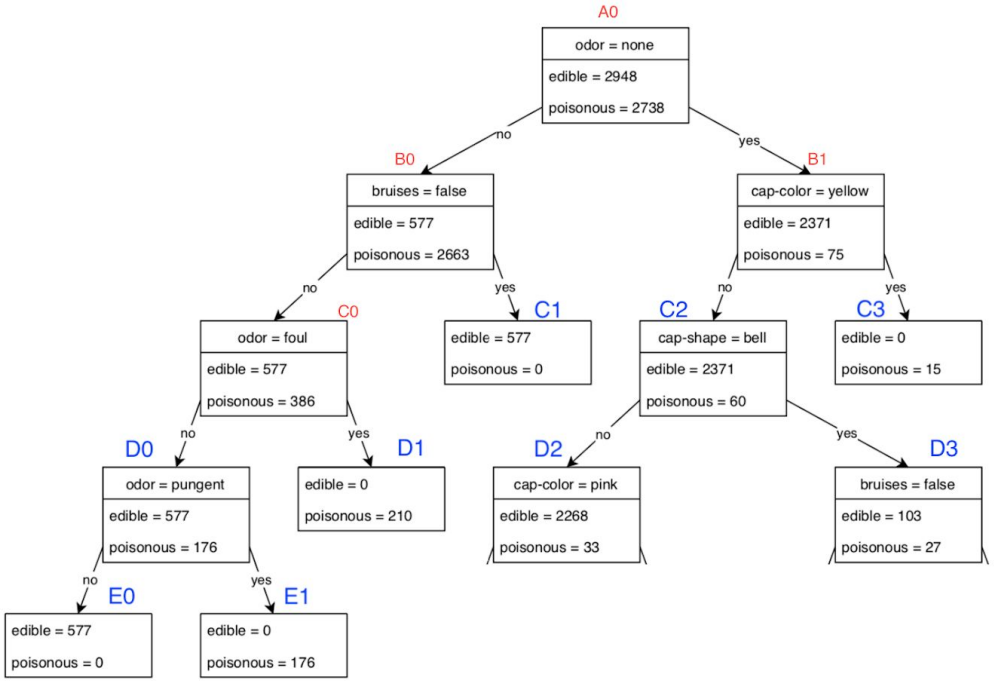
Second, we find the inner node with all leaves (D0, D2, D3) with least error and try to merge its children into a new leaf node.
The least error inner node is D3, because it has 27 errors. Then we pruned E4 and E5 and turned D3 into a new leaf.

Mmerging E4 and E5 into D3	Overall rate	Overall error	C1	C3	D1	E0	E1	E2	E3	D3 (merging E4 and E5)
	1.0552%	60	0	0	0	0	0	24	9	27



We then again started to look for a viable node for pruning, which is D0, D2.
The D2 has the least error. Then we pruned E2 and E3 and turned D2 into a new leaf.

Merging E2 and E3 into D2	Overall rate	Overall error	C1	C3	D1	E0	E1	D2 (merging E2 and E3)	D3 (merging E4 and E5)
	1.0552%	60	0	0	0	0	0	33	27



In the end, we prune two decisions D2 and D3 and turn them into new leaves by pruning leaves E2, E3, E4, and E5.

3 Regression with Decision Trees and kNN

How does the construction of regression trees differ from classification trees? How is a prediction computed in regression trees?

A regression tree can be built through stratification of the feature space, which consists of the following steps:

1. Divide the feature space into several distinct, non-overlapping regions, and use the cut-points/boundaries between regions as values to split the branches.
2. Observations that fall into a certain region are given the same prediction, which is the mean of the response values for the training data.

Determining the regions is done based on the idea of minimizing RSS (Residual Sum of Squares) between the predicted value for a certain region and the true response value, which produces a specific cut-point.

It is different from classification trees with regards to these aspects:

1. *Prediction*. With regression trees, observations that belong to a certain region is predicted as the mean response value of that region. This is not the case with classification trees, where we predict based on the most occurring class label in that region. Regression trees are also limited to quantitative prediction values.

2. *Construction*. RSS is not a suitable criterion to determine cut-points in classification trees, because we do not take the mean of response values for a region.

How can kNN be used for regression?

In using kNN for regression, we can do the following steps for training:

1. Find the k nearest data points to the point of interest, for all training data points
2. Take the average of those points
3. Connect the points (of average values) to build the regression line

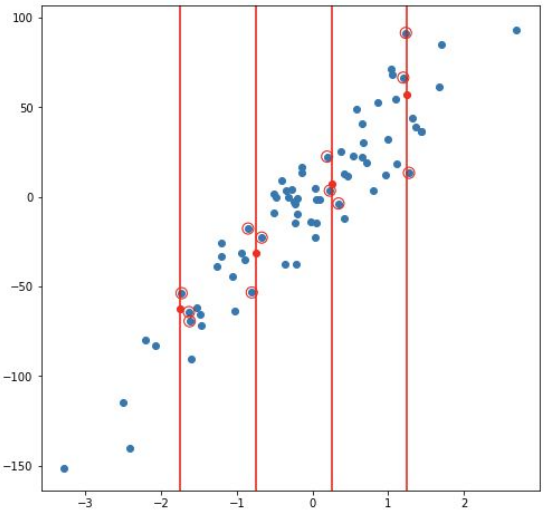


Fig. X. Taking the average value of the k nearest neighbours

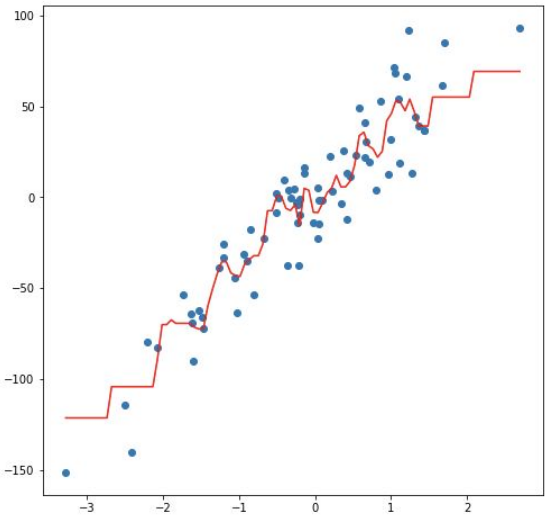


Fig. X. Forming the regression line

References:

1. Gareth James, et al., *An Introduction to Statistical Learning with Applications in R,* Springer, New York, 2013.
2. Max Miller, *The Basics: KNN for classification and regression*, https://towardsdatascience.com/the-basics-knn-for-classification-and-regression-c1e8a6c955