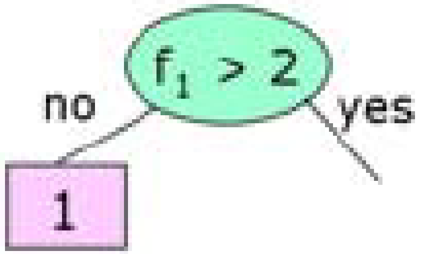
Decision Trees (2)

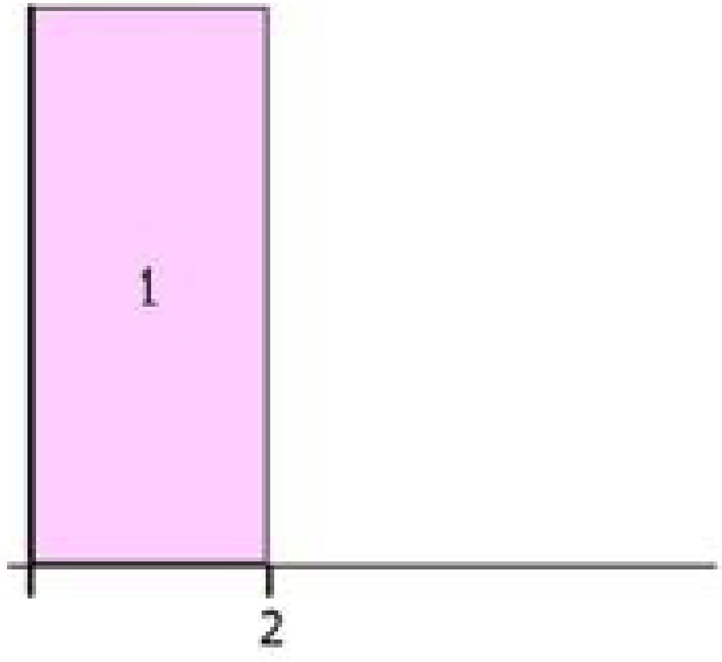
Numeric Attributes

Based on slides by Leslie Kaelbling (MIT)

Numerical attributes

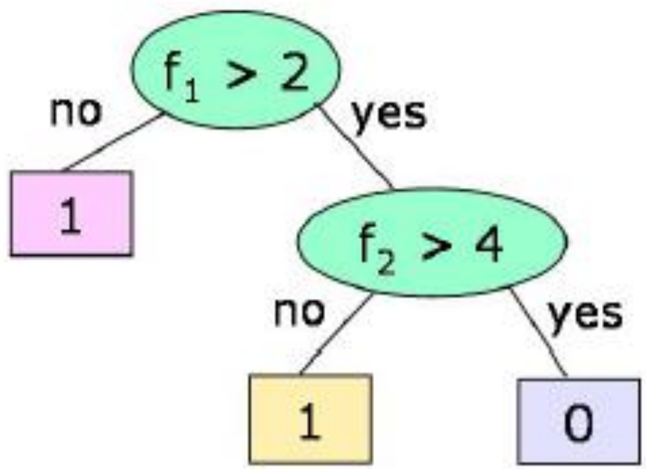
- Tests in nodes are of the form $f_i > constant$

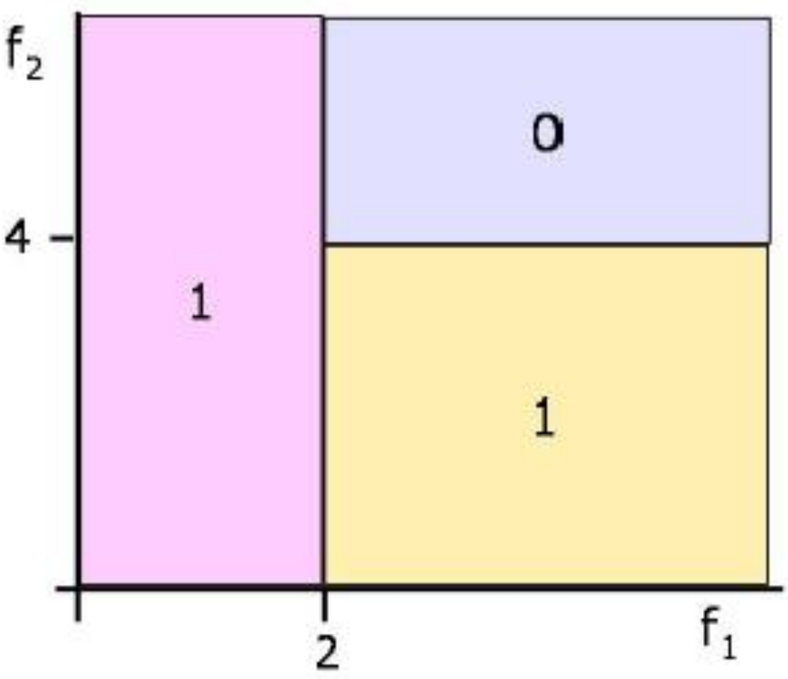




Numerical attributes

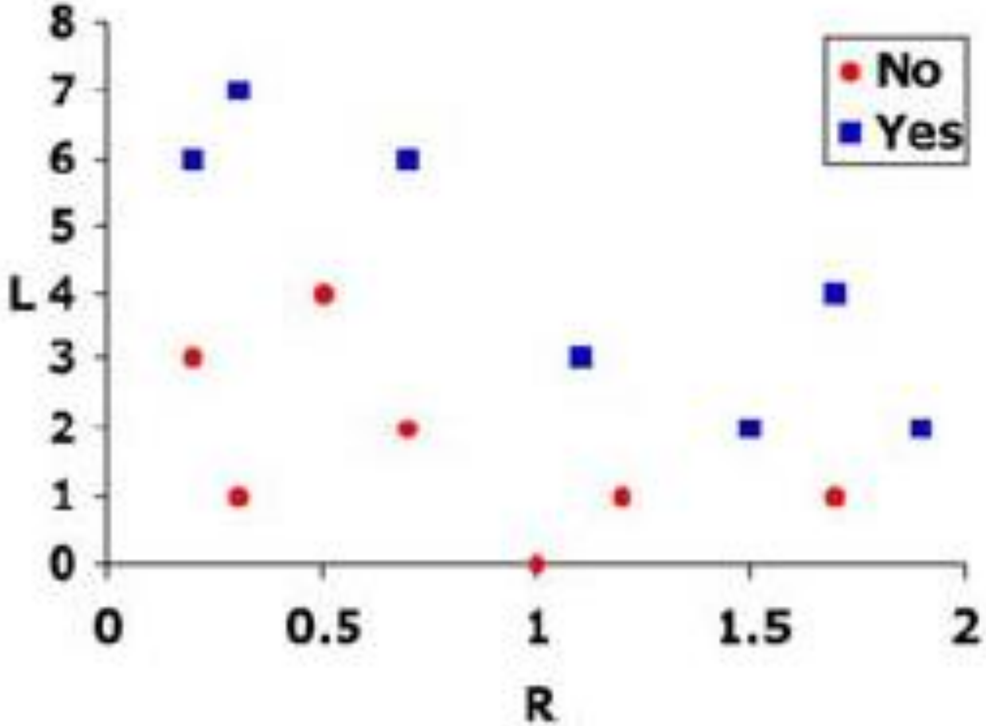
- Tests in nodes can be of the form f_i > constant
- Divides the space into rectangles.





Predicting Bankruptcy

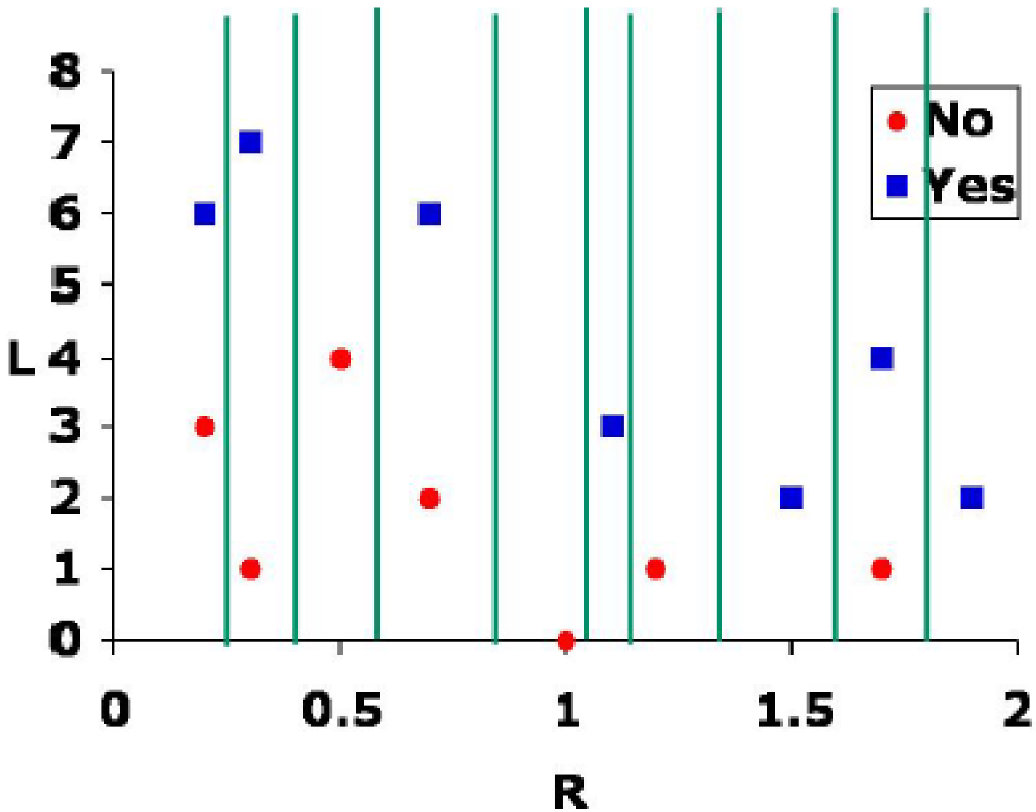
L	R	B
3	0.2	No
1	0.3	No
4	0.5	No
2	0.7	No
0	1.0	No
1	1.2	No
1	1.7	No
6	0.2	Yes
7	0.3	Yes
6	0.7	Yes
3	1.1	Yes
2	1.5	Yes
4	1.7	Yes
2	1.9	Yes



L: #late payments / year
R: expenses / income

Considering splits

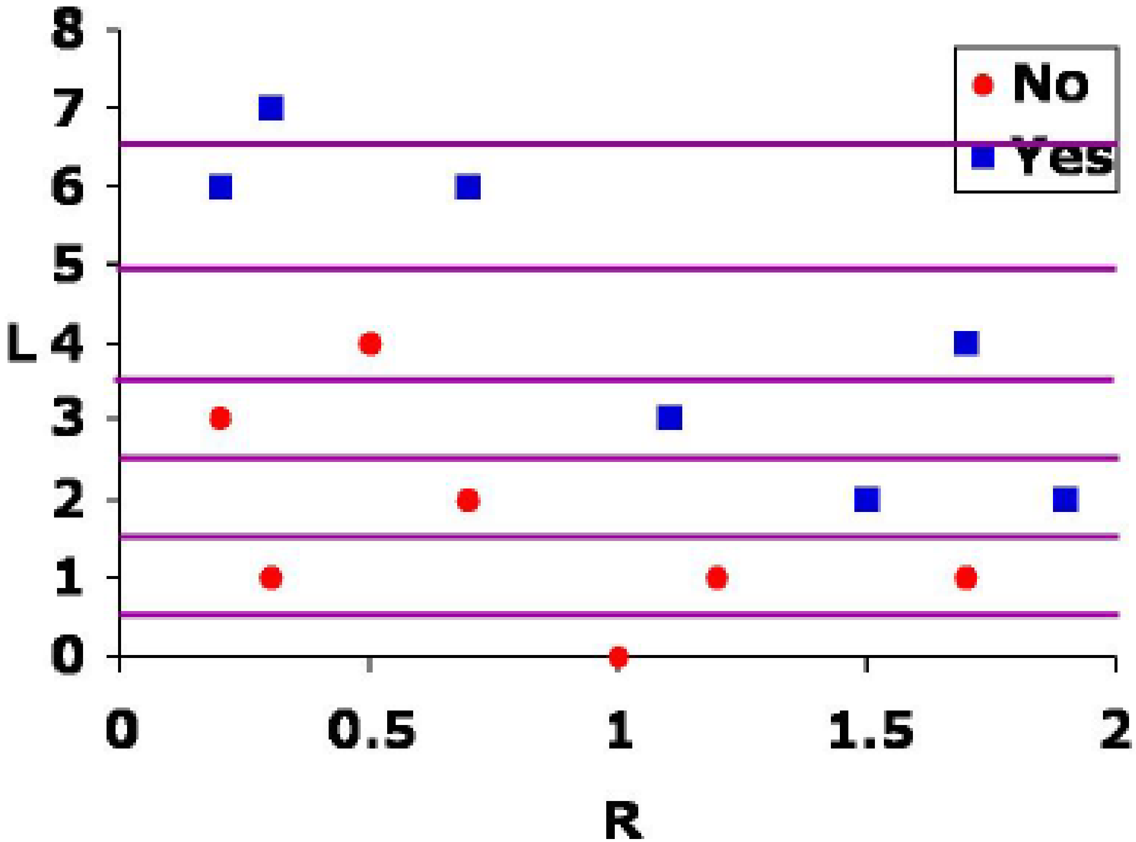
- Consider splitting between each data point in each dimension.



- So, here we'd consider 9 different splits in the R dimension

Considering splits II

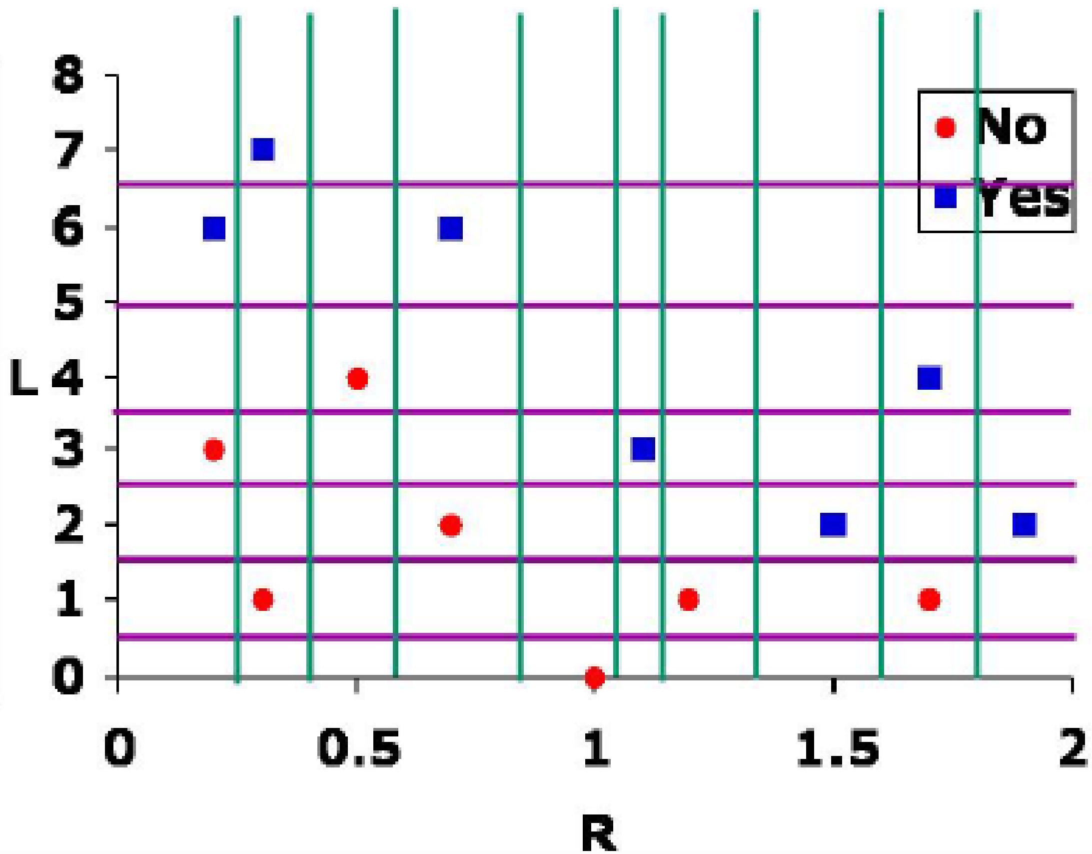
- And there are another 6 possible splits in the L dimension



Bankruptcy Example

L<y	NL	PL	NR	PR	AE
6.5	7	6	0	1	0.93
5.0	7	4	0	3	0.74
3.5	6	3	1	4	0.85
2.5	5	2	2	5	0.86
1.5	4	0	3	7	0.63
0.5	1	0	6	7	0.93

neg to left
pos to left
neg to right
pos to right

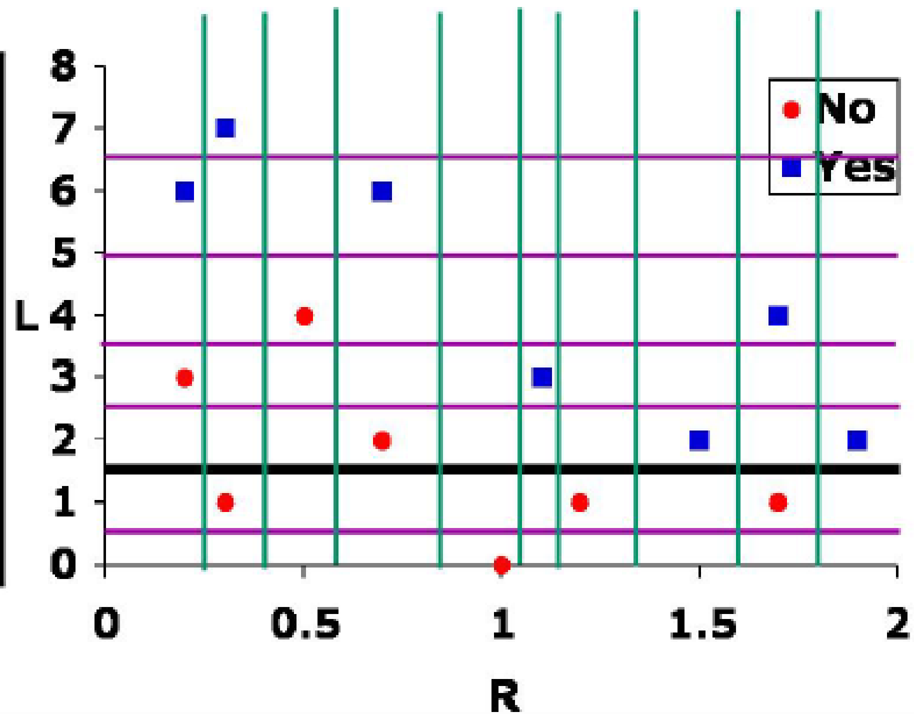


AE	1.00	1.00	0.98	0.98	0.94	0.98	0.92	0.98	0.92
R<x	0.25	0.40	0.60	0.85	1.05	1.15	1.35	1.60	1.80

Bankruptcy Example

- We consider all the possible splits in each dimension, and compute the average entropies of the children.

L<y	NL	PL	NR	PR	AE
6.5	7	6	0	1	0.93
5.0	7	4	0	3	0.74
3.5	6	3	1	4	0.85
2.5	5	2	2	5	0.86
1.5	4	0	3	7	0.63
0.5	1	0	6	7	0.93



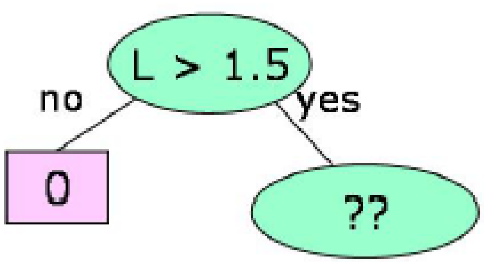
AE	1.00	1.00	0.98	0.98	0.94	0.98	0.92	0.98	0.92
R<x	0.25	0.40	0.60	0.85	1.05	1.15	1.35	1.60	1.80

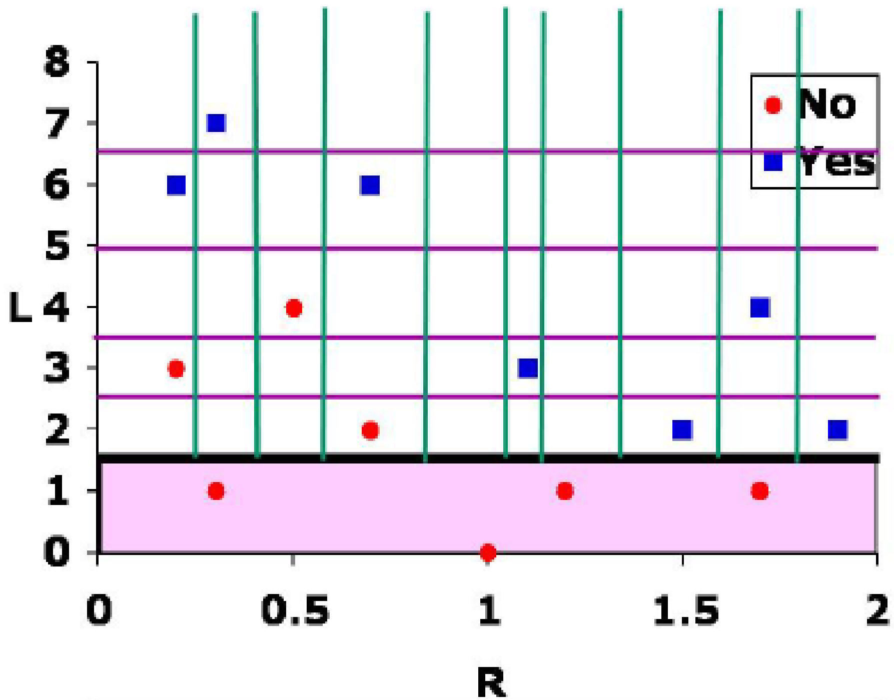
- And we see that, conveniently, all the points with L not greater than 1.5 are of class 0, so we can make a leaf there.

Bankruptcy Example

- Now, we consider all the splits of the remaining part of space.

L<y	NL	PL	NR	PR	AE
6.5	6	3	0	1	0.83
5.0	4	3	0	3	0.69
3.5	3	2	4	1	0.85
2.5	2	1	5	2	0.88



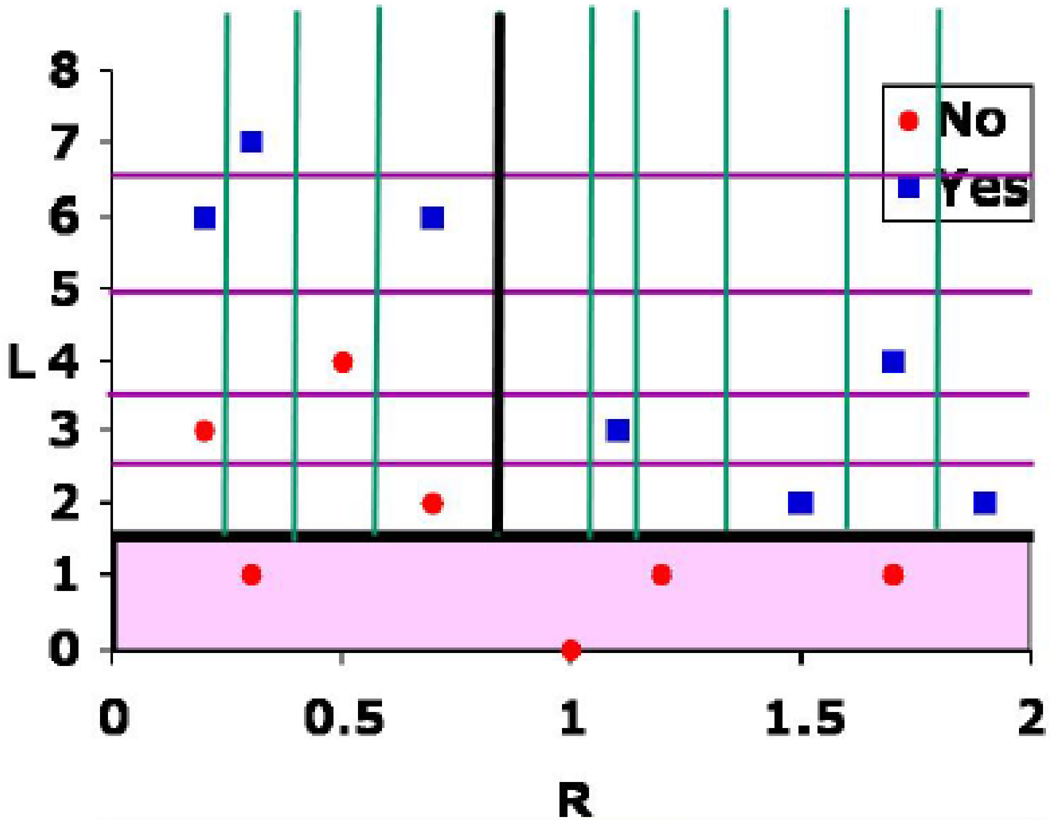


AE	0.85	0.88	0.79	0.60	0.69	0.76	0.83
R<x	0.25	0.40	0.60	0.90	1.30	1.60	1.80

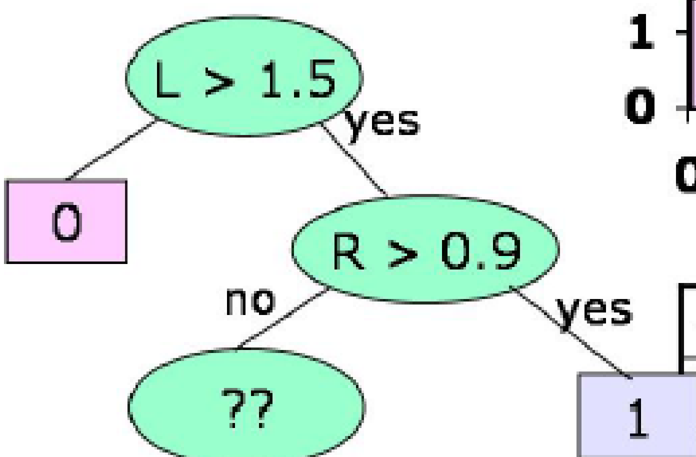
Bankruptcy Example

- Now the best split is at R > 0.9.

L<y	NL	PL	NR	PR	AE
6.5	6	3	0	1	0.83
5.0	4	3	0	3	0.69
3.5	3	2	4	1	0.85
2.5	2	1	5	2	0.88

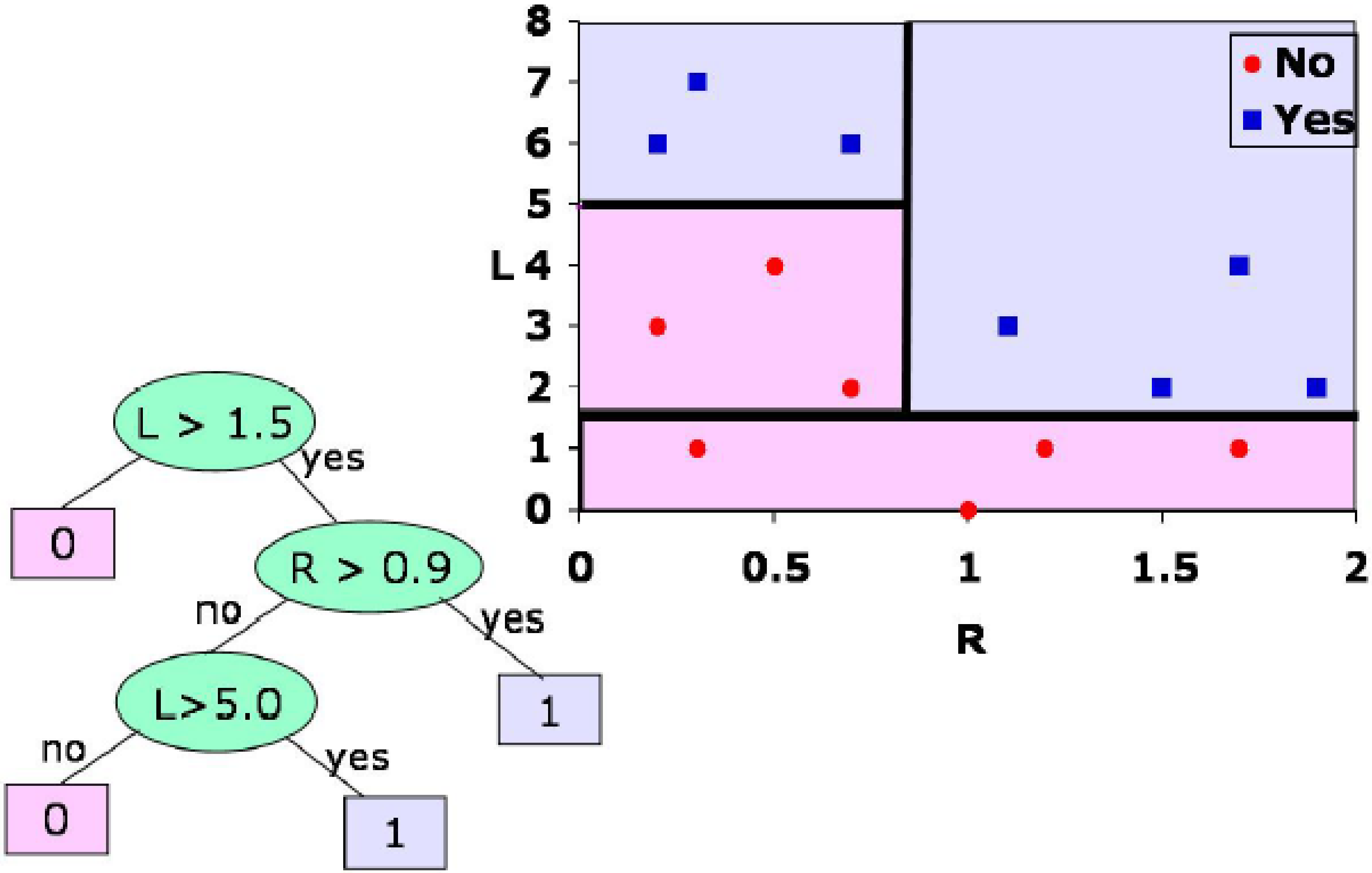


AD	0.85	0.88	0.79	0.60	0.69	0.76	0.83
R<x	0.25	0.40	0.60	0.90	1.30	1.60	1.80



Bankruptcy Example

- Continuing in this way, we finally obtain:



Alternative Splitting Criteria based on GINI

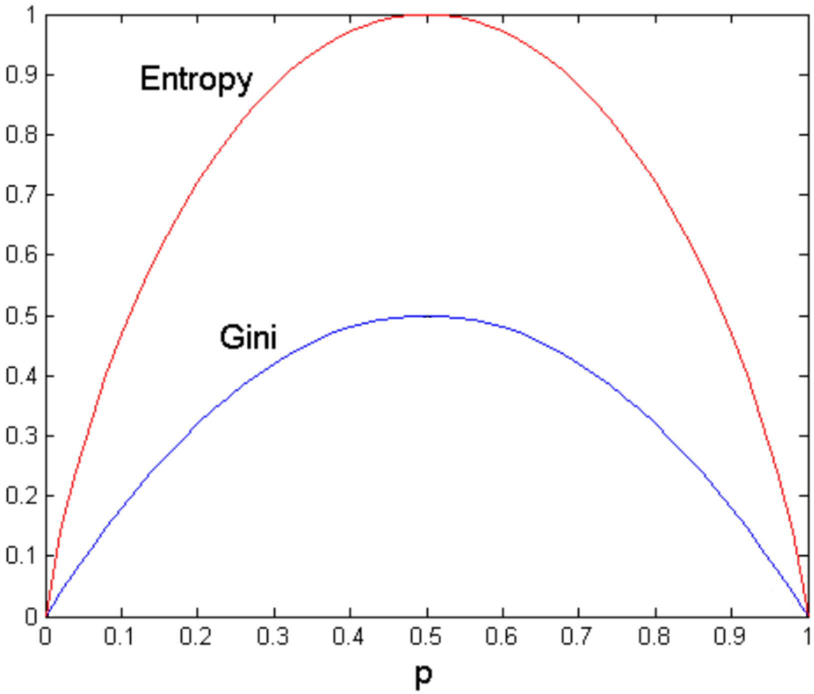
- We have used so far the entropy.
- GINI is an alternative.

$$Entropy(p_1,\ldots,p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

$$GINI(p_1,\ldots,p_n) = 1 - \sum_{i=1}^{n} p_i^2$$

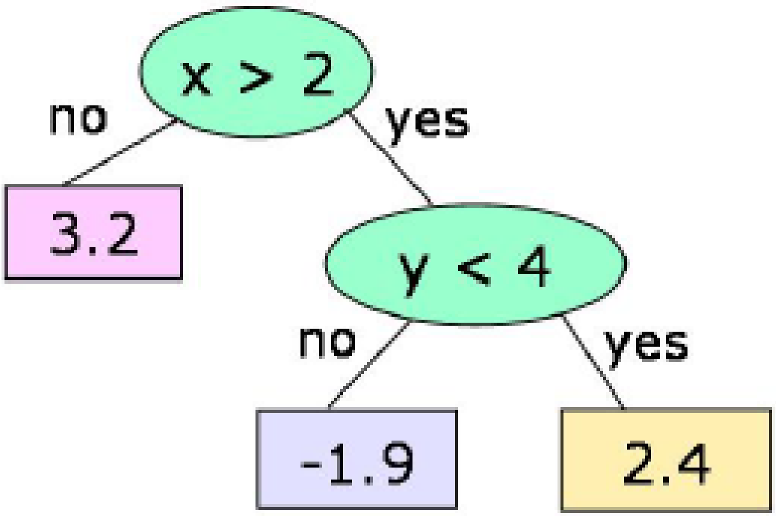
- Both, have:
  - Maximum when records are equally distributed among all classes, implying **least** purity
  - Minimum attained at p equal to 0 or 1, i.e. when all records belong to one class, implying most purity

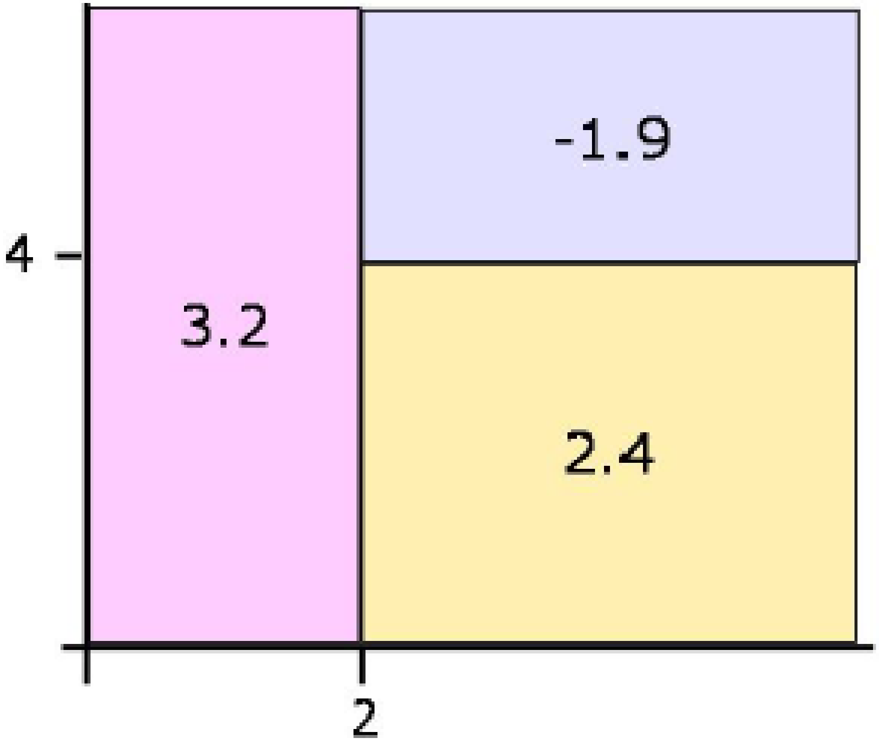
For a 2-class problem:



Regression Trees

- Like decision trees, but with real-valued constant outputs at the leaves.





Leaf values

- Assume that multiple training points are in the leaf and we have decided, for whatever reason, to stop splitting.
  - In the boolean case, we use the majority output value as the value for the leaf.
  - In the numeric case, we'll use the average output value.

- So, if we're going to use the average value at a leaf as its output, we'd like to split up the data so that the leaf averages are not too far away from the actual items in the leaf.

- Statistics has a good measure of how spread out a set of numbers is
  - (and, therefore, how different the individuals are from the average);
  - it's the *variance* of a set.

Variance

- Mean:

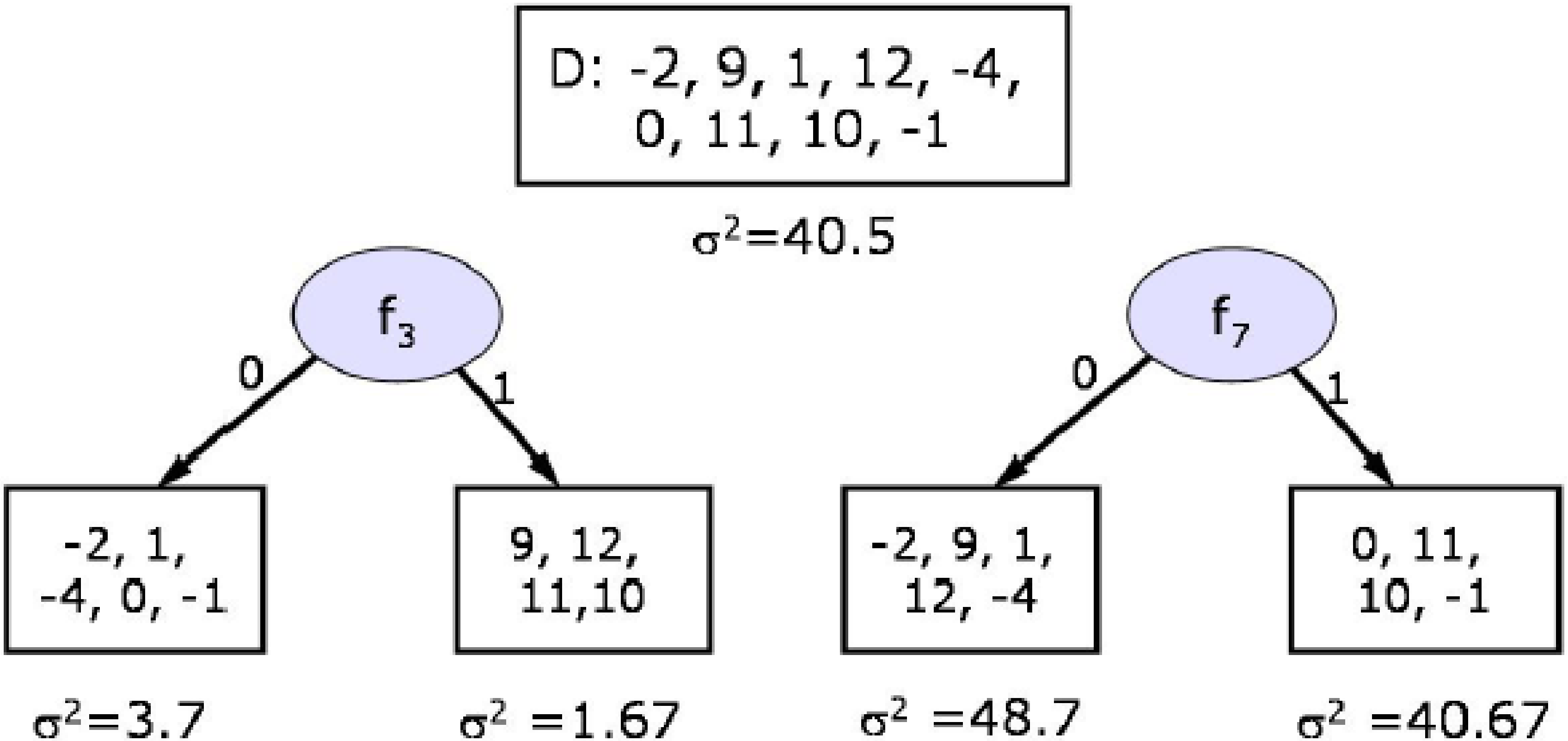
$$\mu = \frac{1}{m}\sum_{k=1}^{m} z_k$$

- Variance (unbiased estimator):

$$\sigma^2 = \frac{1}{m-1}\sum_{k=1}^{m} (z_k - \mu)^2$$

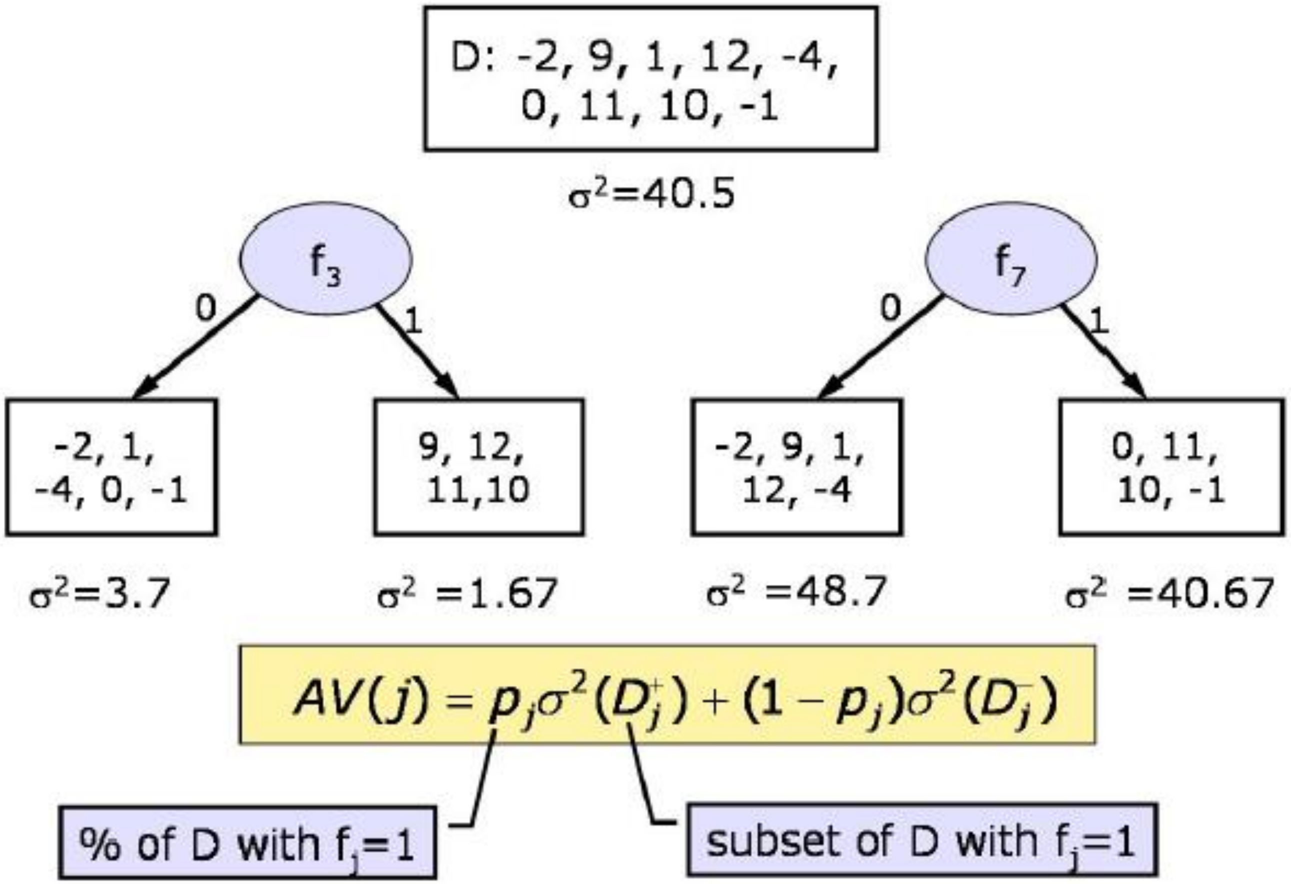
- We will use now the variance instead of entropy.

Splitting



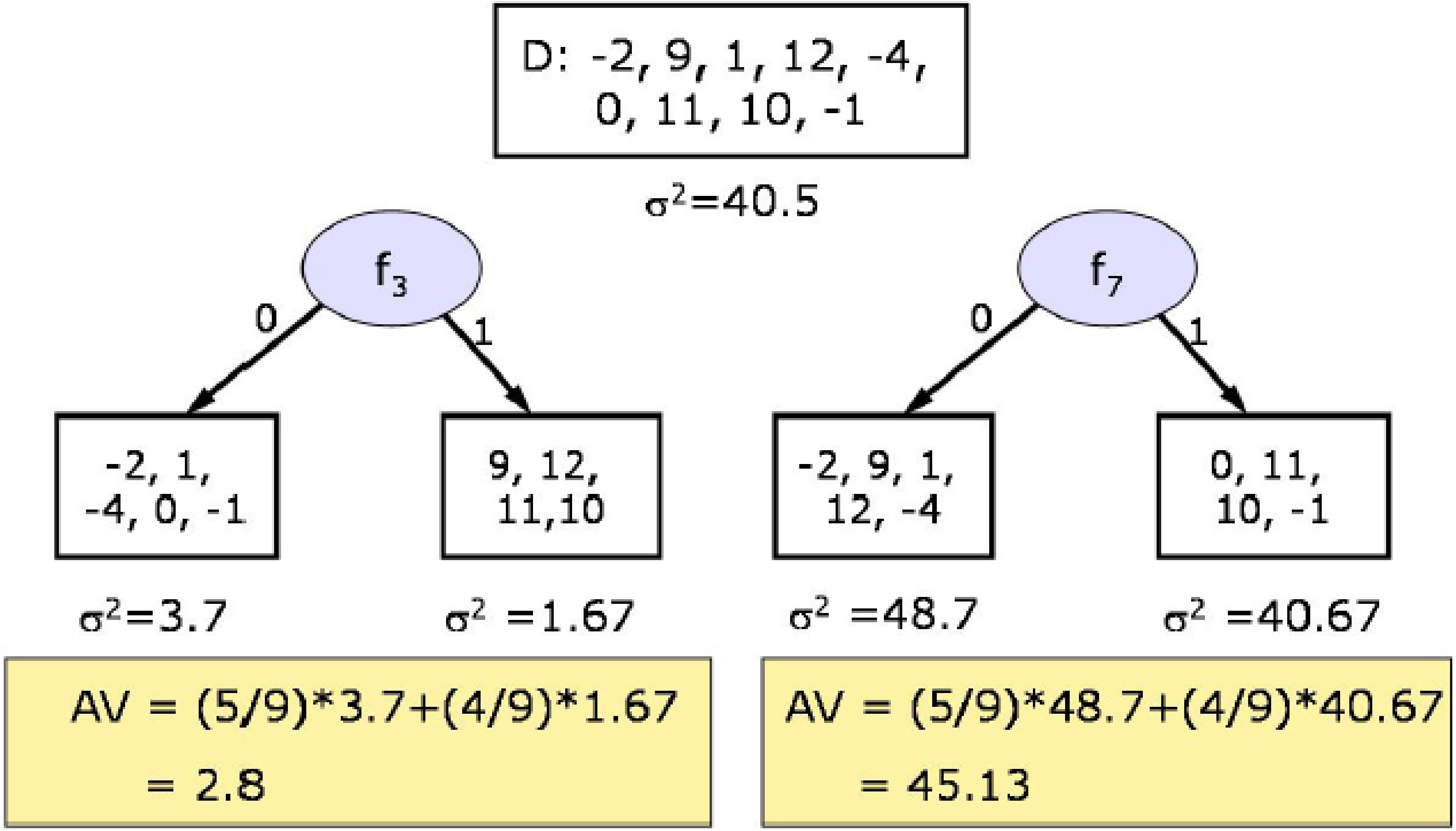
D: -2, 9, 1, 12, -4, 0, 11, 10, -1
σ²=40.5
f₃
0
1
-2, 1, -4, 0, -1
9, 12, 11,10
σ²=3.7
σ² =1.67
f₇
0
1
-2, 9, 1, 12, -4
0, 11, 10, -1
σ² =48.7
σ² =40.67

Splitting



D: -2, 9, 1, 12, -4, 0, 11, 10, -1
σ²=40.5
f₃
0
1
-2, 1, -4, 0, -1
9, 12, 11,10
σ²=3.7
σ² =1.67
f₇
0
1
-2, 9, 1, 12, -4
0, 11, 10, -1
σ² =48.7
σ² =40.67
$AV(j) = p_j\sigma^2(D_j^+) + (1-p_j)\sigma^2(D_j^-)$
% of D with f_j=1
subset of D with f_j=1

Splitting



D: -2, 9, 1, 12, -4,
0, 11, 10, -1
σ2=40.5
f3
0
1
-2, 1,
-4, 0, -1
9, 12,
11,10
σ2=3.7
σ2 =1.67
f7
0
1
-2, 9, 1,
12, -4
0, 11,
10, -1
σ2 =48.7
σ2 =40.67
AV = (5/9)*3.7+(4/9)*1.67
= 2.8
AV = (5/9)*48.7+(4/9)*40.67
= 45.13

Stopping

- Stop when the variance at the leaf is small enough.
- Then, set the value at the leaf to be the mean of the y values of the elements.