Decision Trees and Ensembles

1. In the ID3 family of decision tree algorithms, what is the heuristic we use to choose the next split attribute?

Solution: Maximize information gain, a.k.a., minimize entropy, a.k.a., minimize impurity. That is, at Node m, with N samples in C classes at that node, N_j^i samples belong to class k, and the probability that a sample belongs to class i is

$$p_m^i = \frac{N_m^i}{N_m}$$

and the *impurity* of Node m is its entropy:

$$I'_{ma} = - \sum_{j \in answers(a)} \frac{N_j}{N} \sum_{i \in C} p_j^i \log_2 p_j^i$$

2. Using this heuristic causes a preference bias for the final constructed tree. What is that preference?

Solution: Short trees, i.e., greedy trees, i.e., trees that yield an answer in the least number of steps.

3. Given the following data:

Humidity	Wind	PlayTennis?
High	Weak	No
High	Strong	No
High	Weak	No
Normal	Weak	Yes
Normal	Strong	Yes

using the min-entropy purity criterion, where the entropy after a split at node m on attribute a is given by

$$I'_{ma} = - \sum_{j \in answers(a)} \frac{N_j}{N} \sum_{i \in C} p_j^i \log_2 p_j^i,$$

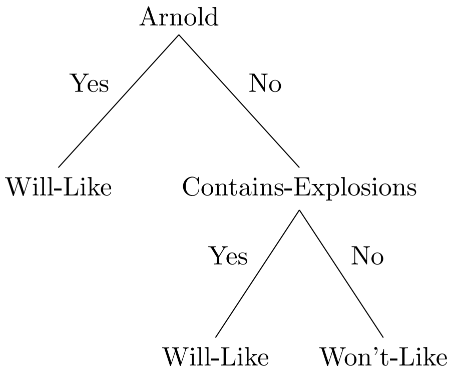
on which attribute should a tree induction algorithm split first? Show all (necessary) calculations.

Solution:

$$\begin{aligned} I'_{humidity} &= -\sum_{j\in\{High,Normal\}} \frac{N_j}{N} \sum_{i\in\{Yes,No\}} p_j^i \log_2 p_j^i \\ &= -[\frac{3}{5}(\frac{0}{3}\log_2\frac{0}{3}+\frac{3}{3}\log_2\frac{3}{3})+\frac{2}{5}(\frac{2}{2}\log_2\frac{2}{2}+\frac{0}{2}\log_2\frac{0}{2})] \\ &= 0 \end{aligned}$$

Since Humidity perfectly classifies the training examples, it has 0 entropy and should be chosen as the split attribute.

4. Given the tree



write down a rule that says whether I will like a movie.

Solution: IF Arnold=Yes OR (Arnold=No AND Explosions=Yes) THEN Will-Like

5. What is the basic idea of an ensemble method?

Solution: Combine multiple learners to reduce the overall bias.

6. What is the core idea in boosting?

Solution: Train base learners on examples they're best suited to classifying correctly.

7. What property must the base learners have for Ada-boost to be effective?

Solution: THe base learners must be weak – classifying just better than random guessing.