# CATEGORICAL FEATURES

- A split on a categorical feature partitions the feature levels:

$$x_j \in \{a, b, c\} \leftarrow \mathcal{N} \rightarrow x_j \in \{d, e\}$$

- For a feature with $m$ levels, there are about $2^m$ different possible partitions of the $m$ values into two groups ($2^{m-1} - 1$ because of symmetry and empty groups).
- Searching over all these becomes prohibitive for large values of $m$.
- For regression with L2 loss and for binary classification, we can define clever shortcuts.

# SURROGATE SPLITS

- Each surrogate split is a decision stump that tries to learn the actual splitting rule
- Consider this tree with the primary split w.r.t. `Sepal.Length` where we perform binary classification (`setosa` vs. `virginica`):


- Our surrogate split should optimize a splitting criterion w.r.t. `Sepal.Length < 5.8`

# SURROGATE SPLITS

- Consider this subsample of the data used to fit the tree:

| | Sepal.Length | ... | Petal.Width | Species | Sepal.Length $< 5.8$ |
|---|---|---|---|---|---|
| 1 | 5.10 | ... | 0.20 | setosa | TRUE |
| 4 | 4.60 | ... | 0.20 | setosa | TRUE |
| 9 | 4.40 | ... | 0.20 | setosa | TRUE |
| 15 | 5.80 | ... | 0.20 | setosa | FALSE |
| 18 | 5.10 | ... | 0.30 | setosa | TRUE |
| 52 | 5.80 | ... | 1.90 | virginica | FALSE |
| 57 | 4.90 | ... | 1.70 | virginica | TRUE |
| 62 | 6.40 | ... | 1.90 | virginica | FALSE |
| 77 | 6.20 | ... | 1.80 | virginica | FALSE |
| 99 | 6.20 | ... | 2.30 | virginica | FALSE |

- Add column that indicates whether `Sepal.Length < 5.8`
- Fit tree of depth 1 using all features but `Sepal.Length` to derive a split that explains `Sepal.Length < 5.8` best $\Rightarrow$ surrogate split
- Typically, software stores the best and a few more surrogate splits