

Decision trees

Oliver W. Layton

CS251: Data analysis and visualization

Lecture 32, Spring 2019

Friday April 26

Plan

- Information entropy and feature selection
- ID3 algorithm

Designing a decision tree: Number of branches

- **What we just learned:** How to judge the quality of *different* split rules for a *fixed* feature. Other tree design considerations:
- Number of branches/splits allowed per node?
 - Often fixed by designer (e.g. always 2). Can vary within tree (e.g. at most M).

Designing a decision tree: Height of tree

- Can implement hard limit on maximum height (e.g. 2 nodes deep).
- Enforce height by purity threshold (node is "pure enough" to become a leaf node).
Example: stop branching if we have 3 or fewer counts in the non-majority classes:

[10, 3, 1]

- We need to update our definition of a **leaf node**.
 - Whether pure or impure, classify data point according to **majority class i** :

[**3**, 0]; [**2**, 1]

- "**Pure enough**" leaf nodes allow decision trees to handle noisy data (Conflicting class labels for same combo of features, Allen's point)

ID3 Algorithm: third iterative dichotomizer

How do we know which feature to put in which node?

1. Compute information gain for each unused feature.
 2. Also use information gain to determine "best" split point **WITHIN** each unused feature.
 3. Assign to the next node the feature with maximal information gain.
 4. Stop when every branch ends with a **leaf node** (pure or "pure enough" node) OR when all the features are used up.
- Let's use ID3 to create a decision tree based on the tennis data.

Tennis data

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No