CSC411 Tutorial #3
Cross-Validation and Decision Trees

February 3, 2016

Boris Ivanovic*

csc411ta@cs.toronto.edu

*Based on the tutorial given by Erin Grant, Ziyu Zhang, and Ali Punjani in previous years.

Outline for Today

- Cross-Validation
- Decision Trees
- Questions

Cross-Validation

Cross-Validation: Why Validate?

So far:

Learning as Optimization

Goal: Optimize model complexity (for the task)
while minimizing under/overfitting

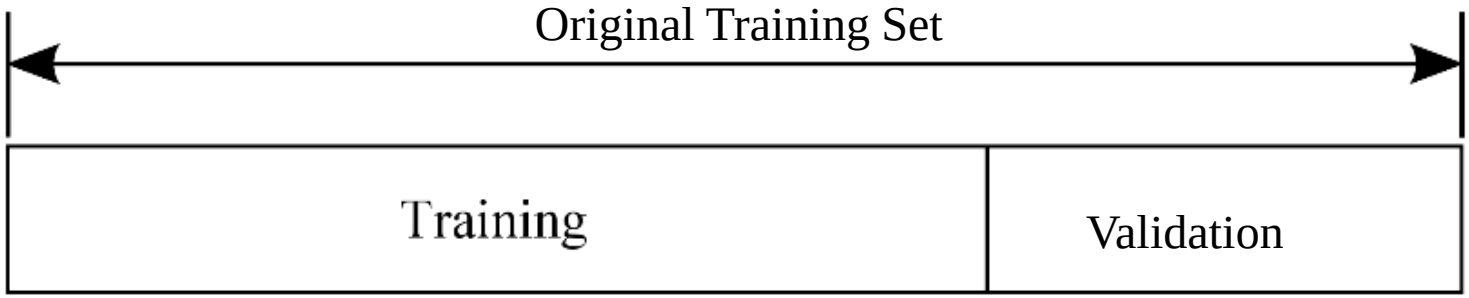
We want our model to **generalize well** without **overfitting**.

We can ensure this by **validating** the model.

Types of Validation

Hold-Out Validation: Split data into training and validation sets.

- Usually 30% as hold-out set.



Problems:

- Waste of dataset
- Estimation of error rate might be misleading

Types of Validation

- **Cross-Validation**: Random subsampling

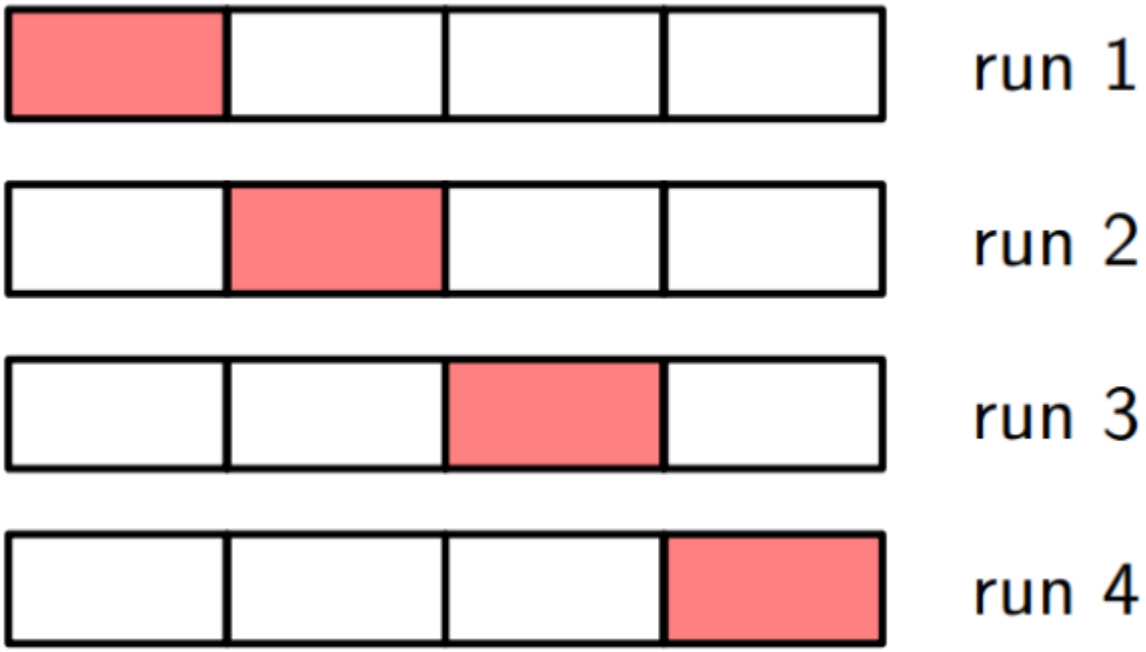


Figure from Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer

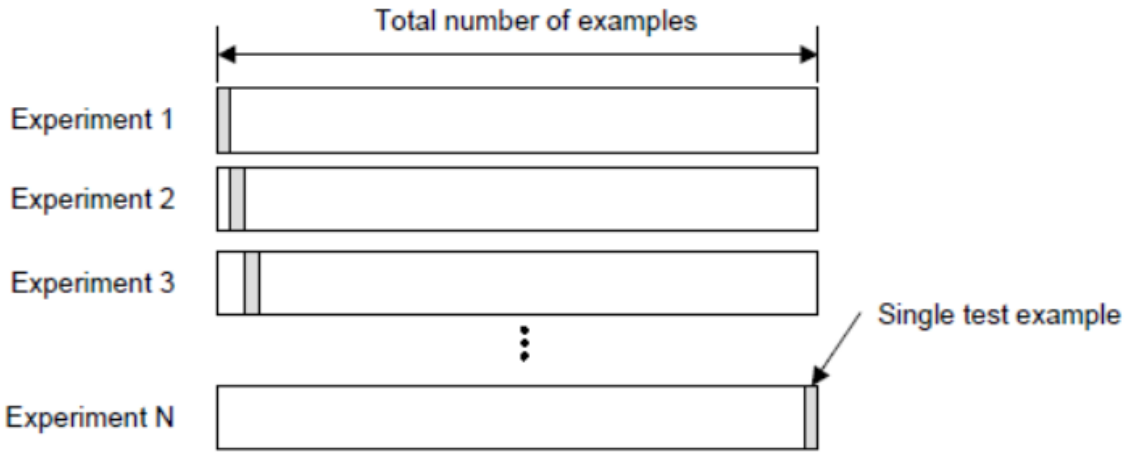
Problem:

- More **computationally expensive** than hold-out validation.

Variants of Cross-Validation

Leave-*p*-out: Use p examples as the validation set, and the rest as training; repeat for all configurations of examples.

e.g., for $p = 1$:

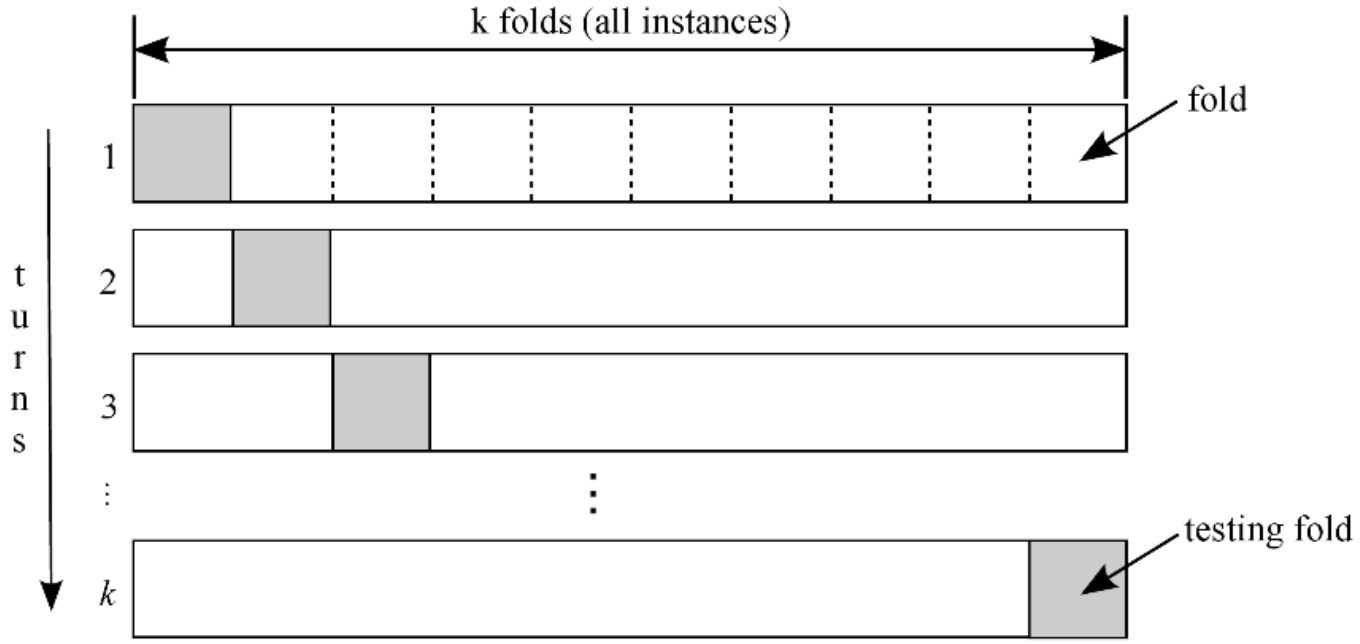


Problem:

- **Exhaustive**. We have to train and test $\binom{N}{p}$ times, where N is the # of training examples.

Variants of Cross-Validation

K-fold: Partition training data into K equally sized subsamples. For each fold, use the other K-1 subsamples as training data with the last subsample as validation.



K-fold Cross-Validation

- Think of it like leave-*p*-out but without combinatoric amounts of training/testing.

Advantages:

- All observations are used for both training and validation. Each observation is used for validation **exactly once**.
- **Non-exhaustive**: More tractable than leave-*p*-out

K-fold Cross-Validation

Problems:

- **Expensive** for large *N, K* (since we train/test *K* models on *N* examples).
  - But there are some efficient hacks to save time…
- Can still **overfit** if we validate too many models!
  - **Solution**: Hold out an additional test set before doing any model selection, and check that the best model performs well on this additional set (*nested cross-validation*). => Cross-Validception

Practical Tips for Using K-fold Cross-Val

Q: How many folds do we need?

A: With **larger** *K*, …

- Error estimation tends to be **more accurate**
- But, computation time will be **greater**

In practice:

- Usually use $\boldsymbol{K \approx 10}$
- BUT, larger dataset => choose **smaller *K***

Questions about Validation

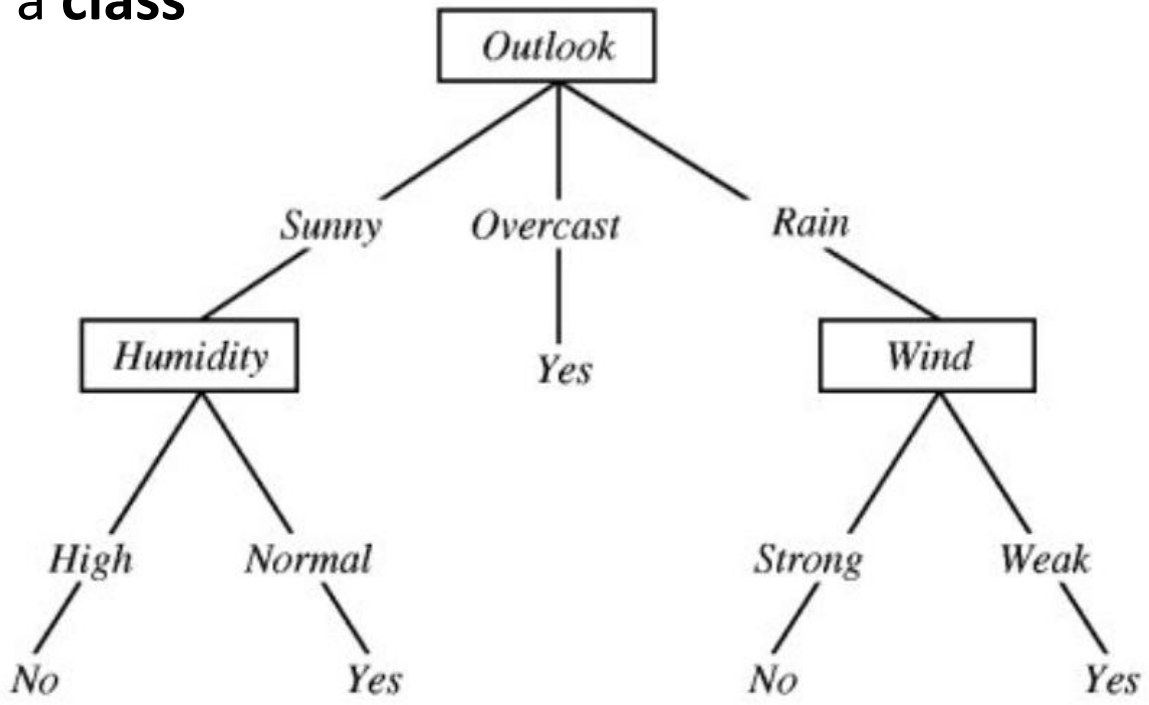
Decision Trees

Decision Trees: Definition

Goal: Approximate a discrete-valued target function

Representation: A tree, of which

- Each internal (non-leaf) node tests an **attribute**
- Each branch corresponds to an **attribute value**
- Each leaf node assigns a **class**



Example from Mitchell, T (1997). *Machine Learning*. McGraw Hill.

Decision Trees: Induction

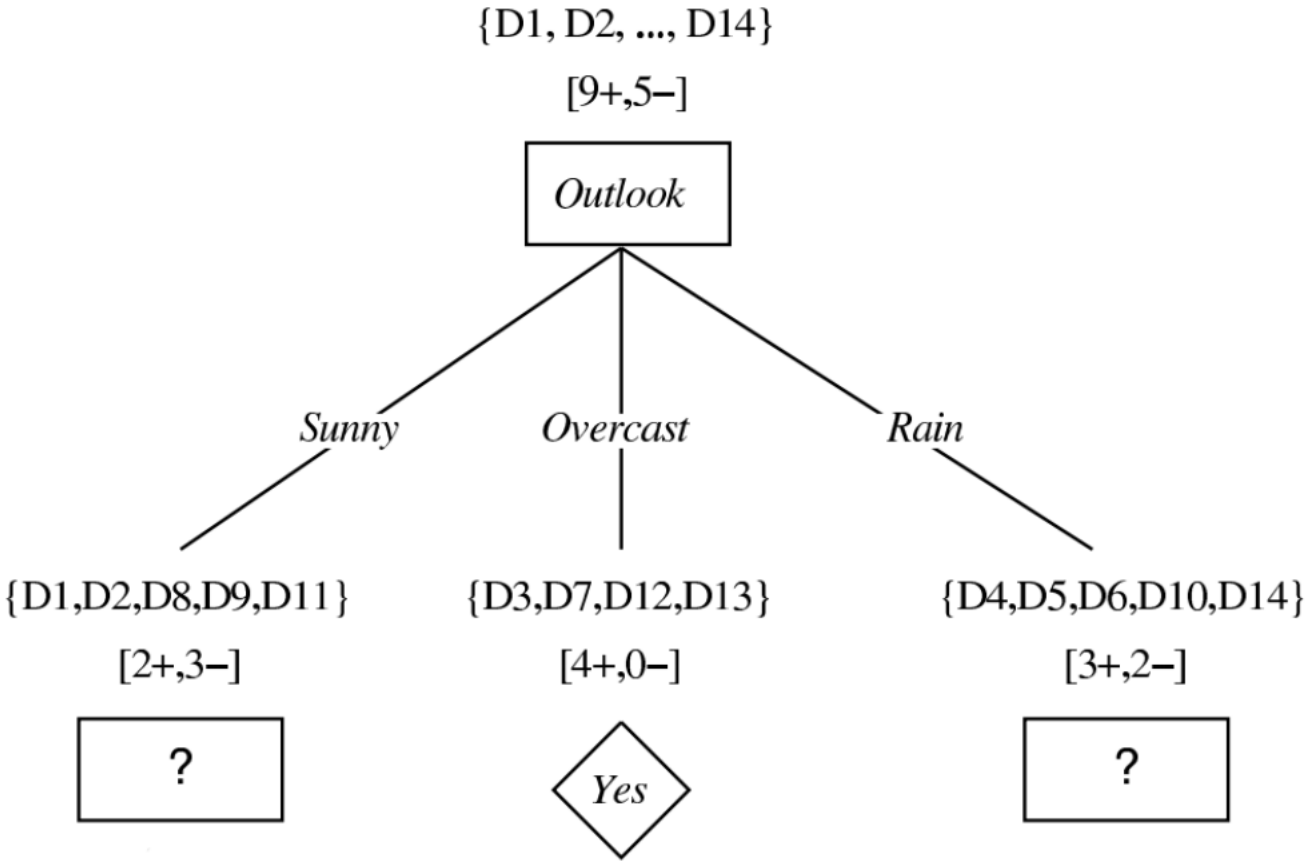
The ID3 Algorithm:

```
while ( training examples are not perfectly classified ) {
        choose the "most informative" attribute θ (that
        has not already been used) as the decision
        attribute for the next node N (greedy selection).

        foreach ( value (discrete θ) / range (continuous θ) )
                create a new descendent of N.

        sort the training examples to the descendants of N
}
```

Decision Trees: Example *PlayTennis*

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	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

After first splitting the training examples on Outlook…



- What should we choose as the next attribute under the branch *Outlook* = Sunny?

Choosing the "Most Informative" Attribute

Formulation: Maximize information gain over attributes *Y*.

$$\text{Information Gain}\,(PlayTennis \mid Y)$$

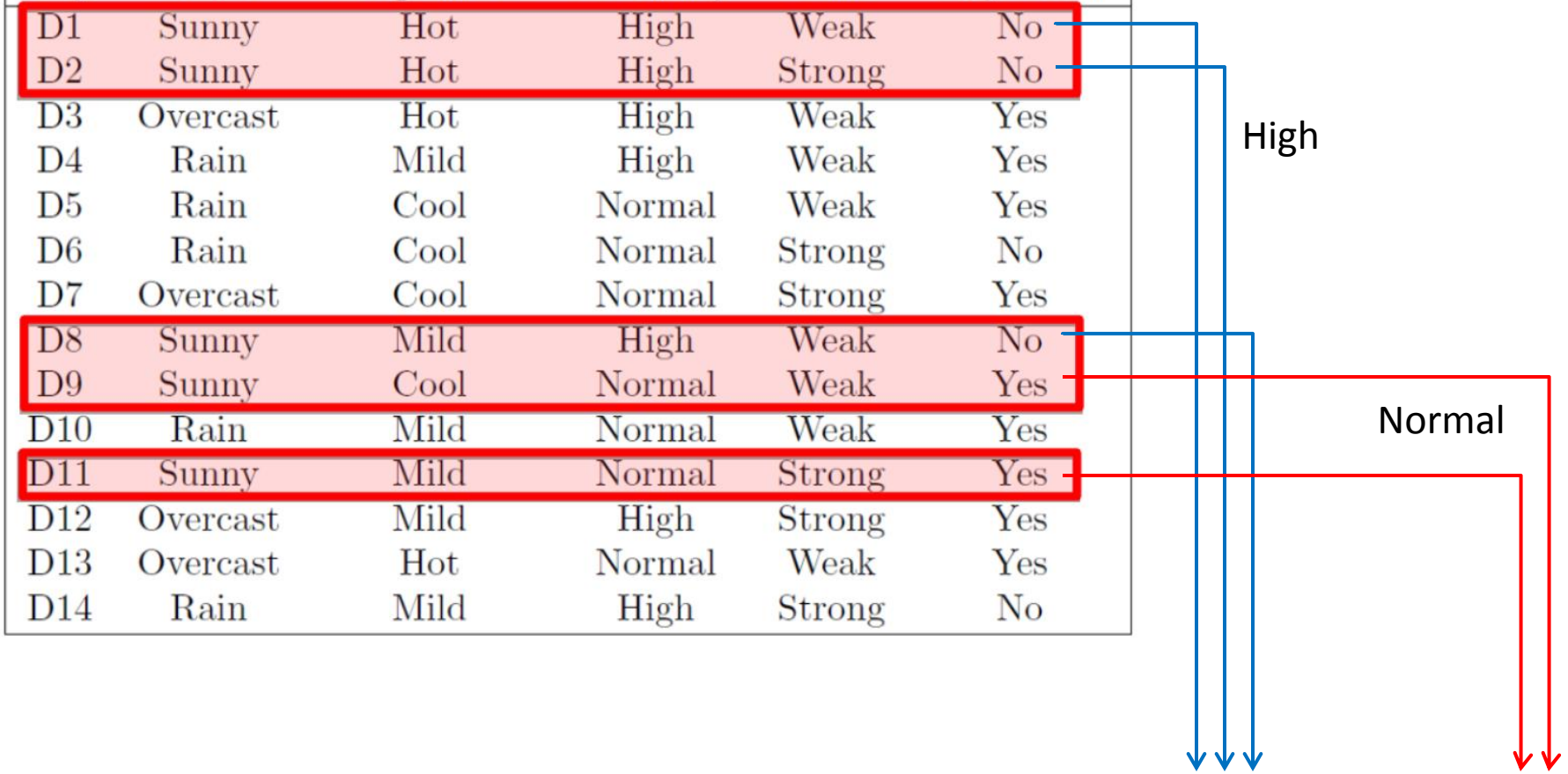
$$= H(PlayTennis) - H(PlayTennis \mid Y)$$

$$= \overbrace{\sum_x P(PlayTennis = x) \log P(PlayTennis = x)}^{H(PlayTennis)}$$

$$\underbrace{- \sum_{x,y} P(PlayTennis = x, Y = y) \log \frac{P(Y = y)}{P(PlayTennis = x, Y = y)}}_{H(PlayTennis \mid Y)}$$

Information Gain Computation #1

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	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



- IG(*PlayTennis* | Humidity) $= 0.970 - \frac{3}{5}(0.0) - \frac{2}{5}(0.0)$
$= 0.970$

Information Gain Computation #2

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	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

3 values b/c Temp takes on 3 values!

- IG(*PlayTennis* | Temp) $= 0.970 - \frac{2}{5}(0.0) - \frac{2}{5}(1.0) - \frac{1}{5}(0.0)$

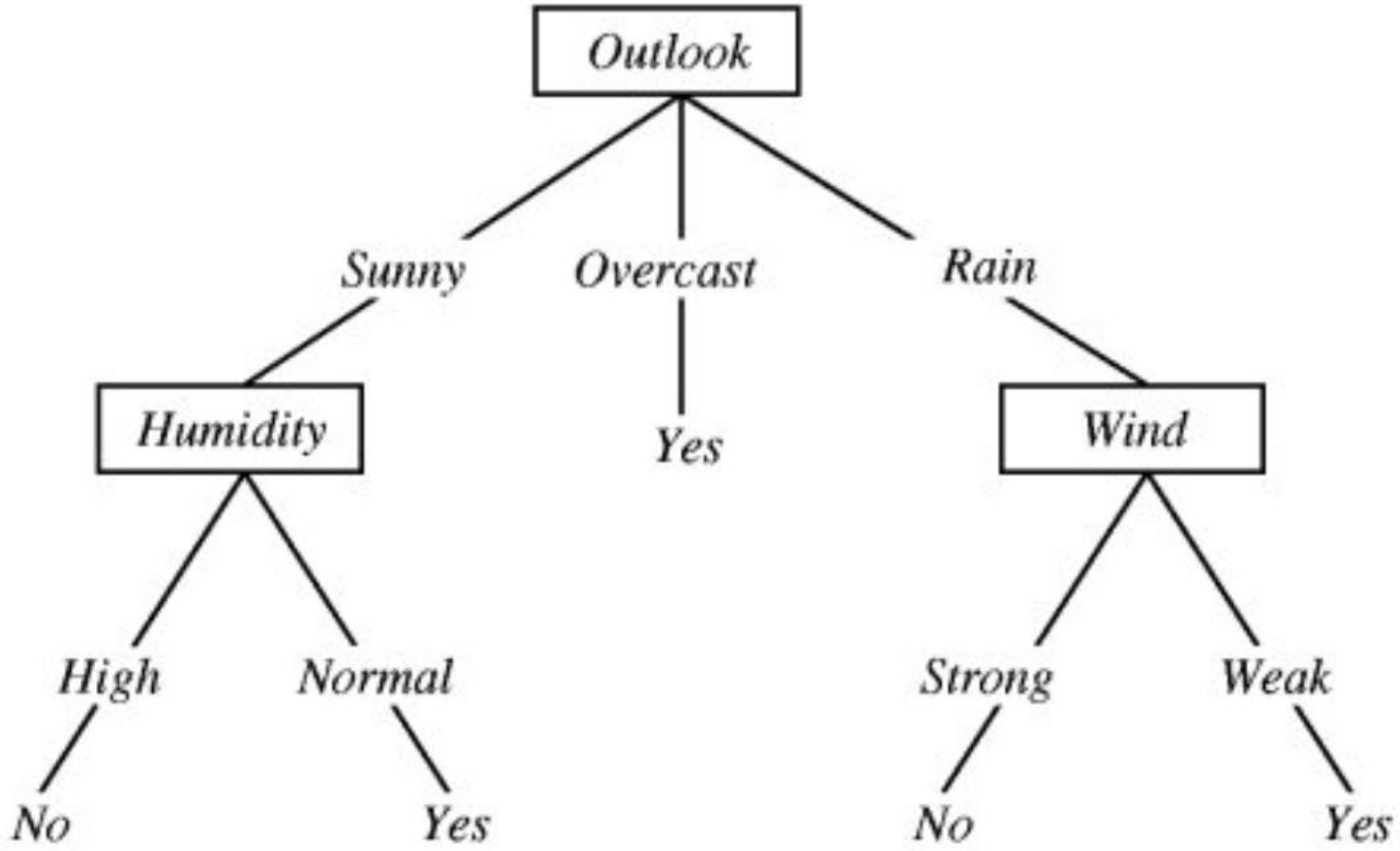
$= 0.570$

Information Gain Computation #3

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	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

- IG(*PlayTennis* | Wind) $= 0.970 - \frac{2}{5}(1.0) - \frac{3}{5}(0.918)$
$= 0.019$

The Decision Tree for *PlayTennis*



Questions about Decision Trees

Feedback (Please!)

boris.ivanovic@mail.utoronto.ca

- So… This was my first ever tutorial!
- I would really appreciate some feedback about my teaching style, pacing, material descriptions, etc…
- Let me know any way you can, tell me in person, tell Prof. Fidler, email me, etc…
- Good luck with A1!