CPSC 340:
Machine Learning and Data Mining

Feature Selection

BONUS SLIDES

Bonus Slide: Causal Discovery with Interventions

- Consider observing two variables dependent ‘i’ and ‘j’.
- In general, we can’t distinguish between these statements:
  - Dependency is due to ‘i’ having a causal effect on ‘j’.
  - Dependency is due to ‘j’ having a causal effect on ‘i’.
  - Dependency is due to a common cause.
- We can distinguish them using interventional data:
  - “Force” variable ‘i’ to have certain values, then measure effect on ‘j’.
  - “Force” variable ‘j’ to have certain values, then measure effect on ‘i’.

Bonus Slide: Causal Discovery with Interventions

- If dependency is only due to common cause:
  - Variables should be independent in the interventional data.
- If dependency is due to a causal effect:
  - Variables should be independent in one direction but not the other.

- This is the basis for randomized control trial:
  - To see if a medical treatment really works.
  - Randomly assign treatment to “forces” value of “treatment” variable.

Feature Selection Approach 1: Hypothesis Testing

- Hypothesis testing ("constraint-based") approach:
  - Performs a sequence of conditional independence tests.

$X_j \perp Y \mid X_S$

feature 'j' — label — other features 'S'

"If I know features in 'S' does feature 'j' tell me anything about label?"

  - If they are independent, say that 'j' is "irrelevant".
- Common way to do the tests:
  - "Partial" correlation (numerical data).
  - "Conditional" mutual information (discrete data).

Hypothesis Testing

- Hypothesis testing ("constraint-based") approach:
  - Performs a sequence of conditional independence tests.

$X_j \perp Y \mid X_S$

"feature 'j'" "label" "other features 's'"

"If I know features in 's' does feature 'j' tell me anything about label?"

  - If they are independent, say that 'j' is "irrelevant".
- Two many possible tests, "greedy" method is for each 'j' do:

First test if $X_j \perp Y$

If still dependent test $X_j \perp Y \mid X_S$ where 's' has one feature

If still dependent test $X_j \perp Y \mid X_S$ where 's' has one more feature

Often choose features to minimize dependence.

⋮

If still dependent when 's' includes all other features, declare 'j' relevant.

Hypothesis Testing Issues

- Advantages:
  - Deals with conditional independence.
  - Algorithm can explain why it thinks ‘j’ is irrelevant.
  - Doesn’t necessarily need linearity.
- Disadvantages:
  - Deals badly with variable dependence: doesn’t select “mom” or “mom2” if both present.
  - Usual warning about testing multiple hypotheses:
    - If you test $p < 0.05$ more than 20 times, you’re going to make errors.
  - Greedy approach may be sub-optimal.
- Neither good nor bad:
  - Allow tiny effects.
  - Says “gender” is irrelevant when you know “baby”.
  - This approach is better for finding relevant factors, not to select features for learning.