Class Project Progress (11/21)

Overview

This week, I successfully implemented H2: TAN (Tree-Augmented Naive Bayes) into the project pipeline. This represents the next advancement after H1 (standard Naive Bayes), allowing the model to capture limited dependencies among genes while still maintaining a simple Bayesian Network structure. TAN relaxes the independence assumptions of Naive Bayes by allowing each gene to have one additional parent gene, selected based on conditional mutual information. This creates a more expressive model without the complexity of a fully learned Bayesian Network.

What was Implemented

1. Tan Structure Learning (H2)

   For each fold, the algorithm

   - Selects the best direction for the Stage -> Y or Y -> Stage relationship using BIC scoring
   - Starts with the Naive Bayes backbone (each gene depends on Y).
   - Computes conditional mutual information for each gene pair given Y.
   - Adds one additional parent to each gene when beneficial.
2. Cross-Validation for TAN
   - 5-fold stratified group cross-validation was used (grouped by patient ID).

   Each fold prints

   - Best Stage/Y direction
   - Accuracy, BIC, and log-likelihood
   - Final mean performance metrics calculated
3. TAN structure visualization

   The network graph that was generated shows

   - Y and Stage as central nodes
   - Genes with arrows from Y
   - Some genes having arrows from other genes (TAN edges)

Output

1. Stage -> Y Direction
   - Across all 5 folds, the model consistently selected Y-> Stage. This means that the class variable (tumor vs normal) is a better parent for Stage than the reverse.

This matches the intuition that cancer stage is conditionally related to tumor/normal status.

2. Accuracy
    - Average accuracy: 0.9500
    - Standard Deviation: 0.0486 This is comparable to H1 (0.9583), showing that while TAN adds complexity, it did not dramatically change performance. This stability is expected because the dataset is small and the class signal is strong.
3. BIC Score
    - Mean BIC: -364.75 Compared to H1:
    - H1 BIC: -311.18 A lower (more negative) BIC suggests that TAN introduces more parameters (penalizes complexity but still fits the data well).
4. Structure Visualization

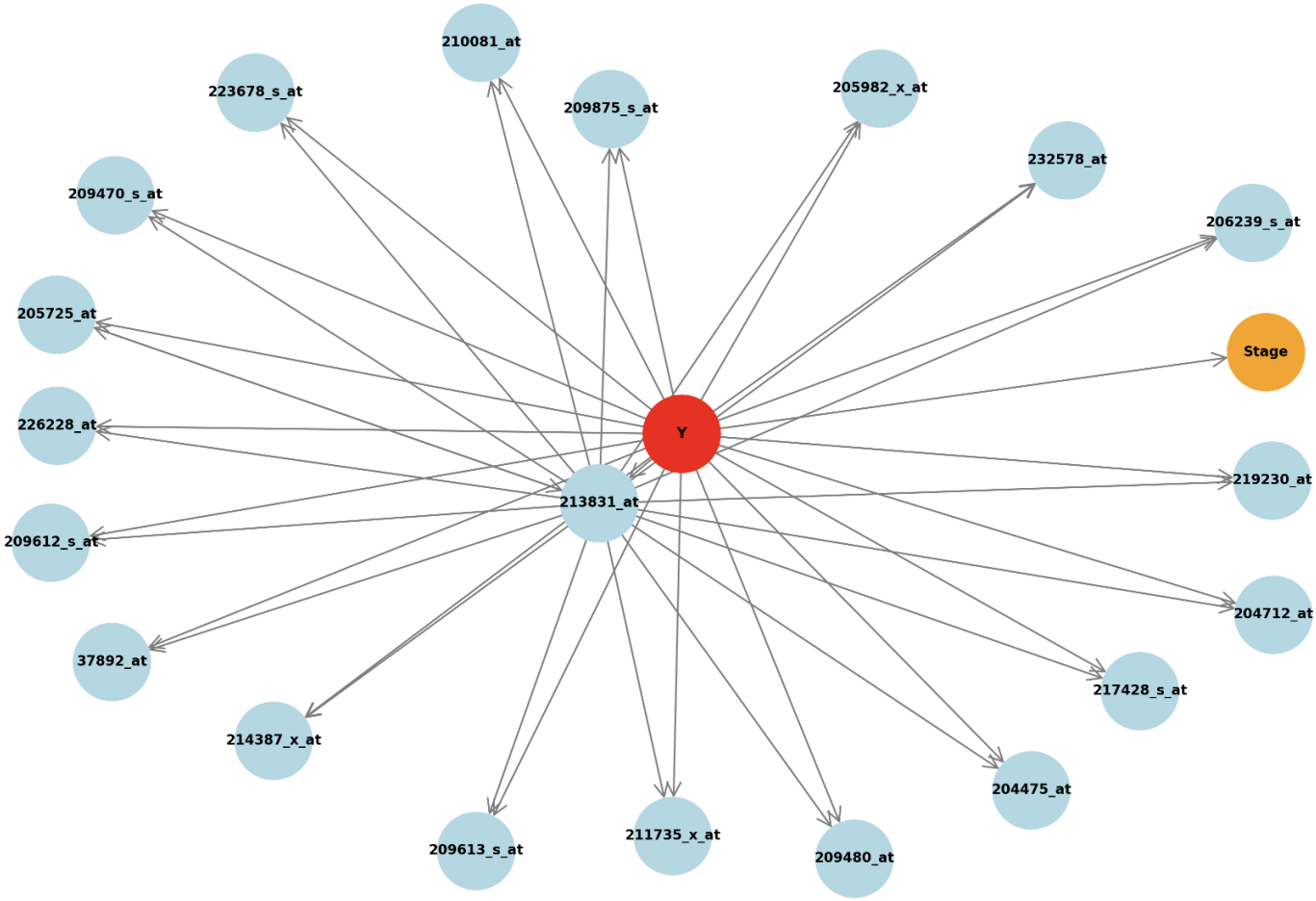
    The TAN network shows:

    - Y as the dominant parent of nearly all genes (as expected in medical classification).
    - Several genes gaining extra edges from other genes, showing mild correlations.
    - Stage connected only to Y, consistent across folds. This visualization confirms that the TAN structure is being learned correctly.

Next Steps

- Implement H3: Fully learned bayesian network structure
- Final comparison of H1, H2, & H3

H2: TAN - Bayesian Network Structure



```
CROSS-VALIDATING H2: TAN

 Fold 1/5

Stage direction for H2
  Stage->Y BIC: -3154.86, Y->Stage BIC: -995.27
  Selected: Y -> Stage
Accuracy: 1.0000, BIC: -358.56, LL: -223.49

 Fold 2/5

Stage direction for H2
  Stage->Y BIC: -3149.61, Y->Stage BIC: -993.89
  Selected: Y -> Stage
Accuracy: 0.9167, BIC: -390.65, LL: -252.41

 Fold 3/5

Stage direction for H2
  Stage->Y BIC: -3196.86, Y->Stage BIC: -1036.08
  Selected: Y -> Stage
Accuracy: 0.9583, BIC: -337.50, LL: -199.26

 Fold 4/5

Stage direction for H2
  Stage->Y BIC: -3203.46, Y->Stage BIC: -1046.76
  Selected: Y -> Stage
Accuracy: 1.0000, BIC: -345.80, LL: -207.55

 Fold 5/5

Stage direction for H2
  Stage->Y BIC: -3230.94, Y->Stage BIC: -1074.25
  Selected: Y -> Stage
Accuracy: 0.8750, BIC: -391.23, LL: -252.98

H2: TAN Results:
  Accuracy: 0.9500 ± 0.0486
  BIC: -364.75 ± 22.41
  LL: -227.14 ± 22.27
Saved: results/H2:_TAN_structure.png

Model Performance Summary (H1 vs H2):
H1: Naive Bayes: Acc=0.9583 (±0.0527), BIC=-311.18
H2: TAN        : Acc=0.9500 (±0.0486), BIC=-364.75
```