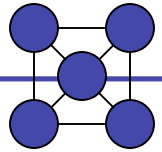
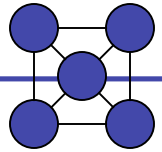


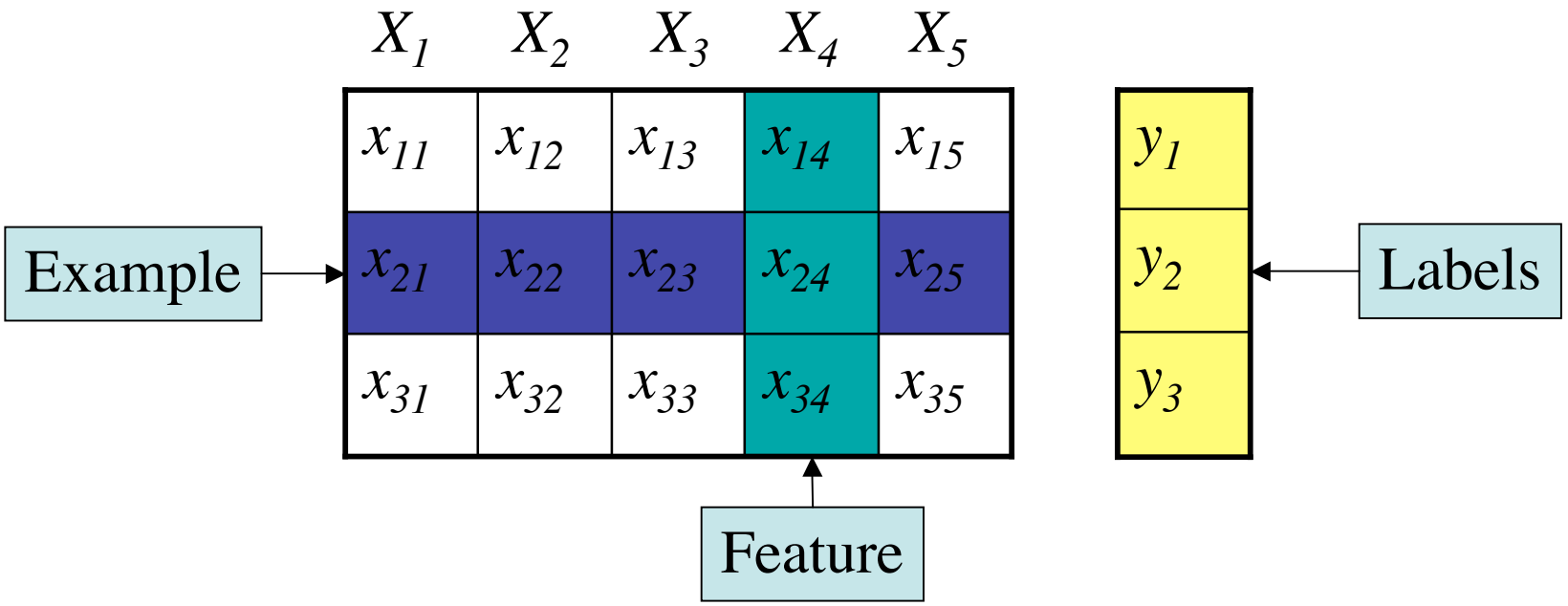
Toward Suboptimal Feature Selection

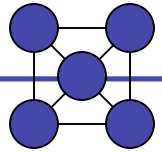
Karen Glocer



Examples, Labels, and Features

- Suppose you have a matrix of labeled data.
- Each row in the matrix is an example and each column is a feature.

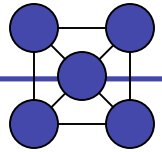




Feature Selection

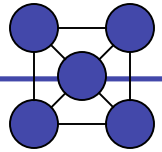
It is often beneficial to only use a subset of features.

- *Curse of Dimensionality*: The number of examples you need to effectively learn a decision surface grows exponentially with the dimensionality of your problem.
- *Overfitting*: If your model is too complex for your data, you end up with a classifier that does not generalize well.



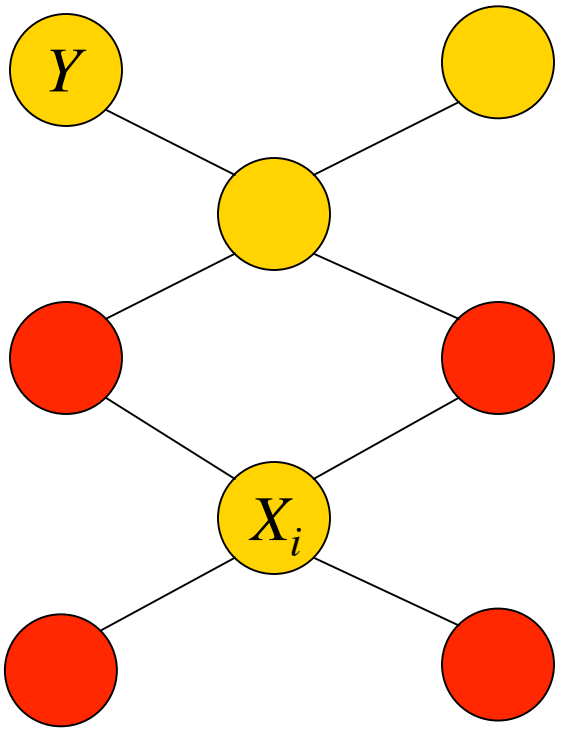
Optimality

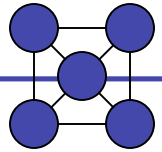
- Feature selection algorithms are generally judged by generalization error.
- Optimality is rarely mentioned in the feature selection literature, partly because it is not well defined.
- Koller and Sahami, the first to propose the Markov blanket criterion for feature selection, used the term “optimal feature selection” in the title. Claim:
  - MB only remove attributes that are really unnecessary.
  - MB remove all attributes that are really unnecessary.
- Tsamardinos and Aliferis make this more explicit. Claim: The Markov Blanket problem is the optimal solution to the feature selection problem when the data is drawn from a faithful BN.



Markov Blankets

- Definition: Let ***M*** be some set of features which does not contain X_i. We say that ***M*** is a Markov Blanket for X_i if X_i is conditionally independent of of $\mathcal{X}$-***M***-$\{X_i\}$ given ***M***.
- In a Markov model, the Markov blanket of a feature is its neighbors.





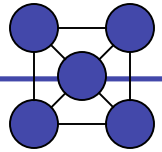
Markov Blanket Approximation Algorithm (Koller and Sahami)

1. Compute correlation matrix for features.
2. For each feature X_i, choose the k features most correlated with X_i to be the Markov blanket $\boldsymbol{M_i}$.
3. Compute the expected relative entropy

$$\delta(X_i \mid M_i) = \sum_{X_{M_i}, X_i} P(M_i = m_i, X_i = x_i) \cdot$$

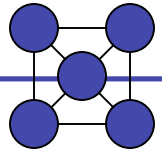
$$D(P(Y \mid M_i = m_i, X_i = x_i), P(Y \mid M_i = m_i))$$

4. Remove the feature that minimizes this quantity.
5. Repeat as until you have removed n features.



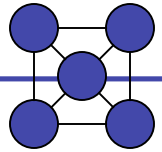
What's Wrong with the MB Approximation Algorithm?

- The approximation algorithm only returns a set of selected features, not a classifier.
- You have to pre-specify the number of features you want to remove and the size of the Markov blankets you want to approximate.
- There is no inherent stopping point: You must specify either the number of features to remove or a threshold for the expected relative entropy.

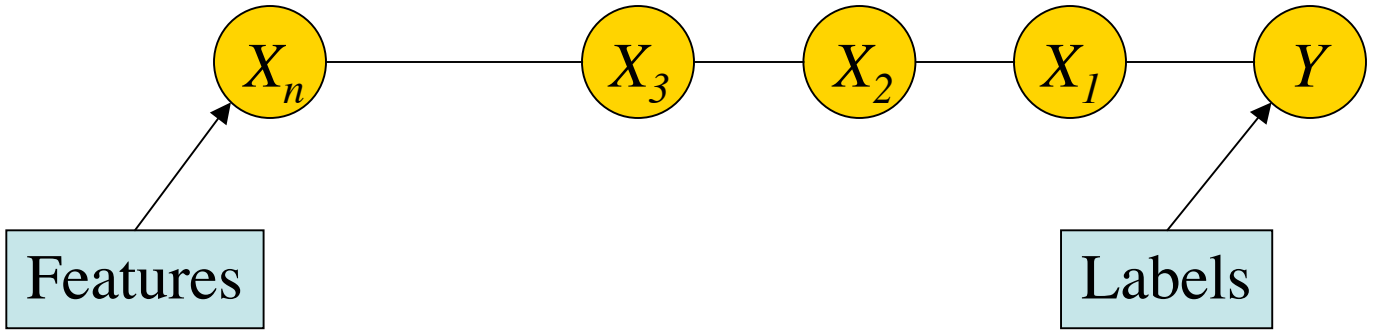


New Goal

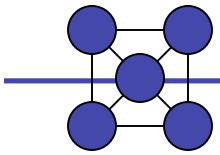
- Assume a hypothetical algorithm removes a feature if and only if it has a Markov Blanket.
- Find a counter-example where this algorithm doesn't eliminate the optimal number of features.



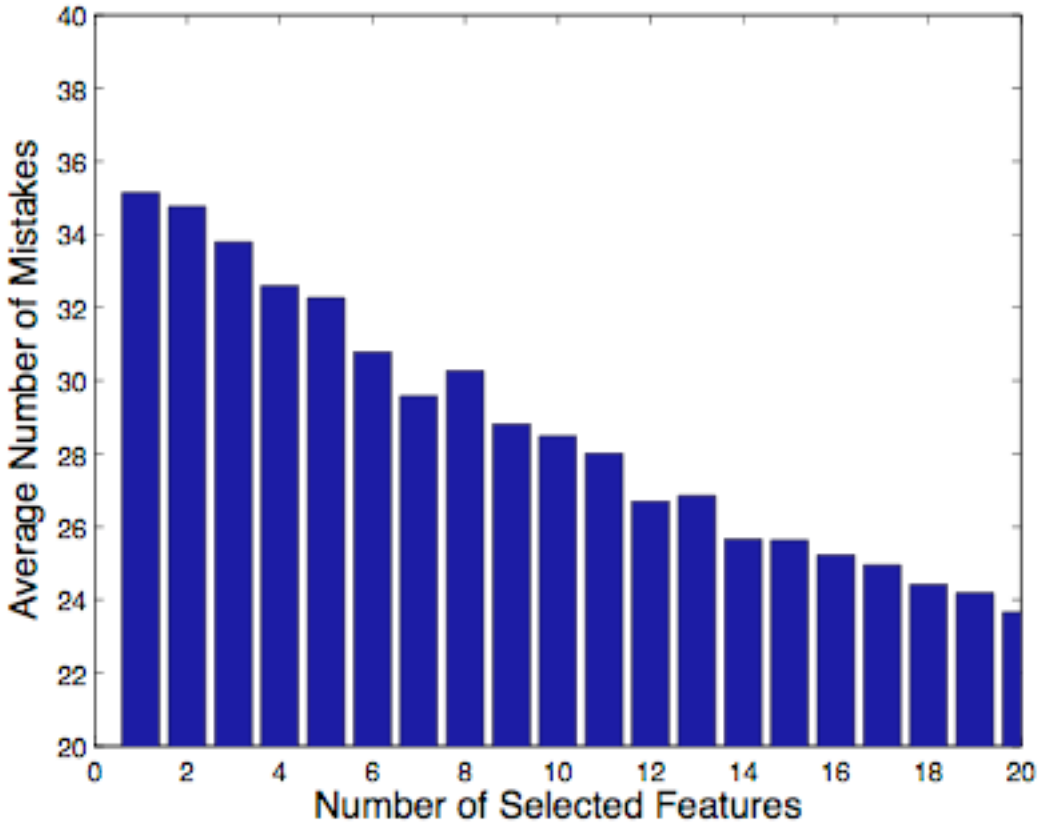
Counter-Example 1

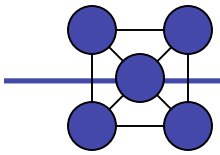


The only feature a Markov Blanket Algorithm won't remove is X_1.

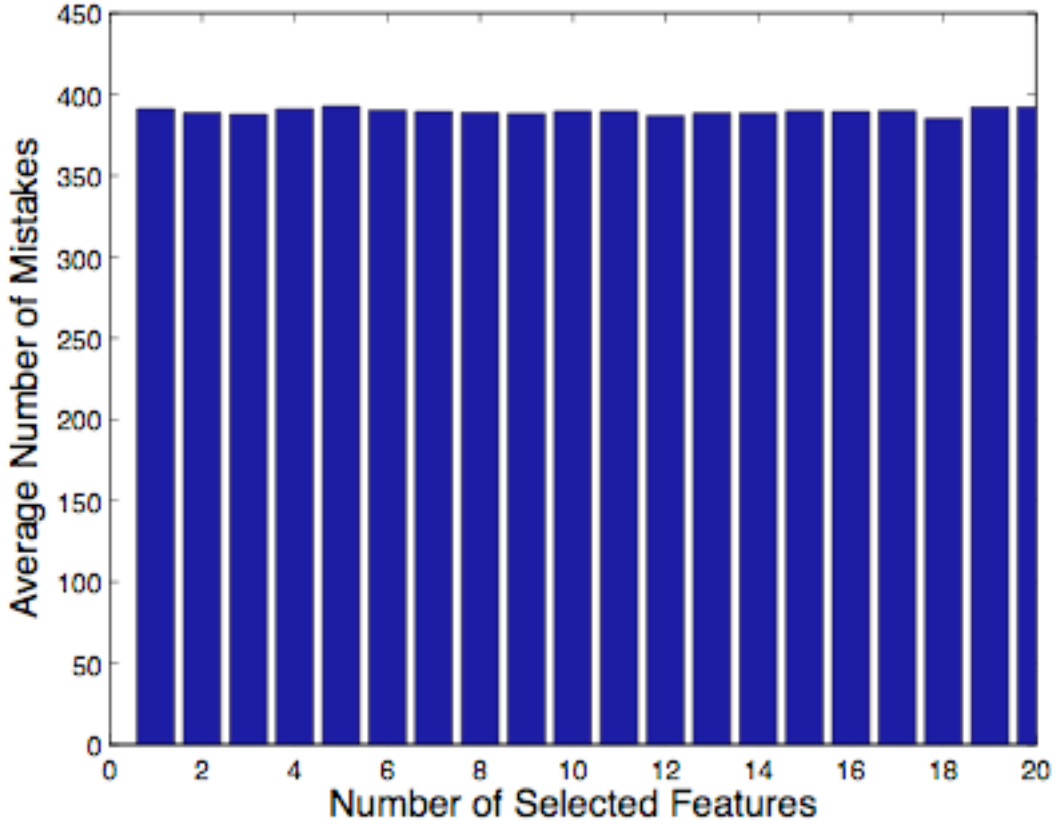


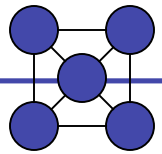
In-Sample Error



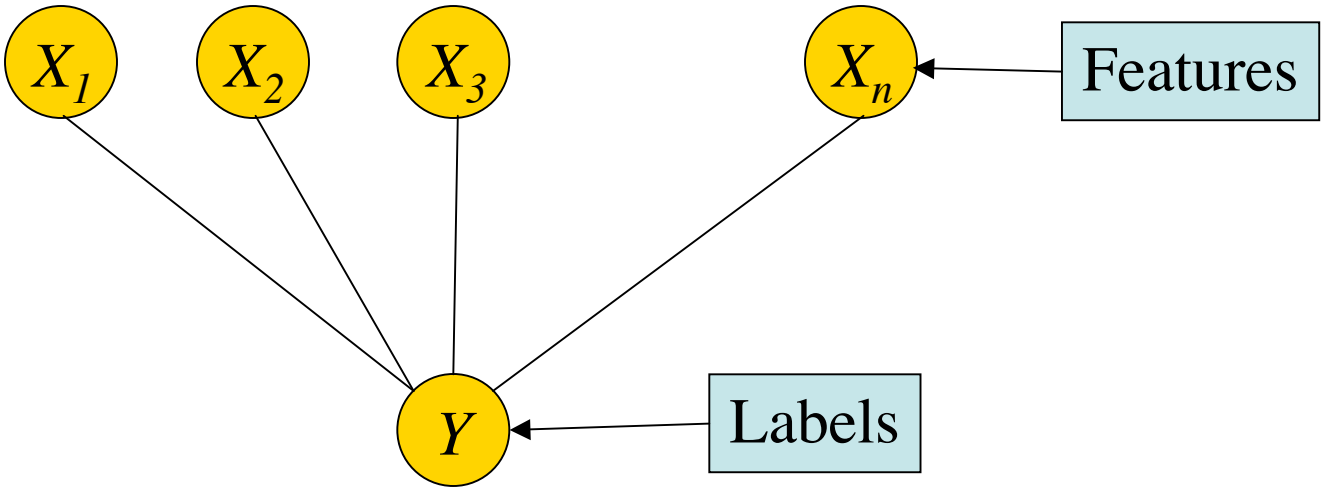


Generalization Error

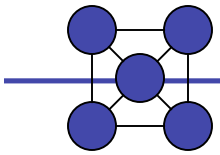




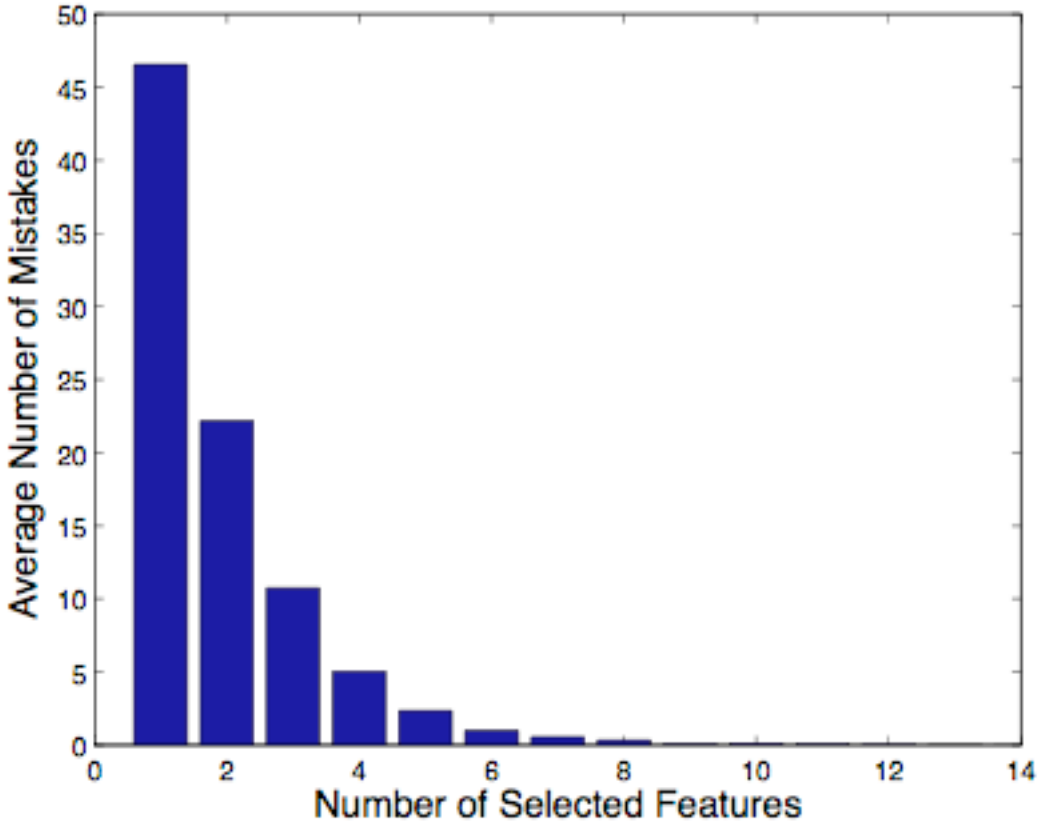
Counter-Example 2

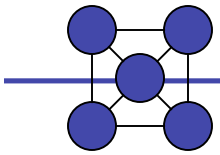


- All features are weakly correlated with Y.
- Features X_i,X_j are independent of each other given Y.
- No features will be removed via Markov Blankets.

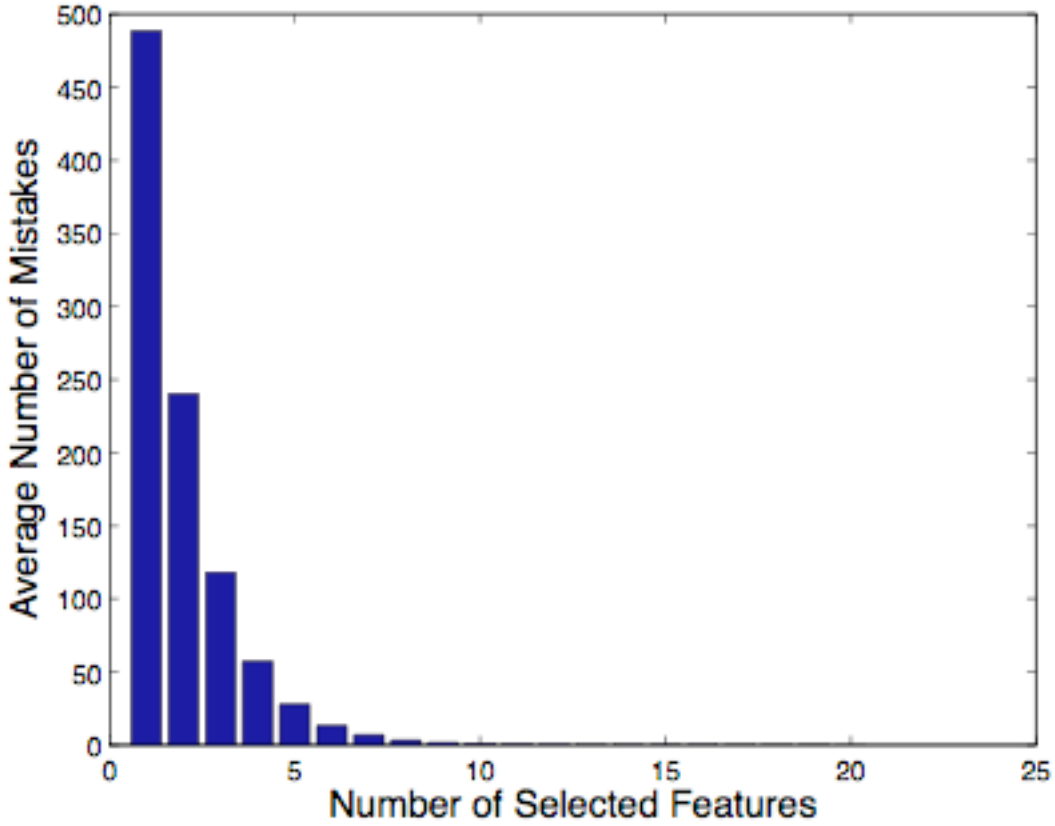


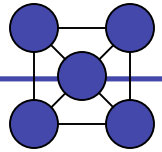
In-Sample Error





Generalization Error





Conclusion

I have provided a proof by counter-example that an algorithm that removes features if and only if it has a Markov blanket is not always optimal.