CPSC 340:
Machine Learning and Data Mining

Feature Selection

Bonus Slides

Bayesian Information Criterion (BIC)

- A disadvantage of these methods:
  - Still prefers a larger ‘p’ as ‘n’ grows.

- Solution: make λ depend on ‘n’.
- For example, the Bayesian information criterion (BIC) uses:

$$\lambda = \frac{1}{2}\log(n)$$

- BIC penalizes a bit more than AIC for large ‘n’.
  - As ‘n’ goes to ∞, recovers “true” model (“consistent” for model selection).
- In practice, we usually just try a bunch of different λ values.
  - Picking λ is like picking ‘k’ in k-means.

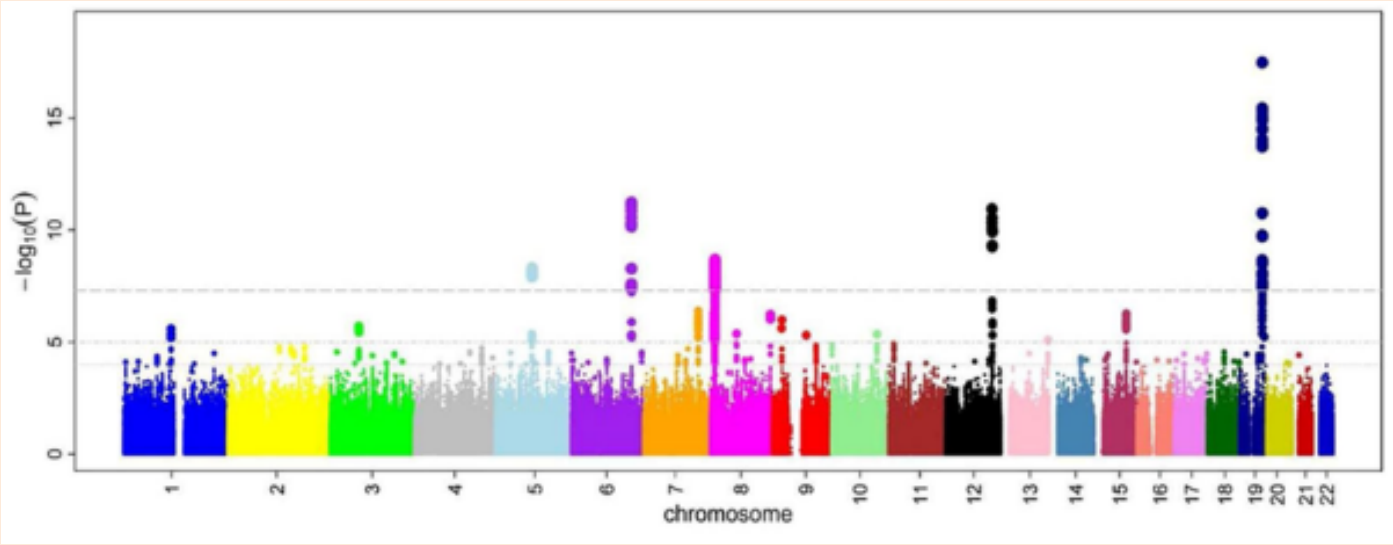
Discussion of other Scores for Model Selection

- There are many other scores:
  - Elbow method (corresponds to specific choice of λ).
    - You could also use BIC for choosing 'k' in k-means.
  - Methods based on validation error.
    - "Take smallest 'p' within one standard error of minimum cross-validation error".
  - Minimum description length.
  - Risk inflation criterion.
  - False discovery rate.
  - Marginal likelihood (CPSC 540).

- These can adapted to use the L1-norm and other errors.

Genome-Wide Association Studies

- Genome-wide association studies:
  - Measure if there exists a dependency between each individual “single-nucleotide polymorphism” in the genome and a particular disease.



  - Has identified thousands of genes “associated” with diseases.
    - But *by design* this has a huge numbers of false positives (and many false negatives).

https://en.wikipedia.org/wiki/Genome-wide_association_study

Backward Selection and RFE

- Forward selection often works better than naïve methods.
- A related method is backward selection:
  - Start with all features, compute score after removing each feature, remove the one that improves the score the most.
- If you consider adding or removing features, it's called stagewise.
- Stochastic local search is a class of fancier methods.
  - Simulated annealing, genetic algorithms, ant colony optimization, etc.
- Recursive feature elimination is another related method:
  - Fit parameters of a regression model.
  - Prune features with small regression weights.
  - Repeat.

Is “Relevance” Clearly Defined?

- Consider a supervised classification task:

gender	mom	dad
F	1	0
M	0	1
F	0	0
F	1	1

SNP
1
0
0
1

- Predict whether someone has particular genetic variation (SNP).
  - Location of mutation is in “mitochondrial” DNA.
    - “You almost always have the same value as your mom”.
  - For simplicity we’ll assume 1950s-style gender and parentage.

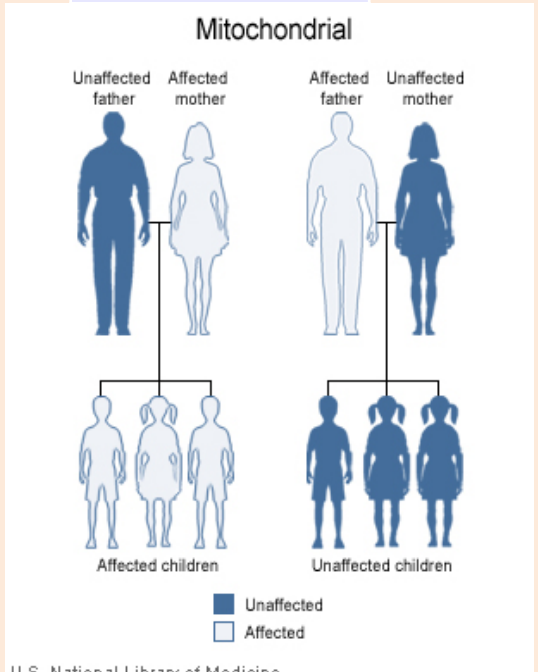
Is “Relevance” Clearly Defined?

- Consider a supervised classification task:

gender	mom	dad
F	1	0
M	0	1
F	0	0
F	1	1

SNP
1
0
0
1

- True model:
  - (SNP = mom) with very high probability.
  - (SNP != mom) with some very low probability.
- What are the “relevant” features for this problem?
  - Mom is relevant and {gender, dad} are not relevant.



https://en.wikipedia.org/wiki/Human_mitochondrial_genetics

Is “Relevance” Clearly Defined?

- What if “mom” feature is repeated?

gender	mom	dad	mom2
F	1	0	1
M	0	1	0
F	0	0	0
F	1	1	1

SNP
1
0
0
1

- Are “mom” and “mom2” relevant?
  - Should we pick them both?
  - Should we pick one because it predicts the other?

Neither of these is "correct", but not picking either is incorrect.

- If features can be predicted from features, can’t know which to pick.
  - Collinearity is a special case of “dependence” (which may be non-linear).

Is “Relevance” Clearly Defined?

- What if we add (maternal) “grandma”?

gender	mom	dad	grandma
F	1	0	1
M	0	1	0
F	0	0	0
F	1	1	1

SNP
1
0
0
1

- Is “grandma” relevant?
  - You can predict SNP very accurately from “grandma” alone.
  - But “grandma” is irrelevant if I know “mom”.
- A feature is only “relevant” in the context of available features.
  - Adding/removing features can make features relevant/irrelevant.

Is “Relevance” Clearly Defined?

- What if we don’t know “mom”?

gender	grandma	dad
F	1	0
M	0	1
F	0	0
F	1	1

SNP
1
0
0
1

- Now is “grandma” is relevant?
  - Without “mom” variable, using “grandma” is the best you can do.

- A feature is only “relevant” in the context of available features.
  - Adding/removing features can make features relevant/irrelevant.

Is “Relevance” Clearly Defined?

- What if we don’t know “mom” or “grandma”?

gender	dad
F	0
M	1
F	0
F	1

SNP
1
0
0
1

- Now there are no relevant variables, right?
  - But “dad” and “mom” must have some common maternal ancestor.
  - “Mitochondrial Eve” estimated to be ~200,000 years ago.
- A “relevant” feature may have a tiny effect.

Is “Relevance” Clearly Defined?

- What if we don’t know “mom” or “grandma”?

gender	dad
F	0
M	1
F	0
F	1

SNP
1
0
0
1

- Now there are no relevant variables, right?
  - What if “mom” likes “dad” because he has the same SNP as her?

- Confounding factors can make “irrelevant” variables “relevant”.

Is “Relevance” Clearly Defined?

- What if we add “sibling”?

gender	dad	sibling
F	0	1
M	1	0
F	0	0
F	1	1

SNP
1
0
0
1

- Sibling is “relevant” for predicting SNP, but it’s not the cause.

- “Relevance” for prediction does not imply a causal relationship.
  - Causality can even be reversed…

Is “Relevance” Clearly Defined?

- What if don’t have “mom” but we have “baby”?

gender	dad	baby
F	0	1
M	1	1
F	0	0
F	1	1

SNP
1
0
0
1

- “Baby” is relevant when (gender == F).
  - “Baby” is relevant (though causality is reversed).
  - Is “gender” relevant?
    - If we want to find relevant causal factors, “gender” is not relevant.
    - If we want to predict SNP, “gender” is relevant.
- “Relevance” may depend on values of certain features.
  - “Context-specific” relevance.

Is “Relevance” Clearly Defined?

- Warnings about feature selection:
  - If features can be predicted from features, you can’t know which to pick.
  - A feature is only “relevant” in the context of available features.
  - A “relevant” feature may have a tiny effect.
  - Confounding factors can make “irrelevant” variables the most “relevant”.
  - “Relevance” for prediction does not imply a causal relationship.
  - “Relevance” may depend on values of certain features.

Is this hopeless?

- We often want to do feature selection we so have to try!
- Different methods are affected by problems in different ways.
- These “problems” don’t have right answers but have wrong answers:
  - Variable dependence (“mom” and “mom2” have same information).
    - But should take at least one.
  - Conditional independence (all “grandma” information is captured by “mom”).
    - Should take “grandma” only if “mom” missing.
- These “problems” have application-specific answers:
  - Tiny effects.
  - Context-specific relevance (is “gender” relevant if given “baby”?).
- See bonus slides for discussion causality and confounding issues.
  - Unless you control data collection, standard feature selection methods cannot address those issues.

Rough Guide to Feature Selection

Method\Issue	Dependence	Conditional Independence	Tiny effects	Context-Specific Relevance
Association (e.g., measure correlation between features ‘j’ and ‘y’)	Ok (takes “mom” and “mom2”)	Bad (takes “grandma”, “great-grandma”, etc.)	Ignores	Bad (misses features that must interact, “gender” irrelevant given “baby”)

Rough Guide to Feature Selection

Method\Issue	Dependence	Conditional Independence	Tiny effects	Context-Specific Relevance
Association (e.g., measure correlation between features 'j' and 'y')	Ok (takes "mom" and "mom2")	Bad (takes "grandma", "great-grandma", etc.)	Ignores	Bad (misses features that must interact, "gender" irrelevant given "baby")
Regression Weight (fit least squares, take biggest $\|w_j\|$)	Bad (can take irrelevant but collinear, can take none of "mom1-3")	Ok (takes "mom" not "grandma", if linear and 'n' large.	Ignores (unless collinear)	Ok (if linear, "gender" relevant give "baby")

Rough Guide to Feature Selection

Method\Issue	Dependence	Conditional Independence	Tiny effects	Context-Specific Relevance
Association (e.g., measure correlation between features 'j' and 'y')	Ok (takes "mom" and "mom2")	Bad (takes "grandma", "great-grandma", etc.)	Ignores	Bad (misses features that must interact, "gender" irrelevant given "baby")
Regression Weight (fit least squares, take biggest $\|w_j\|$)	Bad (can take irrelevant but collinear, can take none of "mom1-3")	Ok (takes "mom" not "grandma", if linear and 'n' large.	Ignores (unless collinear)	Ok (if linear, "gender" relevant give "baby")
Search and Score w/ Validation Error	Ok (takes at least one of "mom" and "mom2")	Bad (takes "grandma", "great-grandma", etc.)	Allows	Ok ("gender" relevant given "baby")

Rough Guide to Feature Selection

Method\Issue	Dependence	Conditional Independence	Tiny effects	Context-Specific Relevance
Association (e.g., measure correlation between features 'j' and 'y')	Ok (takes "mom" and "mom2")	Bad (takes "grandma", "great-grandma", etc.)	Ignores	Bad (misses features that must interact, "gender" irrelevant given "baby")
Regression Weight (fit least squares, take biggest $\|w_j\|$)	Bad (can take irrelevant but collinear, can take none of "mom1-3")	Ok (takes "mom" not "grandma", if linear and 'n' large.	Ignores (unless collinear)	Ok (if linear, "gender" relevant give "baby")
Search and Score w/ Validation Error	Ok (takes at least one of "mom" and "mom2")	Bad (takes "grandma", "great-grandma", etc.)	Allows (many false positives)	Ok ("gender" relevant given "baby")
Search and Score w/ L0-norm	Ok (takes exactly one of "mom" and "mom2")	Ok (takes "mom" not grandma if linear-ish).	Ignores (even if collinear)	Ok ("gender" relevant given "baby")

Feature Selection in Tree-Based Methods

- Decision trees naturally do feature selection while learning:
  - The features used for the splits are the ones that are “selected”.

- There are a variety of ways evaluate features in random forests:
  - Compute proportion of trees that use feature ‘j’.
  - Compute average infogain increase when using feature ‘j’.
  - Permute all values of feature ‘j’, and see how “out of bag” error increases.
- You could use any of above to select features from random forest.

Mallow's Cp

- Older than AIC and BIC is Mallow's Cp:

$$f(w) = \frac{\|Xw - y\|^2}{\frac{1}{n}\|X\hat{w} - y\|^2} - n + 2\|w\|_0$$

least squares weights if we used all features.

- Minimizing this score is equivalent to L0-regularization:

$$f(w) = \frac{1}{2}\|Xw - y\|^2 + \lambda\|w\|_0$$

with $\lambda = \frac{\|X\hat{w} - y\|^2}{n}$

- So again, viewing λ as hyper-parameter, this score is special case.

Adjusted R^2

- Older than AIC and BIC and Mallow's Cp is adjusted R^2:

$$f(w) = 1 - (1 - R^2)\frac{n-1}{n-\|w\|_0-1} \quad \text{where} \quad R^2 = 1 - \frac{\|Xw - y\|^2}{\|X\hat{w} - y\|^2}$$

- Maximizing this score is equivalent to L0-regularization:

$$= \frac{1}{2}\|Xw - y\|^2 + \lambda \|w\|_0$$

$$\text{with} \quad \lambda = \frac{\|X\hat{w} - y\|^2}{2(n-1)}$$

- So again, viewing λ as hyper-parameter, this score is special case.

ANOVA

- Some people also like to compute this "ANOVA" quantity:

$$f(w) = \frac{\|Xw - \bar{y}\|^2}{\|y - \bar{y}\|^2}$$

↑ mean of y_i values repeated 'n' times

- This is based on the decomposition of "total squared error" as:

$$\underbrace{\|y - \bar{y}\|^2}_{\text{"total" error}} = \underbrace{\|Xw - \bar{y}\|^2}_{\text{"explained" error}} + \underbrace{\|Xw - y\|^2}_{\text{"residual" (usual) error.}}$$

- Notice that "explained error" goes up as our usual ("residual") error goes down.
- Trying to find the 'k' features that maximize 'f' ("explain the most variance") is equivalent to L0-regularization with a particular λ (so another special case).

Information Criteria with Noise Variance

- We defined AIC/BIC for feature selection in least squares as:

$$f(w) = \frac{1}{2}\|Xw - y\|^2 + \lambda \|w\|_0$$

- The first term comes from assuming $y_i = w^T x_i + \varepsilon$, where ε comes from a normal distribution with a variance of 1.
  - We'll discuss why when we discuss MLE and MAP estimation.
  - If you aren't doing least squares, replace first term by "log-likelihood".
- If you treat variance as a parameter, then after some manipulation:

$$f(w) = \frac{n}{2}\log(\|Xw - y\|^2) + \lambda \|w\|_0$$

- However, this is again equivalent to just changing λ.

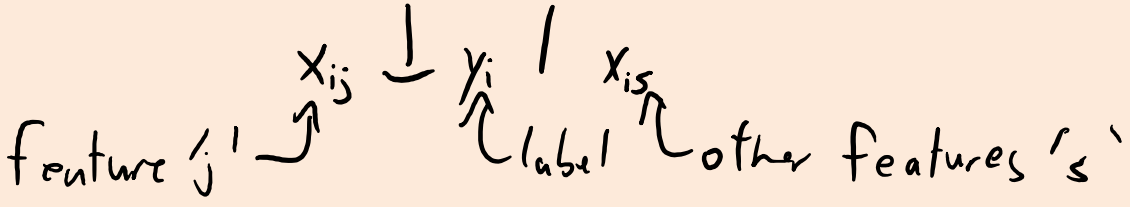
Complexity Penalties for Other Models

- Scores like AIC and BIC can also be used in other contexts:
  - When fitting a decision tree, only split a node if it improves BIC.
  - This makes sense if we're looking for the "true tree", or maybe just a simple/interpretable tree that performs well.

- In these cases we replace "L0-norm" with "degrees of freedom".
  - In linear models fit with least squares, degrees of freedom is number of non-zeroes.
  - Unfortunately, it is not always easy to measure "degrees of freedom".

Alternative to Search and Score: good old p-values

- Hypothesis testing ("constraint-based") approach:
  - Generalization of the "association" approach to feature selection.
  - Performs a sequence of conditional independence tests.



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

  - If they are independent (like "p < .05"), say that 'j' is "irrelevant".
- Common way to do the tests:
  - "Partial" correlation (numerical data).
  - "Conditional" mutual information (discrete data).

Testing-Based Feature Selection

- Hypothesis testing ("constraint-based") approach:
- Two many possible tests, "greedy" method is for each 'j' do:

First test if $x_{ij} \perp y_i$

If still dependant test $x_{ij} \perp y_i \mid x_{is}$ where 's' has one feature

If still dependent test $x_{ij} \perp y_i \mid x_{is}$ where 's' now has two features

Often choose features to minimize dependence.

⋮

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

- "Association approach" is the greedy method where you only do the first test (subsequent tests remove a lot of false positives).

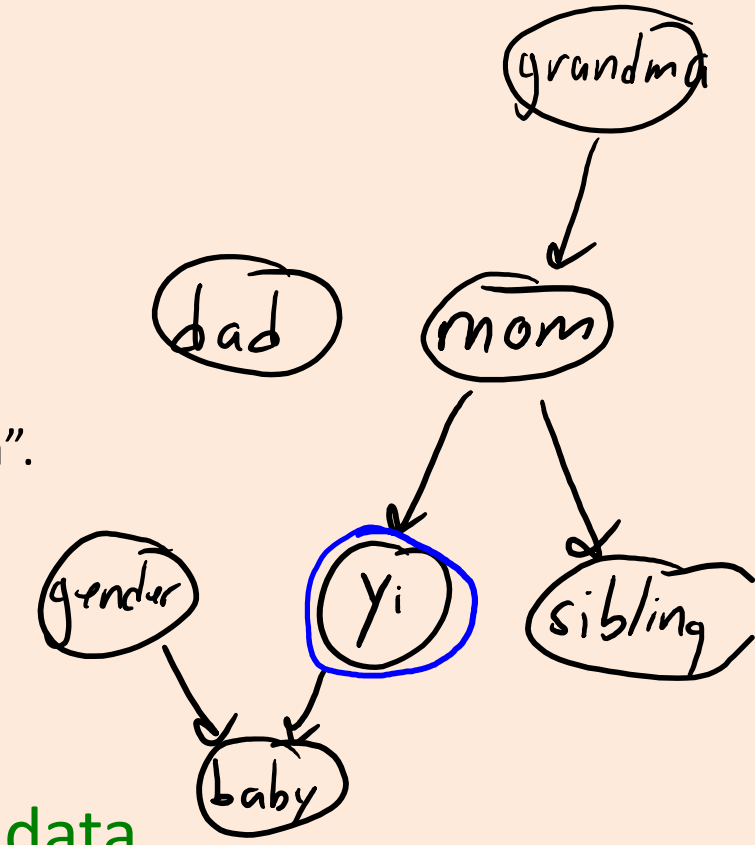
Hypothesis-Based Feature Selection

- Advantages:
  - Deals with conditional independence.
  - Algorithm can explain why it thinks ‘j’ is irrelevant.
  - Doesn’t necessarily need linearity.
- Disadvantages:
  - Deals badly with exact 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:
  - Allows tiny effects.
  - Says “gender” is irrelevant when you know “baby”.
  - This approach is sometimes better for finding relevant factors, not to select features for learning.

Causality

- None of these approaches address causality or confounding:
  - "Mom" is the only relevant direct causal factor.
  - "Dad" is really irrelevant.
  - "Grandma" is causal but is irrelevant if we know "mom".

- Other factors can help prediction but aren't causal:
  - "Sibling" is predictive due to confounding of effect of same "mom".
  - "Baby" is predictive due to reverse causality.
  - "Gender" is predictive due to common effect on "baby".



- We can sometimes address this using interventional data…

Interventional Data

- The difference between observational and interventional data:
  - If I see that my watch says 10:45, class is almost over (observational).
  - If I set my watch to say 10:45, it doesn't help (interventional).

- The intervention can help discover causal effects:
  - "Watch" is only predictive of "time" in observational setting (so not causal).

- General idea for identifying causal effects:
  - "Force" the variable to take a certain value, then measure the effect.
    - If the dependency remains, there is a causal effect.
    - We "break" connections from reverse causality, common effects, or confounding.

Causality and Dataset Collection

- This has to do with the way you collect data:
  - You can't "look" for variables taking the value "after the fact".
  - You need to manipulate the value of the variable, then watch for changes.

- This is the basis for randomized control trial in medicine:
  - Randomly assigning pills "forces" value of "treatment" variable.
    - Randomization means they aren't taking the pill due to confounding factors.
    - Differences between people who did and did not take pill should be caused by pill.
  - Include a "control" as a value to prevent placebo effect as confounding.

- See also Simpson's Paradox:
  - https://www.youtube.com/watch?v=ebEkn-BiW5k

Structure Learning: Unsupervised Feature Selection

- "News" data: presence of 100 words in 16k newsgroup posts:

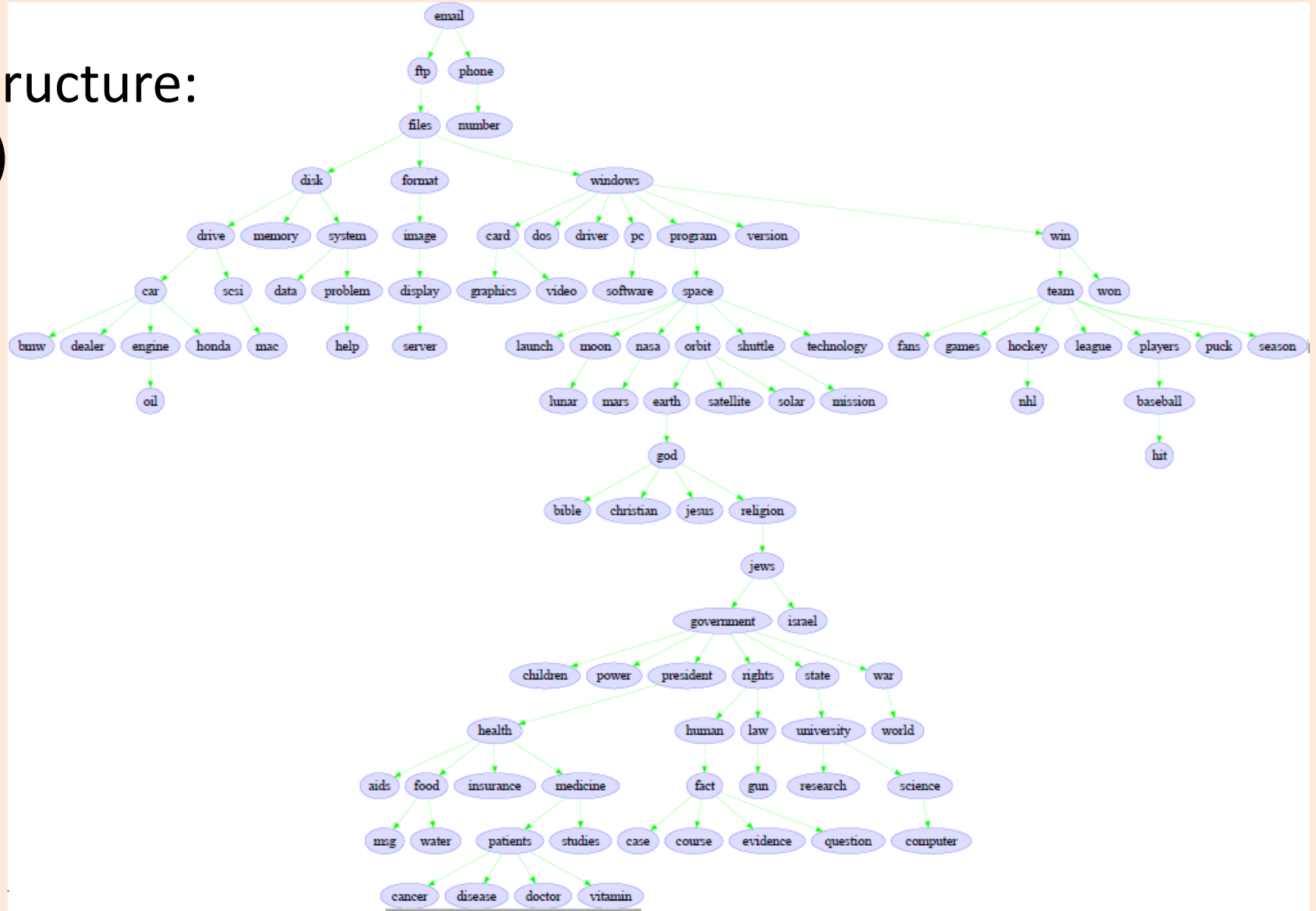
car	drive	files	hockey	mac	league	pc	win
0	0	1	0	1	0	1	0
0	0	0	1	0	1	0	1
1	1	0	0	0	0	0	0
0	1	1	0	1	0	0	0
0	0	1	0	0	0	1	1

- Which words are related to each other?

- Problem of structure learning: unsupervised feature selection.

Structure Learning: Unsupervised Feature Selection

- Optimal tree structure: (ignore arrows)



Naïve Approach: Association Networks

- A naïve approach to structure learning ("association networks"):
  - For each pair of variables, compute a measure of similarity or dependence.

- Using these n^2 similarity values either:
  - Select all pairs whose similarity is above a threshold.
  - Select the "top k" most similar features to each feature 'j'.

- Main problems:
  - Usually, most variables are dependent (too many edges).
    - "Sick" is getting connected to "Tuesdays" even if "tacos" are a variable.
  - "True" neighbours may not have the highest dependence.
    - "Sick" might get connected to "Tuesdays" before it gets connected to "milk".

- (Variation: best tree can be found as minimum spanning tree problem.)

Example: Vancouver Rain Data

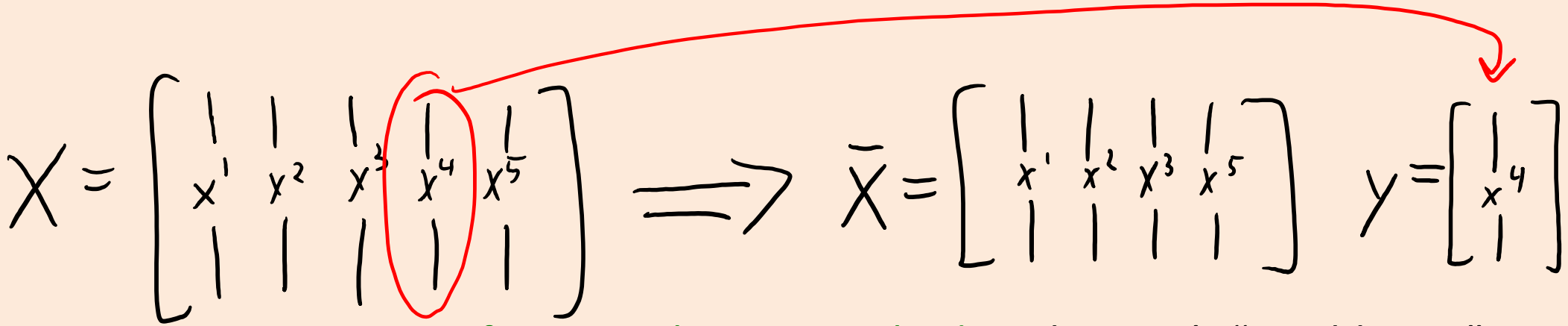
- Consider modeling the "Vancouver rain" dataset.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	…
Month 1	0	0	0	1	1	0	0	1	1	
Month 2	1	0	0	0	0	0	1	0	0	
Month 3	1	1	1	1	1	1	1	1	1	
Month 4	1	1	1	1	0	0	1	1	1	
Month 5	0	0	0	0	1	1	0	0	0	
Month 6	0	1	1	0	0	0	0	1	1	

- The strongest signal in the data is the simple relationship:
  - If it rained yesterday, it's likely to rain today (> 50% chance that $x^{t-1} = x^t$).
  - But an "association network" might connect all days (all dependent).

Dependency Networks

- A better approach is dependency networks:
  - For each variable ‘j’, make it the target in a supervised learning problem.

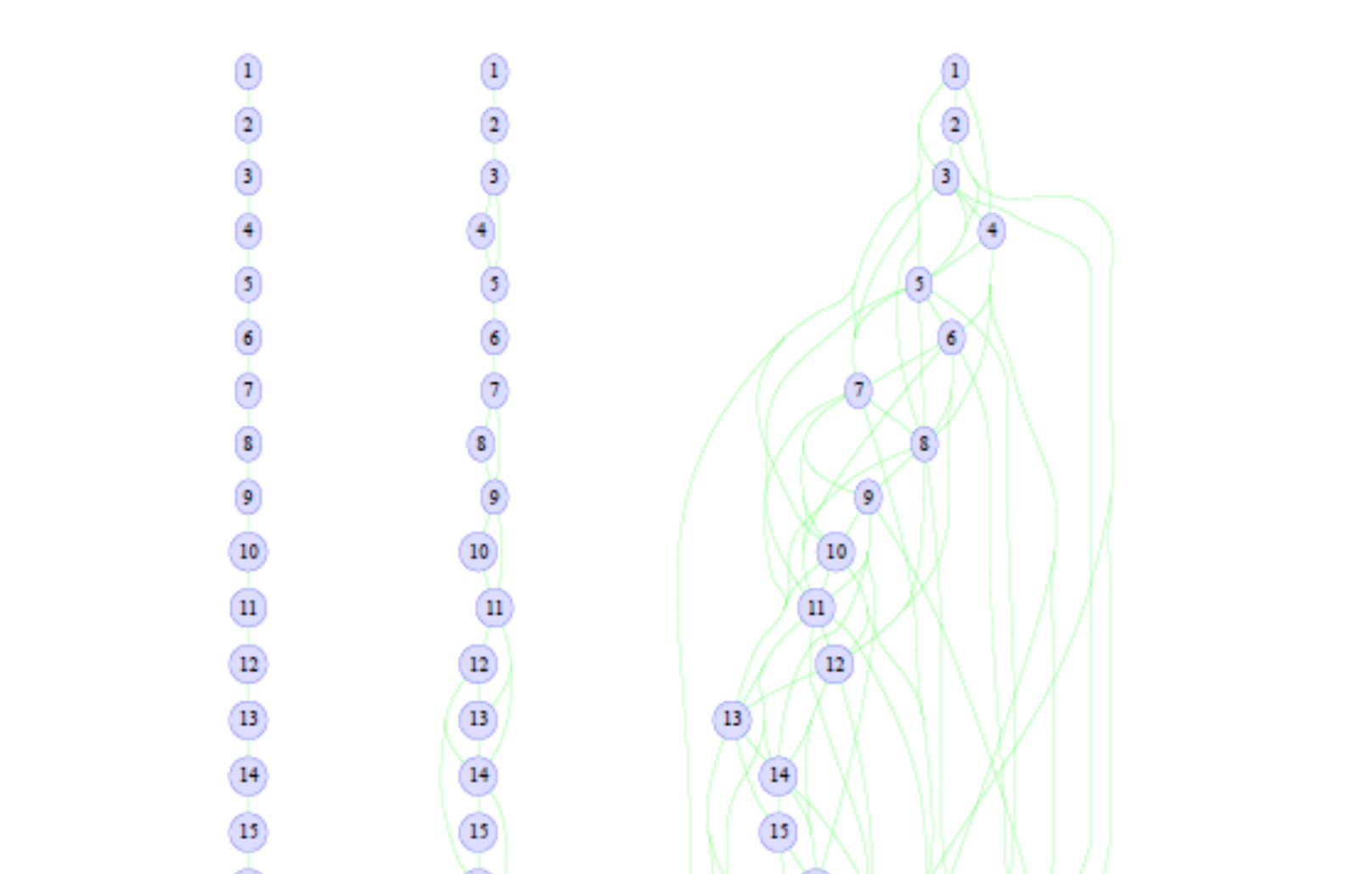


  - Now we can use any feature selection method to choose j’s “neighbours”.
    - Forward selection, L1-regularization, ensemble methods, etc.

- Can capture conditional independence:
  - Might connect “sick” to “tacos”, and “tacos” to “Tuesdays” (w/o sick-tacos).

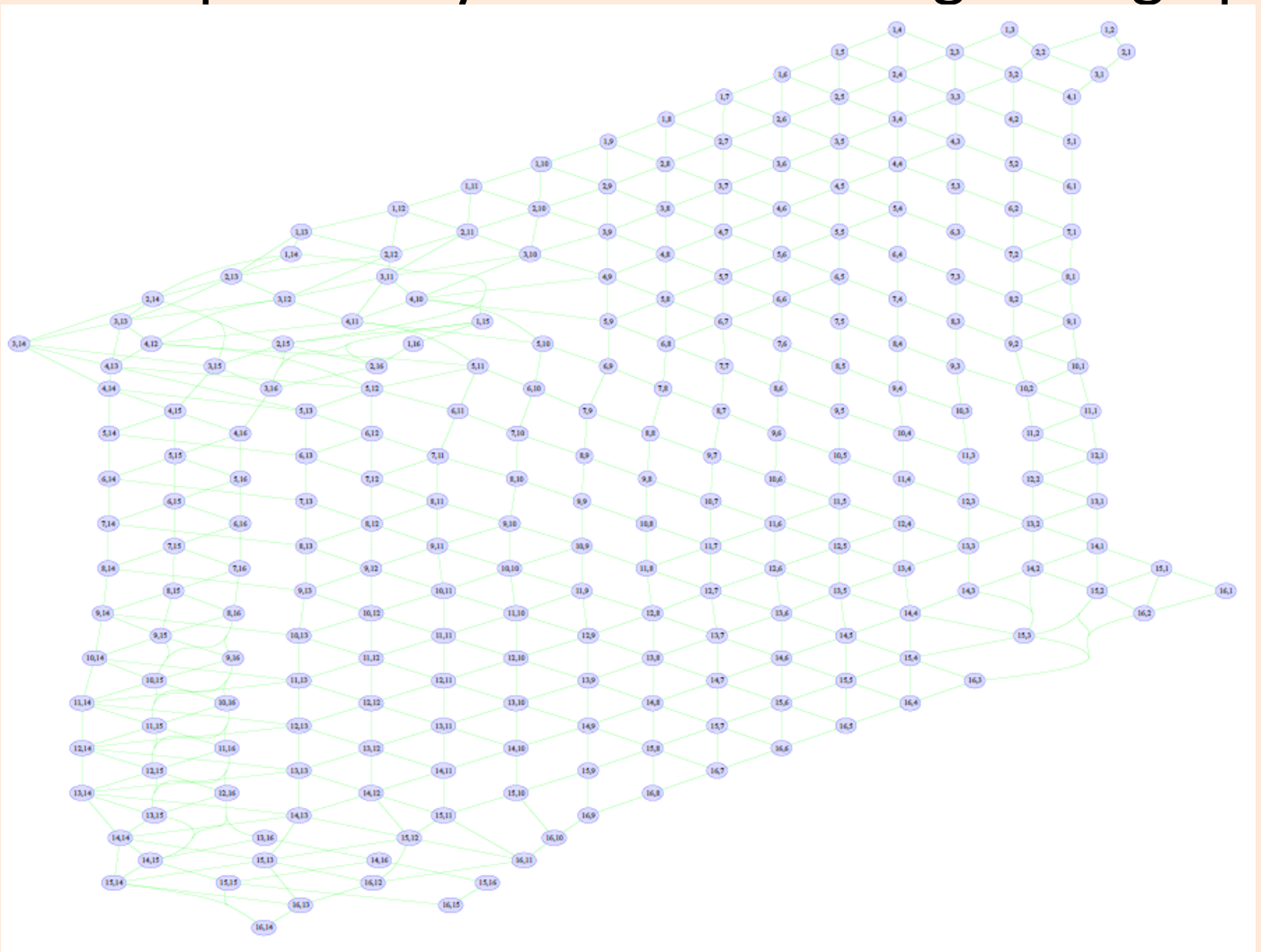
Dependency Networks

- Dependency network fit to Vancouver rain data (different λ values):



Dependency Networks

- Variation on dependency networks on digit image pixels:



Another popular structure learning method is the "PC" algorithm.