Networks - Network Prediction (26s3)

Synthesized Lecture Summary

3/5/26

Color Key:
Red = Source 1 (lecture_summary_26s3_1) yellow = Source 2 (lecture_summary_26s3_2)
Green = Source 3 (lecture_summary_26s3_3) **Blue** = My own additions

Objectives

By the end of this lecture, students should:

1. Understand the basic principles of predicting connectivity between nodes in a network (interactions)
2. Be able to compare and contrast different methods to assess network interaction predictions, including union, intersection, majority vote, and Bayesian integration
3. Understand naive Bayes' Rule and its advantages over other prediction strategies
4. Be able to construct a basic Bayesian Network using Bayesian formalism
5. Evaluate combined predictions using ROC curves (TPR vs FPR) and understand why Bayesian integration outperforms simple heuristics
6. Recognize the limitations of the naive Bayes independence assumption and when feature correlation matters

Definitions and Key Concepts

- **Gold Standard (GSTD):** A benchmark dataset of high-confidence interactions (GSTD+) and non-interactions (GSTD-) used to train and validate prediction models. In the Pol II example, GSTD+ comes from the known crystal structure (13 true contacts) and GSTD- from pairs outside the complex (32 non-contacts).
- **Network assessment terminology:**
  - Union - if any experiment says TRUE, output is TRUE
  - Intersection - if any experiment says FALSE, output is FALSE
  - Majority - output is the most common input, FALSE if tied
  - Weighted voting / supervised classification - scores are given weights and voting occurs in weighted fashion (outcome R = w-vec dot f-vec + w0)
- **Bayes Rule:** P(Y|X) = P(X|Y) * P(Y) / P(X). Basically just a framework for updating what you believe about a hypothesis after seeing new data. Prior P(Y) is your starting belief, likelihood P(X|Y) is how probable the data would be if the hypothesis were true, posterior P(Y|X) is the updated belief.
- **Naive Bayes Assumption:** All experimental features are assumed independent given the class. So p(f1, f2, ... | I) = p(f1|I) * p(f2|I) * ... which is a big simplification. Means you can break a multi-parameter problem into a sum of single log-likelihood ratios, way easier to work with.

- **Likelihood Ratio (LR):** $L_f = p(x_f \mid GSTD+) / p(x_f \mid GSTD-)$. Basically how much more likely you are to see this feature value in true interactions vs non-interactions. > 1 means evidence for interaction, < 1 means evidence against. This is probably the most important thing to understand for te quiz.
- **ROC Curve:** Plots true-positive rate (sensitivity) vs false-positive rate (1 - specificity). Useful because you can evaluate prediction quality without commiting to a single threshold.
- **Directed Acyclic Graph (DAG):** A network structure of nodes and directional edges with no loops, used to represent conditional dependencies in Bayesian Networks.
- **Bayesian Network:** A DAG that specifies a joint distribution over X as a product of local conditional distributions, one per node. The main advantage is it describes the joint probability way more compactly than the full chain rule because the graph structure tells you which variables are conditionally independent.
- **Markov Blanket:** The set of a node's parents, children, and co-parents in a Bayesian network. If you know the state of everything in the Markov blanket, the target node is conditionally independent of all other nodes. Think of it like: fix the connector genes between modules, and each module becomes independent of the rest.

Main Content

Problem Setup: Noisy Evidence and Gold Standards

In order to construct and validate a network, we need to develop methods that integrate information we collect about the network. For example, if we want to understand how the subunits of a protein are interconnected, how do we integrate noisy experimental evidence to accurately characterize subunit interactions? We use RNA Polymerase II (RNAPII) as a case study because its crystal structure is known (Cramer et al., 2000), so we have a ground truth to validate against.

The experimental data comes from multiple assay types: 3 pull-down experiments, cross-linking, and 3 far-western experiments (7 total). Each one gives a binary readout (1 or 0) for each subunit pair. They all have different error rates and none of them are totally reliable on their own. The gold standard from the crystal structure has 13 true contacts (GSTD+) and 32 non-contacts (GSTD-).

Comparing Integration Approaches

The most permissive approach is **union**: if any method says the subunits interact, we say they interact. Obviously this will overestimate connectivity. The opposite is **intersection** where any negative result kills a positive result - avoids false positives but you lose a lot of true positives too. **majority** rule is the middle ground (most common vote wins). Better than either extreme but still vulnerable to correlated experiments.

The problem with all of these is that they treat every experiment equally. In reality we're often more confident about some experiments than others. The fix is to weight them, which is basically supervised classification. But then how do you figure out the right weights? That's where Bayes Rule comes in.

From Weighted Voting to Bayesian Integration

A simple classifier computes R = f1 + f2 + ... + fn (with f = 1 or -1) and classifies by the sign of R. A weighted version is R = w1*f1 + w2*f2 + ... + wn*fn = w-vec dot f-vec. If we have prior knowledge, we add w0: R = w-vec dot f-vec + w0. This has the same form as many other classifiers (SVC, LDA, logistic regression - see ISLR for comparison).

The neat thing is that naive Bayes gives you exactly this linear additive form in log space. After applying Bayes' theorem with the independence assumption, the posterior log-odds breaks down to:

$$\log(P(I \mid f1,f2,...) / P(\sim I \mid f1,f2,...)) = \log(TPR1/FPR1) + \log(TPR2/FPR2) + ... + \log(P/N)$$

Where each log(TPR_k / FPR_k) is a weight w_k for feature k, and log(P/N) is the prior w_0. So basically you start with prior odds and keep adding likelihood ratio terms for each feature until you get your final posterior odds.

Calculating Likelihood Ratios

For each binary feature, you just count how often it fires in GSTD+ vs GSTD- pairs. E.g. for Pull-down 1: L1 = p(x=1 | GSTD+) / p(x=1 | GSTD-) = (6/13) / (11/32) = 1.34. And L0 = (4/13) / (14/32) = 0.70.

The slides walk through all seven assays (slides 18-20). The calculated LRs for each are:

- Pull-down 1: L1=1.34, L0=0.70
- Pull-down 2: L1=1.91, L0=0.31
- Pull-down 3: L1=1.64, L0=2.46
- Cross-linking: L1=3.52, L0=0 (needs smoothing!)
- Far Western 1: L1=1.23, L0=1.23
- Far Western 2: L1=2.95, L0=0.29
- Far Western 3: L1=2.46, L0=2.46

Watch out for cases where a feature count is zero (like cross-linking L0 = 0) - you'll need to add dummy counts (pseudocounts). This is called Laplace smoothing. Source 1 also mentions the m-estimate, which is basically the same idea.

The combined score for a subunit pair is just the product of all its likelihood ratios: L(f1,...,fn) = L(f1) * L(f2) * ... * L(fn). Thats the whole point of the weighted voting - each assay contributes proportional to how good it actually is at telling true from false interactions.

Evaluation: ROC Curves

The worked example runs through slides 17-22. The main takeaway is that the ROC curve shows Bayesian integration beating union, intersection, and majority at basically every threshold.

ROC plots TPR (= TP/P = Sensitivity) on y-axis vs FPR (= FP/N = 1-Specificity) on x-axis. The Bayesian line sits above the other methods at most operating points. Basically you get better sensitivity for any given false positive rate, which is the whole payoff of doing principled weighting.

Feature Correlation and its Consequences

In practice the naive Bayes independence assumption almost never holds. In the RNAPII example we literally have repeat experiments of the same type which are obviously correlated (slides 24-25). If two features carry the same signal, NB double-counts them and you get inflated confidence.

Slide 24 shows this directly. they duplicate a far-western experiment and the duplicate just contributes the same LR again, which is obviously wrong. To deal with this you can use a 'fully connected Bayes' model that computes joint LRs for pairs of correlated features, e.g. w_{4,5} = log P(f4=1, f5=1 | I) / P(f4=1, f5=1 | ~I). Not covered in detail in this lecture but good to know it exists.

Bayesian Networks and Markov Blankets

Bayesian Networks use DAGs to organize parameters and outcomes. Each node's probability depends only on its parents in the graph, and you apply Bayes Rule to describe the relationships.

The big advantage is that the network structure makes the joint probability much more compact. Slide 28 shows this with the Burglary/Earthquake/Alarm example: full chain rule needs 1+2+4+8+16 = 31 parameters for 5 binary variables, but the BN only needs 1+1+4+2+2 = 10. Pretty dramatic reduction.

Worth noting that naive Bayes is just a special case of a Bayesian network; its a star-shaped DAG where the class node points to all features with no edges between features (slide 29).

The slides show a nice example with gene co-expression networks (Gao et al., 2024). The Bayesian network version produces much cleaner module structure than raw co-expression because it picks up direct dependencies instead of indirect correlations. The Markov blanket idea ties in here - fix the connector genes between modules and each module becomes roughly independent.

Discussion

The lecture goes from a concrete biological question (which Pol II subunits touch each other?) to the statistical tools you need to answer it with noisy data. It walks through union/intersection/majority to show why they're not great, then builds up to Bayesian integration as a principled alternative.

Under naive Bayes, the whole thing simplifies to summing log-likelihood ratios. The ROC curves confirm it works better than the simpler approaches.

Some things I think are worth remembering: the prior odds term log(P/N) really matters when true interactions are rare, which they usually are in biology. And the cross-linking L0 = 0 thing is a good example of why you need smoothing.

The last part of the lecture on Bayesian networks and Markov blankets shows how these ideas extend beyond the Pol II toy example to things like gene co-expression networks. The filtering you get from Bayesian network structure helps cut through noisy correlations to find the real relationships.

Suggested Reading

1. James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert. An Introduction to Statistical Learning: with Applications in R [ISLR, 2nd edition]. Chapter 4.4.4 and 4.7.5 give background on Naive Bayes.
2. Edwards et al. (2002). Trends in Genetics, 18(10), 529-536. Bridging structural biology and genomics: assessing protein interaction data with known complexes. (Relates to the worked example.)
3. Jansen R, Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003;302(5644):449-53.
4. Gao J, Gerstein M. Representing core gene expression activity relationships using the latent structure implicit in Bayesian networks. Bioinformatics. 2024;40(8):btae463.
5. Cramer, P., Bushnell, D.A., Fu, J., et al. (2000). Architecture of RNA polymerase II. Science 288(5466):640-649.
6. Spiegelhalter, D. & Brooks, S. (2025). The Art of Uncertainty. (Referenced on slide 15 for the Turing connection to the additive update form.)
7. ISL/ESL reference sections: ISL 4.4.4 and 4.7.5; ESL 6.6.3