CMPSCI 348 **Exam 3 Review Sheet** Spring 2019

Exam 3 will be given on Friday, May 3, 10:30AM-12:30PM, in Goessmann Laboratory Addition, Room 64. The questions below cover the concepts that will appear on the exam. Most topics are drawn from the last third of the course (e.g., d-separation), but a few are drawn from earlier in the course (e.g., conditional probability estimation), and Exam 3 will require you to use some of these concepts in combination. Questions on the exam will be drawn from both the lectures *and from the readings associated with each lecture*. Annotations after each topic refer to lecture dates (L1.24 is the lecture on January 24) and associated readings.

Probability theory (L1.24, L1.29, L2.14, L4.11)

- *Probability distributions* — What is a probability distribution? What is the difference between a marginal, conditional, and joint probability distribution?
- *Independence and conditional independence* — What is the difference between independence and conditional independence? What do independence and conditional independence imply about how you can decompose a joint probability distribution?
- *Bayes' rule* — What is the multiplication rule and how can it be used to derive Bayes Rule?
- *Random variables* — What is a random variable? How are random variables used in typical data science applications?

Data science goals (L1.24, L1.29)

- *Central dogma* — Describe the central dogma of data science. What can go wrong in this process that will interfere with the ability to make valid inferences about the world?
- *Descriptive, predictive, and prescriptive analytics* — Describe the goals of descriptive analytics, predictive analytics, and prescriptive analytics. How do the three types differ? What are some methods typically used to perform each type?
- *Data generating processes and models* — What is a data generating process? What is a model? How do data generating processes and models typically differ? How do purely associational and causal models typically differ?

Modeling conditional probability distributions (L2.12, L2.14, L3.05)

- *Linear models* — What are the components of a linear model? How are those components interpreted to make statistical inferences about a data instance? How can they be used to output a probability distribution rather than a point estimate (e.g., continuous value)? How accurate are those distributions?
- *Simple Bayesian models* — What are the components of a simple Bayesian classifier? How are those components interpreted to make statistical inferences about a data instance? How can they be used to output a probability distribution rather than a point estimate (e.g., categorical value)? How accurate are those distributions?
- *Tree-structured models* — What are the components of a classification tree and a regression tree? How are those components interpreted to make statistical inferences about a data instance? How can they be configured to output a probability distribution rather than a point estimate (e.g., categorical or continuous value)? How accurate are those distributions?

Directed graphical models (L4.11, L4.16)

- *Graphical syntax* — What are the components of a directed graphical model (also known as a "Bayesian network" (BN))? What do those components represent? What constraints are necessary for a set of those components to form a valid BN?

- *Factorization* — How do conditional probability distributions correspond to the classification and regression models we covered earlier in the course? What is the most compact mathematical expression for the joint probability defined by a directed graphical model?
- *Markov condition* — What is the Markov condition for directed graphical models? What does it imply?

d-separation (L4.11, L4.16)

- *Simple sources of dependence* — What simple network structures can produce statistical dependence between two variables? How does conditioning change the statistical dependence produced? What is a collider, a confounder, a fork, and a chain?
- *Conditional independence and directed graphical models* — What is d-separation? Under what circumstances are two variables d-separated by a set of zero or more other variables? What does d-separation imply about conditional independence? How do you reason about d-separation in general network structures?

Causal graphical models (L4.16)

- *Causal semantics of graphical models* — In the context of causal inference, what are “treatments” and “outcomes”? How does causal dependence differ from statistical dependence?
- *do*-calculus — How do we represent perfect intervention in a graphical model? What is the do-calculus? How does it use operations on a Bayesian network to represent the effects of intervening on a variable?

Inferring causal effects (L4.18)

- *Assumptions* — What is the faithfulness assumption? What is the causal sufficiency assumption? What is the positivity assumption?
- *Randomization* — Why is randomization such an effective strategy when designing experiments? What threats to validity does it protect against? How can a graphical model be used to express the effects of randomization?
- *Instrumental variables* — What graphical structure represents a valid instrumental variable design? What are two necessary conditions for a valid “instrument”?
- *Propensity scores* — What is propensity score analysis and how does it work? How can a graphical model be used to express the effects of applying propensity score analysis? What types of conditional probability models are useful in conducting propensity score analysis?

Learning causal structure (L4.23)

- *Learning structure from data* — What are Markov equivalence classes? What are the two basic classes of algorithms for learning causal structure? How does d-separation facilitate learning causal structure from data?
- *Constraint-based learning of causal structure* — What are the two basic phases of the PC algorithm? What do conditional independencies imply about the existence of edges? What is the collider detection rule?
- *Score-based learning of causal structure* — What are the inputs and outputs of BIC? How is used by GES? What search strategies does GES use?