CMPSCI 348 **Exam 1 Review Sheet** Spring 2019

Exam 1 will be given in class on Thursday, February 28. The questions below cover the concepts that will appear on the exam, although the exam questions will require you to use some 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 their associated readings.

Probability Theory (L1.24, L1.29, L2.14)

- *Probability distributions and independence* — What is a probability distribution? What is the difference between a marginal, conditional, and joint probability distribution? What is the difference between independence and conditional independence?
- *Bayes' rule* — What is the multiplication rule and how can you use it to derive Bayes Rule?
- *Random variables* — What is a random variable? How are random variables used in typical data science applications?
- *Probabilistic programs* — What are probabilistic programs? What do probabilistic programs represent? What does a single execution of a probabilistic program represent?

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 goal of descriptive analytics, predictive analytics, and descriptive analytics. How do these three types of analytics differ? What are some methods typically used to perform descriptive and predictive analytics?
- *Data generating processes and statistical models* — What is a data generating process and a statistical model? How do data generating processes and statistical models differ?
- *Parametric and non-parametric models* — What are parametric and non-parametric models? How do these types of models differ? What are some examples of parametric and non-parametric models?

Data Representation and Sampling (L1.29, L1.31)

- *Basic data representation* — What are data instances? What is a “unit of analysis”? Given data with a hierarchical structure, how can you create a single data table in which columns are variables and rows are data instances? What circumstances produce data with dependent instances?
- *Variable types* — What are nominal, ordinal, interval, and ratio scales?
- *Outliers* — What are outliers? What are examples of methods that can be used to identify them?
- *Measures of central tendency* — What are the mean and median? Why would you use one versus the other?
- *Measures of dispersion* — What are standard deviation, variance, and interquartile range? Why would you use one versus the others?
- *Tukey’s ladder of powers* — What does Tukey’s ladder of powers provide the ability to do? Under what circumstances would you apply a particular element of Tukey’s ladder of powers?

Descriptive analytics (L1.31, L2.07)

- Exploratory data analysis — What is exploratory data analysis (EDA)? How does it differ from presentation graphics or modeling?

- *Data-ink ratio* — What does it mean to maximize the data-ink ratio in a graphic?
- *Small multiples* — What are “small multiples” and why would you use them?
- *Visualizing distributions* — What are useful methods of visualizing single-variable distributions? What are useful methods for visualizing conditional distributions? What are useful methods for visualizing joint distributions?
- *Box plots* — What are the parts of a box plot and what do they mean?
- *Kernel density estimators* — How does a kernel density estimator work? What are the relative advantages of different kernel functions?

Models (L1.29, L2.07, L2.14)

- *Parameters and hyper-parameters* — What are model parameters and hyper-parameters? What are examples of hyper-parameters for different types of models?
- *Model capacity* — What is model capacity? Why don’t we always use the highest capacity model?
- *Error components* — What are the two primary components of error for a learned statistical model?
- *Bias-variance tradeoff* — Under what conditions is the bias of a learned model virtually guaranteed to be large? Under what conditions is the variance of a learned model virtually guaranteed to be large? What causes bias and variance to be "traded off" against each other?

Linear models (L2.12)

- *Linear equations as a model representation* — What are the necessary and sufficient components of a linear model? How are those component interpreted to make statistical inferences about a data instance? How do we represent nominal variables with more than two values?
- *Construction* — How are the structure and parameters of a linear model estimated from data? What is the residual sum of squares (RSS)? What is the meaning of a *p*-value? What is correlation coefficient and what does it indicate?
- *Diagnostics* — What are residual plots and why are they useful? What is heteroskedasticity and what assumption does it correspond to? What are outliers and how can they be identified? How do different types of outliers affect the parameter estimates of linear regression?

Simple Bayesian Classifiers (L2.14)

- *Bayesian classifier as a representation* — What are the necessary and sufficient components of an simple Bayesian classifier (SBC)? How are those component interpreted to make statistical inferences about a data instance?
- *Construction* — How are the structure and parameters of an SBC learned?
- *Representational power of an SBC* — Why, in the 2-D graphics we examined in class, does the SBC only reproduce decision boundaries that are ovals? Could it do better? Could it produce an ellipse oriented along the diagonal?

k-Nearest Neighbor Classifiers (L2.14)

- *K nearest neighbor* — How does a KNN classifier work? What is Minkowski distance? What is the shape of possible decision boundaries for KNN? What are the tradeoffs of KNN vs. other classifiers?