Name ______________________

COMPSCI 348 **Practice Exam 1 Answer Sheet** (v.1) Spring 2019

Instructions (for the actual exam)

- *Do not open the exam until directed to by the instructor or TA*
- *Do not use books, notes, electronic devices, or other aids.*
- *Please avoid wrinkling the exam because that makes it difficult to scan.*
- *Your answers must be your own, so keep your eyes on your exam. Do not look at other students' exams.*
- *Answer each question. Note the point values and allocate your time accordingly.*
- *Be clear in marking your answers, and please place your answers in the designated spaces.*
- *Only the final answer in the designated space will be graded. However, other markings and calculations will be reviewed in support of regrade requests.*

Name ________________________

1. **Probability distributions** (6 points)

   Suppose that you gathered data about the characteristics and behavior of citizens in the 2015 presidential election in the fictional country of Sokovia. With respect to this data set, label each of the following questions as being best answered by either a *marginal* probability distribution (M), a *conditional* probability distribution (C), or a *joint* probability distribution (J).

   a. ___C___ Some reporters assumed that voting was easier for city residents than for rural residents. Thus, they wanted to know "*Did city residents vote in higher proportions than rural residents?*", expecting that the difficulty of voting would keep some rural voters away from the polls.

   b. ___M___ Election observers from the U.S. State Department were very concerned about whether the election was fair. State Department officials want to know *"What proportion of all citizens voted?"*, since a low proportion of voters can indicate a widespread belief among citizens that the election is not fair.

   c. ___J___ The U.S. State Department was also very interested in the overall characteristics of voters, including their age group (e.g., 18-22) and voting district. State Department analysts created a large two-dimensional table showing *"What percentage of all voters have a particular combination of age group and voting district?"*

2. **Independence and conditional independence** (4 points)

   Does marginal independence necessarily imply conditional independence? Briefly explain your answer or provide an example.

   No, marginal independence does not necessarily imply conditional independence. It is possible for two variables to be marginally independent, but dependent conditional on a third variable. Consider the equation Z = X + Y. X and Y can be marginally independent. However, if I know the value of Z, then X tells me something about the distribution of Y (indicating dependence).

Name ______________________

3. Bayes rule (6 points)

Derive Bayes rule from the product rule.

The general product rule (also called the "chain rule") states that:

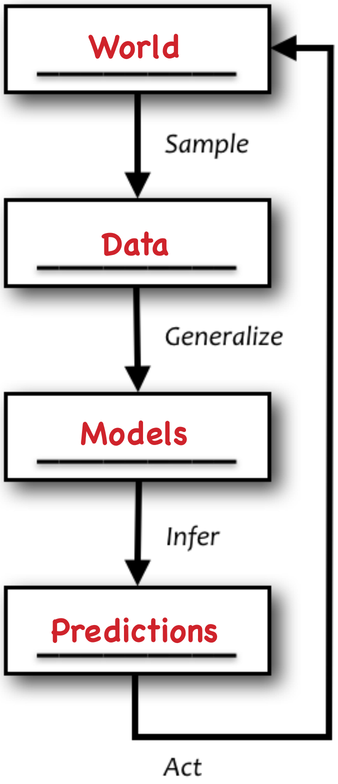
$$P(X,Y) = P(X)P(Y|X) = P(X|Y)P(Y)$$

Rearranging, we can get:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

4. Central Dogma (4 points)

At left is the diagram that was used in lecture to describe the "central dogma" of data science. In each box, fill in the missing word that describes the box. If necessary, use the space below to describe what you mean by the words.



Name ________________________

5. **Units of analysis** (5 points)

Read the description below, and answer the question.

"*Citizens* of Sokovia cast their votes for president via a secret ballot. That means that data analysts cannot know individual *votes*. Instead, votes are reported by *precinct* (geographic units of a few thousand people). Voting in precincts is aggregated at the *state* level, which each state contains may precincts. Each state then casts one vote for a *presidential candidate*, depending on who receives the most votes in that state."

If an analysis aims to discover what factors are associated with voting for a particular presidential candidate, what is the right *unit of analysis*?

______Precinct______

Ideally, we would like to get data about voters (or, equivalently, votes), but that's not available, so precinct is the best we can do. We want precincts because it is the lowest-level (least aggregated) unit that carries information about both voting behavior and other factors (e.g., age, prior voting behavior, party affiliation, etc.).

6. **Analytic tasks** (6 points)

For each question given below, name the most relevant analytic task: *descriptive analytics (desc)*, *predictive analytics (pred)*, and *prescriptive analytics (pres)*. Note that all some analytic tasks will be used more than once and some may not be used at all.

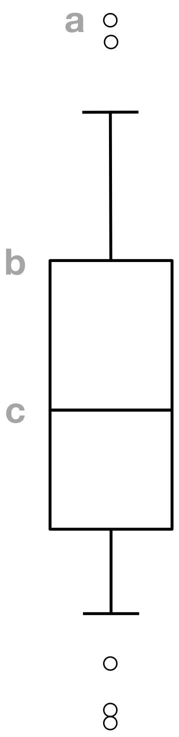
a. __desc__ What is the age distribution of voters in Sokovia?

b. __pres__ What discouraged some citizens from voting in the 2015 presidential election in Sokovia?

c. __pred__ Which types of citizens were most likely to vote for current president rather than the challenger?

Name ________________________

7. Box Plots (6 points)



For the box plot shown at right, indicate the meaning of the parts labeled **a**, **b**, and **c** in the diagram.

a. outlier

b. 75th percentile

c. median or 50th percentile

8. Transformations (6 points)

Briefly describe Tukey's Ladder of Powers, and provide at least one example of when you might wish to use it.

Tukey's Ladder of Powers is a family of transformations that can be applied to variables measured on interval or ratio scales. Specifically, the ladder of powers uses the equation:

$$y = \begin{cases} x^{\lambda}, & \text{if } \lambda > 0 \\ log(x), & \text{if } \lambda = 0 \\ -(x^{\lambda}), & \text{if } \lambda < 0 \end{cases}$$

One common use is to attempt to make a non-linear variable linear so that it can be accurately modeled using linear regression.

Name ________________________

9. Model types (6 points)

Briefly describe the difference between parametric and non-parametric models. Use examples from the model families that we have discussed in class.

Parametric models can be described using a fixed number of parameters. Simple linear regression and simple Bayesian classifiers are examples of parametric models. In contrast, non-parametric models grow in the number of parameters to accommodate the complexity of the data. Kernel density estimators and k nearest neighbor classifiers are examples of non-parametric models.

10. Parameters and hyper-parameters (4 points)

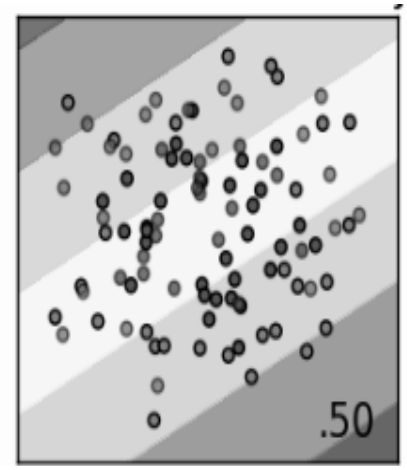
Briefly describe the difference between parameters and hyper-parameters in statistical models. Use one or more examples from models we have discussed in class.

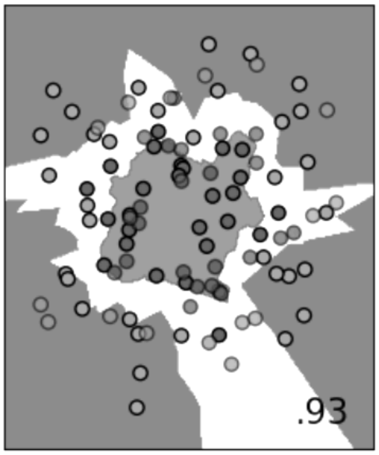
Parameters are numeric or symbolic values that determine how a model maps from input to output. The slope of a linear regression model is an example of a parameter. Hyper-parameters are numeric or symbolic values that control the capacity of the learned model. The value K in a k-nearest neighbor model or the bandwidth parameter of a kernel density estimator are examples of hyper-parameters.

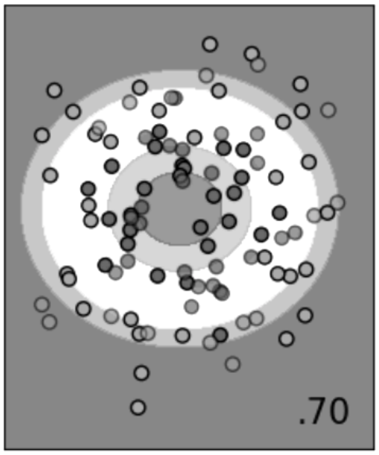
Name ________________________

11. Comparing classifiers (5 points)

Below are graphs showing how three different classifiers assign probabilities to different classes of data points. Each graph shows two input (predictor) variables (shown on the x and y axes) and one output (class) variable (shown as point shading). Label each graph as a simple Bayesian classifier (SBC), K nearest neighbor classifier (KNN), or linear discriminant analysis (LDA).







LDA

KNN

SBC

12. Kernel density estimators (8 points)

a. Briefly describe how a kernel density estimator is constructed from a set of data points.

KDEs are constructed by summing a set of smaller distributions each of which corresponds to a data point and each of which has total area 1/N.

b. Briefly describe the computational advantage of the Epanechnikov kernels over the the Gaussian kernel.

Epanechnikov kernels have bounded range, whereas Gaussian kernels do not. This reduces computational complexity for obtaining estimates from a KDE.

Name ________________________

13. Simple Bayesian classifiers (12 points)

a. A simple Bayesian classifier stores a probability distribution for each of several predictor variables that are used to help estimate the probability of the class variable. For the predictor variable X, specify that probability distribution in probability notation (e.g., $p(X)$).

$P(X|C)$

b. A simple Bayesian classifier also stores a probability distribution for the class variable that is used to help estimate the probability of the class variable. For the class variable C, specify that probability distribution in probability notation (e.g., $p(X)$).

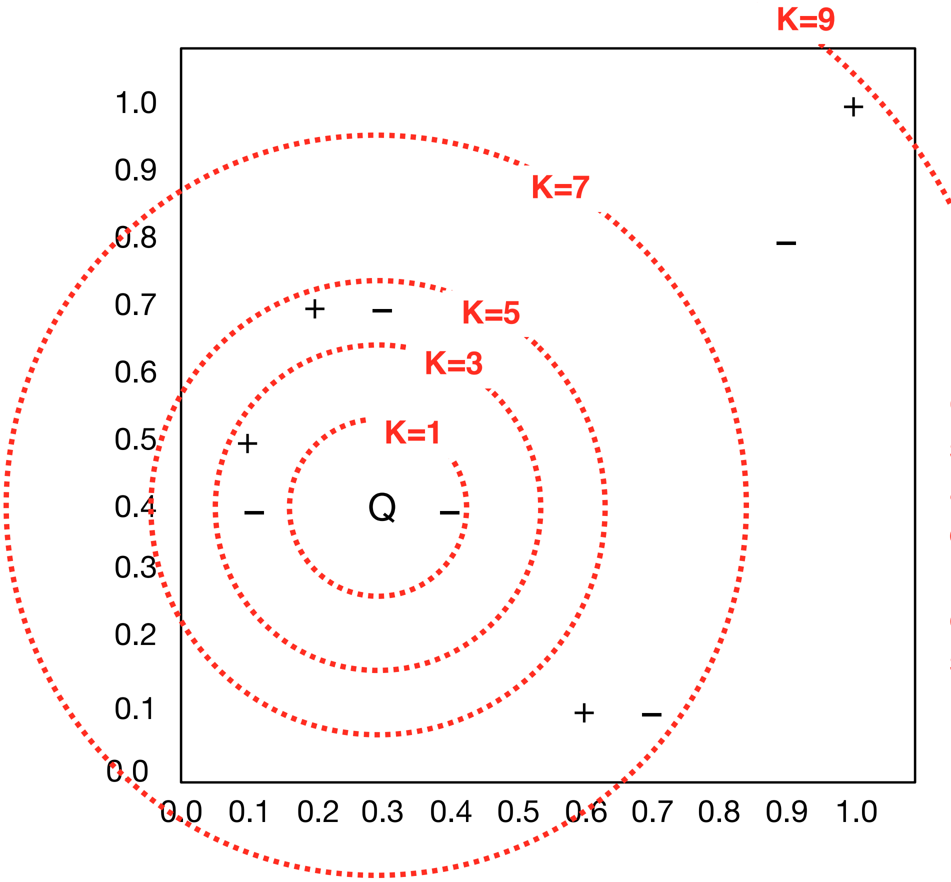
$P(C)$

c. Briefly describe the independence assumption made by a simple Bayesian classifier.

SBCs assume that predictor variables are conditionally independent, given the class.

Name ________________________

14. K nearest neighbor (10 points) — In the data set shown below, the two axes represent features and the symbols {+, –} represent classes. For each value of K shown below, indicate the output for a KNN classifier for the query instance Q assuming Minkowski distance with p=1.



Concentric rings shown at left are for explanation only, and are not part of the answer expected from students.

a. ____–____ K = 1

b. ____–____ K = 3

c. ____–____ K = 5

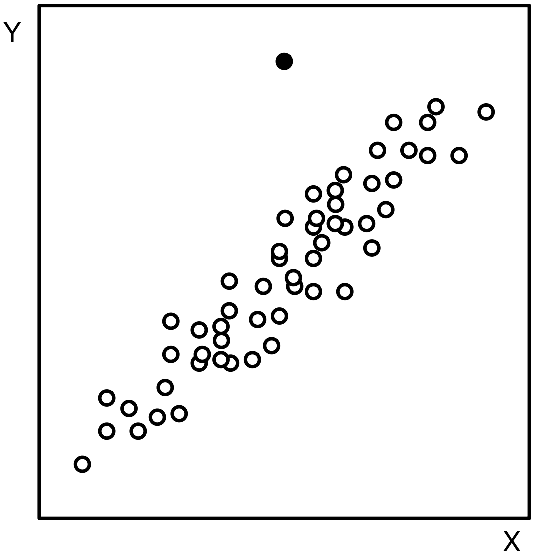
d. ____–____ K = 7

e. ____–____ K = 9

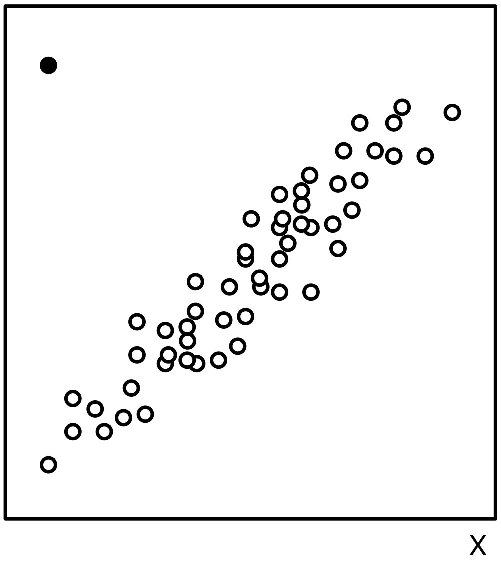
Name ________________________

15. Linear models and outliers (4 points)

Below are two plots with a large number of data points (open circles) and a single outlier (filled circle). In each case, briefly describe the impact of the outlier on the slope and intercept of the linear regression model.



(i)



(ii)

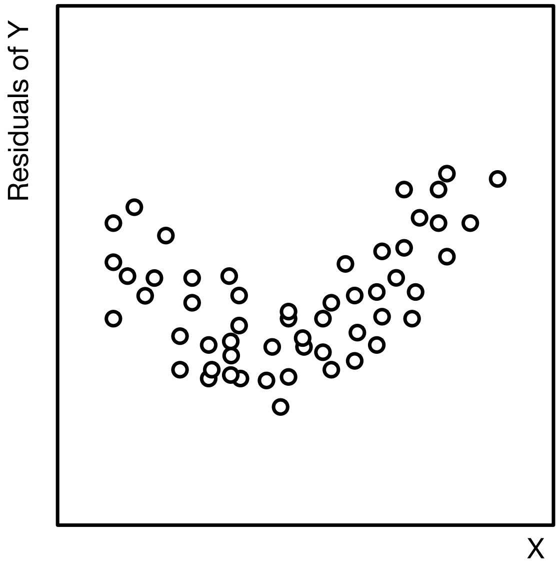
(i) The outlier will affect the y-intercept, but not the slope, of the linear equation produced by simple linear regression.

(ii) The outlier will affect the slope and the intercept of the linear equation produced by simple linear regression.

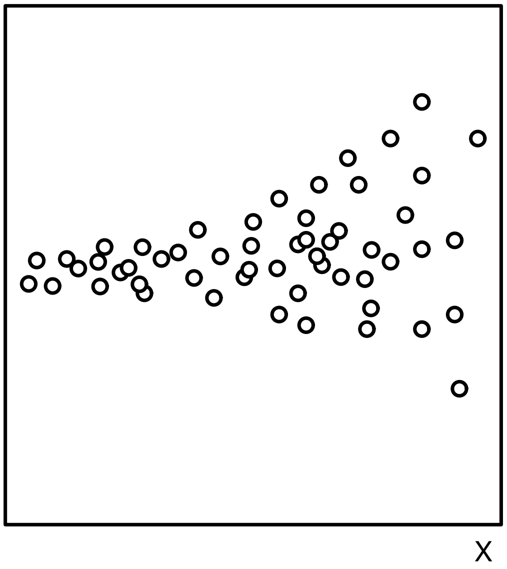
Name ________________________

16. Linear models and residual plots (4 points)

Below are two residual plots for a linear regression model. In each case, name the violation of assumptions that is indicated by the plot.



(i)



(ii)

(i) Linearity (of Y in terms of X)

(ii) Homoskedasticity (or equal variance of Y across the range of X)

Name ________________________

17. Linear regression and nominal variables (4 points)

Briefly describe how a simple linear regression model represents nominal variables with more than two values.

For a nominal variable with n values, a standard approach in simple linear regression is to represent the nominal variable as a set of n-1 binary variables, in which each new variable encodes whether the original variable takes on the ith value, for i = 1….(n-1). The nth value is denoted by making all n-1 new variables equal to zero.