Name _____

COMPSCI 348 **Exam 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 of meals served at the fictional fast-food restaurant RocketFood. Variables about each meal include its price, the time and day of the week it was ordered, the total number of calories, and also variables indicating whether the meal includes various components (e.g., burger, fries, Coke, ice cream, etc.). 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).

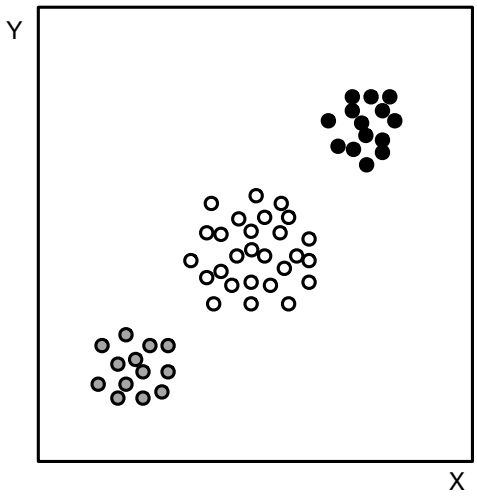
a. J The advertising director suspects that people will pay more for low-calorie meals. He wants to know: "*What proportion of all meals are both low-calorie (<500 calories) and expensive (>$8.00)?"*

b. C The restaurant owner has a theory that customers on weekends (Saturday and Sunday) spend more money than customer on weekdays (Monday-Friday). One of her questions is: *"What is the average cost of a meal ordered on a weekend?"*

c. M The manager of RocketFood wants to understand more about the types of meals that are ordered by the restaurants customers. His question is: *"What proportion of meals include french fries?"*

2. Independence and conditional independence (4 points)

At right is a scatterplot showing the values of three variables. The values of the variables *X* and *Y* are shown by the positions of points on the x and y axes of the plot, and *Z* (a nominal variable with three values) is shown using the color of each point. Mark each statement below as true (T) or false (F).



a. F The variables *X* and *Y* are marginally independent.

b. T The variables *X* and *Y* are conditionally independent given *Z*.

X and Y cannot be marginally independent, because P(X) ≠ P(X|Y). However, within each set of points of a single color (single value of Z), P(X) = P(X|Y).

Name _____

3. Bayes rule (6 points)

Your doctor tells you that you have a serious disease, and that 80% of the people who have recovered from the disease have used his new treatment. That is, *P(Treatment | Recovery)* = 0.80. *Using probability notation*, identify the two questions you need to ask your doctor to accurately estimate your probability of recovery given that you use his treatment (Hint: Be sure to use Bayes' Rule).

P(Recovery)

P(Treatment)

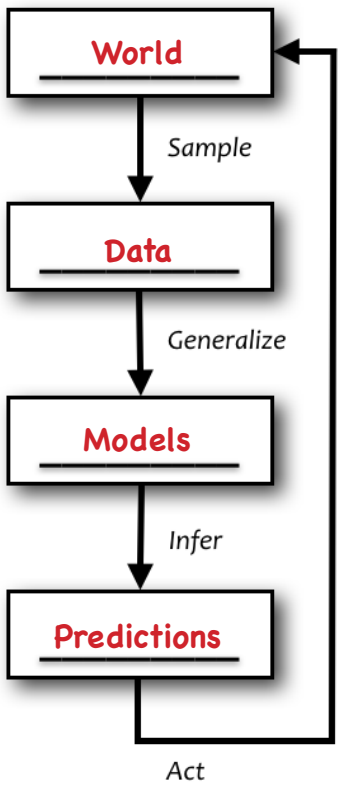
Bayes' Rule states:

$$P(R \mid T) = \frac{P(T \mid R)P(R)}{P(T)}$$

and we already know P(T | R).

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.

"Orders at RocketFood are generally paid for by one person, but those orders often contain meals for multiple customers (e.g., a group of friends or a family group). Even a meal for one person usually consists of multiple items (e.g., a burger, fries, and a drink)."

If an analysis aims to discover the level of customer satisfaction with the meals at RocketFood, what is the right *unit of analysis*? Write the name in the space below.

Customer or Meal

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. pred Which food items are frequently ordered together (e.g., burger and fries or salad and diet drink)?

b. desc What is the distribution of price for meals ordered at RocketFood?

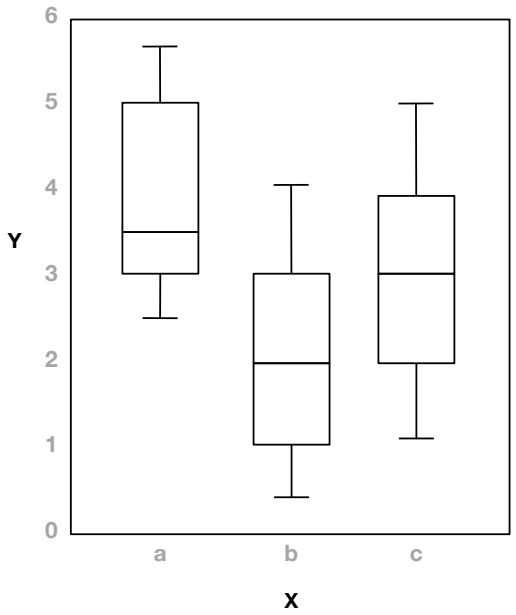
c. pres Do "buy one, get one free" deals increase the total amount that customers spend at RocketFood?

6(a) is predictive because knowing something about one food item will change your estimate of the probability of another item being in the order.

Name _____

7. Box Plots (6 points)

The box plot shown at right show the values (a, b, and c) of a nominal variable X and the corresponding distributions of a continuous variable Y. Answer each question below as true (T) or false (F).



a. ___T___ P(Y>5|X=a) = 0.25.

b. ___F___ P(Y|X)=P(Y).

c. ___T___ If P(X=a) = P(X=b) = P(X=c), then P(Y>3) = 0.5.

8. Transformations (4 points)

Tukey's Ladder of Powers takes two inputs: a value *x* and a parameter λ. It generates an output *y*. In the space below, write down two of the three functions of *x* and λ that produce the value *y*.

log(x) or x^{λ} or $-x^{\lambda}$

log(x) or x^{λ} or $-x^{\lambda}$

Name _____

9. Model types (6 points)

For each type of model below, label it as parametric (P) or non-parametric (NP).

a. P Simple Bayesian classifier

b. NP K nearest neighbor

c. NP Neural network

10. Parameters and hyper-parameters (6 points)

For the k nearest neighbor (KNN) classifier, label each of the items below as a parameter (P) or a hyper-parameter (H).

a. H The value of *p* used in the Minkowski distance measure.

b. H The value of *k* (the number of nearest neighbors used for classification).

c. P The variable values associated with a given point in the training set.

11. Comparing classifiers (5 points)

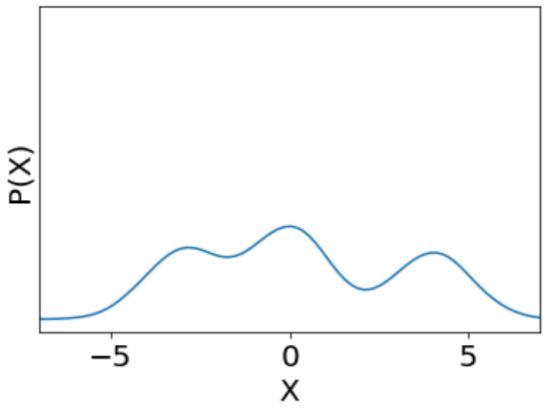
Consider the problem of "data set poisoning". A classification model *M* is learned using a data set *S*. You are given the task of inserting a small number (e.g., 1-3) of new data instances into *S,* with the goal of changing the estimate that *M* will make for a specific test instance. If you want to be successful, which model would you most want to be used for *M*: Linear discriminant analysis (LDA), k nearest neighbor (KNN), or simple Bayesian classifier (SBC)? Write the answer in the space below.

LDA or KNN

Name _____

12. Kernel density estimators (8 points)

Suppose the figure at right is a probability density produced by a kernel density estimator from a training set of size 100 (not shown in the figure). All of those points lie between –5 and 5. Answer the questions below:



a. What is the total area under the probability density?

1

b. Suppose that thirty of the original data points lie between the x-axis locations of 0 and 4. What (approximately) will be P(0<X<4)?

0.30

c. Suppose that we increased the value of the bandwidth parameter. Would we expect that $P(X \leq -5)$ would *increase* (⬆) *decrease* (⬇), or *stay the same* (**0**)?

⬆

d. Is a kernel density estimator a parametric (P) or non-parametric (NP) model?

NP

Name _____

13. Simple Bayesian classifiers (12 points)

Suppose that your goal is to learn a model that classifies email messages as "spam" or legitimate. You decide to use a simple Bayesian classifier. The training data consist of 1000 email messages that you have labeled as either spam or legitimate (*C=spam* or *C=legit*) and for which you know three predictor variables ("features"): (1) the length of the email (e.g., *L=234*); (2) whether the sender is in your address book (*A=True* or *A=False*); and (3) the top-level domain of the sender (e.g., *D=com*, *D=edu*, etc.).

a. Suppose you train your simple Bayesian classifier and then want to classify a new instance where *L=452*, *A=True*, and *D=gov*. Using probability notation, name the six probability values (that carry information about features) that you would need in order to estimate *P(C=spam | L=452, A=True, D=gov).* (Hint: Be specific).

P(L=452 | C=spam) P(A=True | C=spam) P(D=gov | C=spam)

P(L=452 | C=legit) P(A=True | C=legit) P(D=gov | C=legit)

b. Using probability notation, name the two other probability values you would need (that don't carry information about features). Don't worry if one can be derived from the other

P(C=spam)

P(C=legit)

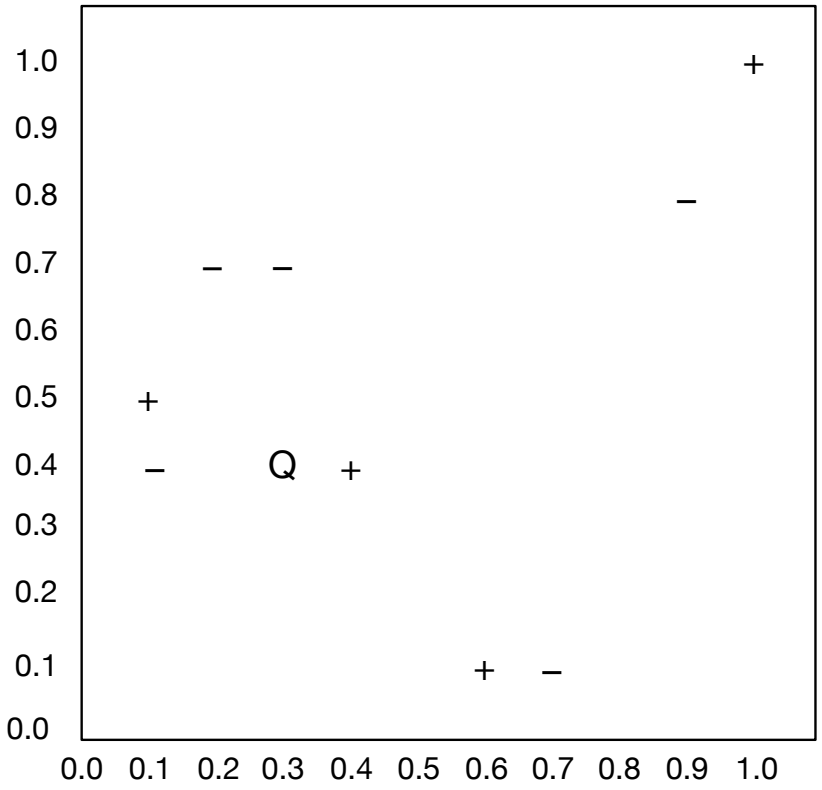
c. Suppose you substitute a new data set of 3000 instances where, for each email message, you only know the class (*C=spam* or *C=legit*) and <u>one</u> of the three feature values, either *L* or *A* or *D* (but not all three). For each message, the feature value that you know is selected uniformly at random. Will the simple Bayesian classifier learned with this data set be approximately more accurate (More), less accurate (Less), or equivalently accurate (Equal) to the SBC above?

Equal

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=2.



a. ___+___ K = 1

b. ___+___ K = 3

c. ___–___ K = 5

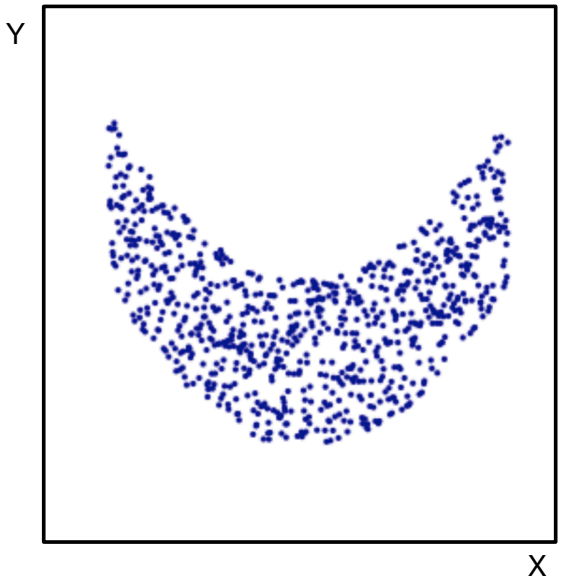
d. ___–___ K = 7

e. ___–___ K = 9

Name _____

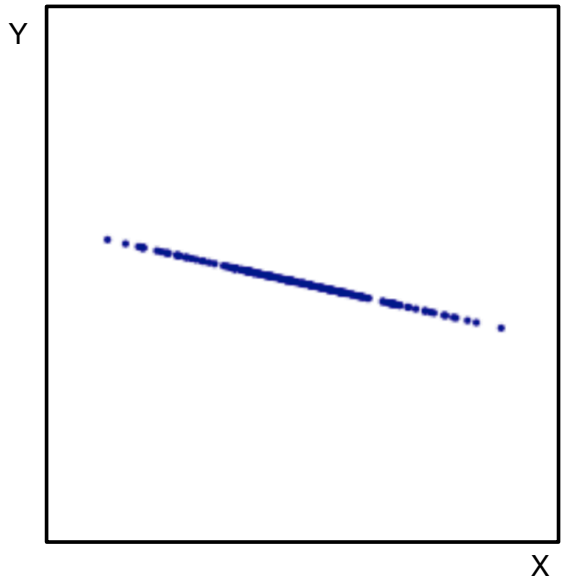
15. Linear models and outliers (4 points)

Below are two plots with a large number of data points. For each plot, estimate the value of the correlation coefficient and write it in the space below the plot.



(i)

0.0



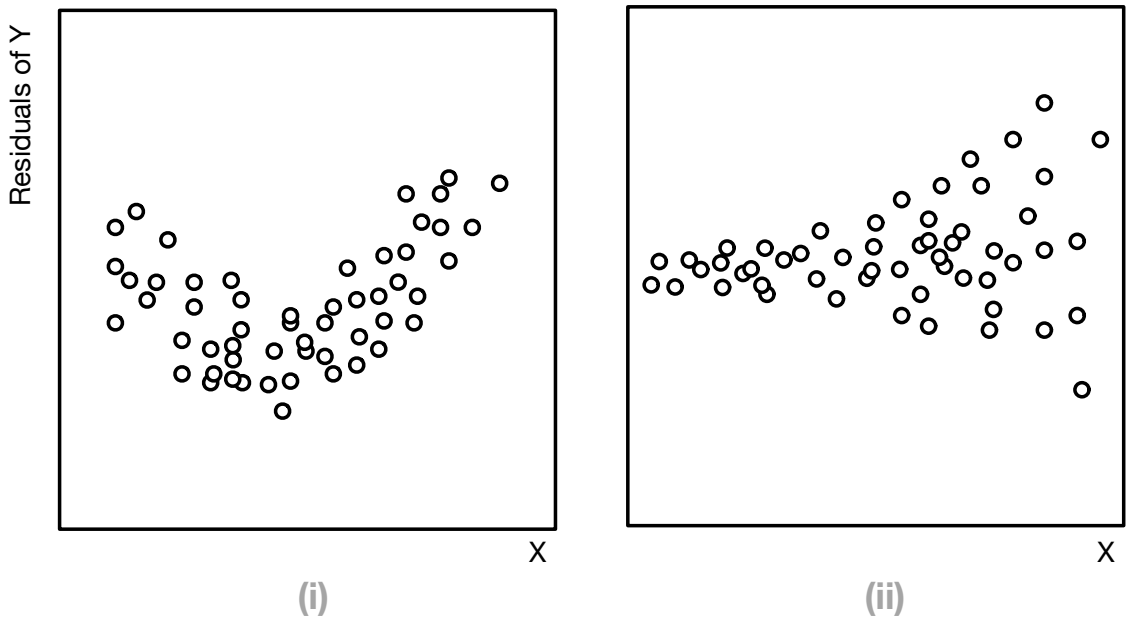
(ii)

-1.0

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 and write it in the space below the corresponding plot.



(i) (ii)

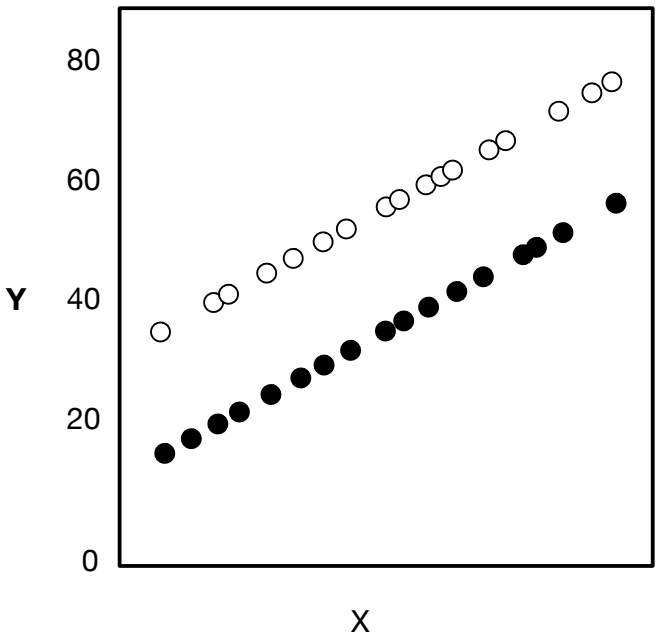
Linearity _____ Homoskedasticity _____

Name _____

17. Linear regression and nominal variables (4 points)

In the plot below, the points represent predictions of a linear regression model with two predictor variables, X_1 and X_2. The variable X_1 is measured on a ratio scale, and values of X_1 are shown by the position of the point on the X axis. X_2 is measured on a nominal scale (with only two values), and values of X_2 are shown based on the color of the point.

_____20_____ Estimate the value of β_2, the coefficient in the regression equation corresponding to X_2.



X_2 is a binary variable that only affects the y-intercept of the model, not the slope. The amount that it changes the y-intercept can be identified by following the two lines to the y-axis and estimating the distance between those two y-intercepts (30-10=20).