Data 102, Fall 2021 Midterm 2- Solutions

- You have 110 minutes to complete this exam. There are 5 questions, totaling 41 points.
- You may use one 8.5×11 sheet of handwritten notes (front and back). No other notes or resources are allowed.
- You should write your solutions inside this exam sheet.
- You should write your name and Student ID on every sheet (in the provided blanks).
- Make sure to write clearly. We can't give you credit if we can't read your solutions.
- Even if you are unsure about your answer, it is better to write down partial solutions so we can give you partial credit.
- We have provided two blank pages of scratch paper, both near the end. No work on these pages will be graded.
- You may, without proof, use theorems and facts that were given in the discussions or lectures, **but please cite them**.
- There will be no questions allowed during the exam: if you believe something is unclear, clearly state your assumptions and complete the question.
- Unless otherwise stated, no work or explanations will be graded for multiple-choice questions.
- Unless otherwise stated, you must show your work for free-response questions in order to receive credit.

Last name	
First name	
Student ID (SID) number	
Calcentral email (`@berkeley.edu`)	
Name of person to your left	
Name of person to your right	

Honor Code

I will respect my classmates and the integrity of this exam by following this honor code.

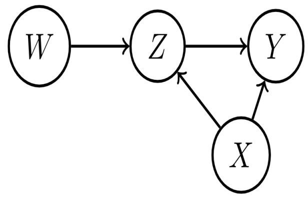
I affirm:

- All of the work submitted here is my original work.
- I did not collaborate with anyone else on this exam.

Signature: ______________________________

1. (9 points) For each of the following, circle either **TRUE** or **FALSE**.

(a) (1 point) (TRUE / FALSE) In the following causal graph, W is a valid instrumental variable for the effect of X on Y.



FALSE

(b) (1 point) (TRUE / FALSE) The UCB algorithm is guaranteed to have a smaller regret than Explore-then-Commit algorithm in a Multi-Armed Bandits problem.

FALSE

(c) (1 point) (TRUE / FALSE) Suppose $X_1, ..., X_n$ are i.i.d. samples from a known but unbounded distribution $\mathcal{P}$. Then, no matter the choice of $\mathcal{P}$, we can use Hoeffding's inequality to produce a guaranteed upper bound on the probability $P(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \geq 1)$.

FALSE

(d) (1 point) (TRUE / FALSE) The coefficients fitted in both ridge and LASSO regression are maximum a posteriori estimates.

TRUE

(e) (1 point) (TRUE / FALSE) The moment generating function $M_X(t) = E[e^{tX}]$ can be expanded as $M_X(t) = \sum_{n=0}^{\infty} a_n * t^n$. The value of $a_n = \frac{E[X^n]}{n!}$.

TRUE

(f) (1 point) (TRUE / FALSE) The role of the link function in a generalized linear model is that it transforms the response such that it is linearly related to the predictors.

TRUE

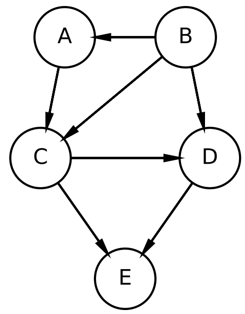
(g) (1 point) (TRUE / FALSE) The common name for a generalized linear model with binomial likelihood is Poisson regression.

FALSE

(h) (1 point) (TRUE / FALSE) For count data, coefficients obtained from Poisson regression are an example of a maximum likelihood estimate.

TRUE

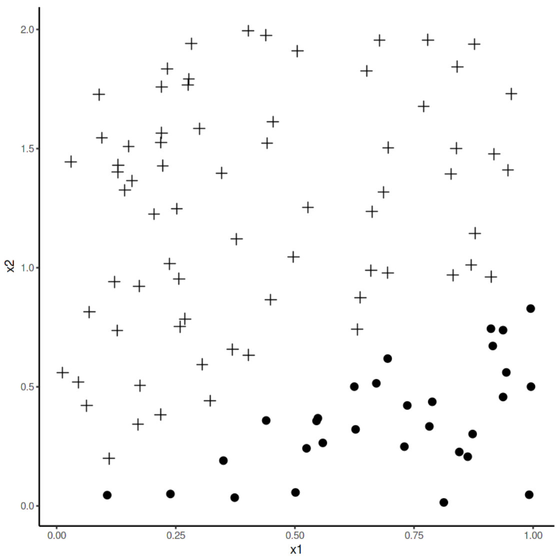
(i) (1 point) (TRUE / FALSE) In the causal diagram below, if we are trying to quantify the causal relationship between B and D, adjusting only for C satisfies the backdoor criterion.



FALSE

2. (8 points) For each of the following plots, answer the corresponding question. You should assume that all training points are visible on each graph. **Select all that apply.**

(a) (2 points) Given only X_1 and X_2 as your features, which of the following classification algorithms can achieve a training accuracy of 1? Select all that apply.



A) Decision Trees

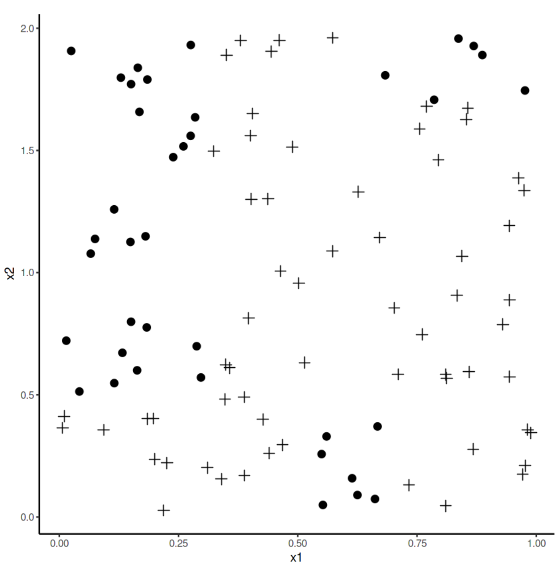
B) Logistic Regression

C) Random Forest

D) Neural Network

Solution: A, B, C, D. Since the two classes are linearly separable, any of the above methods can be used.

(b) (2 points) Given only X_1 and X_2 as your features, which of the following classification algorithms can achieve a training accuracy of 1? Select all that apply.



A) Decision Trees

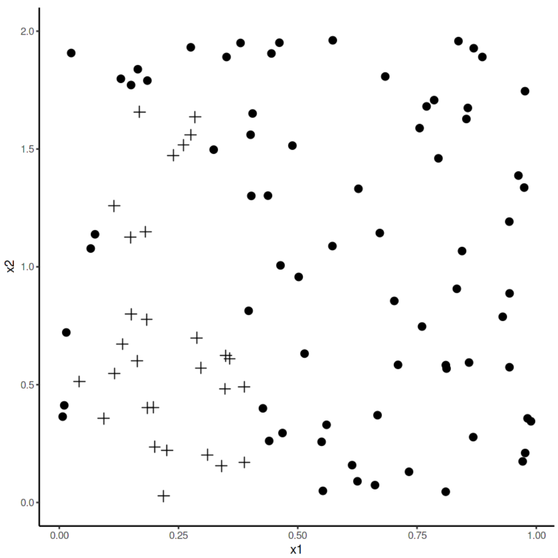
B) Logistic Regression

C) Random Forest

D) Neural Network

Solution: A, C, D. Since this is the two classes are not linearly separable, we can only use methods that are will create boundaries that fully separate the two classes.

(c) (2 points) Given only X_1 and X_2 as your features, which of the following classification algorithms can achieve a training accuracy of 1? Select all that apply.



A) Decision Trees

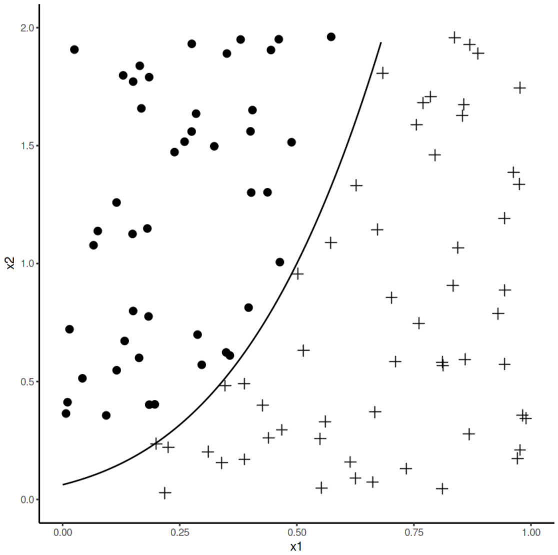
B) Logistic Regression

C) Random Forest

D) Neural Network

Solution: A, C, D. Since this is the two classes are not linearly separable, we can only use methods that are will create boundaries that fully separate the two classes.

(d) (2 points) Given only X_1 and X_2 as your features, which of the following classification algorithms can output the plotted decision boundary? Select all that apply.



A) Decision Trees

B) Logistic Regression

C) Random Forest

D) Neural Network

Solution: D. Only neural networks can generate a curved decision boundary.

3. (4 points) Recall the setup of Markov Decision Processes from lecture: given an agent X, at each time $t = 1, 2, ..$, the agent is in state $s_t \in S$, where S is the set of all possible states. It then takes an action $a_t \in A$, where A is the set of actions it can take. The transition function is given by $T(s, a, s')$, which returns the probability that action a from current state s leads to state s'. In addition to these, there is also a reward function $R(s)$ which denotes the reward the agent gets when it enters state s.

   (a) (1 point) (TRUE / FALSE) The optimal policy mandates that the best move for agent X at time t is always to move to a state s' such that

$$s' = \underset{s \text{ reachable from } s_t}{\text{argmax}} R(s),$$

   that is, the state with the highest reward.

   Solution: False. The optimal policy gives the action to take at each step in order to maximize the total reward we get at the end of the game. We might get larger reward in the end by taking a move with less reward in the current round.

   (b) (1 point) (TRUE / FALSE) The value (utility) function $V^*(s)$ of a state s is the reward you get entering state s, i.e. $R(s)$.

   Solution: False. See the definition of the value function in the lecture slides

   (c) (2 points) Consider the recurrence relation of the value function:

$$V^*(s) = \max_{s} \sum_{s'} T(s, a, s')[R(s') + \gamma V^*(s')]$$

   When solving, one major issue we discussed in lecture is that this equation is circular. How can you avoid this circularity by introducing an additional variable? With this extra variable, how can you recover $V^*(s)$?

   Solution: You can avoid circularity by adding a time component, i.e. $V^*(s, t)$. You can recover $V^*(s)$ by taking $t \to \infty$ (as described in lecture) **OR** through Dynamic Programming.

4. (11 points) **Movie Recommendations**

You are a data scientist working for a brand-new content streaming start-up, tasked with building a movie recommendation system. Since your company is quite new, you don't have much user data to work with. To simplify the problem, you start by developing a movie recommendation system to predict the user's favorite genre, either Horror or Comedy. You have the results of a user survey with the following information:

- Movie genre preference (either Horror or Comedy)
- Total number of Horror movies watched
- Total number of Comedy movies watched
- Average rating for Horror movies (between 1 and 5)
- Average rating for Comedy movies (between 1 and 5)

For parts (a) and (b), you'll use a logistic regression model to predict a user's favorite genre. After some feature engineering, you decide on the following features and predictions:

- First feature: $X_1 = \#$ Horror movies watched $- \#$ Comedy movies watched
- Second feature: $X_2 =$ Average Horror movie rating$-$Average Comedy movie rating
- Output: $Y = \begin{cases} 1 & \text{if user prefers Horror movies} \\ 0 & \text{if user prefers Comedy movies} \end{cases}$

(a) (2 points) Suppose you use frequentist logistic regression.

The results of the model fit are presented below. Is the model a good fit for the data? Justify your answer.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      Y   No. Observations:                  100
Model:                            GLM   Df Residuals:                       97
Model Family:                Binomial   Df Model:                            2
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -4.0585
Date:                Tue, 02 Nov 2021   Deviance:                       8.1171
Time:                        13:16:45   Pearson chi2:                     16.9
No. Iterations:                     9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -1.6757      1.856     -0.903      0.367      -5.314       1.963
X1             0.3712      0.453      0.820      0.412      -0.516       1.258
X2             4.2448      1.928      2.201      0.028       0.465       8.024
==============================================================================
```

Solution: Yes, the model is a good fit for the data. If the model was a good fit for the data, we expect the average log-likelihood to be close to 0. In this case, we can see that the total log-likelihood is around -4, so the average log-likelihood (over 100 observations) is around -0.04 which is indeed close to zero. In addition, the Deviance and Pearson χ^2 are both much smaller than the threshold of $n - p = 97$.

(b) (2 points) *Model Checking for Bayesian GLM*

Suppose instead you used a Bayesian Logistic Regression model. After observing several users' behavior, you find the posterior distribution $q(\beta)$ over regression coefficients $(\beta_0, \beta_1, \beta_2)$. Write a formula for the posterior predictive probability that the next user's favorite movie genre is Horror, given that that user has watched 15 more comedy movies than horror movies, and rates horror movies 1.5 stars higher than comedy movies on average. You do not need to simplify your answer.

Hint 1: Your answer should be given in terms of an integral.

Hint 2: To find the likelihood, consider that Bayesian logistic regression uses the model:

$$\mathbb{E}\left[Y|X, \beta_0, \beta_1, \beta_2\right] = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}$$

Solution: The formula for posterior predictive probability is given by:

$$\mathbb{P}[Y_{rep} = y|\mathbf{X}, Y_{obs}] = \int_{\mathbb{R}^k} \mathbb{P}[Y_{rep}|\mathbf{X}, \beta]\mathbb{P}[\beta|\mathbf{X}, Y_{obs}]d\beta$$

where $\mathbb{P}[\beta|\mathbf{X}, Y_{obs}]$ is the posterior distribution of the regression coefficients β, $\mathbb{P}[Y_{rep}|\mathbf{X}, \beta]$ is the likelihood, and k is the number of regression coefficients within the β vector. From the information in the problem, we want to find

$$\mathbb{P}[Y_{rep} = 1|X_1 = -15, X_2 = 1.5, Y_{obs}]$$

We are told in the problem that the posterior distribution over β is given by $q(\beta)$. To solve for the likelihood, we have that:

$$\mathbb{P}[Y_{rep} = 1|X_1 = -15, X_2 = 1.5, \beta] = \mathbb{E}\left[Y|X_1 = -15, X_2 = 1.5, \beta_0, \beta_1, \beta_2\right]$$
$$\mathbb{P}[Y_{rep} = 1|X_1 = -15, X_2 = 1.5, \beta] = \frac{\exp(\beta_0 - 15\beta_1 + 1.5\beta_2)}{1 + \exp(\beta_0 - 15\beta_1 + 1.5\beta_2)}$$

Plugging the quantities for the likelihood and the posterior distribution in, we get that:

$$\mathbb{P}[Y_{rep} = 1|X_1 = -15, X_2 = 1.5, Y_{obs}] = \int_{\mathbb{R}^3} \frac{\exp(\beta_0 - 15\beta_1 + 1.5\beta_2)}{1 + \exp(\beta_0 - 15\beta_1 + 1.5\beta_2)} q(\beta)d\beta$$

For the remainder of the question, you will solve the problem in an online setting. Specifically, suppose you have a new user who hasn't watched any movies yet, and hasn't filled out your survey. Instead of logistic regression, you decide to take a different strategy to provide movie recommendations: using multi-armed bandits. Assume that you need to make a sequence of T recommendations for this user, where each time you suggest to them either a Horror or Comedy movie. The result of interest is the rating the user provides to the recommended movie.

(c) (1 point) For this multi-armed bandit problem, what is the total number of arms?

> **Solution:** There are 2 total arms, corresponding to either a Horror or Comedy movie recommendation.

(d) (2 points) Formulate this problem in the Multi-Armed Bandit framework. Specifically, provide definitions of the variables A_t and R_t, which are the arm chosen at time t and reward at time t respectively, in the context of the problem.

> **Solution:**
>
> A_t = the genre of the movie recommended to the user at time step t
>
> R_t = the user's rating of the movie recommended to them at time step t

(e) (2 points) *Offline vs. Online Decision Making*

Considering the same new user who hasn't watched any movies, explain why multi-armed bandits would be preferable over logistic regression.

> **Solution:** Since we are told that this user has not watched any movies on the platform yet, we do not have access to their feature variables, which form the input into a Logistic Regression classifier. As a result, we cannot use Logistic Regression to make good recommendations for this user. In contrast, Multi-Armed Bandits algorithms operate in an online setting, where they are able to make recommendations without any prior knowledge of the user's preferences. Over time, MABs will be able to learn the user's preferences and serve good recommendations.

(f) (2 points) *Checking Properties of MABs*

Some of the assumptions of the multi-armed bandits framework are **not** likely to hold for this particular setting. Give one such assumption, and explain why it's not likely to hold.

> **Solution:** The two properties that we expect for MABs are:
>
> - X_t is conditionally independent of $A_1, X_1, A_2, X_2, \ldots, X_{t-1}$ given A_t.

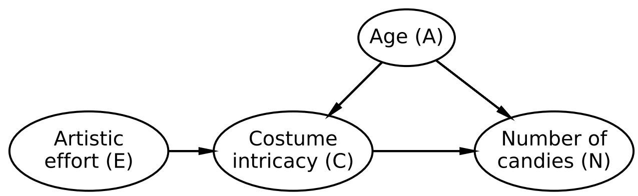
- The learner's selection of A_t is allowed to be probabililistic, but the policy π_t can only depend on the past $(A_1, X_1, A_2, X_2, \ldots, X_{t-1})$ and not the future $(X_t, A_{t+1}, \ldots, A_n, X_n)$.

In our case, the second property seems reasonable to hold, since the next genre we recommend will only rely on the genres we previously picked and ratings we've observed so far. However, the first property may not hold in practice. For example, let's say you provide several sub-optimal recommendations to the user at the beginning of the process. Then, the user may get frustrated and change the way they rate the subsequent movies (i.e. the reward distributions change).

This page intentionally left blank for scratch work. No work on this page will be graded.

5. (9 points) **Causal Halloween Candy**

You are observing the Data 102 Halloween party, where trick-or-treaters are arriving and the GSIs are passing out candy to each one. Some of the trick-or-treaters are children and some of them are adults. Some are wearing more intricate (fancy) costumes than others. You notice that the GSIs give different trick-or-treaters different amounts of candy. You believe that the amount of candy given to a specific trick-or-treater is related to the intricacy of their costume, the effort put into their costume, and their age. You come up with the causal diagram below:



Age (A), Costume intricacy (C), and Artistic effort (E) are all binary random variables (0=low, 1=high). For this question, you should assume that the diagram above represents the true causal relationships, and that all relationships are linear.

Over the course of the party, you count 120 trick-or-treaters. You want to quantify the causal relationship between costume intricacy (C) and number of candies received (N).

(a) (2 points) *Average Treatment Effect*

You decide to investigate the treatment effect of Costume intricacy on Number of candy pieces. Which of the following produce an unbiased estimate of the Average Treatment Effect (ATE)? **Select all that apply and justify your answer**.

(A) $\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} N_i C_i - \frac{1}{120-n_1}\sum_{i=1}^{n} N_i(1-C_i)$, where n_1 is the number of students with $C_i = 1$.

(B) The slope of a line that uses simple linear regression to predict N from C.

(C) The coefficient β_C, where (β_A, β_C) are the result of running least squares on N with covariates A and C.

(D) $P(N = 1|C)$.

Solution: Only option C provides an unbiased estimate of the ATE. Both options A and B fail to account for the confounding effect of A, while option D both fails to account for confounding and also measures the target quantity. C works (assuming the true relationship is linear) because it controls for the confounder A.

(b) (2 points) *Instrumental variables*

Is it reasonable to use Artistic Effort as an instrumental variable when performing 2-stage least squares regression to predict the treatment effect of costume intricacy on Number of candies? Provide at least one reason to justify your answer.

Solution: Yes, because it only affects the treatment (Costume intricacy) and is independent of the confounder (Age).

Inverse propensity weighting and matching. You next consider inverse propensity weighting. From your observations, you notice that 3/4 of the trick-or-treaters are kids (A = 0). Roughly 1/5 of the kids have low quality costumes. Roughly 1/3 of the adults (A = 1) are wearing low quality costumes.

(c) (2 points) Assuming that C is the treatment variable, provide a numerical estimate of the propensity score $\pi(A)$ when $A = 0$ and when $A = 1$.

Solution: $\pi(0) = 1 - 1/5 = 4/5$ ($P(C = 1 \mid A = 0)$ is the probability that kids has high-quality costumes)

Solution: $\pi(1) = 1 - 1/3 = 2/3$ ($P(C = 1 \mid A = 1$ is the probability that adults have high-quality costumes)

(d) (2 points) One problem with propensity scores is that some subjects might have propensity scores very close to 0 and 1. That is, subjects with some covariate values are almost always assigned the control or treatment group, respectively. Explain why this is a problem if you tried to perform inverse propensity weighting in this scenario.

Solution: The formula for inverse propensity weighting is

$$\frac{1}{250}\sum_{i=1}^{250}\frac{C_i N_i}{\pi(A_i)} - \frac{(1-C_i)N_i}{1-\pi(A_i)}$$

In this case, propensity scores close to 0 and 1 would lead to denominators close to zero in the formula above. This would lead to an estimate with very high variance, making it unreliable.

(e) (1 point) (TRUE / FALSE) In this scenario, we can reasonably use matching to construct an artificial control group and estimate the ATE. Justify your answer.

> **Solution:** TRUE. Because there are 250 trick-or-treaters and the propensities aren't too small (or large), we can likely find many pairs of people with matched control covariates.

This page intentionally left blank for scratch work. No work on this page will be graded.