**q**

1. Consider the following sequence of transactions by 3 users along with their timestamps. Predict the next purchases by these users along with probability distributions for each of them. Consider lookback windows of length 3 to calculate the probabilities.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
| U1 | A | B | C |  | A | B |  | B | A |  |
| U2 |  | C | D |  | B | C | A |  | A | C |
| U3 |  | E | A | C |  | C |  | E | A |  |

**Solution sketch:**

For each user, consider the last 3 transactions (because window size=3) ({A,B,\_} in case of U1, {A,C} for U2) etc. Identify the number of times this set has been purchased together in the same window (size 3), and find the item that followed these purchases. Use these to construct the probability distribution for the next item for each user.

For eg, {\_,A,B} in T1-T2 is followed by C, {\_,A,B} in T4-T6 is followed by \_ (no purchase), {A,B,\_} in T5-T7 is followed by \_, etc. Note that the order of purchases within the window need not be considered.

1. Given the following dataset with two features and real-valued output, which of the following models will be most suitable in case of i) linear regression ii) ridge regression, iii) LASSO regression?

X1 1.5 -8.4 4.1 -2.3 5.7

X2 2.7 2.2 2.0 -5.5 -0.5

Y 6.3 -3.7 7.9 -12.9 4.9

Model 1: Y = X1 + 2\*X2 Model 2: Y = 2.5\*X2 Model 3: Y = X1 + 2\*X2 – 0.3

**Solution sketch**:

For each of the three models, the loss function is different

1. Linear Regression: 1/NΣi(y\_i- y\_pred,i)^2
2. Ridge Regression: 1/NΣi(y\_i- y\_pred,i)^2 + 0.5\*||W||^2\_2
3. LASSO regression: 1/NΣi(y\_i- y\_pred,i)^2 + ||W||\_1

Where y\_pred = W.X = w1\*x1 + w2\*x2 + w0

In each case, compare the loss value of the 3 models. The model for which this value is least, is the best. The best model need not be the same for all the loss functions.

1. Two time-series X and Y are provided. Check for (lagged) auto and cross-correlation between them. Using Granger causality, identify if there is any causal relationship between them. Consider a maximum lag of 2.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 |
| X | 2 | 4 | 0 | 4 | 4 | 2 | -4 | -6 | -2 |
| Y | 1 | -2 | 2 | 0 | -1 | -3 | -1 | 2 | 0 |

**Solution Sketch**:

Correlation coefficient between two time-series A and B: Cov(A,B)/sqrt(Var(A)\*Var(B)).

Cov(A,B) = 1/T(ΣtA(t)B(t)) - (1/T(ΣtA(t)))\* (1/T(ΣtB(t))) where T is the number of observations.

Set A = X.

To calculate lagged auto-correlation with a lag of Δ, Consider B(t) = X(t- Δ)

To calculate lagged cross-correlation with a lag of Δ, Consider B(t) = Y(t- Δ). Repeat with B(t) = Y(t+Δ)

The number of observations T will decrease as the lag Δ increases. However, to calculate (1/T(ΣtA(t))) and (1/T(ΣtB(t))) we can use all observations.

In case of Granger causality:

Model 1: X(t) = a1\*X(t-1) + a2\*X(t-2); Model 2: X(t) = a1\*X(t-1) + a2\*X(t-2) + b1\*Y(t-1) + b2\*Y(t-2)

Estimate the coefficients for both models using least square regression, based on a subset of the observations (maybe 5).

See which model has less squared error in predicting the remaining observations.

If Model 2 has less error, then we can say that Y Granger-causes X. Else, Y doesn’t Granger-cause X.

Repeat:

Model 1: Y(t) = a1\*Y(t-1) + a2\*Y(t-2); Model 2: Y(t) = a1\*Y(t-1) + a2\*Y(t-2) + b1\*X(t-1) + b2\*X(t-2)

Estimate the coefficients for both models using least square regression, based on a subset of the observations (maybe 5).

See which model has less squared error in predicting the remaining observations.

If Model 2 has less error, then we can say that X Granger-causes Y. Else, X doesn’t Granger-cause Y.

1. We are trying to predict whether there will be a crop shortage or not in a given year. Several factors are recorded from past experience in earlier years and locations, based on which two decision trees are proposed. On the given dataset, which do you think is more suitable?

Tree 1: If R > 600: if P=Y then predict Y, else predict N If R < 600: predict Y

Tree 2: If P = Y: if R>600 then predict N, else predict Y If P = N: if W<26 then predict Y, else predict N

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Annual rainfall  (R)(cm) | 550 | 437 | 898 | 580 | 754 | 604 | 566 | 725 | 685 | 815 |
| Pest attack  (P) | Y | N | N | Y | Y | N | N | N | Y | N |
| Farming workforce  (W) (million) | 20 | 18 | 25 | 22 | 30 | 28 | 32 | 23 | 21 | 27 |
| Crop  Shortage (C) | Y | Y | Y | Y | Y | N | N | N | N | N |

**Solution sketch**:

Construct the trees as described. Take each data-point, and classify them according to each tree. Calculate the accuracy for both cases.

1. We are trying to predict if the Masters’ Degree marks M of a person is a causal factor for their annual income I. Consider the following dataset which records their Masters’ Degree marks M, inherited income I and age A. What are your observations regarding the causal relationship in question? Can you see any confounders in the dataset?

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| M | 0 | 0 | 65 | 78 | 55 | 0 | 85 | 50 | 70 |
| I | 1 | 2 | 0 | 5 | 2 | 0 | 10 | 3 | 5 |
| A | 40 | 42 | 27 | 31 | 40 | 55 | 25 | 50 | 28 |
| Y | 2 | 3 | 3 | 5 | 6 | 8 | 8 | 10 | 15 |

Solution sketch:

This problem should be approached using the Double-ML formulation. Form two linear regression models: one for the target variable “I” based on intervention variable “M”, using “A” as possible confounder. In another model, regress “M” on “A”. Estimate the parameters as follows:

1. Regress I on A (I=bA), get residuals W and coefficient b by least squares
2. Regress M on A (M=cA), get residuals V and coefficient c by least squares
3. Regress W on V (W=dV), to get the coefficients d by least squares

Interpretation: how significant is the coefficient d compared to b, I.e. b/(b+d)? If high, then b has a causal relation on I. Also, does A have a strong impact on both I and M (I.e. are the residuals W and V low)? If so, A is a confounder that impacts both I and M.

1. We want to check if a particular scholarship can prevent children from dropping out of School. Accordingly, we collected data of several students, including their various attributes. Accordingly, we attempt to carry out a randomized control trial by dividing the population into a control and an intervention group. Discuss how you will do it and what are your observations.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Scholarship  received? | N | N | N | Y | Y | Y | N | Y | Y |
| Parents  educated? | N | N | Y | Y | Y | N | N | Y | Y |
| Parents working? | N | N | Y | N | Y | N | Y | Y | Y |
| Gender | F | M | F | F | M | M | F | M | M |
| Dropped out? | Y | Y | Y | Y | N | N | N | N | N |

Solution sketch:

Target variable: Dropped Out. Intervention: Scholarship received?

Divide the population into two groups (control and intervention), based on who received the intervention (scholarship) and who didn’t.

Check out: are the two groups identical w.r.t. the other 3 attributes, I.e. is the fraction of “Y” equal in both groups for each attribute?

If yes: condition right for RCT. Check if fraction of drop-out is less in intervention group compared to control group.

If no, inspect if any of the other attributes seem to be strongly related to the drop-out outcome

1. In a small country, the chance of any person finding employment in a quarter is found to be related to the GDP growth rate G of the country, which has two states – “rising (R)” and “falling (F)”. The employment status ES(t) of any person (“employed E” or “unemployed U”) in quarter t follows a probability distribution conditioned on this national GDP growth rate G(t), and also their employment status ES(t-1) in the previous quarter. The state transition distribution of GDP and the employment status distribution are provided below. A sequence of the GDP growth status over 5 quarters is provided, along with the number of employed NE(1) and unemployed NU(1) people in the first quarter (t=1). Estimate the number of employed NE(t) and unemployed NU(t) people in the remaining quarters.

GDP sequence (quarters 1 to 5): R R F F R NE(1) =120, NU(1) =80

|  |  |  |
| --- | --- | --- |
| TRANSITION | ES(t) = E | ES(t) = U |
| G(t)=R, ES(t-1) = E | 0.9 | 0.1 |
| G(t)=R, ES(t-1) = U | 0.6 | 0.4 |
| G(t)=F, ES(t-1) = E | 0.7 | 0.3 |
| G(t)=F, ES(t-1) = U | 0.1 | 0.9 |

Solution sketch:

At each quarter, estimate how many people change from employed to unemployed and vice versa using the probability table. For example, at t=2, G(t)=R and there are 120 employed people at t=1. According to the probability table, 90% of them (108) are expected to remain employed in t=2. Out of the 80 unemployed people, 60% of them (48) are likely to find employment. So NE(2)= 108+48=156, NU(2)=200-156=44. In t=3, G(3)=F, so according to the table 30% of the employed persons (156\*0.3=47) may turn unemployed, while 90% of the unemployed people (44\*0.9=40) are expected to remain unemployed. So after t=3, NU(3)=87, NE(3)=200-87=113 etc