**Machine Learning from Problem Solving -HW1**

**Anupam Dewan(adewan@andrew.cmu.edu)**

**QUESTION 1:**

**Online music and movie stores such as Spotify and Netflix have been developing rapidly, and more and more people choose to consume digital entertainment products through these channels. With availability of rich historical data, machine learning can be applied for many tasks and help the operations. For each of the 3 tasks below, specify what type of machine learning problem it is (supervised or unsupervised; and classification, regression, etc.). Explain your reasoning briefly in 1-2 sentences each.**

a. **(2 pts) Predicting the sales of an album.**
   **Supervised -Regression**: This is a classical case of supervised learning, making use of regression. This is because we have x dimensions in X space which help us in predicting the sales(outcome) value in Y space which is a regression model use case.

b. **(2 pts) Organizing the movies for better browsing experience.**

   **Unsupervised-Clustering:** Here the online music and movies stores like Netflix and Spotify are clustering the movies/music(content) based on the viewership of the content and so it is a kind of clustering problem, where different user cluster are formed based on their viewership's.

c. **(2 pts) Estimating the probability that a certain customer will watch a certain movie in January.**
   **Supervised-regression:** This is a prediction problem where they are predicting which customer can watch which movie in January based on the past data of customer's movie watched and time analysis. Hence, this can be treated as a regression problem from past data and so making it supervised regression.

d. **(2 pts) Estimating the probability that the online store contains albums of a given total duration.**
   **Unsupervised-Probability density estimation:** Here again the probability that the online store contains an album of duration is being estimated and so this is the case of density estimation as given x in X space we are learning about the f(x) or the distribution of total duration of albums.

**QUESTION 2: EXPLORATORY DATA ANALYSIS**

a. **(2 pts) Give an example (other than the ones shown in the lecture slides) for each of the following types of attributes: nominal, ordinal, interval, and ratio. Would it be suitable to calculate the following quantities for these attributes? frequency distribution, median, mean, and coefficient of variation. Why (not)?**

| Variables | Example: | frequency distribution | Median | Mean | coefficient of variation |
|---|---|---|---|---|---|
| **NOMINAL:** | Hair color (White, Black, Blonde, Brown, grey, other) | Possible, as various frequency of people with respective hair color can be plotted in histogram. | No, we cannot specify anything greater or less, or mid-point in types of hair color, they are just labels | No, we cannot calculate the mean of labels as the mean of hair colors does not makes sense | No, we cannot meaningfully comment on variation as mean of the nominal data is meaningless and so the standard deviation |
| **ORDINAL:** | Restaurant rating at yelp (Worst (0), Bad (1), Average (2), Good (3), Great (4), Awesome! (5)) | Possible, as various rating for a restaurant can be plotted on histogram. | Possible, median makes sense, as now the available parameter occurs in a ordered fashion. Hence, we can calculate median | No, mean does not makes sense, as some of ratings generally do not make sense. | No, again ratings cannot tell the difference between two specific ratings as a result, standard deviation and mean is meaningless, hence, we cannot comment on coefficient of variation |
| **INTERVAL:** | Date of births of students in class | Possible, as various date of births can be plotted on histogram depicting the frequency of the date of births | Possible, as date of birth can be arranged in chronological fashion to determine the median date of birth from the population of students. | Possible, mean of days between two dates can be calculated as the difference between dates is operable. | No, since one can only divide by differences, one cannot define measures that require some ratios like coefficient of variation |
| **RATIO:** | Student marks in Heinz college. | Possible, as we can plot marks on histogram and report on the occurrence of | Possible, student marks can be arranged in a ordered fashion and | Possible, student marks mean can be found from the data | Possible, we can calculate differences in student marks and so can actually calculate coefficient of variation. |

| | | | | | |
|---|---|---|---|---|---|
| | | frequency of marks scored. | then the median can be reported | | |

**b. (2 pts) List two problems that might be caused by discarding data records with missing data.**

1) **Shortening of dataset:** If we discard all the rows in dataset that happen to have missing values we might end up reducing a lot of size of our dataset making it too small to perform any kind of analysis and machine learning predictions**.**

2) **Loosing out some important patterns and information:** If we discard all the rows from our dataset that happen to have missing values we might also loose some important patterns occurring in our dataset or some other key information in other dimensions that had data but were eliminated due to missing values in other dimensional cells.

**c. (3 pts) Duplicate data:**

**1. (1 pt) What is the main issue with doing data analytics using exact-duplicate records?**

Duplicate records lead to making data unclean, redundant and unusable for performing any kind of analytics. The two-same record would serve any purpose in our dataset and make the data redundant leading to incorrect analytical result.

**2. (2 pts) Now instead imagine a retail store database (such as Macy's) with near-duplicate records where we have records with different names for a brand (e.g., 'Williams Sonoma' and 'Williams– Sonoma'). This is called the entity resolution problem. List one application which would be negatively affected by analysis on such a database. (Hint: database with near-duplicate would treat one entity as multiple separate entities.)**

If the database has near duplicate records (e.g., 'Williams Sonoma' and 'Williams– Sonoma') they will be treated as two separate entity and can negatively affect analysis on database. One Such scenario could be if some users are sent promotional vouchers by retail store like Macy's. So, this could result in William Sonoma receiving 2 vouchers as in database it is identified as two different entities. This could lead to Macy into incurring extra promotional cost due to duplicate (nearly duplicate records).

**c. (2 pts) Describe the difference between a linear model, a generalized linear model (GLM), and a generalized additive model (GAM), and give an example equation for each of the models.**

**Linear model: Linear model is** simply a class of statistical Models in which the usual Linear relationship between the Response and Predictors is set. Using various combination of predictor variables with a weighted average of them on the response is calculated. This equation help in predicting the response variable with the combination of observed features.

$$y=\beta 0+\sum \beta iXi+\epsilon i$$

**GLM:** GLM generalizes the model by allowing response variable to be related with the linear model through a linking function and allowing the variance of each coefficient to be the function of the predicted value**.** It model's the expected value of a continuous variable, *Y*, as a linear function of the continuous predictor

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

**GAM :** GAM basically removes all the linear relationship between the response and the predictors by a nonlinear smoothened function to model and capture the non-linear behavior in model to better account for the variation.

$$f(x) = yi = \alpha + f1(xi1) + f2(xi2) + \ldots fp(xip) + \epsilon i$$

where the functions $f_1, f_2, f_3, \ldots f_{pf1,f2,f3,\ldots fp}$ are different Non Linear Functions on variables $X_{pXp}$ .

**e. (4 pts) Given dataset X = {4,5,5,6,12,14,15,15,16,17}, use Parzen windows to estimate the density p(x) at y = 3,10,15. Use a boxcar kernel with Δ = 2.**

dataset X = {4,5,5,6,12,14,15,15,16,17}

Boxcar k(x)=1/2I(x) so its ½ if x<=1 and 0 otherwise

$$P(y) = \frac{1}{\Delta} \frac{\left[\sum_{i=1}^{n} k\left(\frac{x_i - y}{\Delta}\right)\right]}{n}$$

| Xi | $k\left(\frac{x_i-y}{\Delta}\right)$ at Y=3 | $k\left(\frac{x_i-y}{\Delta}\right)$ at Y=10 | $k\left(\frac{x_i-y}{\Delta}\right)$ at Y=15 |
|---|---|---|---|
| 4 | K(4-3/2)=k(1/2)=1/2 | K(4-10/2)=k(-3)=1/2 | K(4-15/2)=k(-11/2)=1/2 |
| 5 | K(5-3/2)=k=1/2 | K(5-10/2)=k(-5/2)=1/2 | K(5-15/2)=k(-5)=1/2 |
| 5 | K(5-3/2)=k=1/2 | K(5-10/2)=k(-5/2)=1/2 | K(5-15/2)=k(-5)=1/2 |
| 6 | K(6-3/2)=k(3/2)=0 | K(6-10/2)=k(-2)=1/2 | K(6-15/2)=k(-9/2)=1/2 |
| 12 | K(12-3/2)=k(9/2)=0 | K(12-10/2)=k(1)=1/2 | K(12-15/2)=k(-3/2) =1/2 |
| 14 | K(14-3/2)=k(11/2)=0 | K(14-10/2)=k(2)=0 | K(14-15/2)=k(-1/2) =1/2 |
| 15 | K(15-3/2)=k(6)=0 | K(15-10/2)=k(5/2)=0 | K(15-15/2)=k(0) =1/2 |
| 15 | K(15-3/2)=k(6)=0 | K(15-10/2)=k(5/2)=0 | K(15-15/2)=k(0) =1/2 |
| 16 | K(16-3/2)=k(13/2)=0 | K(16-10/2)=k(3)=0 | K(16-15/2)=k(1/2) =1/2 |
| 17 | K(17-3/2)=k(7)=0 | K(17-10/2)=k(7/2)=0 | K(17-15/2)=k(1) =1/2 |

| | | | |
|---|---|---|---|
| $\frac{1}{\Delta n}\sum_{i=1}^{n} k\left(\frac{x_i - y}{\Delta}\right)$ | **1/16*3/2=3/32** | **5/2*1/16=5/32** | **5*1/16=5/16** |

Answer:

**At y=3:**

**P(y)=3/32**

**At y=10:**

**P(y)=5/32**

**At Y=15:**

**P(y)=5/16**

**f. (7 pts) True or False. For the ones you answer 'False', briefly explain why it is false.**

**1. For categorical attributes, more fine-grained data is always more desirable.**

Ans: False, because if the granularity is so fine grained it can lead to too much splitting and too many variables, leading to high dimensionality and so the data could suffer from ***curse of dimensionality***

**2. Semantic accuracy is more difficult to check than syntactic accuracy.**

Ans: True

**3. In kernel density estimation, the choice of the kernel bandwidth is often more important than the choice of the kernel function.**

Ans: True

**4. In kernel density estimation, a bin width that is too large will under-smooth the estimate and yield a spiky estimate.**

Ans: False, because a bin width that is too large will over-smooth

**5. If Spearman's rho of two variables X and Y equals to 1, then X and Y have perfect positive linear correlation.**

Ans: False, if value of spearman's rho for two variables is 1 we can say that there is a positive correlation between the two(which may or may not be linear).This is because spearman correlation coefficient access the monotonic relationship between the variable.

**6. $y = \beta 0 + \beta 1X_1 + \beta 2X_2^2 + X\beta_3^3$ , where y is a numerical outcome, Xs are features and βs are parameters, is a linear model.**

Ans: False, No as the equation of the model is not of degree 1 it is a non-linear model.

**7. Principal Component Analysis can be performed using one of two different objectives: Maximum variance subspace, and minimum reconstruction error. The solutions to these two objectives are similar in most cases, but they are not identical.**

Ans: True

**3 Conceptual: Linear Regression [10 points]**

**What did housing prices look like in the 'good old days'? The median sales prices for new single-family houses are given in the accompanying table for the years 1972 through 1979.1 Letting y denote the median sales price and x the year (using integers 1, 2, . . . , 8)**

| Year(x) | Median Sales Price (×1000) (y) |
|---|---|
| 1972(1) | $ 27.6 |
| 1973(2) | $ 32.5 |
| 1974(3) | $ 35.9 |
| 1975(4) | $ 39.3 |
| 1976(5) | $ 44.2 |
| 1977(6) | $ 48.8 |
| 1978(7) | $ 55.7 |
| 1979(8) | $ 62.9 |

**a. (2 pts) Fit the linear model $Y = \beta 0 + \beta 1x + e$: Report the coefficient estimates β0 and β1. (Do not use simple linear regression command in statistical software for this question. Do it by hand and show your calculations. You will be asked to implement linear regression using statistical software in Problem 6.)**

Lets take any two points:

X=1 y=27.6

$27.6 = \beta_0 + \beta 1 *_1 + e$

$27.6 = \beta_0 + \beta_1 + e$------(1)

X=2,y=32.5

$32.5 = \beta_0 + 2\beta_1 + e$ ---------(2)

(2)-(1)

$\beta_1 = 4.9$

1): $27.6 - 4.9 = \beta_0 + e = 22.7 = \beta_0 + e$

2): $32.5 - 9.8 = 22.7 = \beta_0 + e$

$\beta_0$ +e(residual) = 22.7

y=22.7 + 4.9x

**b. (1 pts) Based on the model you fit in part a, what is your estimated value of the median sales price for year 1980?**

**In year 1980 x=10**

So putting vaues as obtained above: **y=22.7 + 4.9x**

Y=22.7+4.9*10

Y= 22.7+49

y=71.7

**c. (2 pts) Calculate the RSS (Residual Sum of Squares) and estimate the residual variance σ2.**

**(I) RSS =(sum of squares of difference in residuals)**

| Year(xi) | Median sale price(yi) | $B_1X_i$ | Predicted = $B0+B_1X_i$ | Delta = Yi- $B0+B_1X_i$ | $(Yi- B0+B_1X_i)^2$ |
|---|---|---|---|---|---|
| 1972(1) | $ 27.6 | 4.9 | 27.6 | 0 | 0 |
| 1973(2) | $ 32.5 | 9.8 | 32.5 | 0 | 0 |
| 1974(3) | $ 35.9 | 14.7 | 37.4 | 0.5 | 0.25 |
| 1975(4) | $39.3 | 19.6 | 42.3 | 3.0 | 9 |
| 1976(5) | $ 44.2 | 24.5 | 47.2 | -3.0 | 9 |
| 1977(6) | $ 48.8 | 29.4 | 52.1 | -3.3 | 10.89 |
| 1978(7) | $ 55.7 | 34.3 | 57 | -1.3 | 1.69 |
| 1979(8) | $ 62.9 | 39.2 | 61.9 | 1.0 | 1.0 |
| | | | | **RSS=** | **31.83** |

**II) Residual variance=**

Residual variance=RSS/(n-2)

Residual of Variance $\sigma^2$ =31.83/6=**5.305**

**d. (1 pts) Estimate the standard error for ˆ β1.**

| Year(xi) | (xi-xbar) | $(xi-xbar)^2$ | |
|---|---|---|---|
| 1972(1) | **-3.5** | **12.25** | |
| 1973(2) | **-2.5** | **6.25** | |
| 1974(3) | **-1.5** | **1.25** | |
| 1975(4) | **-0.5** | **0.25** | |
| 1976(5) | **0.5** | **0.25** | |
| 1977(6) | **1.5** | **1.25** | |
| 1978(7) | **2.5** | **6.25** | |
| 1979(8) | **3.5** | **12.25** | |
| | **Sum=** | **40** | **$\sigma^2$/sum(xi-xbar)$^2$ = 5.305/40= sqrt(0.1326)= 0.3641** |

**e. (1 pts) Is there sufficient evidence to indicate that the median sales price for new single-family houses increased over the period from 1972 through 1979 at the 0.01 level of significance?**

**Here we need to test hypothesis**

$H_0$: $\beta_1 <= 0$

$H_1$: $\beta_1 >$ **0**

$t = \beta_1(HAT)/SE(\beta_1(HAT))$

t=4.9/0.3641 =13.457

p value = 0.00000089

As p value<<<0.01 we can say that the value is significant and we cannot reject null and there isn't enough evidence to to indicate that the median sales price for new single-family houses increased over the period from 1972 through 1979 at the 0.01 level of significance

**f. (1 pts) Give a 95% confidence interval for β1.**

95% CI = [β1+2 se(β1), β1-2 se(β1)]
Putting in values
We get,
[4.9+2*0.3641, 4.9-2*0.3641] =[5.6282,4.1718]

g**. (1 pts) Which of the following have closed-form analytical solutions ?:**

**(I) Unregularized Linear Regression,**

**(II) Ridge Regression (i.e. linear regression with l2 regularization),**

**III)Lasso Regression (i.e. linear regression with l1 regularization)**

**Answer:** All 3 of the above options can have closed form analytical solution as all three can be expressed and solved in finite number of steps

**H. (1 pts) Consider a dataset which has millions of features. Describe two challenges one would face in calculating the closed-form multiple linear regression solution $\beta* = (XTX)-1XTy$. (Hint: Consider the time and space complexities.)**

If we have million features in our dataset: We might face two bug issues on computing closed form equation:

a. **Time complexity:**
   If we compute million features log transformations and then solve it, by computing all the inverse matrices. This will take a significant amount of time and is not time efficient. We need to think about dimensionality reduction techniques.
b. **Space Complexity:**
   Also storing those million features log transformations and inverse matrices require a lot of space making it space inefficient as well.

***QUESTION 4(HANDWRITTEN)***