

# EMATMoo44 Introduction to AI

## Coursework

Student number: 2038626

### Part 1

#### Q1

##### 1.1 Introduction

###### 1.1.1 Problem Formulation

The dataset presents massive data from nine sensors around the gas power station and the corresponding nitrous oxide (NOX) values. The problem is to establish the model to predicting nitrous oxide emissions under the condition that nine sensor data are available.

###### 1.1.2 Problem Analysis

In this section, the optimal solution is determined for this problem from the perspective of three methods: regression, classification, and clustering. The regression task is characterized by labelled data sets with numerical target variables. Specifically, each observed sample has a numerical labelled truth value to supervise the algorithm. The predicted values of nitrous oxide emissions are continuous and fall squarely into this type of regression problem. However, the classification method is a supervised learning algorithm for modelling or predicting discrete random variables. Although many regression algorithms have their corresponding classification algorithms, classification algorithms are usually suitable for predicting a class (or probability of a class) rather than a continuous value. This kind of output is not in accordance with the requirements of the task. At last, clustering is an unsupervised learning task where the algorithm finds natural communities (i.e., clusters) of observed samples based on the internal structure of the data. Because clustering is a form of unsupervised learning (i.e., the data is not labelled) and the results are usually evaluated using data visualization. This approach does not correspond to the requirements of the task. As a result, regression is determined as the approach for this task.

###### 1.1.3 Approach Analysis

First, initial observations of the data were made to try to explore correlations between the data. Secondly, linear regression was selected as the model to analyse the data for this task. Simultaneously, in order to avoid the over-fitting problem of the model, ridge regression and Lasso regression are adopted. Ridge regression is a biased regression method specifically designed for the analysis of covariance data. It is essentially a modified least squares estimation method that gives up the unbiased nature of least squares to obtain more realistic and reliable regression coefficients at the expense of losing some information

and is a better fit for pathological data than least squares method. While Lasso regression focuses on dimension reduction and feature selection. By applying Lasso regression, the weight of sparse features can be set to be zero, which means that the main features that affect the predicted values can be enhanced.

## 1.2 Methods

### 1.2.1 Performance metric

Mean squared error (MSE) is inferred to the square of the difference between the true value and the predicted value then summed and averaged, which is shown in Formula (). The smaller the mean squared error is, the better the model performance is. It is very intuitive to be employed as a loss function for the linear regression task and  $MSE$  is convenient to derive it through the squared form. However, there is no uniform scale for  $MSE$  as an evaluation metric across different prediction tasks and different datasets.

$$MSE(\hat{y}, y) = \sum_i (\hat{y}^{(i)} - y^{(i)})^2$$

Therefore, another metric called  $R Squared$  is introduced. Formula () represents  $R Squared$ .  $R Squared$  converts the prediction results to accuracy, which is between  $[0, 1]$ , and the prediction accuracy for different problems can be compared and the metric can be used to determine which problem the model is better suited to predict.

$$R^2 = 1 - \frac{\sum_i (\hat{y}^{(i)} - y^{(i)})^2}{\sum_i (\bar{y} - y^{(i)})^2} = 1 - \frac{MSE(\hat{y}, y)}{Var(y)}$$

$R Squared$  is selected as the performance metric in the regression task. As a result, following inferences can be deduced:

- (1)  $R^2 \leq 1$
- (2) The larger the  $R^2$  is, the better the model is. When the prediction model is free of any errors,  $R^2 = 1$
- (3) When the model is equal to the baseline:  $R^2 = 0$
- (4) If  $R^2 < 0$ , the prediction model is not as good as the baseline.

### 1.2.2 Baseline

A baseline is generated by *sklearn.dummy* for models' comparison. The strategy of the baseline to generate predictions is "mean". It always predicts the mean of the training set. Then, the metrics of baseline can be obtained. The mean squared error is 110.11. The  $R Squared$  is 0.

### 1.2.1 Algorithm Details

Our goal is to find a linear equation that fits the relationship between nitrous oxide emissions and the values of nine sensors. The linear equation is shown as below,

$$h(x) = h_{\theta}(x) = \sum \theta_i x_i = \theta^T x$$

Ridge regression and Lasso regression emerged to address the problem of overfitting that occurs in linear regression, and both regressions achieve their purpose by introducing a regularization term into the loss function. By choosing the optimal  $\theta$  such that  $h(x)$  is

nearest into the true value. This problem then translates into solving for the optimal  $\theta$  such that the loss function  $J(\theta)$  takes the minimum value; a comparison of the loss functions of the three is shown in the following formulas.

The loss function of linear regression:

$$J(\theta) = \frac{1}{2}(X\theta - Y)^T(X\theta - Y)$$

According to the least square method, the linear regression coefficient  $\theta$  is obtained as:

$$\theta = (X^T X)^{-1} X^T Y$$

As is state in Chapter 1.1.3, Ridge regression and Lasso regression are adopted in order to avoid over-fitting and equipped the model with greater generalization ability. In Ridge regression, the cost function is altered by adding penalty equivalent to square of magnitude of the coefficient, which is  $L_2$  norm.  $\lambda$  is called the regularisation parameter. The loss function of the Ridge regression is shown as:

$$J(\theta) = \frac{1}{2}(X\theta - Y)^T(X\theta - Y) + \frac{1}{2}\lambda||\theta||_2^2$$

The major difference between Ridge regression and Lasso regression is that Ridge regression introduces an  $L_2$  norm, while Lasso regression introduces an  $L_1$  norm. Lasso regression can make many  $\theta$  in the loss function become zero, while Ridge regression requires all  $\theta$  to exist. So, the computational workload of Lasso regression will be much smaller than that of Ridge regression. The loss function of Lasso regression is shown as:

$$J(\theta) = \frac{1}{2m}(X\theta - Y)^T(X\theta - Y) + \lambda||\theta||_1$$

If  $\lambda$  is chosen too large, it will minimize all parameters  $\theta$ , resulting in underfitting, and if  $\lambda$  is chosen too small, it will result in an improper solution to the overfitting problem. Therefore, the selection of the appropriate regularisation parameter will be the focus of the following discussion.

## 1.3 Hyperparameters

$\lambda$  is the hyperparameters. As indicated above, the regularisation parameters of the Lasso regression and Ridge regression help to identify the models with better performance.

The most suitable model is found by comparing the performance metrics *MSE* and *R Squared* under different values of regularisation parameters.

$\lambda$ (Ridge)	0.1	1	10	100	1000	10000	100000
<i>MSE</i>	49.25	49.25	49.26	49.30	49.83	51.85	56.46
<i>R Squared</i>	0.170	0.170	0.169	0.162	0.125	0.014	-0.583

Table 1.1 Metrics Sampling for Ridge Regression

$\lambda$ (Lasso)	0.0001	0.001	0.01	0.1	1	10	100
<i>MSE</i>	49.43	49.43	49.46	50.04	52.98	61.77	110.10
<i>R Squared</i>	0.161	0.161	0.156	0.122	-0.015	-1.196	-5.45

Table 1.2 Metrics Sampling for Lasso Regression

Table 1.1 and 1.2 shows sampling values of  $MSE$  and  $R Squared$  with regularization terms  $\lambda$  changing in Ridge regression and Lasso regression. As  $\lambda$  increases, the model error becomes larger. The optimal regularisation parameter chosen is 0.001 for the Lasso regression and 1 for Ridge regression.

## 1.4 Results

Models	mean squared error (MSE)	$R^2$
Baseline	110.11	0
Linear Regression	49.25	0.171
Ridge ( $\lambda = 1$ )	49.25	0.170
Lasso ( $\lambda = 0.001$ )	49.43	0.161

Table 1.3 Metrics and Performance

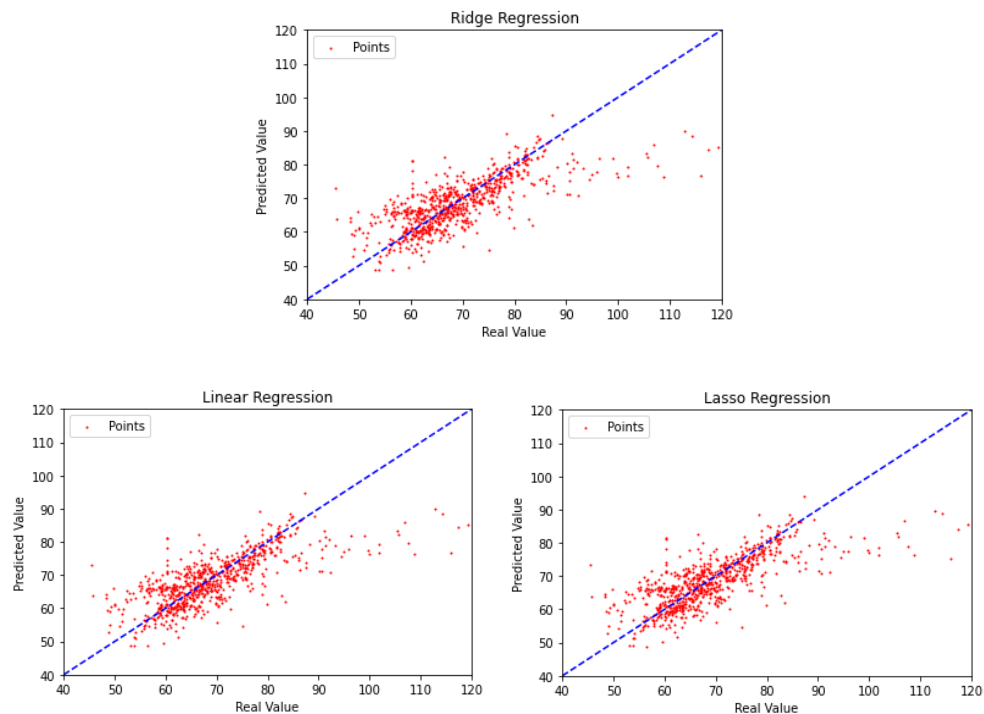


Fig. 1.1 Performance of the models

### Result Analysis:

- (1) The test results of the three different regression models are shown in Table 1.3 and visualized in Fig. 1.1, using the true value as horizontal coordinates and the predicted value as vertical coordinates. Through the selection of hyperparameters, all three regression models were able to fit the data reasonably well for prediction of nitrous oxide emissions. The models have good robustness.
- (2) The difference in performance between the Ridge regression and the linear regression is small. This suggests that the addition of the regularisation term L2 noun to the Ridge regression has little effect on the model. the R Squared metric for the Lasso regression is smaller than that of the linear regression, suggesting that the regularisation term L1 noun introduced by the Lasso regression does not improve the model. The reason for this may be that there is very little sick data in the dataset.

## Q2

### Data representation

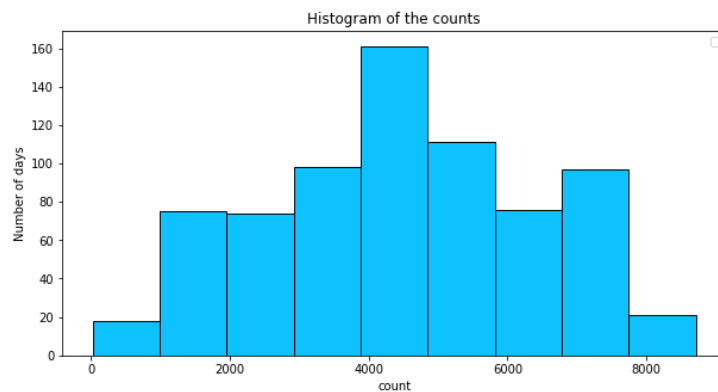


Fig. 2.1 Histogram of the counts

Fig. 2.1 shows the histogram of the counts giving the distribution of number of bicycles and days. The number of bicycles used each day ranges from 22 to 8714. Three intervals divide the counts into three categories 'high', 'medium' and 'low'. Therefore, a new feature in the dataset called 'usage' is created representing values 'high', 'medium', 'low', which corresponds to the following three intervals of count data: [6000,8714], [2000, 6000], [22, 2000].

The attributes we focus on are 'season', 'workingday', 'weathersit' and a new feature 'count'. Therefore, the data is represented in Fig. 2.2.

	season	workingday	weathersit	usage
0	1	0	2	low
1	1	0	2	low
2	1	1	1	low
3	1	1	1	low
4	1	1	1	low
...	...	...	...	...
726	1	1	2	low
727	1	1	2	low
728	1	0	2	low
729	1	0	1	low
730	1	1	2	low

Fig. 2.2 Data Table

**Information gain** The attribute 'season' has the highest information gain.

**Decision Tree** The decision tree is shown in Fig. 2.3.

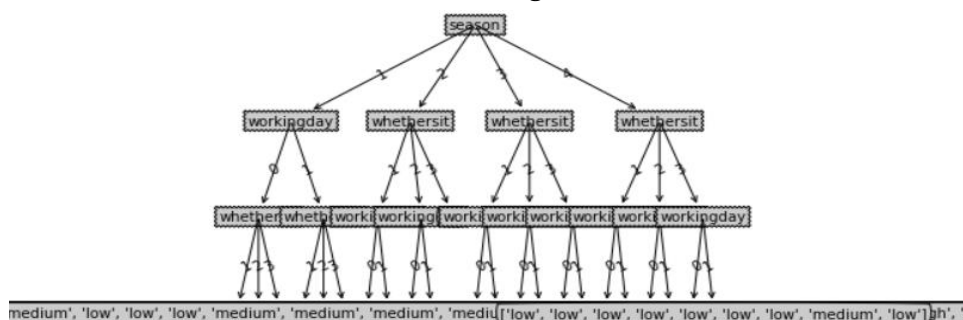


Fig. 2.3 Decision tree

**Techniques 1: Pessimistic Error Pruning**

In order to reduce the complexity of decision tree and the probability of over fitting, Pessimistic Error Pruning (PEP) as a post-pruning approach is proposed. After applying the algorithm, a relevance check is carried out on each internal node for classification error rate, which decides whether to drop the node. This approach improves the generalization ability of the decision tree.

**Techniques 2: Discretization of continuous attributes**

The dataset cannot be divided when attributes are continuous values. It is necessary to discretize the continuous attribute values and then a decision tree analysis can be analyzed.

**Techniques 3: Attributes based on Information gain rate**

Through obtaining the information gain rate, this approach avoids the effect of taking too many values for an attribute.

$$Information\ gain\ rate = \frac{information\ gain}{attribute\ entropy}$$

**New attribute ‘tempbins’**

The data is shown in Fig 2.4. And the decision tree is shown in Fig 2.5.

	season	workingday	weathersit	tempbins	usage
0	1	0	2	medium	low
1	1	0	2	medium	low
2	1	1	1	low	low
3	1	1	1	low	low
4	1	1	1	low	low
...	...	...	...	...	...
726	1	1	2	low	low
727	1	1	2	low	low
728	1	0	2	low	low
729	1	0	1	low	low
730	1	1	2	low	low

Fig 2.4 Data Table with ‘tempbins’

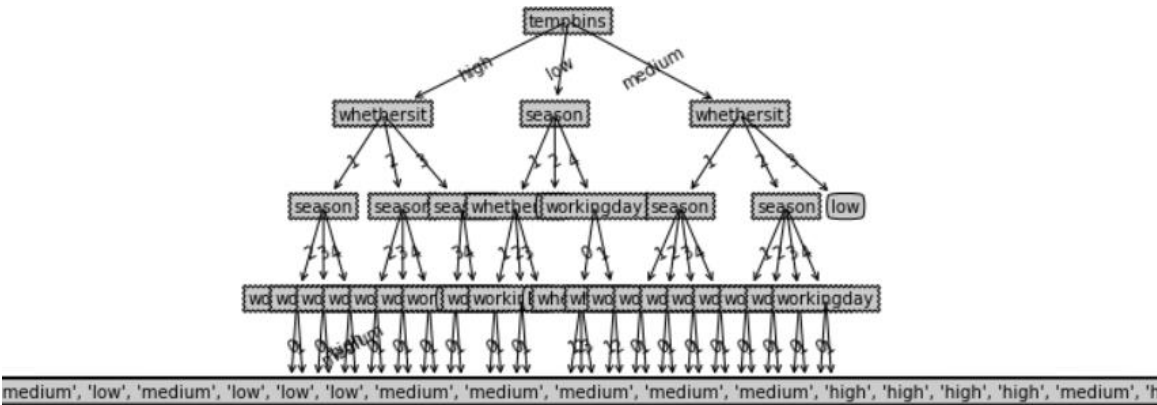


Fig 2.5 Decision Tree with ‘tempbins’

## Q3

### Motivation

**1. *For what purpose was the dataset created?***

Provide large-scale, accurate and diversified datasets of art (mostly paintings) which has annotations for emotions evoked in the observer.

**2. *Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?***

Saif M. Mohammad and Svetlana Kiritchenko, National Research Council Canada

**3. *Who funded the creation of the dataset?***

The dataset was funded by National Research Council Canada (NRC).

### Composition

**1. *How many instances are there in total (of each type, if appropriate)?***

There are 4,105 pieces of art containing emotional annotations for emotions evoked in the observer. The artworks are a collection of twenty-two categories (Impressionism, Realism, etc.) from four Western styles (Renaissance Art, Post-Renaissance Art, Modern Art, and Contemporary Art). The annotations consist of one or more of twenty emotion categories (including neutral), including the depiction of a face and how much the observers like the art.

**2. *What data does each instance consist of?***

Art and corresponding annotations.

The annotations include:

all emotions that the image of art brings to mind;

all emotions that the title of art brings to mind;

all emotions that the art as a whole (title and image) brings to mind;

a rating on a scale from -3 (dislike it a lot) to 3 (like it a lot);

whether the image shows the face or the body of at least one person or animal;

whether the art is a painting or something else (e.g., sculpture).

**3. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?***

Yes. The dataset consists of artworks and annotations for emotions evoked in the observer. However, the artworks, as a prime part of the artwork, have a great diversity of styles and artistic languages, and are therefore very likely to bring up emotions such as offensive, insulting, threatening, etc. for the viewer. For example, a scene of an angry mother elephant defending her calf from a predator, which might cause anxiety or pensiveness. All negative emotions such as arrogance, fear, remorse, and sadness are worthy annotations to pursue.

**4. *Are there any errors, sources of noise, or redundancies in the dataset?***

Yes. Human error can never be fully eliminated in large-scale annotations, though the researchers take several measures to ensure high-quality and reliable data annotation (e.g., multiple annotators, clear and concise questionnaires, etc.). Expect a small number of clearly wrong entries such as wrong emotions annotated and duplicate and confusing labels.

### ***5. Does the dataset relate to people?***

Yes. From the perspective of dataset content, the entity of the artwork originates from the person. The artistic language is based on the human being. The expression of emotion emanates from the human being. The dataset is closely related to people.

## **Collection Process**

### ***1. How was the data associated with each instance acquired?***

The emotion annotation was associated by the annotators and specifically the annotators are supposed to identify the emotions that the art evokes in three scenarios:

Scenario I: only the image (no title) is presented, and the annotator is asked to identify the emotions it evokes;

Scenario II: only the title of the art (no image) is presented, and the annotator is asked to identify the emotions it evokes;

Scenario III: both the title and the image of the art are presented and the annotator is asked to identify the emotions that the art as a whole evokes.

Besides, emotions are integrated as options and the annotator is guided through the annotation task to avoid cross-influence due to complex information sources in the annotation process. Five instances are showed in scenario I in a random order, followed by five instances in scenario II in a different random order, followed by five in scenario III in another random order.

### ***2. What mechanisms or procedures were used to collect the data(e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?***

Data obtained by means of an internet questionnaire on crowdsourcing. Links to the art and the annotation questionnaires were uploaded on the crowdsourcing platform. Respondents were free to annotate as many instances as they wished to. 19 options of closely-related emotion sets and a final neutral option are presented. A text box was also provided for the annotators to capture any additional emotions. None of the proposed additional emotions was used which indicates that the predefined set of the 19 emotions provided covered the art emotion space well. As a result, these mechanisms or procedures are validated.

### ***3. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?***

Data was obtained directly from individuals through designed mechanisms and procedures in the form of an internet questionnaire on the crowdsourcing platform.



- 4. Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)?**

The data was directly observable, and the data was reported by survey responses.

- 5. Did the individuals in question consent to the collection and use of their data?**

All annotators for the tasks had already agreed to the CrowdFlower terms of agreement. They chose to do our task among the hundreds available, based on interest and compensation provided. Respondents were free to annotate as many instances as they wished to. The annotation task was approved by the National Research Council Canada's Institutional Review Board, which reviewed the proposed methods to ensure that they were ethical. Special attention was paid to obtaining informed consent and protecting participant anonymity.

## **Pre-processing/cleaning/labelling**

- 1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

A generous aggregation criteria is applied for data processing: if at least 40% of the responses (four out of ten people) indicate that a certain emotion applies, then that label is chosen. We also created two other versions of the labeled dataset by using an aggregation threshold of 30% and 50%, respectively. The dataset simply captures the perceptions of the majority group among the annotators.

- 2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**

Information not available from original paper.

## **Uses**

- 1. Has the dataset been used for any tasks already? If so, please provide a description.**

Yes. The WikiArt dataset is used extensively for various task. Firstly, it contributes to studying the correlation between different concepts in the domain of fine art images. Secondly, the dataset is used to train deep neural networks for the tasks of aesthetic, sentiment or memorability prediction [2]. Thirdly, a method of completing paintings intended for recovery of damaged works of art is developed with the help of WikiArt dataset [3].

- 2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.**

No repository created. The available link to papers which cite this dataset is as below.

<https://scholar.google.co.uk/scholar?cites=12035432133931996159>

### **3. *What (other) tasks could the dataset be used for?***

The dataset can be used to develop deep learning algorithms for art generation; for instance, to create systems that can transform a given piece of art (especially abstract paintings) to alter the affective reaction it evokes. Besides, the dataset can help develop an interactive visualization that allows users to search for paintings with desired attributes such as style, genre, emotion, and average art ratings [1].

### **4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?***

The dataset simply captures the perceptions of the majority group among the annotators. The applied aggregation criteria indicates that different socio-cultural groups can perceive art differently and taking the majority vote can have the effect of only considering the perceptions of the majority group. When these views are crystalized in the form of a dataset, it can lead to the false perception. Meanwhile, emotion annotations of specific works may lead to discomfort and prejudice among some groups in society. Future users should therefore, firstly, be explicit about the purpose of use and avoid using perceptions of emotion associations to discuss acute social issues. Secondly, the data set should be pre-processed artificially for the purpose of the study to avoid misinterpretation.

### **5. *Are there tasks for which the dataset should not be used? If so, please provide a description.***

Yes, the dataset is not supposed to be used to automatically detect emotions of people from their facial expressions. And the dataset cannot be used for drawing inferences about individuals through the perceptions of emotion associations.

## **Distribution**

### **1. *When will the dataset be distributed?***

May 2018

### **2. *How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?***

This dataset can be downloaded from their official website. URLs to the images are publicly available at <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.html>

**3. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**

No. It is available free for research purposes. But the organization disclaims any responsibility for the use of the lexicons and does not provide technical support.

**4. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

Information not available from original paper.

**5. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

Yes. Current restrictions show that this study has been approved by the NRC Research Ethics Board (NRC-REB) under protocol number 2017-98. REB review seeks to ensure that research projects involving humans as participants meet Canadian standards of ethics. Different restrictions apply to users in different areas.

## **Maintenance**

**1. Who is supporting/hosting/maintaining the dataset?**

Saif M. Mohammad and Svetlana Kiritchenko.

**2. How can the owner/curator/manager of the dataset be contacted?**

Contact Saif Mohammad through email: [saif.mohammad@nrc-cnrc.gc.ca](mailto:saif.mohammad@nrc-cnrc.gc.ca)

**3. Is there an erratum?**

Information not available.

**4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**

No, emotion associations can change with time, but the entries here are largely fixed. They pertain to the time they are created and to the people who annotated them.

**5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**

No, the formed dataset is retained permanently. The only mechanism about data discarding is the gold instances mechanism in annotation process. If the worker's accuracy on the gold instances (annotated internally beforehand) falls below 70%, they are refused further annotation, and all of their annotations are discarded. This serves as a mechanism to avoid malicious annotations.

## Part 2

### Question 1

- (i) In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. If  $k = 1$ , then the object is simply assigned to the class of that single nearest neighbor. The system is prone to under-fitting. The value of accuracy is high for training datasets but low for testing datasets. The system is prone to over-fitting.
- (ii) The consequence of large  $k$  value is that training instances that are more distant from the input instances also play a role in the predictions. if  $k$  is equal to the number of datapoints, whatever the input instance is, it will simply be predicted to belong to the class with the most instances in the training. In this model, all prediction results are "cat". The value of accuracy is low.
- (iii) A smaller value of  $k$  results in a more complex model that is prone to overfitting, the estimation error of learning increases, and the predictions are very sensitive to the instance points in the immediate neighborhood. Larger values of  $k$  reduce the estimation error of learning, but the approximation error of learning will increase, and training instances that are farther away from the input instances will also play a role in the prediction, making the prediction incorrect.

### Question 2

2(a) Inferring gender by height and shoe size. The pair of height and shoe size are often not independent. Taller heights tend to be associated with larger shoe sizes.

2(b)

(i)

	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$
Species 1	1.978	0.1367	2.242	0.190
Species 2	2.382	0.3256	2.444	0.3680
Species 3	1.414	0.8062	1.044	0.0755

$$\hat{c} = \operatorname{argmax}_i \log(P(c_i)) + \sum_{j=1}^n \log(P(x_j|c_i))$$

$$P(x_i|c_j) = \frac{1}{\sigma\sqrt{2\pi}} * \exp\left(\frac{1}{2}\left(-\frac{x_i - \mu_j}{2\sigma_j}\right)^2\right)$$

Species 1:

$$P(c_1) = \frac{1}{3}$$

$$P(x_{Altitude}|c_1) = \frac{1}{0.137\sqrt{2\pi}} * \exp\left(\frac{1}{2}\left(-\frac{2.1 - 1.98}{0.137}\right)^2\right) = 1.98$$

$$P(x_{Height}|c_1) = \frac{1}{0.19\sqrt{2\pi}} * \exp\left(\frac{1}{2}\left(-\frac{2.3 - 2.24}{0.19}\right)^2\right) = 1.998$$

$$\log P(c_1) + \log P(x_{Altitude}|c_1) + \log P(x_{Height}|c_1) = 0.12$$

Species 2:

$$P(c_2) = \frac{1}{3}$$

$$P(x_{Altitude}|c_2) = \frac{1}{0.326\sqrt{2\pi}} * \exp\left(\frac{1}{2}\left(-\frac{2.1 - 2.38}{0.326}\right)^2\right) = 0.846$$

$$P(x_{Height}|c_2) = \frac{1}{0.368\sqrt{2\pi}} * \exp\left(\frac{1}{2}\left(-\frac{2.3 - 2.44}{0.368}\right)^2\right) = 1.008$$

$$\log P(c_2) + \log P(x_{Altitude}|c_2) + \log P(x_{Height}|c_2) = -0.55$$

Species 3:

$$P(c_3) = \frac{1}{3}$$

$$P(x_{Altitude}|c_3) = \frac{1}{0.806\sqrt{2\pi}} * \exp\left(\frac{1}{2}\left(-\frac{2.1 - 1.41}{0.806}\right)^2\right) = 0.343$$

$$P(x_{Height}|c_3) = \frac{1}{0.368\sqrt{2\pi}} * \exp\left(\frac{1}{2}\left(-\frac{2.3 - 1.04}{0.368}\right)^2\right) = 0$$

$$\log P(c_3) + \log P(x_{Altitude}|c_3) + \log P(x_{Height}|c_3) = -60.7$$

The point (2.1, 2.3) is classified to Species 1 since the probability is the highest.

(ii)

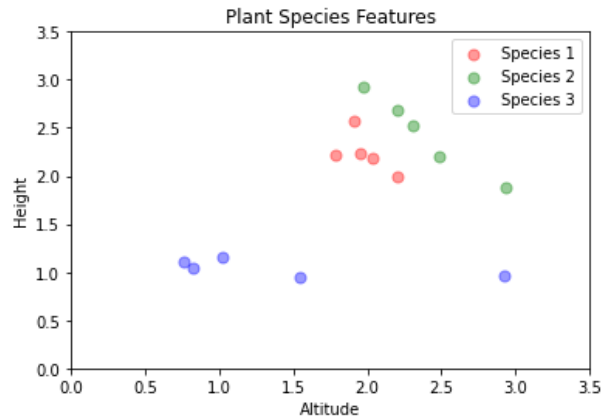


Fig 2.1 Plant Species Features

Fig 2.1 presents the scatter of altitude and height of plant species. Species 1 and Species 3 fit the independence assumption. Because there is no obvious relationship between their altitude and height of the two species.

Species 2 has an obvious inverse correlation between altitude and height. As altitude decreases, the height of Species 2 decreases. Species 2 violates the independence assumption.

2(c)

No.  $k - means$  clustering assigns individual data point to the nearest centroid and, when the assignment is complete, moves the centroid to the mean centroid position of the clusters represented.

- (1) The data distributions of Species 1 and Species 2 are extremely similar. It is likely that they will be grouped into the same cluster.
- (2)  $k - means$  can only generate spherical clusters, however the distribution of Species 2 features is correlated and difficult to fit.

Therefore,  $k - means$  will have poor performance on this dataset.

### Question 3

- (1) The K-means algorithm starts by selecting a random sample from the dataset as the initial cluster centroid. Meanwhile, initial values of the log-likelihood function can be obtained for Gaussian mixture model. The initial values contribute significantly to the clustering results.
- (2) The covariance matrix of each component is set to a small multiple of the identity matrix. Therefore, GMM generates the same spherical clusters as  $k - means$ .
- (3) The Gaussian mixture model is a weighted accumulation of multiple Gaussian distributions. If the weight of each component is set to  $\frac{1}{k}$ , each Gaussian distribution generates samples with equal probability. This is consistent with the idea that the Euclidean sum of the distance of each sample from the cluster center for each cluster should be as small as possible in the k-means objective equation.

As a result, the clustering performances of the two algorithms are approximately equal.

## Question 4

4(a)

Breadth-first search: HCABUQG

Depth-first search: HCUG

Best first greedy search: HAQG

Uniform cost search: HAQG

Best first greedy search and Uniform cost search.

4(b)

Not admissible: HBAQG. The following formula can be met in this route.

$$h(n) \leq h^*(n)$$

Not consistent: The following formula can be met.

$$h(n) \leq \text{Cost}(n, a) + h(n')$$

4(c)

The uniform cost search cannot find the optimal path.

The edge between the student union (node U) and the city museum (node G) now has a negative cost of -1. Therefore, the optimal path will be HCUG. And the total distance is 2.2.

However, the distance of the path from node H to node U is 3.2 in this situation. The distance is longer than the distance of the path HAQG, which is 2.9.

4(d)

The uniform cost search cannot find the optimal path.

Due to an edge between B and Q at negative cost of -2, the optimal path will be HBQG and the total distance of the path is 0.7.

According to the triangle rule, the distance between node B and node Q should meet the following condition:

$$0.8 < \text{Distance}(BQ) < 1.2$$

As a result,

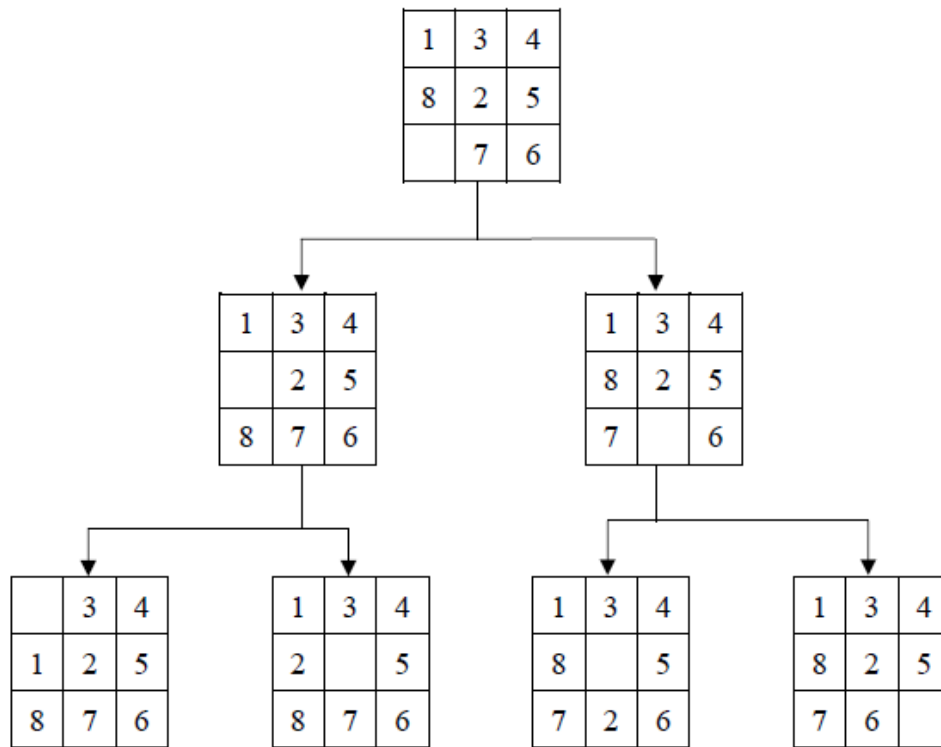
$$\text{Distance}(HBQG) > 3.5$$

$$\text{Distance}(HBQG) > \text{Distance}(HAQG) = 2.9$$

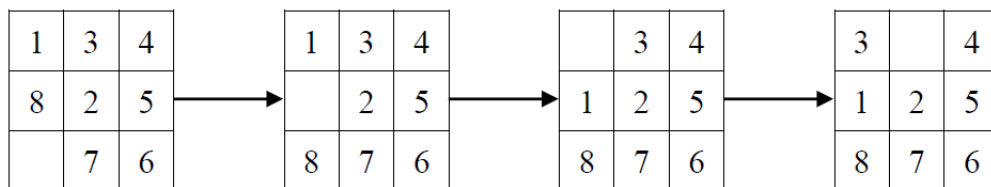
The path HBQG cannot be the optimal route in the real-world situation. As a result, the optimal path cannot be found by the uniform cost search.

5.

5(a)



5(b)



5(c)

$A^*$  takes into account the cost from the root node to the current node and estimates the path cost from the current node to the goal node.

$$f(n) = g(n) + h(n)$$

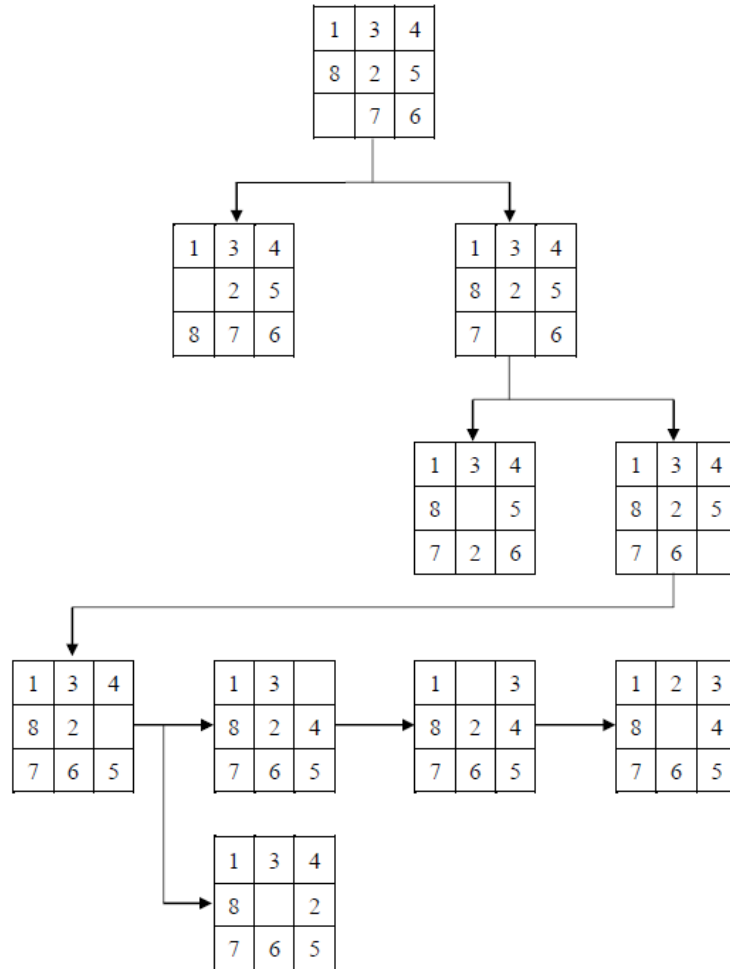
where,

$g(n)$ : path cost from the start node to node  $n$

$h(n)$ : estimated cost of the cheapest path from  $n$  to the goal

$f(n)$ : estimated cost of the cheapest solution through  $n$





5(d)

No. Breadth-first search is an algorithm for traversing or searching tree or graph data structures. However, it cannot find the optimal solution. Because it needs evaluation metrics to define the optimal solution.

## Question 6

- (i) The probability that there is a hole in the garden:

$$P(h) = P(h|d) \cdot P(d) + P(h|\neg d) \cdot P(\neg d) = 0.3 \cdot 0.5 + 0.7 \cdot 0.1$$

- (ii) The probability that the dog is loose, given that there is a hole in the garden.

$$P(h) = 0.22$$

$$P(d|h) = \frac{P(h|d) \cdot P(d)}{P(h|d) \cdot P(d) + P(h|\neg d) \cdot P(\neg d)} = \frac{0.5 \cdot 0.3}{0.5 \cdot 0.3 + 0.1 \cdot 0.7}$$

- (iii) The probability that the dog is loose, given that your cake is eaten, and your flatmate is home.

$$P(d|f, c) = \frac{P(f, c|d) \cdot P(d)}{P(f, c)} = \frac{P(c|d, f) \cdot P(d, f)}{P(f, c, d) + P(f, c, \neg d)} = \frac{P(c|d, f) \cdot P(d) \cdot P(f)}{P(f, c, d) + P(f, c, \neg d)}$$

$$P(d|f, c) = \frac{P(c|d, f) \cdot P(d) \cdot P(f)}{P(c|d, f) \cdot P(d, f) + P(c|\neg d, f) \cdot P(\neg d, f)}$$

$$P(d|f, c) = \frac{1 * 0.3 * 0.4}{1 * 0.3 * 0.4 + 0.9 * 0.7 * 0.4} = 0.32$$

- (iv) The probability that your flatmate is home, given that your cake is eaten and there is a hole in the garden?

$$P(f|c, h) = \frac{P(f, c, h)}{P(c, h)}$$

$$= \frac{P(f, c, h, d) + P(f, c, h, \neg d)}{P(c, h, f, d) + P(c, h, f, \neg d) + P(\neg c, h, f, d) + P(\neg c, h, f, \neg d)}$$

$$P(f|c, h) = \frac{P(f) \cdot P(d) \cdot P(h|d) \cdot P(c|d, f) + P(f) \cdot P(\neg d) \cdot P(h|\neg d) \cdot P(c|\neg d, f)}{P(c, h, f, d) + P(c, h, f, \neg d) + P(\neg c, h, f, d) + P(\neg c, h, f, \neg d)}$$

$$P(f|c, h) = 0.528$$

## Question 7

7(a)

$$U_{i+1}(s) = R(s) + \gamma \max_{a'} \sum_{s'} P(s'|a, s) U_i(s')$$

The utility of the states consists of the reward and the expected utility of subsequent states.

Since all cells have utility 0, the predicted utility of subsequent states is 0 in the first iteration. After the first iteration, each cell's utility is only related to the reward associated with that cell. As a result, the utility of each state is equal to the reward.

7(b)

After two iterations:

3.58	6.61	10.0
0.85	-0.95	6.61

After three iterations:

7.157	7.21	10
2.065	4.931	7.21

The optimal policy is as followed according to the utilities after three iterations.

7.157	7.21	10
2.065	4.931	7.21

7(c).

The optimal policy is as followed according to the utilities after convergence,

17.34	14.38	10.0
14.38	12.25	10.13

No, the agent cannot reach the goal. As the agent has reached the cell with coordinates (1, 2), the utilities around the cell with coordinates (1, 2) are all smaller than itself, which means the agent is not able to move.


7(d)

Suppose the discount factor is reduced to 0.01.

The optimal policy can be obtained after iterations, which is shown in the following figures.

2.01	-0.42	10
-0.48	-0.50	-0.42

Then, the optimal policy can be obtained as

2.01 	-0.42	10
-0.48	-0.50	-0.42

The optimal policy is the same as when the discount factor is 0.9.

The goal cannot be reached. The reward value of (1, 2) is too large. Therefore, the utility of the cell with coordinates (1, 2) is too large after convergence.

## References

- [1] Mohammad S, Kiritchenko S. Wikiart emotions: An annotated dataset of emotions evoked by art[C]//Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). 2018.
- [2] Cetinic E, Lipic T, Grgic S. Learning the Principles of Art History with convolutional neural networks[J]. Pattern Recognition Letters, 2020, 129: 56-62.
- [3] Jboor NH, Belhi A, Al-Ali AK, Bouras A, Jaoua A (2019) Towards an inpainting framework for visual cultural heritage. In: 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT). IEEE, pp 602–607