MONASH University

MONASH INFORMATION TECHNOLOGY

程序代写代做 CS编程辅导

Introduction to Machine Learning

WeChat: cstutorcs

Assignment Project Exam Help

Developed by Prajwol Sangat
Updated by Chee-Ming Ting (3 April 2021)

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

GROUP OF EIGHT AUSTRALIA

Last week

程序代写代做 CS编程辅导

Parallel Aggregation
- Parallel Sort
- Parallel Group-By

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

This week

程序代写代做 CS编程辅导

What is Machine Le
Machine Learning B
Types of Machine Learning
Feature Engineering

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

程序代写代做 CS编程辅导

According to McKinsey study, 35% of what consumers purchase on Amazon and 75% of what they watch on Netflix is driven by machine learning–based product recommendations.

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

What is Machine Learning?

程序代写代做 CS编程辅导

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

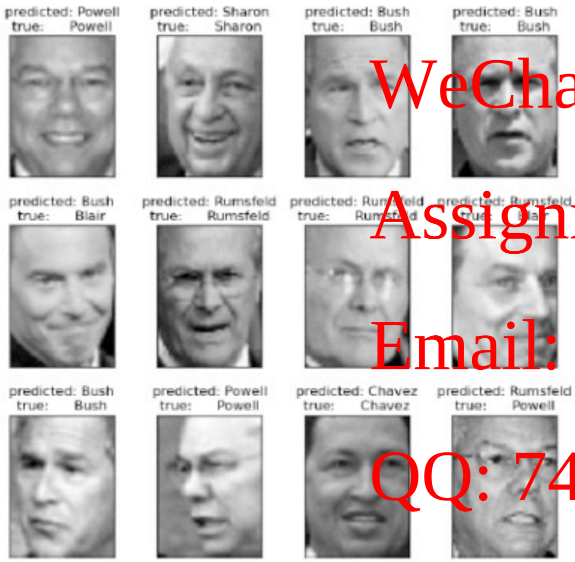
https://tutorcs.com

What is Machine Learning?

程序代写代做 CS编程辅导

A computer program is said to learn fr[...] *E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as m*[...]*nproves with experience E*", (Tom Mitchell, 1997)

Face recognition



Experience E	Task T	Performance P
databases of thousands of known faces	given a new photo, recognise the name of the face	how accurate the recognition is

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Examples

程序代写代做 CS编程辅导

Watson, Ken Jennings, and Brad Rutter in their *Jeopardy!* exhibition match.

Detecting Spam Emails

Detect credit card fraud

[illegible]e E	Task T	Performance P
[illegible] of [illegible]nswer pairs	given an question, find the best answer	how accurate the answer is
Examples of spam emails and not-spam email	To assign a label "spam" or "not-spam" to an email	how accurate spam email can be detected
Data collected for credit-card transactions deemed as fraud and not-fraud	To assign a label 'fraud' or "not fraud" to a given credit-card transaction	how accurate a credit-card fraud transaction can be detected.

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Elements of machine learning

程序代写代做 CS编程辅导

① Data

feature $x \in \mathbb{R}^d$
label $y \in Y$

Dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$

el

$: X \to Y$

Y is label space
θ: model parameter

③ Assessment

How well is f_θ doing w.r.t data $\mathcal{D}$?

WeChat: cstutorcs

Data processing

feature extraction, feature selection, feature transformation, feature reduction, feature scaling, feature normalization

Predictive Model $Y = f_\theta(X)$

Assignment Project Exam Help

Model Learning (Training)

- Find an optimal model f_θ (by estimating model parameters θ) using training data
- Based on loss function (e.g., minimize error between true and predicted labels)

Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Model Testing $\hat{y} = f_\theta(x_{test})$

- Test the learned model in predicting unseen test data
- Performance metrics to assess model accuracy

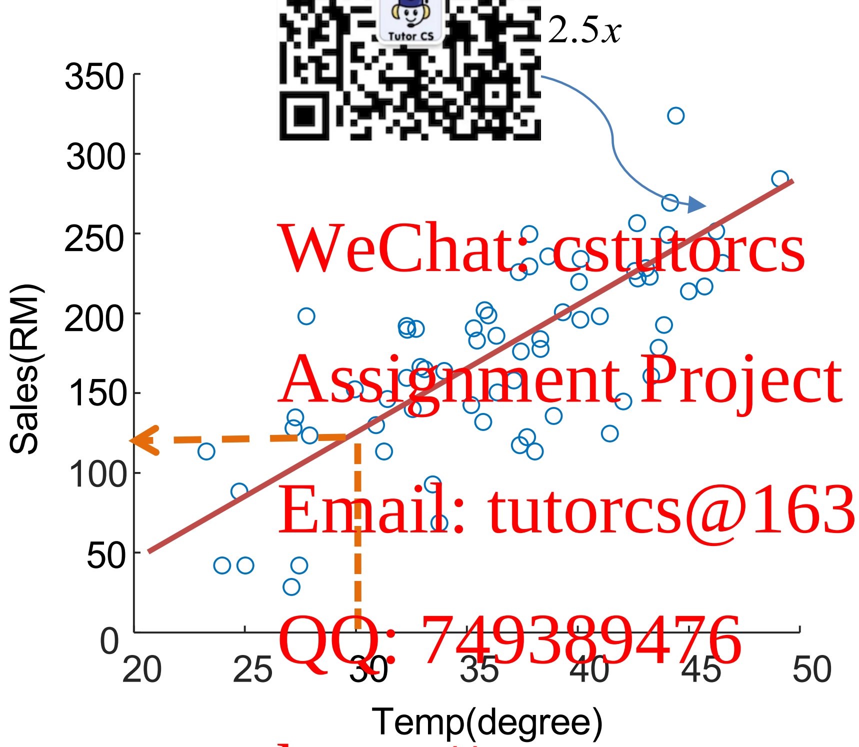
Illustration: Linear Regression model

程序代写代做 CS编程辅导

Problem: Predict ice cream sales g[illegible]ature

Data

Day	Temp	Sales
i	x_i	y_i
1	36	200
2	31	100
3	24	50
⋮	⋮	
100	38	250



Predictive Model:

- What is good model *f*(.) to maps *x* to *y*?

$$f(x) = \theta_o + \theta_1 x$$

Model Learning/Estimation:

- How to choose parameters θ_o, θ_1 ?
  - ❑ Define loss function
  - ❑ Estimate using learning algorithm

Estimated parameters: $\theta_o = 50, \; \theta_1 = 2.5$

Prediction:

- Given new input, predict *y* with learned model $\hat{y} = f(x_{new}) = \theta_o + \theta_1 x_{new}$

Predicted output $\hat{y} = 50 + 2.5(30) = 125$

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Overview of machine learning

程序代写代做 CS编程辅导

...odel θ
...ata

Training Data
$D_{train}=\{x_i, y_i\}_{i=1}^{n}$

...orithm

WeChat: cstutorcs

Training Stage

Test Data
x_{test}

Learned Predictive Model
$Y = f_\theta(X)$

Assignment Project Exam Help

Predicted Label
$\hat{y} = f_\theta(x_{test})$

Classification /Prediction Stage

Email: tutorcs@163.com

$f_\theta(.)$ *maps X to Y*

QQ: 749389476

Assessment of prediction performance

https://tutorcs.com

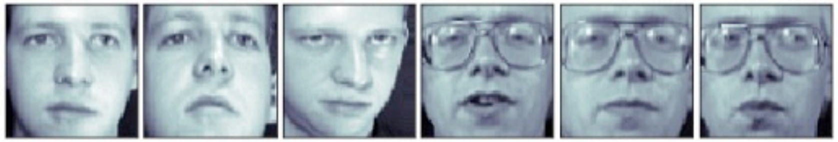
Data

程序代写代做 CS编程辅导

Features: x_i

- a set of attributes, each is a vector or matrix.
- E.g., represent each email as a bag-of-word vector (features) into a real-valued matrix.

data points ... $\{x_i\}$



WeChat: cstutorcs

Labels: y_i

- values, categories, classes, assigned to data points.
- E.g., 0 = non-spam, 1 = spam;

data points with labels ... $\{x_i, y_i\}$

Assignment Project Exam Help

Data points (aka instances, samples) $\{x_i\}$ or $\{x_i, y_i\}$

- these are items or instances of data used for training and evaluating ML models.
- E.g., labelled emails in spam detection; transaction data in credit card fraud detection; a photo in face recognition.

Email: tutorcs@163.com

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

x_i features

y_i 0 = Jack, 1 = John, etc ...

QQ: 749389476

https://tutorcs.com

Dataset with n samples: $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$

Machine Learning: Data Types

程序代写代做 CS编程辅导

- **Vector**
  - A mathematical
  - *dense vectors*, where every entry is stored, and
  - *sparse vectors*, where only the nonzero entries are stored to save space.
- **Labeled Point**
  - A labeled data point for supervised learning algorithms such as classification and regression.
  - Includes a feature vector and a label (which is a floating point value).

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

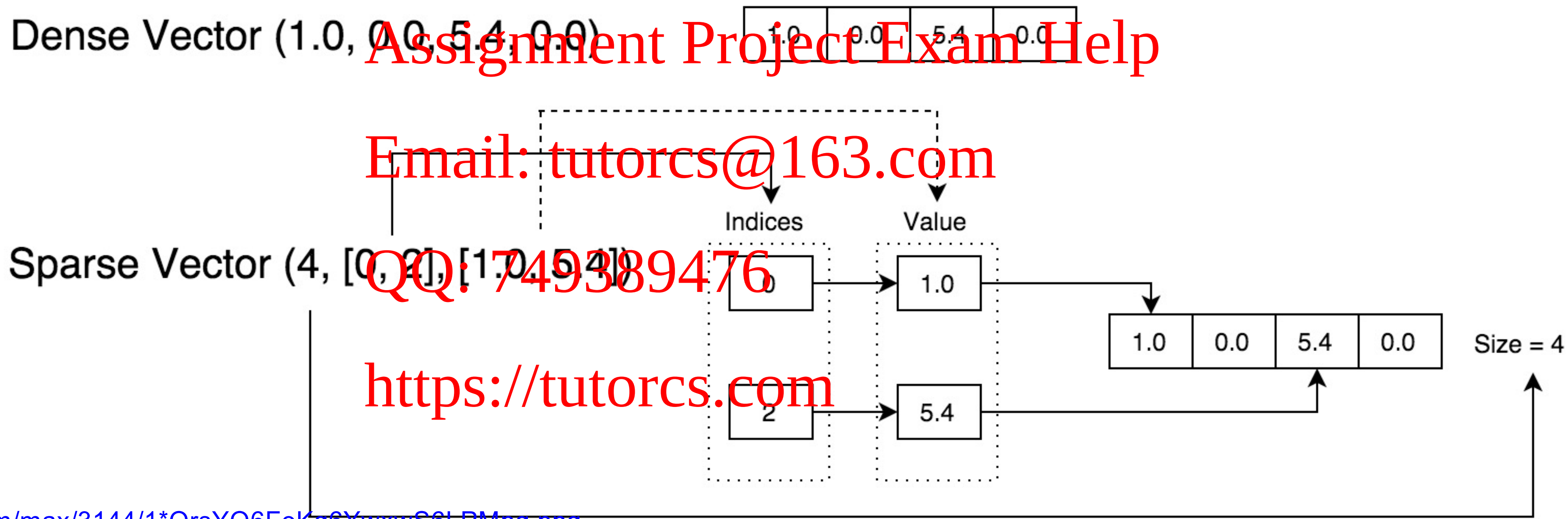
Machine Learning: Data Types

程序代写代做 CS编程辅导

Vector

- A mathematical
- *dense vectors*, where every entry is stored, and
- *sparse vectors*, where only the nonzero entries are stored to save space.

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com



https://miro.medium.com/max/3144/1*OrsYQ6FoKq6YwxwS6LPMpg.png

Features

程序代写代做 CS编程辅导

All learning algorithms require defining a set of *features* for each item, which will be fed into the learning function.

- *For example, for an email, some features might include the server it comes from, or the number of mentions of the word free, or the color of the text.*

WeChat: cstutorcs

In many cases, defining the right features is the most challenging part of using machine learning.

Assignment Project Exam Help

- *For example, in a product recommendation task, simply adding another feature (e.g., realizing that which book you should recommend to a user might also depend on which movies she's watched) could give a large improvement in results.*

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Machine Learning Fundamentals

程序代写代做 CS编程辅导

Supervised and Unsupervised Models

Bias and Variance

Underfitting and Overfitting

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

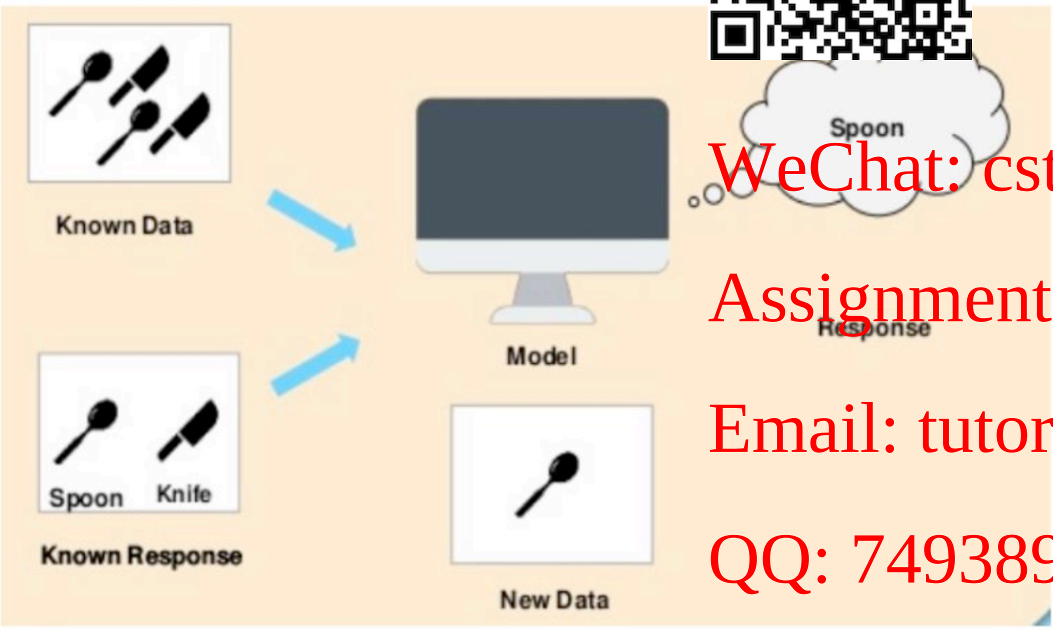
https://tutorcs.com

Model: Types of Model Learning

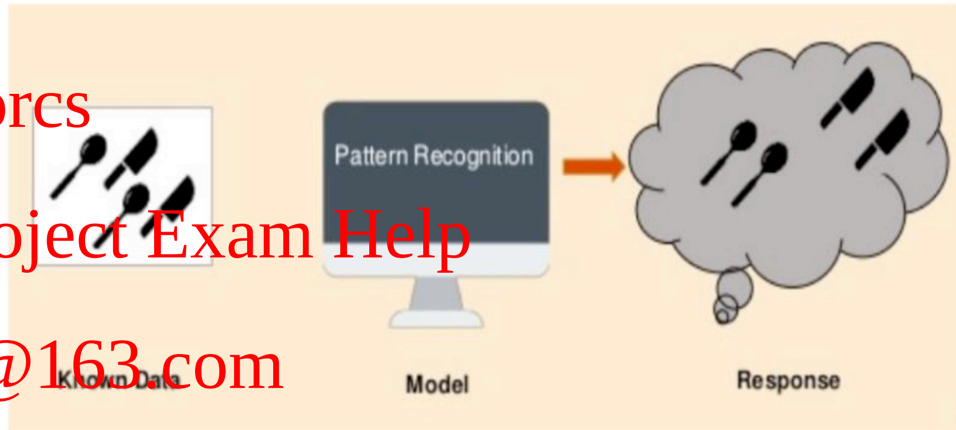
程序代写代做 CS编程辅导

Supervised

Unsupervised



WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com



https://www.simplilearn.com/tutorials/machine-learning-tutorial/supervised-and-unsupervised-learning

Types of Model Learning: Supervised

程序代写代做 CS编程辅导

- Goal: Learn a funct[illegible]abelled training data to predict the output label(s) give[illegible]unlabeled input.

- Training data consists of input features and output information (labels)

- Two types of supervised learning:
  - ❑ Classification
  - ❑ Regression

Data: $(x_1, y_1), \ldots, (x_n, y_n)$

Function: $f: X \rightarrow Y$

x = feature
y = a discrete label (classification),
y = a continuous value (regression)

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Supervised Machine Learning: Classification

程序代写代做 CS编程辅导

Classification problem: To s… puts into a discrete set of classes or labels.

- Binary classification
- Multinomial (Multi-cl… cation

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

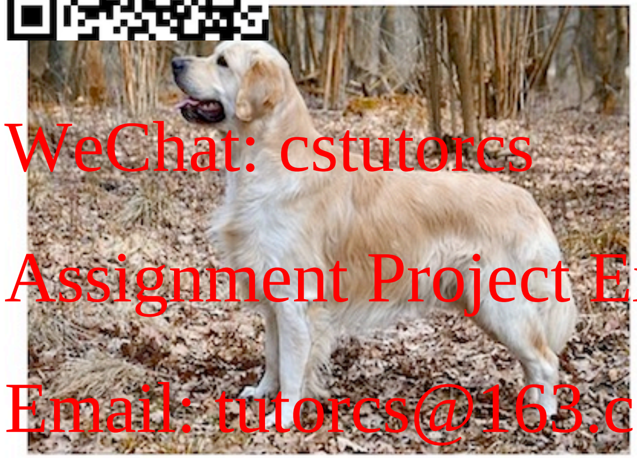
QQ: 749389476

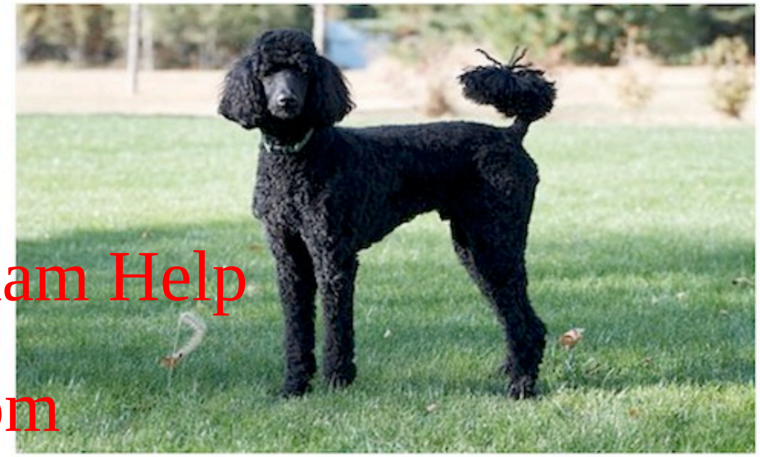
Binary classification example: dog or not dog

https://tutorcs.com

Supervised Machine Learning: Classification

程序代写代做 CS编程辅导





WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

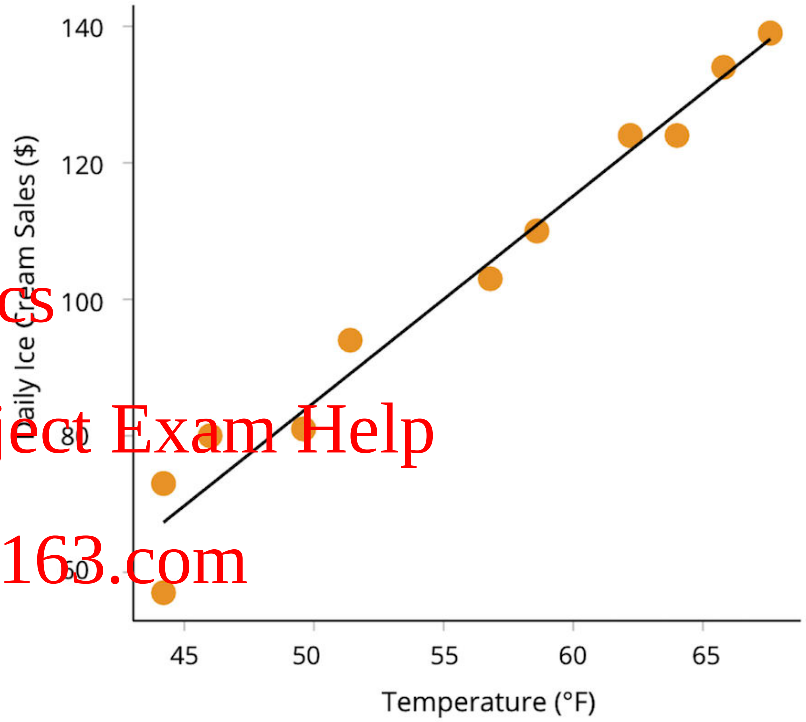
Multinomial classification example: Australian shepherd, golden retriever, or poodle

https://tutorcs.com

Supervised Machine Learning: Regression

程序代写代做 CS编程辅导

- A regression pro when the output e is a real value, such as “dollars” or “weight”.



Regression example: predicting ice cream sales based on temperature

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Supervised Machine Learning in Apache Spark

程序代写代做 CS编程辅导

Algorithm	Typical usage
Linear regression	Regression
Logistic regression	Classification (we know, it has regression in the name!)
Decision trees	Both
Gradient boosted trees	Both
Random forests	Both
Naive Bayes	Classification
Support vector machines (SVMs)	Classification

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Types of Model Learning: Unsupervised

程序代写代做 CS编程辅导

- Goal: Explore the ur structure of the data to extract meaningful informat ut guidance of known output info.
- Deals with unlabelled data (no output labels)
- Two types of unsupervised learning:
  - ❑ Clustering
  - ❑ Association

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

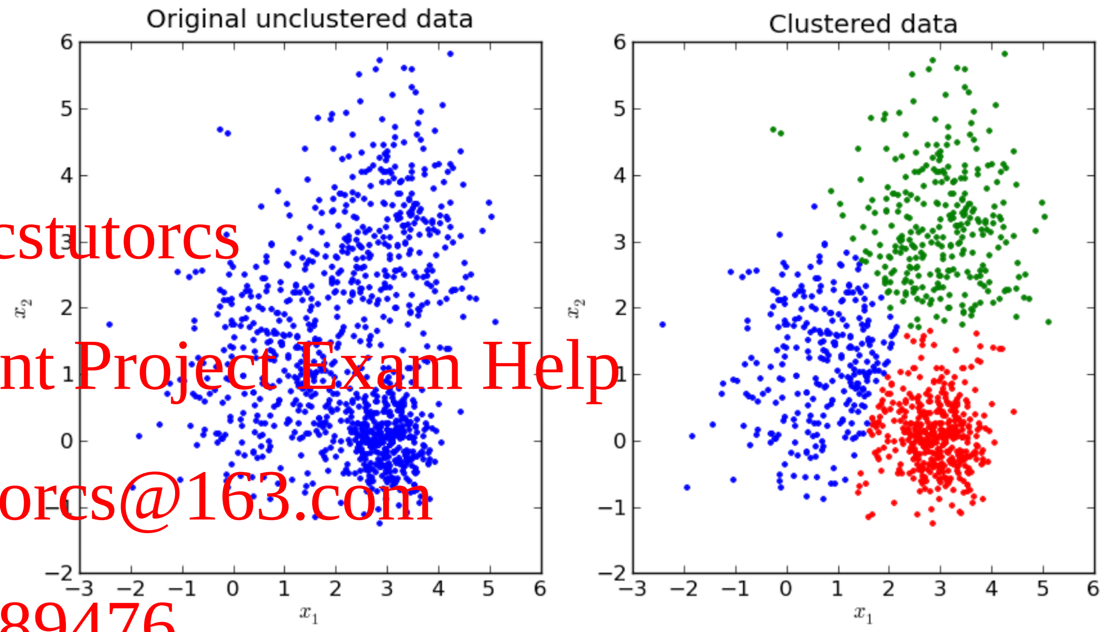
https://tutorcs.com

Unsupervised Machine Learning: Clustering

程序代写代做 CS编程辅导

- Clustering problem: Div… …to clusters which are simil… …n them and are dissimilar … …a belonging to another cluster.
- Where you want to discover the inherent groupings in the data, eg grouping customers by purchasing behaviour

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com



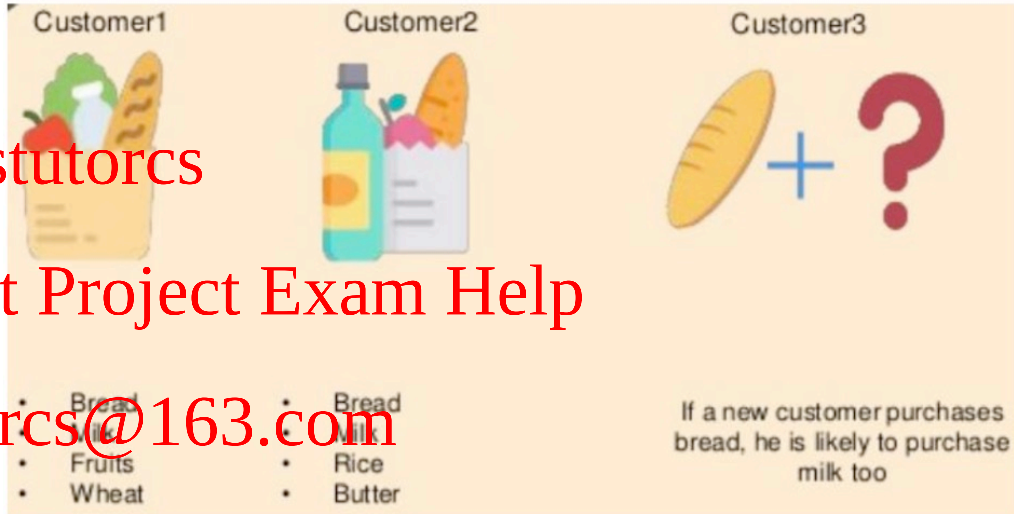
Clustering example

Unsupervised Machine Learning: Association

程序代写代做 CS编程辅导

- Association rule learning
  Discover the probability of occurrence (association) between items in a large dataset

- Where you want to discover rules that describe large portions of your data, e.g., people who buy X also tend to buy Y.



WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Unsupervised Machine Learning in Apache Spark

程序代写代做 CS编程辅导

- *k*-means,
- Latent Dirichlet Allocation (LDA), and
- Gaussian mixture models.

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Machine Learning: Assessment

程序代写代做 CS编程辅导

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

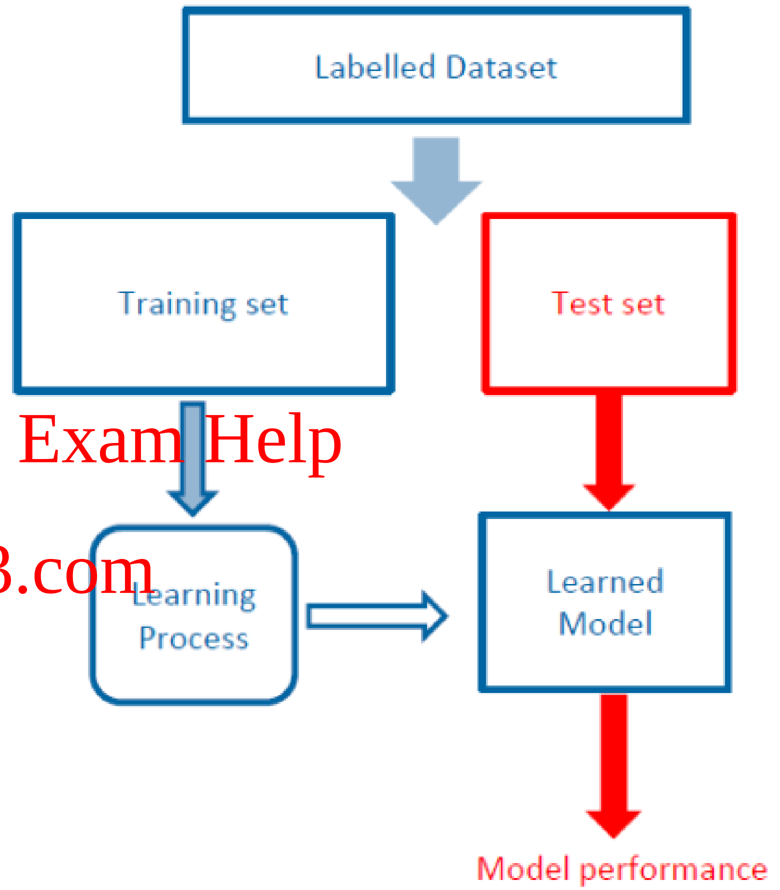
https://tutorcs.com

How to prepare the data?

- Train-Test split
- K-fold cross-validation

How to measure performance?

- TP, FP, TN, FP, confusion matrix
- Accuracy, Recall, Precision, F1-score



Machine Learning: Performance Metrics

程序代写代做 CS编程辅导

Example: Email Spam Detect

In test set: 10 spam, 20 non-sp

Positive = spam

True labels

Predicted labels	SPAM (1)	NON-SPAM (0)
SPAM (1)	7	5
NON-SPAM (0)	3	15

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

actual class

predicted class	positive	negative
positive	true positives (TP)	false positives (FP)
negative	false negatives (FN)	true negatives (TN)

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Machine Learning: Bias and Variance

程序代写代做 CS编程辅导

Bias is the gap between the averaged predicted value by the model and the actual value of the data.

Variance measures the distance of the predicted values in relation to each other.

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Machine Learning: Bias and Variance

程序代写代做 CS编程辅导

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Machine Learning: Overfitting and Underfitting

程序代写代做 CS编程辅导

Overfitting (high var[illegible]w bias) is a model that performs well on the training data [illegible]ralizes poorly to any new data.

Underfitting (low variance, high bias) is an overly simple model that does not perform well even on the training data.

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

程序代写代做 CS编程辅导

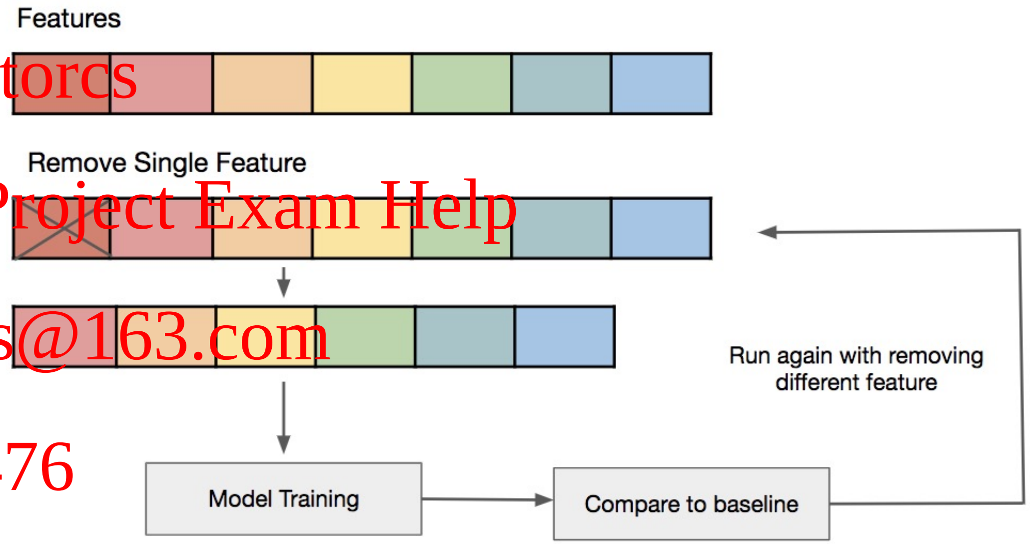
- **Preventing Overfitting**
  - **Train with more data**

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Machine Learning: Overfitting and Underfitting

程序代写代做 CS编程辅导

- **Preventing Overfitting**
  - **Train with more data**
  - **Remove features**



WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

程序代写代做 CS编程辅导

WeChat: cstutorcs

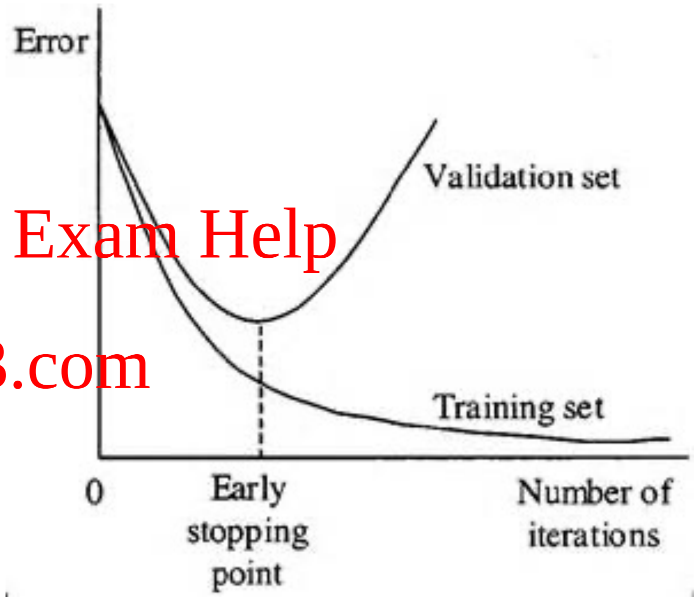
- **Preventing Overfitting**
  - **Train with more data**
  - **Remove features**
  - **Early stopping**

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com



程序代写代做 CS编程辅导

WeChat: cstutorcs

- **Preventing Overfitting**
  - Train with more data
  - Remove features
  - Early stopping
  - Cross validation

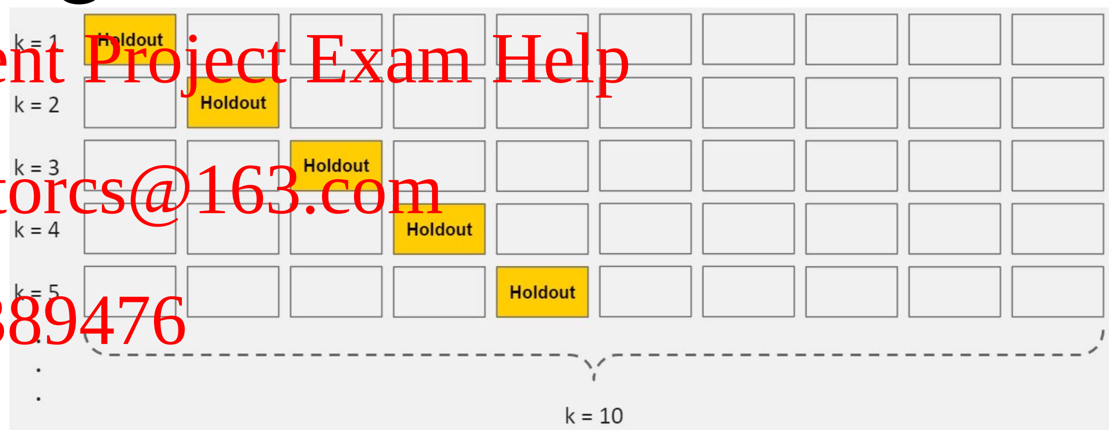
Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

K-Fold Cross-Validation



To be continued..

程序代写代做 CS编程辅导

Next topic -> Featurizatio

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

MONASH University

程序代写代做 CS编程辅导

MONASH INFORMATION TECHNOLOGY

Machine Learning- Featurization

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

GROUP OF EIGHT AUSTRALIA

Machine Learning: Pipeline

程序代写代做 CS编程辅导

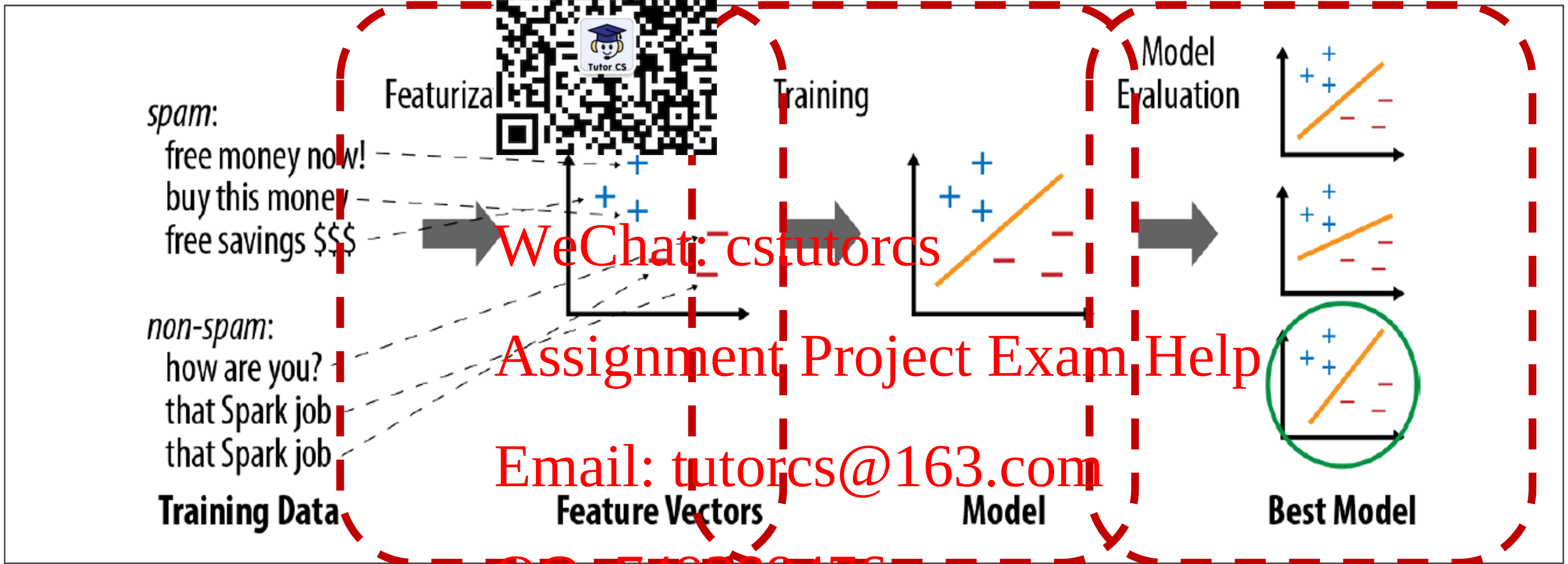
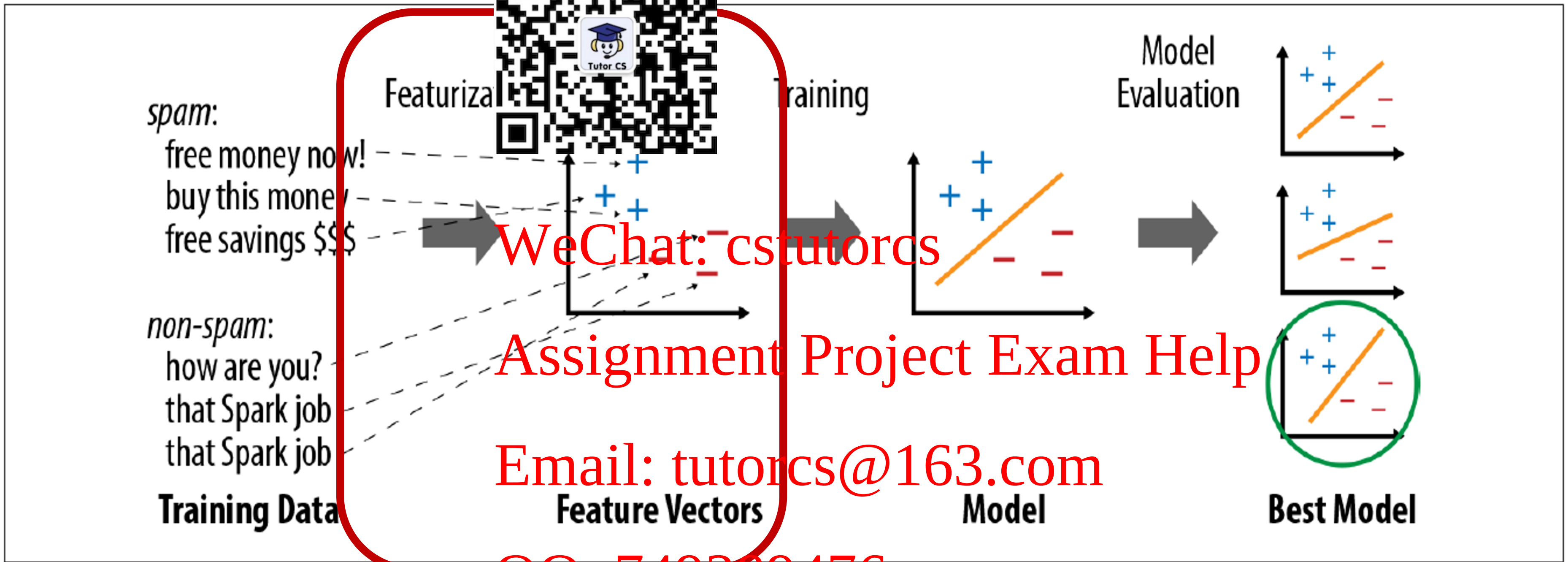


Figure 11-1. Typical steps in a machine learning pipeline

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Machine Learning: Pipeline

程序代写代做 CS编程辅导



WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476

Figure 11-1. Typical steps in a machine learning pipeline

https://tutorcs.com

Featurization: Extraction, transformation and selection

程序代写代做 CS编程辅导

Extraction
- Extracting featu[illegible]raw” data

Transformation
- Scaling, converting, or modifying features

Selection
- Selecting a subset from a larger set of features

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Featurization: Feature Extraction and Transformation

程序代写代做 CS编程辅导

Features

– Any machine learni[illegible]n requires some training data. In training data we have value[illegible]ures for all historical records. Consider this simple data set

Height	Weight	Age	Class
165	70	22	Male
160	58	22	Female

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com

– We can prepare training data by following two techniques

Feature Extraction

Feature Selection

QQ: 749389476
https://tutorcs.com

Featurization: Feature Extraction and Transformation

程序代写代做 CS编程辅导

Feature extractors

- CountVectorizer
- TF-IDF
- Word2Vec
- FeatureHasher (in tutorial)

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Featurization: Feature Extractors

程序代写代做 CS编程辅导

Count Vectorizer

- Convert a collection of text documents to vectors of token counts.
- During the fitting process, *Count Vectorizer* will select the top *vocabSize* words ordered by term frequency across the corpus.

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

```
 id | texts
----|----------------------------------
 0  | Array("a", "b", "c")
 1  | Array("a", "b", "b", "c", "a")
```

Featurization: Feature Extractors

程序代写代做 CS编程辅导

Term Frequency–I[illegible]ocument Frequency, or TF-IDF,

- A simple way t[illegible]e feature vectors from text documents (e.g., web pages).
- It computes two statistics for each term in each document:

  The term frequency (TF), which is the number of times the term occurs in that document, and

  The inverse document frequency (IDF), which measures how (in)frequently a term occurs across the whole document corpus.

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Featurization: Feature Extractors

程序代写代做 CS编程辅导

Term Frequency–Inverse Document Frequency, or TF-IDF,

- Denote a term by *t*, a document by ***d***, and the corpus by ***D***.

- Term frequency ***TF(t,d)*** is the number of times that term ***t*** appears in document ***d***, while document frequency ***DF(t,D)*** is the number of documents that contains term ***t***.

WeChat: cstutorcs

Assignment Project Exam Help

- Inverse document frequency is a numerical measure of how much information a term provides:

Email: tutorcs@163.com

QQ: 749389476

$$IDF(t, D) = \log \frac{|D| + 1}{DF(t, D) + 1},$$

where **|D|** is the total number of documents in the corpus.

https://tutorcs.com

Featurization: Feature Extractors

程序代写代做 CS编程辅导

Term Frequency–I[illegible]ocument Frequency, or TF-IDF,

- The product of th[illegible]s, TF × IDF, shows how relevant a term is to a specific docum[illegible]t is common in that document but rare in the whole corpus).
- The TF-IDF measure is simply the product of TF and IDF:

WeChat: cstutorcs

Assignment Project Exam Help

$$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D).$$

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Featurization: Feature Extractors

程序代写代做 CS编程辅导

Term Frequency–Inverse Document Frequency, or TF-IDF,

Suppose that we have term count tables of a corpus consisting of only two documents, as listed on the right.

Calculate TF-IDF for the term "this".

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Featurization: Feature Extractors

程序代写代做 CS编程辅导

TF-IDF (Solution),

Calculating TF for “thi

TF (“this”, d1) = 1/5 = 0.2
TF (“this”, d2) = 1/7 = 0.14
(Approx.)

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

程序代写代做 CS编程辅导

TF-IDF (Solution),

Calculating IDF for “th

|D| = 2

DF(t, D) = 2

IDF (“this”, D) = log (3/ 3) = 0

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$IDF(t, D) = \log \frac{|D| + 1}{DF(t, D) + 1},$$

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Featurization: Feature Extractors

程序代写代做 CS编程辅导

TF-IDF (Solution),

Calculating TF-IDF for “thi

TF-IDF (“this”, d1, D) = 0.2 * 0 = 0

TF-IDF (“this”, d2, D) = 0.14 * 0 = 0

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D).$$

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Featurization: Feature Extractors

程序代写代做 CS编程辅导

Exercise: Calculate [illegible] for the term “example”.

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Featurization: Feature Extractors

程序代写代做 CS编程辅导

Word2Vec

- maps each word [illegible] e fixed-size vector.
- transforms each document into a vector using the average of all words in the document.
- this vector can then be used as features for prediction, **document similarity calculations**, etc.

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Home Work: Do some research and write a program to find the document similarity using Word2Vec.

Featurization: Extraction, transformation and selection

程序代写代做 CS编程辅导

Extraction
- Extracting featu[illegible]raw” data

Transformation
- Scaling, converting, or modifying features

Selection
- Selecting a subset from a larger set of features

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Featurization: Feature Extraction and Transformation

程序代写代做 CS编程辅导

Feature Transform

- Tokenization
- Stop Words Re
- String Indexing
- One Hot Encoding
- Vector Assembler (Implement In tutorial)

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Featurization: Feature Transformers

程序代写代做 CS编程辅导

Tokenization

- It is the process [illegible] text (such as a sentence) and breaking it into individual terms (usually words).

WeChat: cstutorcs

Text

“The cat sat on the mat.”

Assignment Project Exam Help

↓

Tokens

Email: tutorcs@163.com

“the”, “cat”, “sat”, “on”, “the”, “mat”, “.”

QQ: 749389476

https://tutorcs.com

Featurization: Feature Transformers

程序代写代做 CS编程辅导

Stop Words are wo… should be excluded from the input, typically because th… appear frequently and don't carry as much meaning.

Some words contain more information than others

WeChat: cstutorcs

stopwords — [the, in, for, you, will, have, be]

Assignment Project Exam Help

Quiz: How many words will be removed when we remove stopwords from "Hi Katie the machine learning class will be great best Sebastian"

Email: tutorcs@163.com

QQ: 749389476

3

https://tutorcs.com

Featurization: Feature Transformers

程序代写代做 CS编程辅导

Stop Words Remover

Takes as input a sequence of strings (e.g. the output of a Tokenizer)

Drops all the stop words from the input sequences.

WeChat: cstutorcs

```
id | raw                           | filtered
----|-----------------------------|--------------------
0   | [I, saw, the, red, balloon]  |  [saw, red, balloon]
1   | [Mary, had, a, little, lamb]|[Mary, little, lamb]
```

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Featurization: Feature Transformers

程序代写代做 CS编程辅导

String Indexing

Encoding a string column of labels to a column of label indices.

```
id | category | categoryIndex
----|----------|--------------
0  | a        | 0.0
1  | b        | 2.0
2  | c        | 1.0
3  | a        | 0.0
4  | a        | 0.0
5  | c        | 1.0
```

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Featurization: Feature Transformers

程序代写代做 CS编程辅导

One Hot Encoding

Maps a categorical feature represented as a label index to a binary vector..

A single one-value indicates the presence of a specific feature value from among the set of all feature values.

For string type input data, it is common to encode categorical features using String Indexing first.

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Featurization: Feature Transformers

程序代写代做 CS编程辅导

Why One Hot Enco

For categorical when there is no ordinal relationship, the string indexing is not enough...

Using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results.

A one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Featurization: Feature Transformers

程序代写代做 CS编程辅导

Why One Hot Enco

Example: Let's s ve 3 data instances with attributes of **Preferred Progra anguage** and **OS of Choice.**

WeChat: cstutorcs

Preferred Programming Language	OS of Choice
Javascript	OSX
Python	Linux
Scala	OSX

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Featurization: Feature Transformers

程序代写代做 CS编程辅导

Why One Hot Enco

Preferred Programming Language	OS of Choice
Javascript	OSX
Python	Linux
Scala	OSX

String Indexing

Preferred Programming Language	OS of Choice
0	0
1	1
2	0

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Featurization: Feature Transformers

程序代写代做 CS编程辅导

- **Why cant we ST[illegible]e?**

The Problem Of Or[illegible]

Machine learning algorithms treat the ordinality of numbers in an attribute with some significance: *a higher number "must be better" than a lower number*.

String Indexing

Preferred Programming Language	OS of Choice
0	0
1	1
2	0

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

程序代写代做 CS编程辅导

Featurization: Feature Transformers

- **String Indexing**

Preferred Programming Language	OS of Choice
0	0
1	1
2	0

- **One Hot Encoding**

Javascript	Python	Scala	OSX	Linux
1	0	0	1	0
0	1	0	0	1
0	0	1	1	0

Homework: Have a look at [Principal Component Analysis (PCA)]

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Featurization: Feature Transformers

程序代写代做 CS编程辅导

Why One Hot Encoding

For categorical variables when there is no ordinal relationship, the string indexing is not enough...

Using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results.

A one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Featurization: Extraction, transformation and selection

程序代写代做 CS编程辅导

Extraction

- Extracting featu raw” data

Transformation

- Scaling, converting, or modifying features

Selection

- Selecting a subset from a larger set of features

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

Featurization: Feature Selectors

程序代写代做 CS编程辅导

Feature selection

- This process tri[illegible] most important features that are contributing to decide the label.

Vector Slicer

WeChat: cstutorcs

- It takes a feature vector and outputs a new feature vector with a sub-array of the original features.

Assignment Project Exam Help

- It is useful for extracting features from a vector column.

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com

```
 userFeatures      | features
-------------------|--------------
 [0.0, 10.0, 0.5]  | [10.0, 0.5]
```

Lecture Demo

程序代写代做 CS编程辅导

Spam Classification [illegible]

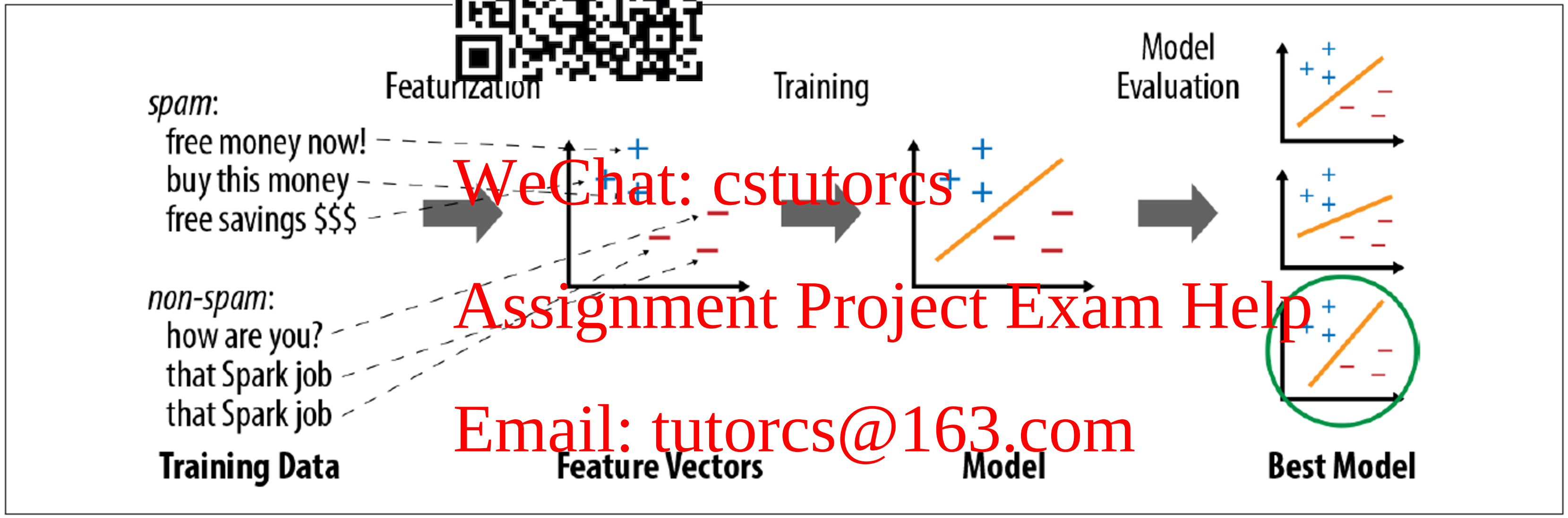


Figure 11-1. Typical steps in a machine learning pipeline

WeChat: cstutorcs
Assignment Project Exam Help
Email: tutorcs@163.com
QQ: 749389476
https://tutorcs.com

Thank You

程序代写代做 CS编程辅导

See you next week

WeChat: cstutorcs

Assignment Project Exam Help

Email: tutorcs@163.com

QQ: 749389476

https://tutorcs.com