
Unit 5
Machine Learning

Part 1

Reading & Translating

Section A: Decision Tree in Machine Learning

Decision Tree in Machine Learning is used for supervised learning [classification and **regression**]. Decision Tree exploits correlation between features and non-linearity in the features.

Wondering what a Decision Tree would be? You might have come across the **programmatic** representation of a decision tree which is a nested if-else.

Let us consider the following pseudo logic, where we are trying to classify the given living-thing into either human, bird or plant:

```
if(displacement is present){
  if(wings are present AND feathers are present){
    living-thing is bird
  } else if(hands are present){
    living-thing is human
  }
} else if(displacement is absent){
  living-thing is plant
}
```

In the above pseudo code, output variable is category of living-thing whose value could be human or bird or plant. Input variable is living-thing. Features of input data taken into consideration are **displacement** [whose values are present/absent], wings [whose values are present/absent], feathers [whose values are present/absent] and hands [whose values are present/absent]. So we have four features whose values are **discrete**.

In the traditional programs, the above if-else-if code is hand written. Efforts put by a human being in identifying the rules and writing this piece of code where there are four features and one input are relatively less.

But could you imagine the efforts required if the numbers of features are in hundreds or thousands. It becomes a tedious job with nearly impossible timelines. Decision Tree could learn these rules from the training data. Despite other classifiers like Naive Bayes Classifier or other linear classifiers, Decision Tree could capture the non-linearity of a feature or any relation between two or more features.

Regarding the capturing relation among features in the above example, the

features (wings and feathers) are co-related. For the considered example (or data set), their values are related in a way **such that** their collective value is deciding on the decision flow.

In machine learning, input dataset for the Decision Tree algorithm would be the list of feature values with the corresponding categorical value. A sample of the dataset is as shown in the Table 5-1.

Table 5-1 A sample of the dataset

Input	Output	Features			
living-being	category	wings	hands	feathers	displacement
Joe	human	absent	present	absent	present
Parrot	bird	present	absent	present	present
Jean	human	absent	present	absent	present
Hibiscus	plant	absent	absent	absent	absent
Eagle	bird	present	absent	present	present
Rose	plant	absent	absent	absent	absent

Each row in the Table 5-1 represents an observation/experiment.

In practical scenarios, the number of features could be from single digit number to thousands, and the data set would contain single digit number to millions of **entries**/observations/experiments.

The common way to build a Decision Tree is to use a greedy approach. Consider you are greedy on the number of Decision Nodes. The number of Decision Nodes should be minimal. By testing a feature value, the Dataset is broken into sub-Datasets, with a condition that the split gives maximum benefit to the classification i.e., the feature value considered(among all the possible feature value combinations) is the best available to categorize the given data set into two subsets. In each sub-Dataset, a new feature value combination is chosen, as in the former split, to divide it into smaller sub-Datasets, with the same condition that the split gives maximum benefit to the classification. The process is repeated until a Decision Node is not required to further split the sub-Dataset, and almost all of the samples in that sub-Dataset belong to a single category.

The graphical representation of Decision Tree for the Dataset mentioned above would be as shown in the Figure 5-1.

From the Figure 5-1, it is evident that the Decision Tree has made use of only two features [displacement, wings] as the other two features are redundant. Thus it needs to reduce the number of Decision Nodes.

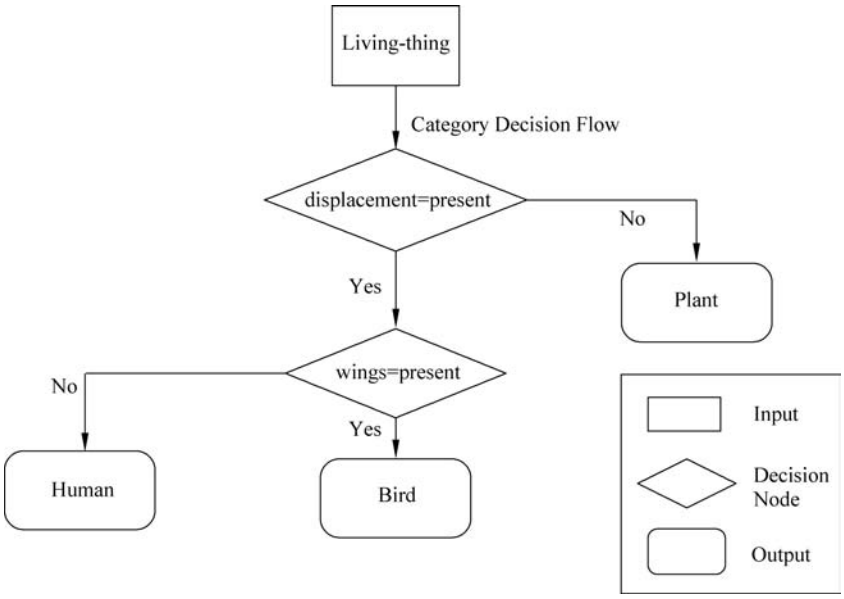

Figure 5-1 Flowchart representation of Decision Tree

Words

regression[ri'greʃn] *n.* 回归
programmatic[ˌprəugrə'mætik] *adj.* 有计划的,按计划的
displacement[dis'pleismənt] *n.* 位移
discrete[di'skri:t] *adj.* 离散的,不连续的
living-being 有机体,生物
parrot['pærət] *n.* 鹦鹉
hibiscus[hi'biskəs; hai'biskəs] *n.* 木槿,芙蓉花
entry['entri] *n.* 条目

Phrases

such that 如此……以致

Exercises

I. Read the following statements carefully, and decide whether they are true (T) or false (F) according to the text.

____ 1. The common way to build a Decision Tree is to use a SVM approach.
____ 2. Regarding machine learning, input dataset for the Decision Tree algorithm would be the list of feature values with the corresponding categorical value.
____ 3. Decision Tree could take the non-linearity of a feature or any relation between two or more features.

____ 4. In the Figure 5-1, the Decision Tree has used only two features [displacement, feathers].

____ 5. Decision Tree in Machine Learning is used for unsupervised learning.

II. Choose the best answer to each of the following questions according to the text.

1. Which of the following is right about the Decision Tree? ( )

A. Decision Tree could take the linearity of a feature or any relation between two or more features.

B. Decision Tree exploits correlation between features and non-linearity in the features.

C. Decision Tree in Machine Learning is used for unsupervised learning.

D. The common way to build a Decision Tree is to use a SVM approach.

2. How many features are mentioned in the Figure 5-1? ( )

A. One

B. Two

C. Three

D. Four

3. Which of the two features has Decision Tree used in the Figure 5-1? ( )

A. [displacement, feathers]

B. [displacement, hands]

C. [displacement, wings]

D. None of the above

III. Fill in the numbered spaces with the words or phrases chosen from the box. Change the forms where necessary.

variable	independent	can	define	call
many	dependent	estimate	what	common

Linear Regression

Linear regression is a basic and ____1____ used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which ____2____ in particular are significant predictors of the outcome variable, and in ____3____ way do they—indicated by the magnitude and sign of the beta estimates—impact the outcome variable? These regression ____4____ are used to explain the relationship between one dependent variable and one or more ____5____ variables. The simplest form of the regression equation with one dependent and one independent variable is ____6____ by the formula y = c + b * x, where y = estimated ____7____ variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

There are ___8___ names for a regression's dependent variable. It may be ___9___ an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables ___10___ be called exogenous variables, predictor variables, or regressors.

IV. Translate the following passage into Chinese.

Support Vector Machine (SVM)

A support vector machine is a supervised learning algorithm that sorts data into two categories. It is trained with a series of data already classified into two categories, building the model as it is initially trained. The task of an SVM algorithm is to determine which category a new data point belongs in. This makes SVM a kind of non-binary linear classifier.

An SVM algorithm should not only place objects into categories, but have the margins between them on a graph as wide as possible.

Section B: K-means Clustering Algorithm and Example

"I'm clueless"

You say, looking at **an ocean of** unlabeled data, waving in front of you. It is true that the lack of labels can sometimes **freak** us **out**, leaving us wondering how to group the data together. But luckily, k-means **clustering** algorithm is here to rescue, one of the simplest algorithms for unsupervised clustering (dealing with data without defined categories). Assigning data points into k **clusters** based on the minimum distance, k-means clustering is simple, helpful, and effective for finding the **latent** structure in the data.

Here we provide some basic knowledge about k-means clustering algorithm and an illustrative example to help you clearly understand what it is.

K-means clustering algorithm is an unsupervised machine learning algorithm for determining which group a certain object really belongs to. What it means by "being unsupervised" is that there are no **prescribed** labels in the data denoting its structure. The main idea is to assign each observation into the cluster with the nearest **mean** (**centroid** [1]), **serving as** a prototype of the cluster.

Here are five simple steps for the k-means clustering algorithm and an example for illustration:

- Step 1: Visualize n data points and decide the number of clusters (k). Choose k random points on the graph as the centroids of each cluster. For this example, we would like to divide the data into 4 clusters, so we pick 4 random centroids (Figure 5-2).
- Step 2: Calculate the **Euclidean** distance between each data point and chosen clusters' centroids. A point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other ones (Figure 5-3).
- Step 3: After assigning all **observations** to the clusters, calculate the clustering score, by **summing up** all the Euclidean distances between each data point and the corresponding centroid.

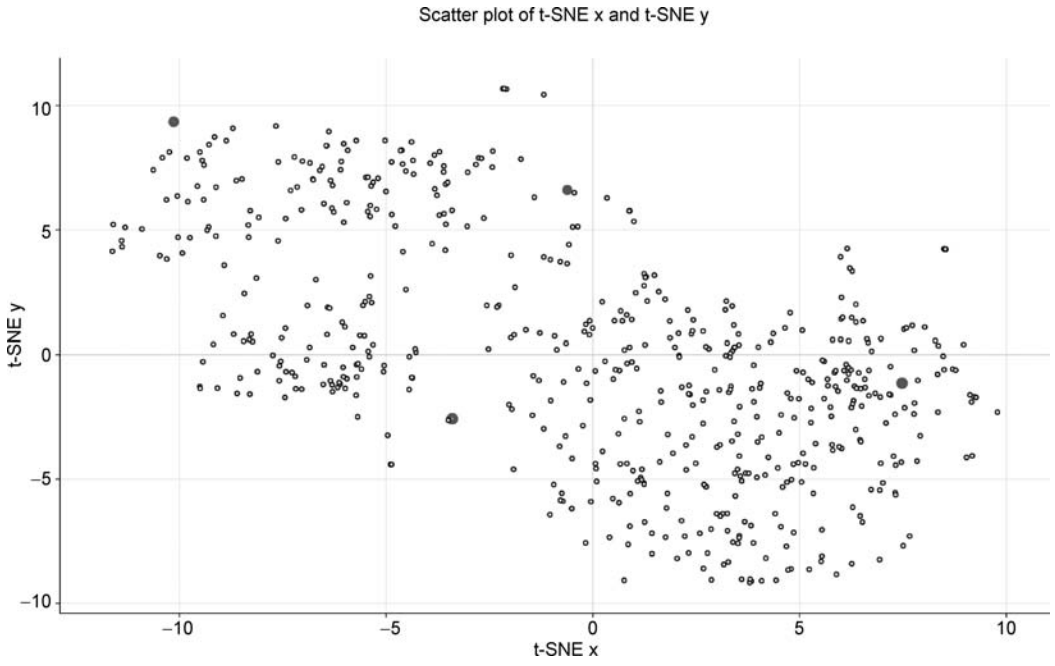

Figure 5-2 Visualize the data and pick the random centroids (which is 4 in this example)

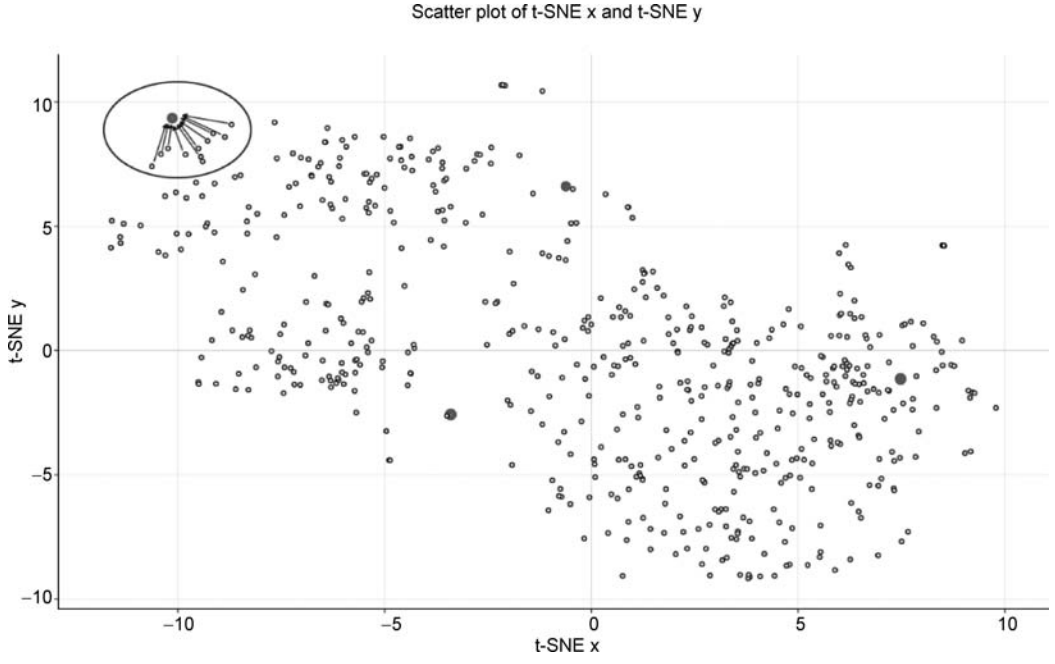

Figure 5-3 Assign each point into the cluster with the nearest centroid

$$\text{Total distances} = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2$$

Where:

k: the number of clusters

n：the number of points belonging to cluster j

c_j：the centroid of cluster j

• Step 4：Define the new centroid of each cluster by calculating the mean of all points assigned to that cluster. Here's the formula（n is the number of points assigned to that cluster）：

$$= \frac{\sum_{i=1}^{n} x_i}{n}$$

• Step 5：Repeat from step 2 until the positions of the centroids no longer move（Figure 5-4）and the assignments stay the same（Figure 5-5）.

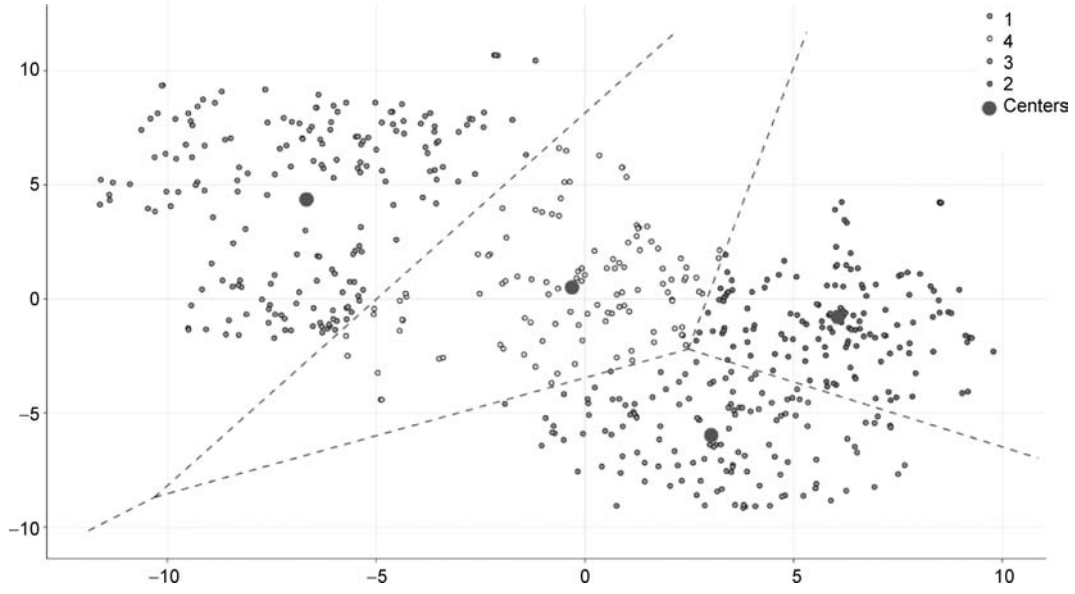

Figure 5-4　Final iteration：distances are minimized and centroids no longer move

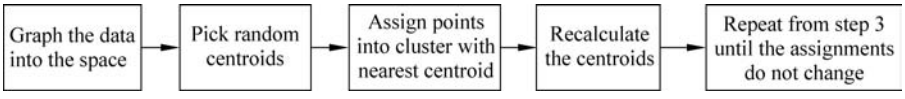

Figure 5-5　Flow chart of k-means clustering algorithm

There you go：data points are now grouped into 4 different clusters. Using a simple idea of minimizing distances between data points to group them together，k-means clustering algorithm is extremely helpful for understanding the structure of the data，how observations are classified，and interpreting the story behind. K-means clustering has been widely used in data analysis，especially in life sciences，in analyzing thousands to millions of data points in single-cell **RNA-seq** and bulk RNA-seq experiments.

Note that the Euclidean **metric** measures the distance based on the vector connecting two points，and will cause some biases for data with different scales. For example，in RNA-seq data，**gene** expression values can range from as little as 0.001 to a thousand，

stretching the data points along an **axis**. That is, the variable with the smaller scale will be easily **dominated** and play little in the **convergence**, as clusters will scatter along an axis only. For this reason, it is necessary to make sure that the variables are at the same scale before using k-means clustering.

Note that before determining the number of clusters to assign the data into (the variable k), you should have an overview of the data and on what basis you want to group them. You can even apply a hierarchical clustering on the data first to briefly understand the structure of the data before choosing k by hand.

A well-known method to validate the number of clusters is the Elbow method [2], that is to run k-means clustering several times for a range of values of k (usually from 2 to 10) and pick out the value of k that causes sudden drop in the sum of **squared distances**. More specifically, for each value of k, we calculate the sum of squared distances (between each point and the corresponding centroid) and graph the results on a line chart. Choose the value where the sum of squares drops, giving an angle in the graph (**a. k. a.** an elbow)—that is the optimal value of k (Figure 5-6).

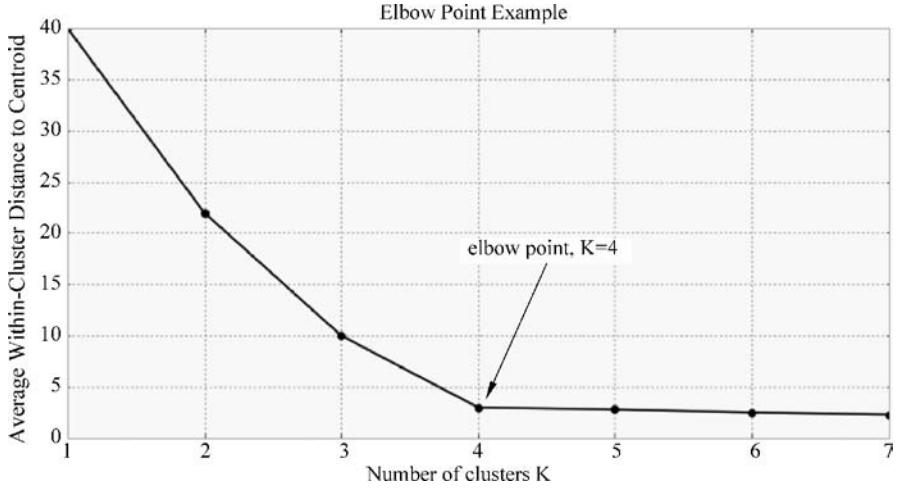

Figure 5-6 Elbow point example

Words

clustering['klʌstəriŋ] *n.* 聚类
cluster['klʌstə(r)] *n.* 群集，簇，集群
latent['leitnt] *adj.* 潜在的，潜伏的
prescribe[pri'skraib] *v.* 规定
mean[mi:n] *n.* 平均数，平均值
centroid['sentrɔid] *n.* 质心，形心
Euclidean[ju:'klidiən] *adj.* 欧几里得几何学的，欧几里得的
observation[ˌɔbzə'veiʃn] *n.* 数据点
RNA-seq 转录组测序技术（RNA sequencing）
dominate['dɔmineit] *v.* 支配，控制
convergence[kən'vɜ:dʒəns] *n.* 趋同，融合，一体化

metric['metrik] *n.* 度量标准
gene[dʒi:n] *n.* 基因
axis['æksis] *n.* 轴，轴线

Phrases

an ocean of 极多的，无穷无尽的
freak out 崩溃，使处于极度兴奋中
serve as 用作，充当
sum up 计算……的总数
squared distances 距离平方

Abbreviations

a.k.a. 亦称，又名(also known as)

Notes

[1] k-means 是一种数据聚类算法，质心(centroid)是指各个类别的中心位置，质心的维数等同于单条数据的维数。比如说，你有 1000 条数据，每条数据 100 维。如果使用 k-means 算法将这 1000 条数据聚为 10 个类别，就会得到 10 个质心。每个类别的质心是该类别所有数据点的均值。比如第一次确定了 10 个质心，同时也将元数据分别归类到这 10 个质心，那么接下来可继续调整质心以致最后达到最优：

(1) 将各个示例 sample 分配到距离最近的质心；

(2) 对于各个类别，计算其所包含的 sample 的平均值，作为该类别新的质心。

[2] 肘部法则(Elbow method)，此种方法适用于 K(簇的数量) 值相对较小的情况，当选择的 k 值小于真正的 K 时，k 每增加 1，cost 值就会大幅地减小；当选择的 k 值大于真正的 K 时，k 每增加 1，cost 值的变化就不会那么明显。这样，正确的 k 值就会在这个转折点，类似 elbow 的地方。

Exercises

I. Read the following statements carefully, and decide whether they are true (T) or false (F) according to the text.

____ 1. K-means clustering algorithm is a supervised machine learning algorithm.

____ 2. The main concept of k-means is to assign each observation into the cluster with the nearest mean (centroid), serving as a prototype of the cluster.

____ 3. To find the latent structure in the data k-means clustering is a simple way to assign data points into k clusters based on the minimum distance.

____ 4. "Being unsupervised" is that there are some prescribed labels in the data denoting its structure.

____ 5. Elbow method is a well-known method which validates the number of clusters.

II. Choose the best answer to each of the following questions according to the text.

1. Which of the following is not mentioned in the text? ( )

A. ID3

B. Centroid

C. Euclidean

D. K-means

2. How many steps are mentioned for the k-means clustering algorithm and an example for illustration? ( )

A. Two

B. Three

C. Four

D. Five

3. Which of the following is right? ( )

A. The main concept of k-means is to assign each observation into the cluster with the nearest mean (centroid), serving as a prototype of the cluster.

B. To find the latent structure in the data k-means clustering is a simple way to assign data points into k clusters based on the minimum distance.

C. Elbow method is a well-known method which validates the number of clusters.

D. All of the above

III. Fill in the numbered spaces with the words or phrases chosen from the box. Change the forms where necessary.

understand	labor	deal	advantage	like
base	method	as	reflect	use

Clustering Algorithms

Clustering algorithms can automatically recognize the pattern inside the data so ___1___ to analyze the collected data without their labels. Using this advantage, three clustering-based fault diagnosis methods are presented to ___2___ with some diagnosis cases of rotating machinery in which the labeled data are limited. In the first method, compensation distance evaluation technique and the weight K nearest neighbor are ___3___ to recognize the mechanical faults, harnessing the merits that the computation of feature weights is simpler and the weights are easier to ___4___. The second method is presented ___5___ on weight fuzzy c-means, which is robust to the local structure of the data and ___6___ the level of uncertainty over the most appropriate assignment.

Finally, a Hybrid clustering algorithm-based fault diagnosis ___7___ is introduced, considering the problems ___8___ the sample influence for clustering and the automatic setting of the cluster number. The results of the diagnosis cases verify that these diagnosis methods take full ___9___ of unlabeled data and reduce the human ___10___ in fault diagnosis.

IV. Translate the following passage into Chinese.

Ensemble Learning

Many ensemble learning tools can be trained to produce various results. Individual algorithms may be stacked on top of each other, or rely on a "bucket of models" method of evaluating multiple methods for one system. In some cases, multiple data sets are aggregated and combined. For example, a geographic research program may use multiple methods to assess the prevalence of items in a geographic space. One of the issues with this type of research involves making sure that various models are independent, and that the combination of data is practical and works in a particular scenario.

Ensemble learning methods are included in different types of statistical software packages. Some experts describe ensemble learning as "crowdsourcing" of data aggregation.

Part 2

Simulated Writing: Developing Reports and Proposals (I)

报告和提案是在工作中最常写的长文档。这两者都回答了某个主题或项目的问题，或者针对某个问题提供解决方案。读者将会研究作者的报告，并且运用其中的结论和分析来帮助他们进行决策。除了商业企业之外，非盈利机构和政府机构也会撰写报告来总结或者分析研究状况。有时，组织会雇佣专业的撰稿人撰写提案以赢得合同，或获得销售机会。学会写作这些重要的文档是一项很有价值的专业技能。

1. 了解报告和提案

报告是一种针对特定主题交流信息而设计的书面文档。虽然有些报告可以包含分析或建议，但撰写的报告往往很客观。提案与报告很相似，但其目的在于说服和通知。提案提供了有关产品、服务或者想法的信息，并且试图说服读者接纳所建议的解决方案。报告与提案的一个关键区别在于它们被写作的时间。提案通常在制定决策过程的早期进行，此时它能够影响决策。报告通常在已经采取一些行动之后撰写。当一项活动或项目发生的时候，一些报告可以记录它们的状态。当活动或项目完结时，可以撰写其他的报告。报告和提案的类型参见图 5-7。

在开始撰写报告或提案前，请回答下面的问题：

撰写的目的是什么？

撰写报告的第一步是明确地定义目的。首先分析想要达到的目标，目标是通知、更新、分析，还是说服？目标将帮助决定应该使用的形式。

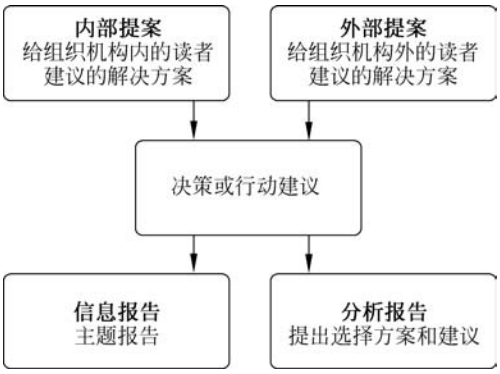

图 5-7 报告和提案的类型

读者是谁？

与其他类型的文档相同，撰写报告或提案时，要考虑读者。为了更好地满足读者的需求，要辨别他们理解报告或提案主旨的程度。他们想要通过阅读报告或提案了解什么？他们有可能怎样阅读？应该怎样撰写才能使信息清晰，并且使读者易懂？一定要考虑主要读者和次要读者，以及包括那些可能会阅读该文档的任何人。

应该撰写报告还是提案？

撰写报告是为了与他人分享信息。撰写提案是为了说服读者采纳想法、产品或者解决方案。这两者与分析报告很类似，但区别是只是这里只呈现一个建议。表 5-2 给出了何时应该撰写报告或提案的建议。

表 5-2 何时撰写报告或提案

场　景	报告	提案	其他
参加一场贸易展示会，希望通告本公司的竞争对手的产品	√		
需要为公司流程撰写文档	√		
分析是购买新的计算机设备还是升级现有设备	√		
提议购买新的计算机设备		√	
为规划职员资源提议一种新方案		√	
为个人或组织提供公司的服务		√	
在所参加的一场会议上为之后的查阅总结所做的笔记			非正式笔记或大纲
为一般的受众推销公司的服务			广告
为潜在的顾客描述公司产品，并且提供样品			展示

报告中会展示信息还是分析话题？

报告可以是下述两种类型中的一个。信息报告以清晰、客观的形式展示信息。当想为读者书面总结针对某个主题的信息时，使用信息报告比较合适。意见和建议不应写在一个信息报告之中。分析报告一般会呈现数据、分析和结论。分析报告通常会提供不同的选择，鉴别优劣以得到替代方案，以及包含具体的建议。

提案是为内部还是外部的读者而撰写？

提案也有两种类型。内部提案建议如何在一个组织内解决问题，例如，通过改变一个程

序或者使用商家的不同产品或服务。外部提案被设计来销售产品或服务于客户，并且通常为响应请求而撰写。

回答这些问题有助于决定报告应该有多长，包含什么样的信息，以及适当形式。

2. 规划报告或提案

有条理地组织业务报告和提案，以便使信息容易阅读和理解。在写第一句话之前，就应该有针对如何组织报告或提案的好的思路。将一般的思路组合在一起，并遵循逻辑顺序。该顺序能够满足目的，并有助于读者明白所写的内容。有逻辑地组织信息的方式应依时间、重要性以及类别，例如位置或产品来决定。撰写正式或非正式的大纲可以有助于规划有效的报告。表 5-3 总结了撰写大纲的注意事项。

表 5-3　撰写大纲的注意事项

要　素	适合提到	尽量避免
主要思路	• 以头脑风暴开始，列出想要包含的所有思路 • 选择一个作为主要的思路 • 写在大纲开头	• 保留所有的思路，而不是只保留那些服务于报告或者提案的目的和那些服务于读者的思路 • 表述主要思路超过两句
大标题和章节	• 选择议题并写出相应的标题 • 使用标准标题，例如介绍（Introduction）和结论（Conclusion） • 按逻辑顺序列出标题 • 在正式大纲中，使用罗马数字标注大标题	• 偏离如下的标准模式：（1）介绍（Introduction）；（2）事实和发现（Facts or findings）；（3）结论（Conclusion） • 包含没有足够细节和证据的议题
子标题	• 用子标题将大议题分解为子议题 • 以逻辑顺序列出子议题，例如，时间、重要性，或者类别 • 在正式大纲中，第一级子标题使用大写字母，下一级使用数字，最后一级使用小写字母	• 以任意顺序列出子议题 • 使用难以解释的子标题

1）首先确定主要的思路

开始撰写大纲可以通过在页面顶端用一两个句子描述主要思路来开始。如果主要的思路太长，可以精简所写的内容。在页面的上方说明主要的思路，有助于在制定大纲的其余部分时专注于自己的目标。许多报告和提案的主要思路是要描述一个解决问题的办法。

2）为重要的思路使用标题

复查报告的思路和主题，并选择最重要的部分。这些都应作为大纲的主要标题。这些标题要按照逻辑顺序列出，比如从最重要的到最不重要的，或按时间顺序（如果报告强调了时间）。这些标题将成为报告的主要部分。图 5-8 展示了正式和非正式大纲中的标题，包括使用罗马数字、大写字母、数字和小写字母的规范。

3）为子议题创建子标题

可以将每一个主要议题分为几个思路，以便详细地讨论它们。在大纲中列出这些思路将其作为子标题。可以为每个大标题提供两个或两个以上的子标题，如图 5-8 所示。如果正在写一个很长的或者很复杂的报告，可以将子议题分解为更小的部分。

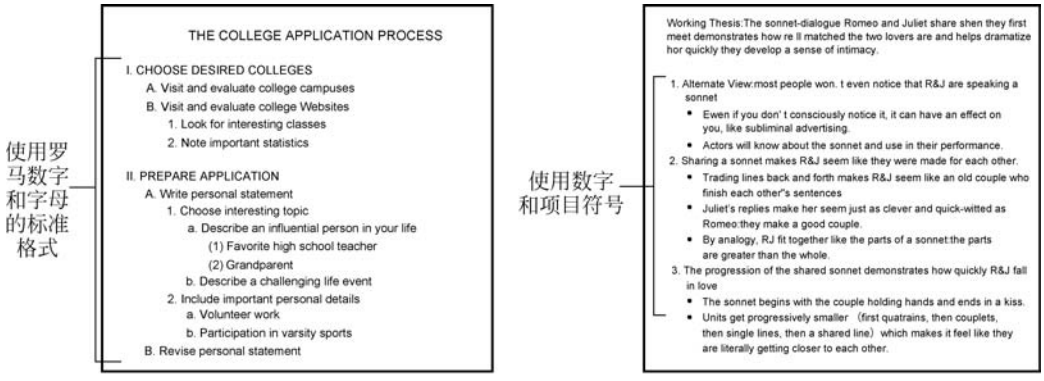

图 5-8 正式和非正式的大纲

4）将合适的章节添加进来

大多数报告和提案包括标准章节，如介绍、背景、现状、事实、提出的解决方案、总结、结论、建议、利弊、参考清单和附录。选择能够服务于报告或提案目的章节。

5）复查大纲

复查大纲的完整草案以便回答以下问题：思路是否按照逻辑顺序安排？如果大声读大纲给自己听，听起来是否有意义？标题和子标题是否具有逻辑性和平衡性？它们的重要性是否差不多？如果有必要则重新排列顺序。议题是否已经有了足够的细节或证据来支持主要的思路？如果不是，那就应该将它们添加到大纲中或者重组大纲。

转 109 页

Part 3

Listening & Speaking

Dialogue: Machine Learning


在线音频

(*Before the first lesson of Machine Learning, Mark met with Henry and Sophie in front of their classroom*)

Mark: Excuse me, Henry and Sophie. Could you help me? [1]

Henry: Sure. What's the problem?

[1] Replace with:
1. Can you give me a hand?
2. Could you please do me a favor?
3. Could you do me favor?

Mark: I'm a little bit confused with machine learning? Exactly [2] what is machine learning?

Henry: Well, machine learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on models and inference instead. It is seen as a subset of artificial intelligence.

Sophie: To my knowledge, machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in the applications of email filtering, detection of network intruders, and computer vision, where it is **infeasible** to develop an algorithm of specific instructions for performing the task.

Mark: Does machine learning have some relationships with other areas?

Henry: Of course. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

Mark: And are there any classifications for the machine learning?

Sophie: Absolutely. Machine learning tasks are classified into several broad categories. In supervised learning, the algorithm builds a mathematical model of a set of data that contains both the inputs and the desired outputs.

[2]Replace with:

1. Accurately
2. Correctly
3. Definitely
4. Truly
5. Precisely

For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each image would have a label (the output) **designating** whether it contained the object.

Henry: In special cases, the input may be only partially available, or restricted to special feedback. Semi-supervised learning algorithms develop mathematical models from incomplete training data, where a portion of the sample inputs are missing the desired output.

Mark: Could you please name a few algorithms for supervised learning?

Henry: Sure. Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to **file** the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false.

Sophie: And regression algorithms are named for their continuous outputs, meaning they may have any value within a range. Examples of a continuous value are the temperature, length, or price of an object.

Mark: So, how about unsupervised learning?

Sophie: Well, in unsupervised learning, the algorithm builds a mathematical model of a set of data which contains only inputs and no desired outputs. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points.

Henry: Moreover, unsupervised learning can discover patterns in the data, and can group the inputs into categories, as in feature learning. Dimensionality reduction is the process of reducing the number of "features", or inputs, in a set of data.

Mark: OK, so what else?

Henry: Well, **reinforcement learning** algorithms are given feedback in the form of positive or negative reinforcement in a dynamic environment, and are used in autonomous vehicles or in learning to play a game against a human opponent.

Sophie: And other specialized algorithms in machine learning include topic modeling, where the computer program is given a set of natural language documents and finds other documents that cover similar topics. Machine learning algorithms can be used to find the unobservable probability density function in density estimation problems, **so on and so forth**.

Mark: So much knowledge I'm interested in! Thank you very much!

Exercises

Work in a group, and make up a similar conversation by replacing the statements with other expressions on the right side.

Words

infeasible[in'fi:zib(ə)l] *adj.* 不可行的，不可实行的	designate['dezigneit] *v.* 指定，指派
	file[fail] *v.* 把……归档

Phrases

reinforcement learning 强化学习
so on and so forth 等等

Listening Comprehension: Supervised Learning

在线音频

Listen to the article and answer the following 3 questions based on it. After you hear a question, there will be a break of 15 seconds. During the break, you will decide which one is the best answer among the four choices marked (A), (B), (C) and (D).

Questions

1. Which of the following is right? ( )
   (A) Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
   (B) Supervised learning infers a function from labeled training data consisting of a set of training examples.
   (C) A supervised learning algorithm analyzes the training data and produces an inferred function.
   (D) All of the above

2. Regarding the hand-written digit recognition problem, which of the following is right? ( )
   (A) A reasonable data set for this problem is a collection of images of hand-written digits.
   (B) A reasonable data set for this problem is for each image, what the digit actually is.
   (C) A set of examples of the form (image, digit) should be considered.
   (D) All of the above

3. Which of the following can't supervised learning do? ( )
   (A) Supervised learning is the machine learning task of learning a function that maps an output to an input based on example output-input pairs.
   (B) Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
   (C) Supervised learning infers a function from labeled training data consisting of a set of training examples.
   (D) A supervised learning algorithm analyzes the training data and produces an inferred function.

Words

map[mæp] *v.* 映射
entirety[in'taiərəti] *n.* 全部，完全
outset['autset] *n.* 开始，开端

在线音频

Dictation: Unsupervised Learning

This article will be played three times. Listen carefully, and fill in the numbered spaces with the appropriate words you have heard.

Unsupervised learning is a ___1___ of machine learning that learns from test data that has not been ___2___, classified or categorized. Instead of ___3___ to feedback, unsupervised learning identifies **commonalities** in the data and reacts based on the presence or ___4___ of such commonalities in each new piece of data. ___5___ include supervised learning and reinforcement learning.

In the unsupervised ___6___, the training data does not contain any output information at all. We are just given input examples $X_1, \cdots, X_N$. You may wonder how we could possibly learn anything from mere inputs. Consider the coin ___7___ problem. Suppose that we didn't know the **denomination** of any of the ___8___ in the data set.

We still get similar ___9___, but they are now ___10___ so all points have the same "color". The decision regions in unsupervised learning may be ___11___ to those in supervised learning, but without the labels. However, the correct clustering is less ___12___ now, and even the number of clusters may be ___13___.

___14___, this example shows that we can learn something from the inputs by themselves. Unsupervised learning can be ___15___ as the task of **spontaneously** finding ___16___ and structure in input data. For instance, if our task is to ___17___ a set of books into topics, and we only use ___18___ properties of the ___19___ books, we can identify books that have similar ___20___ and put them together in one category, without naming that category.

Words

commonality [kɔməˈnæliti] *n.* 公共，共性

denomination [diˌnɔmiˈneiʃn] *n.* 面额

spontaneously [spɔnˈteɪniəsli] *adv.* 自发地，自然地