答案

第1章 机器学习概念

Exercise 1.1

Express each of the following tasks in the framework of learning from data by specifying the input space X, output space Y, target function $f: X \to Y$.and the specifics of the data set that we will learn from.

(a) Medical diagnosis: A patient walks in with a medical history and some symptoms, and you want to identify the problem.
(b) Handwritten digit recognition (for example postal zip code recognition for mail sorting) .
(c) Determining if an email is spam or not.
(d) Predicting how an electric load varies with price, temperature, and day of the week.
(e) A problem of interest to you for which there is no analytic solution,but you have data from which to construct an empirical solution

标识	输入空间	输出空间	目标函数
a	病史，症状	是否患病	判别函数(医疗资料至是否患病的映射)
b	手写数字	判别出来的数字	判别函数(手写数字至数字的映射)
c	邮件内容	是否为垃圾邮件	判别函数(邮件内容至是否为垃圾邮件的映射)
d	电力负载量，电力价格，温度，日期	电量的预测	预测函数
e	数据	处理结果	经验解决方案(empirical solution)

Exercise 1.2

Which of the following problems are more suited for the learning approach and which are more suited for the design approach?

(a) Determining the age at which a particular medical test should be performed
(b) Classifying numbers into primes and non-primes
(c) Detecting potential fraud in credit card charges
(d) Determining the time it would take a falling object to hit the ground
(e) Determining the optimal cycle for traffic lights in a busy intersection

learning approach是指机器学习的方法，design approach是指推导的方法，适合能够直接分析推导的问题

标识	适合的方式
a	learning approach
b	design approach
c	learning approach
d	design approach
e	learning approach

质数和自由落体问题可以直接得出结果，所以使用design approach

Exercise 1.3

For each of the following tasks, identify which type of learning is involved (supervised, reinforcement, or unsupervised) and the training data to be used. If a task can fit more than one type, explain how and describe the training data for each type.

(a) Recommending a book to a user in an online bookstore
(b) Playing tic tac toe
(c) Categorizing movies into different types
(d) Learning to play music
(e) Credit limit: Deciding the maximum allowed debt for each bank customer

(a) 推荐系统的问题，非监督学习
(b) 游戏输赢是确定的，所以是监督学习，然后学习的过程是强化学习
(c) 标准的分类问题，监督学习
(d) 学习音乐的结果没有直接标志，非监督学习，然后学习的过程是强化学习
(e) 标准的预测问题，监督学习

Exercise 1.4

We have 2 opaque bags, each containing 2 balls. One bag has 2 black balls and the other has a black and a white ball . You pick a bag at random and then pick one of the balls in that bag at random. When you look at the ball it is black. You now pick the second ball from that same bag.What is the probability that this ball is also black?
Hint: Use Bayes' Theorem:$\mathbb{P}[A\ and\ B]=\mathbb{P}[A|B]\mathbb{P}[B]=\mathbb{P}[B|A]\mathbb{P}[A]$

这题还是比较基础的，有两个包，一个包里有两个黑球，另一个包里有一个黑球和一个白球，先随机取一个包，再随机从这个包里随机取一个球，当发现第一取出的球是黑球时，再从这个包里再取一次球，现在问第二次取出的球也是黑色的概率？利用贝叶斯公式即可，记第一次拿到黑球的事件为A，连续两次拿到黑球的事件为B，所以概率为

$$\begin{aligned} p &= \mathbb{P}(B|A) \\ &= \frac{\mathbb{P}(AB)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(B)}{\mathbb{P}(A)} \\ &= \frac{\frac{1}{2}\times\frac{1}{2}}{\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}} \\ &= \frac{2}{3} \end{aligned}$$

1.5

ABC

1.6

在保持样本均衡的前提下，正例和反例在训练和测试集中的数量相同。在正例中的采样结果为

$$C_{100}^{500}$$

，所以总的采样结果为

$$(C_{100}^{500})^2$$

1.7

混淆矩阵

	（预测值）正例	**（预测值）反例**
（真实值）正例	200	100
（真实值）反例	200	500

欠拟合

因为误差很大，如果是过拟合的话，在预测正例（真实值）时，应该值很大，且真实值为正例预测成反例的比例会很小。但是 从混淆矩阵上我们看到，真实值为正例 预测成反例的比例很大，占百分之33.3%。