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

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

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

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?

Exercise 1. 5

假设您正在使用垃圾邮件分类器，其中垃圾邮件是正例（y=1），非垃圾邮件是反例（y=0）。您有一组电子邮件训练集，其中99%的电子邮件是非垃圾邮件，另1%是垃圾邮件。以下哪项陈述是正确的？选出所有正确项

A. 一个好的分类器应该在交叉验证集上同时具有高精度precision和高召回率recall。

B. 如果您总是预测非垃圾邮件（输出y=0），那么您的分类器在训练集上的准确度accuracy将达到99%，而且它在交叉验证集上的性能可能类似。

C. 如果您总是预测非垃圾邮件（输出y=0），那么您的分类器的准确度accuracy将达到99%。

D. 如果您总是预测非垃圾邮件（输出y=0），那么您的分类器在训练集上的准确度accuracy将达到99%，但在交叉验证集上的准确率会更差，因为它过拟合训练数据。

Exercise 1. 6

数据集包含1000个样本，其中500个正例、500个反例，将其划分为包含80%样本的训练集和百分之20%样本的测试集用于留出法评估，试估算共有多少种划分方式?

Exercise 1. 7

如果有个1000个样本的数据集，其中300个正例，700个反例，有个线性分类模型对这些样本进行分类，得到正例中有200个正确分类成正例，反例中有500个被分为反例，请画出此时模型的混淆矩阵？并分析此时模型是过拟合还是欠拟合？如果是欠拟合，请说明原因。