第1章 机器学习基础

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

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 1.6

哪一个是机器学习的合理定义?

A. 机器学习从标记的数据中学习
B. 机器学习能使计算机能够在没有明确编程的情况下学习
C. 机器学习是允许机器人智能行动的领域
D. 机器学习是计算机编程的科学

Exercise 1.7

一个计算机程序从经验E中学习任务T，并用P来衡量表现。并且，T的表现P随着经验E的增加而提高。假设我们给一个学习算法输入了很多历史天气的数据，让它学会预测天气。什么是P的合理选择?

A. 天气预报任务
B. 正确预测未来日期天气的概率
C. 计算大量历史气象数据的过程
D. 以上都不

Exercise 1.8

回归问题和分类问题的区别是什么?

A. 回归问题输出值是连续的，分类问题输出值是离散的
B.回归问题输出值是离散的，分类问题输出值是连续的
C.回归问题与分类问题在输入属性值上要求不同
D.回归问题有标签，分类问题没有

Exercise 1.9

哪些机器学习模型经过训练，能够根据其行为获得的奖励和反馈做出一系列决策?

A. 无监督学习
B. 监督学习
C. 强化学习
D.以上全部

Exercise 1.10

谷歌新闻每天收集非常多的新闻，并运用( )方法再将这些新闻分组，组成若干类有关联的新闻。于是，搜索时同一组新闻事件往往隶属同一主题的，所以显示到一起。

A.分类
B.聚类
C.回归
D.关联规则

Exercise 1.11

下列哪种方法可以用来缓解过拟合的产生( )

A. 增加更多的特征
B. 正则化
C. 增加模型的复杂度
D. 以上都是

Exercise 1.12

下列的哪种方法可以用来降低深度学习模型的过拟合问题？ （）

①增加更多的数据

②使用数据扩增技术(data augmentation)

③使用归纳性更好的架构

④ 正规化数据

⑤ 降低架构的复杂度

A、1 4 5
B、1 2 3
C、1 3 4 5
D、所有项目都有用

Exercise 1.13

假如使用一个较复杂的脊回归模型 (Ridge Regression)，来拟合样本数据时，通过调整正则化参数λ，来调整模型复杂度。当λ较大时，关于偏差（bias）和方差（variance），下列说法正确的是 （ ）

A、当λ增大时，偏差减小，方差减小
B、当λ增大时，偏差减小，方差增大
C、当λ增大时，偏差增大，方差减小
D、当λ增大时，偏差增大，方差增大

Exercise 1.14

假如你用logistic Regression (逻辑回归) 算法去预测用户在网上的购买项目，然而，当你在新的用户集上验证你的假设时，你发现预测值有很大的偏差。并且你的假设在训练集上表现也很差，下面那些步骤你应该采纳，选择出正确的选项（ ）

A、尝试着减小正则项 λ
B、尝试增加交叉特征
C、减小样本量
D、尝试更小的测试集或者特征

Exercise 1.15

[**多选题**] 下列方法中，解决欠拟合的方法有哪些（）

A、正则化方法
B、集成学习方法
C、添加新特征
D、减少正则化系数

Exercise 1.16

模型参数与学习算法的超参数之间有什么区别?

Exercise 1.17

如果你的模型在训练数据上表现很好，但是应用到新的实例上的泛化结果却很糟糕，是怎么回事？能提出三种可能的解决方案吗?

Exercise 1.18

什么是交叉验证？它为什么比验证集更好？