ML Assignment 1

10185101210 陈俊潼

1.1

Express each of the following tasks in the framework of learning from data by specifying the input space , output space , target function .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

Sol.

#	Input Space	Output Space	Target Function	Specifics
a	Patient's medical history & symptoms	Patient's problem	Patient $\rightarrow$ problem	Medical history, symptoms, basic information about the patient
b	X is a matrix with picture's pixel data	$Y = \{y \vert y \in Z^* \wedge y < 10\}$	Picture $\rightarrow$ digit	Locations of each pixels for each digit
c	X is a set of emails with all the content	$Y = \{1, 0\}$	Email $\rightarrow$ Boolean	Email's content, length, keywords, etc.
d	X is the price, temperature and date data together with the electric load	$Y = \{y \vert y \in Q^*\}$	{Temperature, Price, Day of the week} $\rightarrow$ Electric load	The connections between electric load and environment data
e	$X = \{x \vert y \in R\}$	$Y = \{y \vert y \in R\}$	X $\rightarrow$ Y	The trend and features of Y when X varies.

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

Sol.

Problem A, C, E is more suitable for solving with Machine Learning. Because firstly, they are uncertain problems and have no analytic solutions, nor specified formula or calculation method to solve. Besides, they are all problems with loads of data to training, for instance, problem A can yield a ML model by learning from millions of real patient experiment data and can be used to determine future patient's best age for medical testing.

While problem B & D can be solved by classic algorithms and physical laws. Since we can simply get accurate answers by basic calculations, there's no need to train a model for them.

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

	Type	Training data
a	Supervised	Users's purchase history, view history, age, gender, etc.
b	Reinforcement	Computer's own simulated match results.
c	Unsupervised	Movie's type, tag, author, comments, etc.
d	Supervised	Existed music MIDI data and basic music theories, etc.
e	Supervised	Customer's bill history, credibility, crime records, etc.

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?

Sol.

Since if we choose the bag with two black balls, the probability is P(A | E) = 1,
P(E) = 1/2,
Since all bags are opague and the probability of picking each ball is equal, so P(A) = 3/4,
From Bayes theorem we get:

$$P(E|A) = \frac{P(A|E)P(E)}{P(A)} = \frac{2}{3}$$

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?

训练样本集的个数：1000 * 80% = 800 ，测试集的个数为 200 个。为了保证既能测试到正例也能测试到反例，所以测试集中的正例个数应当为 1 - 199 个，剩余的所有样本都作为训练集进行训练。所以合理的划分方式一共有 199 种。