Pattern Recognition and Machine Learning: Homework 1, Zhengzuo Liu

Problem 1

Answer:

- T_1: Playing Go game – E_1: Total Go games played – P_1: Winning rate
- T_2: Making medical decisions by CT images – E_2: Total CT images seen through training – P_2: Decision accuracy
- T_3: Controlling a robot walk – E_3: Total time/distance the robot has walked through training – P_3: Average walking distance without falling/ Falling rate within certain distance/ Fastest walking speed within a desired falling rate
- T_4: Autonomous driving – E_4: Total miles (a rough estimation of circumstances encountered) cars driven using the autopilot engine through training – P_4: Decision accuracy when encountering certain upcoming circumstances on the road
- T_5: Generate realistic images – E_5: Total number of realistic images (with a description) that a neural network have seen – P_5: Rating (with a fixed standard) of new realistic images generated according to a certain description
- T_6: ChatGPT chatting with human – E_6: Total dialogues ChatGPT made with human through training – P_6: Average score/satisfaction rate given by ChatGPT users

Problem 2

Answer:

2.1

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} = \frac{999}{999+1} = 99.9\%$$

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} = \frac{980}{980+20} = 98.0\%$$

2.2

Let A represent "Alice actually got a cancer", B represent "Alice tests positive using this method". Then

$$P(A) = 0.1\%$$

$$P(B|A) = Sensitivity = 99.9\%$$

$$P(B) = P(A)P(B|A) + P(\overline{A})P(B|\overline{A}) = 0.1\% \times 99.9\% + (1 - 0.1\%) \times (1 - 98.0\%) = 2.0979\%$$

Therefore

$$P(A|B) = P(A)P(B|A)/P(B) = 4.76\%$$

i.e. the real probability that Alice actually got a cancer is 4.76%.

2.3

The reason why Alice's real chance of getting cancer is low, is that the false positive (FP) rate of the test method is as high as 2.0%, 20 times the 0.1% cancer ration. The high FP rate lowers the real possibility of Alice getting cancer.

Let β denote specificity in this experiment, then $\beta = 98\%$. "The chance of Alice really got a cancer" is equivalent to

$$\frac{p\alpha}{p\alpha + (1-p)(1-\beta)} > 99\%$$

the solution gives $\alpha > 99(1-\beta)\frac{(1-p)}{p}$. For this test method with $(\alpha, \beta) = (99.9\%, 98.0\%)$, only when $p > 66.5\%$, meaning the probability for getting a cancer is larger than 66.5%, would the chance of Alice really got a cancer after the positive test result be more than 99%.

Problem 3

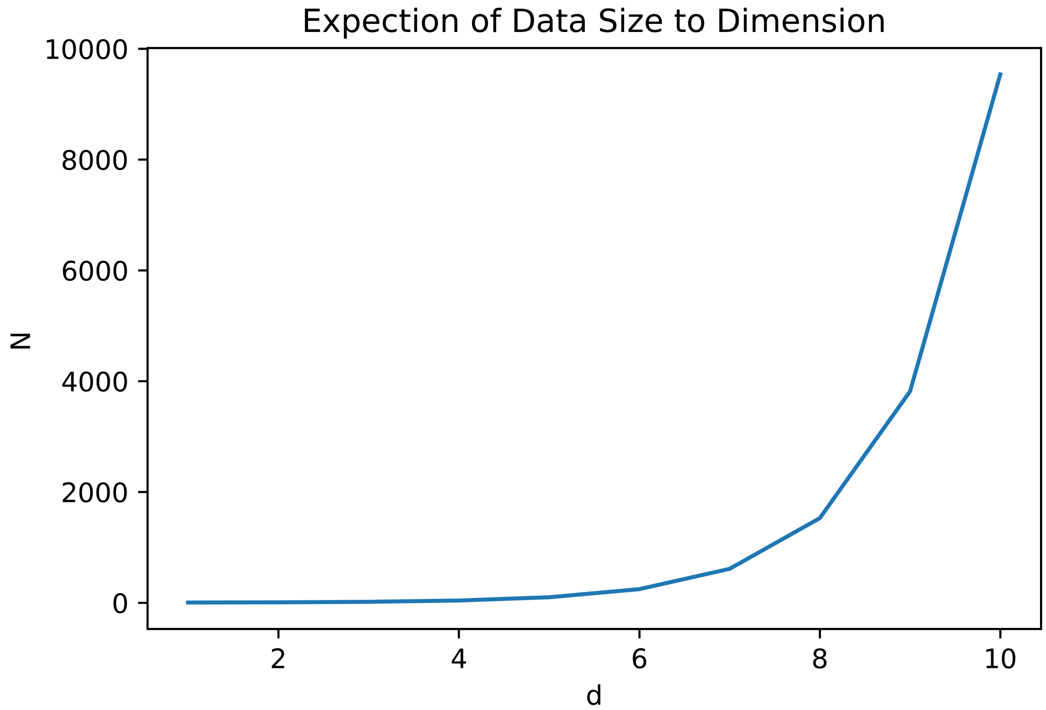
3.1

Let P(d) denote the possibility required in this problem when dimension is d. Then $P(d) = \frac{V_d(0.2)}{V_d(0.5)} = (\frac{2}{5})^d$. The numeric answers are as follows:

d	P(d)
2	0.16
5	0.01024
10	0.0001048576

3.2

It is easy to get that $E(N) = \frac{1}{p} = 2.5^d$. The figure is as follow:



When $d = 100$, $log(N) = 100 \times log(2.5) = 91.6$. It tells that with the growth of number of dimensions, the data points needed for a reliable k-nearest neighbor approach increases exponentially. When dealing with a feature space of a very high dimension, the data size needed for using k-nearest neighbor may become overwhelming.