Pattern Recognition: Homework 1

Due date: 2023.2.28

Problem 1

In the class we have learned many examples of machine learning and pattern recognition. Actually, a more formal and rigorous definition for machine learning [1] goes like

Definition: A machine learning computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

For example, if task T is image classification, then E would be the total number of images that a neural network has seen through training, and P would be classification accuracy. Specify the corresponding E and P for these tasks T. Note that both E and P metric should be **specific and quantitatively measurable**. (Answer maybe not unique)

- T_1: Playing Go game.
- T_2: Making medical decisions by CT images.
- T_3: Controlling a robot walk.
- T_4: Autonomous driving.
- T_5: Generate realistic images.
- T_6: ChatGPT chatting with human.

Problem 2

Suppose a medical company invents a new method for diagnosing the cancer. It is claimed that their method only required several drops of blood. They test 2000 people in the hospitals, of which half people have cancer and the rest don't. Among the people who have the cancer, 999 are tested positive and 1 negative. For the people who do not have cancer, 980 are tested negative and 20 are positive.

2.1 (10 pt)

Calculate the sensitivity and specificity of this method.

2.2 (10 pt)

The population ratio of getting a cancer is around 0.1%, treat it as prior. Alice use this method and test positive. Calculate the real probability that Alice actually got a cancer. (**Hint**: use Bayesian equation $P(A|B) = P(A) \cdot P(B|A)/P(B)$.)

2.3 (Bonus, 5 pt)

Explain the seemingly paradoxical phenomenon. (**Hint:** Denote the precision as α and the prior probability for getting a cancer in population as p. Find the condition requirement for α such that the chance of Alice really got a cancer is more than 99%.)

错误率要比人群中患病的概率再低一点

Problem 3

In this problem we will experience the curse of dimensionality in another way.

An algorithm uses k-nearest neighbor to learn pattern in space $\mathbb{R}^d$. The dataset $\mathcal{D}$ has N data points, each point x is sampled from distribution $\mu = U[0,1]^d$, which means each coordinate is uniformly and independently sampled within interval $[0,1]$.

To get an accurate answer, suppose the result is only reliable when the nearest neighbor $\hat{x} \in \mathcal{D}$ found in dataset is close enough to the query data sample x, which means $\|x - \hat{x}\|_2 \leq \epsilon$. In this problem, we set $\epsilon = 0.2$.

3.1 (10 pt)

For $d = 2, 5, 10$, calculate the probability of sampling a data $x \sim \mu$ that successfully becomes a reliable neighbor for query point $x_c = (0.5, 0.5, \ldots, 0.5)^\top \in \mathbb{R}^d$. **hint**: You just need to give numeric answer for this question. The volume of a d-dimensional sphere with radius r is $V_d(r) = \frac{(\sqrt{\pi}r)^d}{\Gamma(d/2+1)}$, where $\Gamma(n) = \int_0^\infty e^{-t} \cdot t^{n-1} \mathrm{d}t$ is the Gamma function.

3.2 (20 pt)

Suppose we keep sampling data points to create a dataset until we get a reliable neighbor for x_c in $\mathbb{R}^d$. Compute the expectation of dataset size N with respect to d. Plot a figure of N against d for $d = 1, \ldots, 10$. Estimate the $\log N$ for $d = 100$. What does it tell you?

References

[1] Tom Michael Mitchell et al. *Machine learning*, volume 1. McGraw-hill New York, 2007.