BIM3008-Assignment1

Junyang Deng (120090791)

October 8, 2022

Question 1

Please conduct a literature survey on the experimental principles and the performance of the Nucleic Acids Test or Antigen Test for COVID-19 SARS-CoV-2 infections. Briefly describe the principles and the predictive model performance using some measures, including precision, sensitivity, and specificity.

Answer

The principle of Nucleic Acids Test: The test is based on a real-time fluorescent PCR platform. With RNA transcription, polymerase chain reaction and TaqMan probe technology to target the highly conserved regions of the ORF1ab gene and N gene of SARS-CoV-2. When both genes can be detected in the sample with cycle threshold (Ct) value ≥ 35, the sample will be labeled positive.

According to Jarrom et al. (2022), the sensitivity of a reverse-transcriptase PCR test is estimated to be 87.8%. The specificity ranges from 95% to 100%. However, the sensitivity is subject to the way of collecting samples, for example, oropharyngeal swab is less reliable than nasal swab.

Reference:

- Jarrom D, Elston L, Washington J, et al. (2022). Effectiveness of tests to detect the presence of SARS-CoV-2 virus, and antibodies to SARS-CoV-2, to inform COVID-19 diagnosis: a rapid systematic review. BMJ Evidence-Based Medicine 2022;27:33-45.

Question 2

Let us say we are given the task of building an automated taxi. Define the constraints. What are the inputs? What is the output? How can we communicate with the passenger? Do we need to communicate with the other automated taxis, that is, do we need a "language"? (Question 4 of Chapter One #41)

Answer

Constraints are:

1. taxis should run on roads,
2. taxis should not hit other cars and pedestrians,
3. taxis should obey the traffic rules (i.e. stop when traffic lights turn red).

Inputs are:

1. the traffic conditions of nearby region,
2. the location of surrounding objects, like pedestrians, cars, signs.

Outputs are: Direction and speed of taxis.

Ways to communicate with the passanger: buttons, video and audio etc.

Question 3

Problem 6 of Chapter 3

Somebody tosses a fair coin and if the result is heads, you get nothing; otherwise, you get \$5. How much would you pay to play this game? What if the win is \$500 instead of \$5?

Answer

There are two outcomes of tossing a fair coin, with each of them have the probability of 1/2. We can draw the probability table:

Events	Head	Tail
Money (X)	X=5	X=0
P	1/2	1/2

The expectation is $\mathbf{E}(x) = 5 \times 1/2 + 0 \times 1/2 = 5/2$. This means I won't pay more than 2.5\$ to play this game.
If the win is 500\$ instead of 5\$, the expectation is $\mathbf{E}(x) = 500 \times 1/2 + 0 \times 1/2 = 250$. However, I won't take the risk of paying 250 dollars for this game since the variance is very high.

Problem 11 of Chapter 3

Show example transaction data where for the rule X $\rightarrow$ Y:
(a) Both support and confidence are high;
(b) Support is high and confidence is low;
(c) Support is low and confidence is high;
(d) Both support and confidence are low.

Answer

(a) Pencil and rubber: many people buy pencils and rubbers together, so the support is high; Given that a person buys pencils, he may also buy rubbers as well, so the confidence is high too.
(b) Tissue and fruit: many people will buy tissue and fruit together, since they are all neccessies, so the support is high. However, when a person buys tissue, he may or may not buys fruit, so the confidence is low.
(c) A book about toxiology and a book about pharmacology: Not many people buy these two books, so the support is low. But a person who buys a toxiology book (who may be a medical student), he or she is likely to buy a pharmacology book, so the confidence is high.
(d) Sweather and fan: Almost no one will buy sweathers and fans together, so the support is low. When a person buys a sweather, he or she is very not likely to buy a fan, so the confidence is low.

Question 4

In a two-class, two-action problem, if the lost function is $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = 8$, and $\lambda_{21} = 4$, write the optimal decision rule. How does the rule change if we add a third action of reject with $\lambda = 1$?

Answer

In order to reach the optimal decision rule, we first calculate the expected risks of two actions.

$$\begin{aligned}
R(\alpha_1|x) &= \lambda_{11} \cdot P(C_1|x) + \lambda_{12} \cdot P(C_2|x) \\
&= 0 \cdot P(C_1|x) + 8 \cdot P(C_2|x) \\
&= 8 \cdot P(C_2|x) \\
&= 8 \cdot (1 - P(C_1|x))
\end{aligned}$$

$$
\begin{aligned}
R(\alpha_2|x) &= \lambda_{21} \cdot P(C_1|x) + \lambda_{22} \cdot P(C_2|x) \\
&= 4 \cdot P(C_1|x) + 0 \cdot P(C_2|x) \\
&= 4 \cdot P(C_1|x)
\end{aligned}
$$

In this case, we choose α_1 if

$$R(\alpha_1|x) < R(\alpha_2|x) \Rightarrow P(C_1|x) > 2/3$$

Choose α_2 if

$$R(\alpha_1|x) > R(\alpha_2|x) \Rightarrow P(C_1|x) < 2/3$$

In conclusion,

$$\begin{cases} P(C_1|x) > 2/3 & \text{choose } \alpha_1 \\ P(C_1|x) < 2/3 & \text{choose } \alpha_2 \end{cases}$$

When a third action of reject with $\lambda = 1$ is added, we choose α_1 if

$$R(\alpha_1) < \lambda \Rightarrow P(C_1|x) > 1 - \frac{\lambda}{8}$$

$$P(C_1|x) > \frac{7}{8}$$

we choose α_2 if

$$R(\alpha_2) < \lambda \Rightarrow P(C_1|x) < \frac{\lambda}{4}$$

$$P(C_1|x) < \frac{1}{4}$$

In conclusion, with $\lambda = 1$,

$$\begin{cases} P(C_1|x) < \frac{1}{4}, & \text{choose } \alpha_2 \\ \frac{1}{4} < P(C_1|x) < \frac{7}{8}, & \text{reject} \\ P(C_1|x) > \frac{7}{8}, & \text{choose } \alpha_1 \end{cases}$$

Question 5

Provide three examples of machine learning applications to biological or biomedical data sets. Citations and brief descriptions of the references are required in the report.

Answer

Machine learning has been widely used in biological studies for prediction and discovery.

1. Machine learning can be used to predict DNA accessibility based on DNA sequence data and histone modification information. For example, Kelly et al. (2016) build an open-source package Basset which applies CNNs to learn the functional activity of DNA sequences from genomics data.

2. Machine learning can be used to predict urinary stones in CT scans. For example, Babajide et al. (2022) construct a neural network for the identification and measurement of urinary stones on NCCT images. The model reaches 100% specificity and 100% sensitivity.

3. Machine learning approaches have also been used to predict breast cancer therapy response. Sammut et al. (2021) collected clinical, digital pathology, genomic and transcriptomic profiles of pre-treatment biopsies of breast tumours from 168 patients treated with chemotherapy with or without HER2 (encoded by ERBB2)-targeted therapy before surgery. Based on these data, their model has excellent performance with AUC equals 0.87.

References:

- Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research, 26(7), 990–999. https://doi.org/10.1101/gr.2005/35.115
- Babajide, R., Lembrikova, K., Ziemba, J., Ding, J., Li, Y., Fermin, A. S., ... & Tasian, G. E. (2022). Automated Machine Learning Segmentation and Measurement of Urinary Stones on CT Scan. Urology.
- Sammut, SJ., Crispin-Ortuzar, M., Chin, SF. et al. (2022). Multi-omic machine learning predictor of breast cancer therapy response. Nature 601, 623–629. https://doi.org/10.1038/s41586-021-04278-5