CSE 5243 Autumn'22 Assignment #1 – Intr Data Mining

Neng Shi

2022/09/18

1 Problem 1.

(1) The probability that I win is $.1 + .2 + .3 = .6$.

So it is better than a fair die.

(2)

$$\begin{aligned} E[\delta(X=8)] &= p_2 \times \delta(X=2) + p_4 \times \delta(X=4) + p_8 \times \delta(X=8) \\ &= p_8 \end{aligned}$$

(3) We are going to maximum $L = p^k \times (1-p)^{n-k}$.

$log(L) = k \times log(p) + (n-k) \times log(1-p)$.

And

$$\begin{aligned} &\frac{\partial log(L)}{\partial p} \\ =&\frac{k}{p} - \frac{n-k}{1-p} \end{aligned}$$

Let the partial derivative equals to 0, and we can get that p making the outcome most likely is $\frac{k}{n}$.

2 Problem 2.

(a) $X \in \{+, -\}$: the outcome of swine flu test.

$C \in \{Y, N\}$: the patient has swine flu or not.

We want to know $P(C = Y|X = +)$.

Apply Bayes rule:

$$
\begin{aligned}
&P(C=Y|X=+)\\
=&\frac{P(X=+|C=Y)P(C=Y)}{P(X=+)}\\
=&\frac{P(X=+|C=Y)P(C=Y)}{P(X=+|C=Y)P(C=Y)+P(X=+|C=N)P(C=N)}\\
=&\frac{0.97\times 0.0003}{0.97\times 0.0003+0.02\times 0.9997}\\
=&1.43\%
\end{aligned}
$$

(b) Now

$$
\begin{aligned}
&P(C=Y|X=+)\\
=&\frac{P(X=+|C=Y)P(C=Y)}{P(X=+)}\\
=&\frac{P(X=+|C=Y)P(C=Y)}{P(X=+|C=Y)P(C=Y)+P(X=+|C=N)P(C=N)}\\
=&\frac{0.97\times 0.01}{0.97\times 0.01+0.02\times 0.99}\\
=&32.9\%
\end{aligned}
$$

3 Problem 3.

1. $X \in \{+,-\}$: the outcome of the flu test.
   $C \in \{Y,N\}$: the patient has the flu or not.
   We want to know $P(C=Y|X=+)$.
   Apply Bayes rule:

$$
\begin{aligned}
&P(C=Y|X=+)\\
=&\frac{P(X=+|C=Y)P(C=Y)}{P(X=+)}\\
=&\frac{P(X=+|C=Y)P(C=Y)}{P(X=+|C=Y)P(C=Y)+P(X=+|C=N)P(C=N)}\\
=&\frac{0.97\times 0.003}{0.97\times 0.003+0.01\times 0.997}\\
=&22.6\%
\end{aligned}
$$

2. $X_1 \in \{+, -\}$: the outcome of the first flu test.
$X_2 \in \{+, -\}$: the outcome of the second flu test.
We want to know $P(C = Y|X_1 = X_2 = +)$.
Apply Bayes rule:

$$
\begin{aligned}
& P(C = Y|X_1 = X_2 = +) \\
=& \frac{P(X_1 = X_2 = +|C = Y)P(C = Y)}{P(X_1 = X_2 = +)} \\
=& \frac{P(X_1 = X_2 = +|C = Y)P(C = Y)}{P(X_1 = X_2 = +|C = Y)P(C = Y) + P(X_1 = X_2 = +|C = N)P(C = N)} \\
=& \frac{0.97^2 \times 0.003}{0.97^2 \times 0.003 + 0.01^2 \times 0.997} \\
=& 96.6\%
\end{aligned}
$$

3. We want to know $P(C = Y|X_1 = +, X_2 = -)$.
Apply Bayes rule:

$$
\begin{aligned}
& P(C = Y|X_1 = +, X_2 = -) \\
=& \frac{P(X_1 = +, X_2 = -|C = Y)P(C = Y)}{P(X_1 = +, X_2 = -)} \\
=& \frac{P(X_1 = +, X_2 = -|C = Y)P(C = Y)}{P(X_1 = +, X_2 = -|C = Y)P(C = Y) + P(X_1 = +, X_2 = -|C = N)P(C = N)} \\
=& \frac{0.97 \times 0.03 \times 0.003}{0.97 \times 0.03 \times 0.003 + 0.01 \times 0.99 \times 0.997} \\
=& 0.88\%
\end{aligned}
$$

4 Problem 4.

1. Mean: $\frac{69+74+68+70+72+67+66+70+76+68+72+79+74+67+66+71+74+75+75+76}{20} = 71.45$

Median: since the samples after ranking is 66, 66, 67, 67, 68, 68, 69, 70, 70, 71, 72, 72, 74, 74, 74, 75, 75, 76, 76, 79, (71 + 72) / 2 = 71.5

Mode: 74 is the mode since it appears three times in the list, which is the most.

2. $\mu = 71.45$ and $\sigma = 3.72$

The probability density function is:

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$
$$=\frac{1}{3.72\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-71.45}{3.72})^2}$$

5 Problem 5.

1. For example, Euclidean distance between x and x_1 is $\sqrt{(1.4-1.5)^2+(1.6-1.7)^2}$. Euclidean distance is 0.14, 0.67, 0.28, 0.22, 0.60 for x_1, x_2, x_3, x_4, x_5, so the ranking is x_1, x_4, x_3, x_5, x_2.

For example, Manhattan distance between x and x_1 is $|1.4-1.5|+|1.6-1.7|$. Manhattan distance is 0.2, 0.9, 0.4, 0.3, 0.7 for x_1, x_2, x_3, x_4, x_5, so the ranking is x_1, x_4, x_3, x_5, x_2.

For example, Jaccard similarity between x and x_1 is $\frac{1.4\times 1.5+1.6\times 1.7}{1.4^2+1.6^2+1.5^2+1.7^2-(1.4\times 1.5+1.6\times 1.7)}$. Jaccard similarity is 0.996, 0.928, 0.985, 0.988, 0.909 for x_1, x_2, x_3, x_4, x_5, so the ranking is x_1, x_4, x_3, x_2, x_5.

For example, cosine similarity between x and x_1 is $\frac{1.4\times 1.5+1.6\times 1.7}{\sqrt{(1.4^2+1.6^2)(1.5^2+1.7^2)}}$. Cosine similarity is 0.99999, 0.99575, 0.99996, 0.99990, 0.96536 for x_1, x_2, x_3, x_4, x_5, so the ranking is x_1, x_3, x_4, x_2, x_5.

2. For example, to normalize x to make its Euclidean norm equal to 1, $x' = (\frac{1.4}{\sqrt{1.4^2+1.6^2}}, \frac{1.6}{\sqrt{1.4^2+1.6^2}})$.

$x' = (0.659, 0.753)$
$x'_1 = (0.662, 0.750)$
$x'_2 = (0.725, 0.689)$
$x'_3 = (0.664, 0.747)$
$x'_4 = (0.625, 0.781)$
$x'_5 = (0.832, 0.555)$

Euclidean distance is 0.004, 0.092, 0.008, 0.044, 0.263 for $x'_1, x'_2, x'_3, x'_4, x'_5$, so the ranking is $x'_1, x'_3, x'_4, x'_2, x'_5$.