EN.601.448/648 Computational genomics: Problem set 0

JitongCai (jcai14)

Instructions

We have provided this LaTeX document for turning in Problem set 0. We give you one or more boxes to answer each question. The question to answer for each box will be noted in the title of the box.

Other than your name, do not type anything outside the boxes. Leave the rest of the document unchanged.

For written answers, replace the **\TextRequired** (**Place Answer Here**) command with your answer. For the following example *Question 0.1* , you would place your answer where **\TextRequired** (**Place Answer Here**) is located,

Question 0.1

Place Answer Here

Do not change the height or title of the box. If your text goes beyond the box boundary, it will be cut off. We have given sufficient space for each answer, so please condense your answer if it overflows. The height of the box is an upper bound on the amount of text required to answer the question - many answers can be answered in a fraction of the space. Do not add text outside of the boxes. We will not read it.

For True/False or Multiple Choice questions, place your answers within the defined table. To mark the box(es) corresponding to your answers, replace **\Unchecked** (☐) commands with the **\Checked** (☒) command. Do not make any other changes to the table. For example, in *Question 0.2* ,

Question 0.2

☐	Logistic Regression
☒	Perceptron

For answers that require a single equation, we will provide a specific type of box, such as in the following example *Question 0.3* . Please type the equation where **\EquationRequired** (**Type Equation Here**) without adding any $ signs or **\equation** commands. Do not put any additional text in this field.

Question 0.3

$\mathbf{w} =$ **Type Equation Here**

For answers that require multiple equations, such as a derivation, place all equations within the specified box. You may include text short explanations if you wish (as shown in *Question 0.4*). You can put the equations in any format you like (e.g. within $ or $$, the **\equation** environment, the **\align** environment) as long as they stay within the box.

Question 0.4

$x + 2$ x is a real number

the following equation uses the variable y

$y + 3$

Do not change any formatting in this document, or we may be unable to grade your work. This includes but is not limited to the height of textboxes, font sizes, and the spacing of text and tables. Additionally, do not add text outside of the answer boxes. Entering your answers are the only changes allowed.
We strongly recommend you review your answers in the generated PDF to ensure they appear correct. We will grade what appears in the answer boxes in the submitted PDF, NOT the original latex file.

1. Expected value and likelihood (1 point)

Question 1.1

$P(30 \text{ tosses all yield heads}) = 2.211 \times 10^{-7}$

Question 1.2

$\mathbb{E}(\text{One toss}) = 0.4$

Question 1.3

$\mathbb{E}(\text{sum of 4 coin tosses}) = 1.6$

Question 1.4 (a)

Given $\theta = 0.5$, Likelihood $= 9.766 \times 10^{-4}$

Question 1.4 (b)

Given $\theta = 0.8$, Likelihood $= 1.678 \times 10^{-3}$

Question 1.4 (c) Which model do you prefer

- ☐ $\theta = 0.5$
- ☒ $\theta = 0.8$

Question 1.4 (c) Justification

The likelihood of getting the sequence given $\theta = 0.8$ is larger than given $\theta = 0.5$. $\theta = 0.8$ gives the result more support.

Question 1.5 (d) Is there better setting?

Yes. $\theta = 0.7$
$L(\theta) = \theta^7(1-\theta)^3$
$ln(L(\theta)) = 7ln(\theta) + 3ln(1-\theta)$
$\frac{dln(L(\theta))}{d\theta} = \frac{7}{\theta} - \frac{3}{1-\theta} = \frac{7-10\theta}{\theta(1-\theta)} = 0$
$\theta = 0.7$
$\frac{d^2ln(L(\theta))}{d\theta^2} = \frac{-7}{\theta^2} - \frac{1}{(1-\theta)^2} < 0$
So, when $\theta = 0.7$, $ln(L(\theta))$ will reach the maximum value.

2. Probability and independence (0.5 points)

Question 2.1 Prove

$P(A|B) > P(A)$
$SinceP(B) > 0, P(A) > 0$
$P(A|B)P(B) > P(A)P(B)$
$P(A, B) > P(A)P(B)$
$P(B|A)P(A) > P(A)P(B)$
$P(B|A) > P(B)$

Question 2.2 (a.1) are these two events independent

☐ Yes

☒ No

Question 2.2 (a.2) Justification

$P(Steel) = \frac{10+5}{25} = \frac{15}{25}$ $P(Wooden) = \frac{8+2}{25} = \frac{10}{25}$
$P(Red) = \frac{8+5}{25} = \frac{13}{25}$ $P(Wooden) = \frac{10+2}{25} = \frac{12}{25}$

$Expected$	$Steel$	$Wooden$
Red	7.8	5.2
$White$	7.2	4.8

$chisq - stat = \sum_1^4 \frac{(Observation-Expect)^2}{Expect} = 5.23$
$pvalue = 0.02 < 0.05$

Question 2.2 (b)

$P = \frac{5}{13}$

Question 2.2 (c). which is more likely

☐ Steel

☒ Wooden

3. Bayes' Theorem (0.5 points)

Question 3.1 Probability of actually have the disease (write out the derivation)

Assume: Event D means getting disease. Event P means the test result is positive.

$$\begin{aligned}
¶(D|P) &= \frac{P(P|D)P(D)}{P(P|D)P(D) + P(P|D^c)P(D^c)} \\
&= \frac{0.96 \times \frac{1}{250{,}000}}{0.96 \times \frac{1}{250{,}000} + 0.005 \times (1 - \frac{1}{250{,}000})} \\
&= 7.674 \times 10^{-4}
\end{aligned}$$

4. Gaussian data and likelihood

Question 4.1

$\mathbb{E}[\sum_{i=1}^{10} x_i] = 0$

Question 4.2

$E[\sum_{i=1}^{10} x_i^2] = 10$

Question 4.3

Actual ten values (3 signif digits): [-0.15061494, -0.90395363, 0.05517485, 1.89915051, -1.5110287,0.1230296 , -0.89229144, -1.51606782, -0.11841514, -0.31234444]
Sum = -3.327361140920997
Sum of squares = 9.954204778387222

Question 4.4

$p(x_1)$ =0.44013973865784517
$p(x_2)$ =0.18300999421203262
$p(x_3)$ =0.5220004182832698

Question 4.5

Likelihood$(x_1, x_2, ...x_{10})$ =$3.019330746945851 \times 10^{-6}$

Question 4.6

$\log$ Likelihood$(x_1, x_2, ...x_{10})$ =-12.710475358107521

Question 4.6 Why might people prefer to work in log space?

The log will turn multiplication into sum, so it's easier to deal with. In addition, the log scale can turn the value which is close to zero to a very small negative value, which will be in a larger scale, and it's easier for people to find the difference between the likelihood.

5. Linear algebra (0.5 points)

Question 5.1 A is orthonormal

$A^T A$ is I

Question 5.2 (a) If A is invertible write the inverse of A

$A^{-1} =$

$$\begin{bmatrix} 0.3333 & 0.0000 & 0.3333 \\ -0.1667 & 0.5000 & -0.1667 \\ -0.1667 & -0.5000 & 1.6667 \end{bmatrix}$$

Question 5.2 (a) If A is not invertible write the justification

A is invertible.

Question 5.2 (b) If A is invertible write the inverse of A

$A^{-1} =$

$$\begin{bmatrix} \textbf{Place Answer Here} & \textbf{Place Answer Here} & \textbf{Place Answer Here} \\ \textbf{Place Answer Here} & \textbf{Place Answer Here} & \textbf{Place Answer Here} \\ \textbf{Place Answer Here} & \textbf{Place Answer Here} & \textbf{Place Answer Here} \end{bmatrix}$$

Question 5.2 (b) If A is not invertible write the justification

A is not invertible, because $|A| = 0$.

Question 5.3 Implement the function in Python

6. Matrix derivatives. (1 point)

Question 6.1 Matrix derivative (write out the derivation)

$$\frac{dQ}{d\beta} = \frac{d(Y - X\beta)^T(Y - X\beta)}{d\beta} \tag{1}$$

$$= \frac{d(Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta)}{d\beta} \tag{2}$$

$$= 2\beta^TX^TX - 2Y^TX \tag{3}$$

$$= 2(\beta^TX^T - Y^T)X \tag{4}$$

Question 6.2 Optimal β

$\beta = (\mathrm{X}^TX)^{-1}X^TY$

Question 6.3 Unique solutions in linear regression

If we can get a unique solution, there must be $N \geq K$, but if $N \geq K$, we don't necessarily get a unique solution.
We want the equation $X^T X\beta = X^T Y$ only have a unique solution.
The equation only has one unique solution indicates that $rank(X^T X) = rank(X^T X|X^T Y) = K$. Since $rank(X^T X) \leq min(N, K)$ so N must be larger than K, aka N $\geq K$.