DDA2001: Introduction to Data Science

Final Exam
SDS, CUHK(SZ)

May 15, 2024

Instructions

1. The time duration is **120 minutes**.
2. The total mark is **100 points**.
3. There are **7 questions** in total, please check.
4. Please write down your calculations and statements **clearly and in detail**.
5. Give answers up to **2 decimal places** where appropriate.
6. You should answer the questions in the answer sheet in **sequential order**.
7. Please submit the test paper and answer sheet **together** at the end.

Declare: I have read the instructions carefully. I will fully comply with the exam rules and instructions.

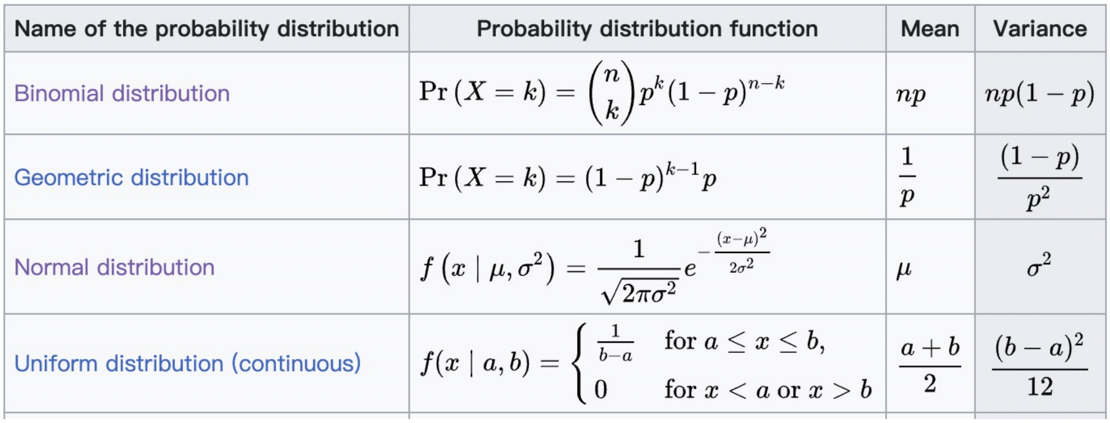
Name: ________________________ Student ID: ______________________

Useful Formulas.

- Let X be a random variable. Define $f(y) = \mathbb{E}[\min(y, X)]$. Then

$$f'(y) = \mathbb{P}(X \geq y).$$

- Mean and variance for common distributions



Name of the probability distribution	Probability distribution function	Mean	Variance
Binomial distribution	$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Geometric distribution	$\Pr(X = k) = (1-p)^{k-1} p$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$
Normal distribution	$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Uniform distribution (continuous)	$f(x \mid a, b) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

1. **(9 points) KNN**

   (a) **(3 points)** Explain the process of classifying a new data point using KNN.

   (b) **(3 points)** Describe how the choice of K affects the performance of the KNN algorithm.

   (c) **(3 points)** Give one solution to determine a proper value of K.

Solution

(a) As long as the provided answer covers the following steps, we can give the full credit. Calculate Distances: Compute the distance between the new data point and all other data points in the dataset using a chosen distance metric, such as Euclidean distance. Select Nearest Neighbors: Identify the K nearest data points (neighbors) to the new data point based on the computed distances. Majority Voting and assigning class labels: Determine the class labels of the K nearest neighbors.

(b) Small K: Smaller values of K can lead to overfitting. The model might be too sensitive to noise and outliers in the data, resulting in poor generalization to unseen data. Large K: Larger values of K result in smoother decision boundaries and may lead to underfitting. The model might oversimplify the classification and fail to capture the complexities of the data, resulting in lower training and testing accuracy. (give the credit as long as they mention overfitting and underfitting).

(c) Cross-Validation: and give the credit if the students mention cross-Validation and can roughly describe the main procedure.

2. **(14 points) Logistic Regression**

(a) **(2 points)** Consider a binary classification problem. Explain the concept of logistic regression in supervised learning, in other words, what does logistic regression try to model?

(b) Let y_i denote the class label (0 or 1) for the i-th data point $x_i \in \mathbb{R}^3$ ($x_i = (x_{i1}, x_{i2}, x_{i3})$), and $\theta \in \mathbb{R}^3$ ($\theta = (\theta_1, \theta_2, \theta_3)$) is model parameters. In other words,

$$\mathbb{P}(y_i = 1 \mid x_i, \theta) = \frac{1}{1 + e^{-\sum_{j=1}^{3} \theta_j x_{ij}}}. \tag{1}$$

i. **(4 points)** What is the log-likelihood function for n data points, i.e., (x_i, y_i), $i = 1, \ldots, n$. Explain in detail.

ii. **(4 points)** Show whether the log-likelihood function, denoted as $\ell(\theta)$, is a convex or concave function with respect to θ. Explain in detail.

(c) **(4 points)** Describe how to estimate the optimal value of θ^* given (x_i, y_i), $i = 1, \ldots, n$, by maximizing the log likelihood function $\ell(\theta)$. Please briefly provide the algorithm, including initialization, updating formula, and stopping rules. Note that, in your answer, you don't need to explicitly compute $\frac{\partial \ell(\theta)}{\partial \theta}$ using the mathematical formula of $\ell(\theta)$. When you want to indicate the gradient, please directly use $\frac{\partial \ell(\theta)}{\partial \theta}$. Also, let's assume that the step size is a constant, denoted as η.

Solution

(a) We try to model the conditional probability $P(y = 1 \mid x, \theta)$, which represents the probability that the output y is 1 given the input feature x and the model parameters θ. This probability is modeled using the sigmoid function, which maps any real-valued input to the range $(0, 1)$ and transforms the linear combination of input features and parameters into a probability.

(b) The log-likelihood is

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \left[y_i \log P\left(y = 1 \mid x_i, \theta\right) + \left(1 - y_i\right) \log\left(1 - P\left(y = 1 \mid x_i, \theta\right)\right) \right]$$

As for the proof. The likelihood for one point (x_i, y_i) is

$$P\left(y = 1 \mid x_i, \theta\right)_i^y \left(1 - P\left(y = 1 \mid x_i, \theta\right)\right)^{1 - y_i}$$

Since all points are independent, the joint likelihood is

$$\prod_{i=1}^{n} P\left(y=1 \mid x_i, \theta\right)_i^y \left(1-P\left(y=1 \mid x_i, \theta\right)\right)^{1-y_i}$$

Then, we can take the log and get the above expression.

(c) Proof: $\log\left(1+\exp\left(-\theta^\top x\right)\right)$ is convex in θ. Consider a point θ_0 and a direction vector e

$$\begin{aligned} h(t) &= \log(1+\exp(-(\theta_0+te)^\top x)) \qquad (2) \\ &= \log(1+\exp(-(\theta_0^\top x + te^\top x)) \qquad (3) \end{aligned}$$

Let's denote $C_1 = -\theta_0^\top x$ and $C_2 = -e^\top x$. Therefore, we have

$$h(t) = \log(1+\exp(C_1+tC_2)) \tag{4}$$

We can show

$$h^{''}(t) = \frac{C_2^2}{(1+\exp(C_1+tC_2))^2}\exp(C_1+tC_2) \geq 0 \tag{5}$$

Obviously, $h^{''}(0) \geq 0$.

(d) Gradient descent (it's ok to answer such as stochastic gradient descent or Newton's method...)

- Start with an initial point $\theta^{(0)}$
- For each iteration, update by

$$\theta^{t+1} = \theta^t + \eta \left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=\theta^t}$$

- Stop the algorithm when $|\theta^{t+1}-\theta^t| \leq \epsilon$ or $\left|\left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=\theta^t}\right| < \epsilon$ where ϵ is some tolerance and is a very small number. (It is ok to say something like two consecutive θ is very close to each other; or the norm of the gradient is close to 0.) Either of the two conditions is ok.

3. **(10 points) Gradient Descent Method**
Let's maximize the function $f(x) = -x^2/2 + x$ through the gradient descent method. We fix the learning rate in the algorithm as a constant, α. Assume that the initial value of x is $x^{(0)} = 2$ and $x^{(t)}$ is the value of x after the t-th iteration of gradient descent. The stoping rule is $|x^{(t+1)} - x^{(t)}| \leq \epsilon$, where ϵ is a non-negative constant.

(a) Let $\epsilon = 1/1000$. In each of the following situations will the final output be the optimal solution? Explain the reasons.

i. **(1 point)** $\alpha = 0$.

ii. **(1 point)** $\alpha = 1$.

iii. **(1 point)** $\alpha = 2$.

iv. **(2 points)** $\alpha = 1/2$.

(b) **(5 points)** Let $\epsilon = 0$. For what range of α will $\lim_{t\to\infty} x^{(t)}$ converge to the optimal solution? Explain the reasons.

Solution

(a) The gradient of f is $-x+1$. Given a constant learning rate, since we are maximizing the function, the updating rule is as follows:

$$x^{(t)} = x^{(t-1)} + \alpha(1 - x^{(t-1)}).$$

Thus

$$x^{(t)} - 1 = (1-\alpha)(x^{(t-1)} - 1) = (1-\alpha)^t(x^{(0)} - 1) = (1-\alpha)^t.$$

We also notice that by the optimality condition of the concave function, the optimal solution shall be 1.

i. If $\alpha = 0$, $x^{(t)} = 2$ for any t. Thus the answer is no.

ii. If $\alpha = 1$, $x^{(1)} = x^{(2)} = 1$. Thus the answer is yes.

iii. If $\alpha = 2$, $x^{(t)}$ oscillates between 2 and 0. Thus the answer is no.

iv. If $\alpha = 1/2$, let t^* be the smallest t such that $|(1-\alpha)^t - (1-\alpha)^{t-1}| \leq \epsilon$. Clearly, t^* is finite. and $x^{(t^*)} \neq 1$. Thus the answer is no.

(b) As $x^{(t)} - 1 = (1-\alpha)^t$, then $\lim_{t\to\infty} x^{(t)}$ exists only when $1 - \alpha \in [0,1]$, namely $\alpha \in [0,1]$. When $\alpha = 1$, $\lim_{t\to\infty} x^{(t)} = 1$; When $\alpha = 0$, $\lim_{t\to\infty} x^{(t)} = 2$; When $\alpha \in (0,1)$, $\lim_{t\to\infty} x^{(t)} = 1$. Thus the range shall be $(0,1]$.

4. **(15 points) Convex Sets**

(a) **(5 points)** Let $f(\boldsymbol{x})$ be a convex function and S a convex set. Then the optimization problem $\min\limits_{\boldsymbol{x}\in S} f(\boldsymbol{x})$ is a convex optimization problem. Let $C \subseteq S$ be the set of points such that $\boldsymbol{x} \in C$ if and only if $\boldsymbol{x}$ is an optimal solution to the problem. Show that C is a convex set.

(b) **(5 points)** Show that the set S given by $S = \{(x,y)|x^2 \leq 5 - 2y^2\}$ is a convex set.

(c) **(5 points)** Let $S_1 \subseteq \mathbb{R}^n$ and $S_2 \subseteq \mathbb{R}^n$ be two convex sets. Show that $S = \{\boldsymbol{x} + \boldsymbol{y} | \boldsymbol{x} \in S_1, \boldsymbol{y} \in S_2\}$ is a convex set.

Solution

(a) Given any two points $\boldsymbol{x}$ and $\mathbf{y}$ such that $\boldsymbol{x} \in C$ and $\mathbf{y} \in C$. For any $\lambda \in [0,1]$, we can obtain that

$$\begin{aligned} f(\lambda \boldsymbol{x} + (1-\lambda)\mathbf{y}) &\leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\mathbf{y}) \\ &= f(\boldsymbol{x}), \end{aligned}$$

where the equality holds because both $\boldsymbol{x}$ and $\mathbf{y}$ are optimal solutions to the problem. Hence, $f(\lambda \boldsymbol{x} + (1-\lambda)\mathbf{y})$ also achieves the minimum and $\lambda \boldsymbol{x} + (1-\lambda)\mathbf{y} \in C$ which completes the proof.

(b) In the class, we have shown that S is the sub-level set of a convex function $f(x,y) = x^2 + 2y^2$ and therefore is a convex set.

(c) Given any two points $\boldsymbol{x}, \boldsymbol{z} \in S_1$ and other two points $\mathbf{y}, \mathbf{w} \in S_2$, for any $\lambda \in [0,1]$, we can obtain that

$$\begin{aligned} &\lambda \boldsymbol{x} + (1-\lambda)\mathbf{y} \in S_1, \text{ and} \\ &\lambda \boldsymbol{z} + (1-\lambda)\mathbf{w} \in S_2. \end{aligned}$$

Therefore, $\lambda(\boldsymbol{x} + \mathbf{y}) + (1-\lambda)(\boldsymbol{z} + \mathbf{w}) = (\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{z}) + (\lambda \mathbf{y} + (1-\lambda)\mathbf{w}) \in S$.

5. **(10 points) Convex Functions**

(a) **(5 points)** Given a vector $\boldsymbol{x} \in \mathbb{R}^n$, let $|\boldsymbol{x}|$ denote the vector with $|\boldsymbol{x}|_i = |\boldsymbol{x}_i|$ (i.e., $|\boldsymbol{x}|$ is the absolute value of $\boldsymbol{x}$, componentwise). Let $|\boldsymbol{x}|_{[i]}$ denote the i-th largest component of $|\boldsymbol{x}|$. In other words, $|\boldsymbol{x}|_{[1]}, |\boldsymbol{x}|_{[2]}, \ldots, |\boldsymbol{x}|_{[n]}$ are the absolute values of the components of $\boldsymbol{x}$, sorted in non-increasing order. Show that $f(\boldsymbol{x}) = \sum_{i=1}^{r} \alpha_i |\boldsymbol{x}|_{[i]}$ is a convex function of $\boldsymbol{x}$, where $\alpha_i = 2n - i$ and $r \leq n$ is a positive integer. Explain the reasons.

(b) **(5 points)** For what range of a will function $f(x) = ax^3 + 6\ln x$ be a real-valued convex function within interval (a^2, ∞)? Explain the reasons.

Solution

(a) In the class, we have proven that $f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i |\boldsymbol{x}|_{[i]}$ is convex if $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq 0$. In this question, the α_i's satisfy this condition, thus we complete the proof.

(b) Observe that we must have $\alpha > 0$, otherwise, this is a concave function, as $f''(x) = 6ax - 6/x^2 < 0$ when $x > 0$. $x^2 f''(x) = 6ax^3 - 6$ has the same sign of $f''(x)$. If f is a convex function, we have that $6ax^3 - 6 \geq 0$ for all $x \in (a^2, \infty)$. As $6ax^3 - 6$ is increasing in x for all $x \in (a^2, \infty)$, thus the condition reduces to $6ax^3 - 6|_{x=a^2} = 6(a^4 - 1) \geq 0$. Thus, considering $a > 0$, $a \geq 1$.

6. **(26 points) Probability + Statistics + Optimization + Machine Learning** An airline company would like to calculate the demand when it sets the price of the flight from Beijing to Shanghai as p. Suppose an arriving potential customer would purchase the ticket if and only if his willingness-to-pay (WTP) is higher than the posted price p. Also assume that a potential customer's WTP X is random, following a distribution with a continuous cumulative distribution function $F(x) = \mathbb{P}(X \leq x)$. It is known that $F(x) = 0$ if $x \leq 1$.

   (a) **(2 points)** For the random experiment of an arriving potential customer buying the ticket or not at price $p \geq 1$, describe the sample space and the probability associated with each outcome.

   (b) **(2 points)** Suppose there is a total of n arriving potential customers, and each customer's purchase decision is independent of each other. Consider the random variable D denoting the total number of sales from n arriving customers under posted price $p \geq 1$ (assume that the total number of flight tickets to sell is larger than n). What is the distribution of D? Give its mean and variance.

   (c) **(3 points)** Consider the mean of D in the previous question. Obviously, it depends on the posted price p. We denote the expected demand as $d(p)$. Suppose for each ticket sold, the airline company incurs a cost of $c \geq 0$ to prepare the flight service. If the airline company wants to set its posted price $p \geq 1$ to maximize the expected profit, write down the formulation of this optimization problem.

      i. What's the decision variable?

      ii. What's the objective function?

      iii. What's the constraint?

   (d) **(3 points)** The profit of the previous part depends on the customers' WTP distribution F. Suppose F is twice continuously differentiable. Give the second order condition under which the profit-maximization problem's objective is concave in $p \in [1, \infty)$.

   (e) The company would like to estimate the relationship between mean demand $d(p)$ and price $p \geq 1$ using the constant elasticity demand function $d(p) = ap^b$, where $a > 0$ and $b < 0$ are parameters to be determined.

      i. **(4 points)** Given that $d(p) = ap^b$ for $p \geq 1$, calculate the value of n and the customer's WTP distribution F.

      ii. **(4 points)** If the price is set as p, the expected consumer surplus of a representative consumer is defined as $\mathbb{E}[\max\{X - p, 0\}]$. Calculate the expected consumer surplus for one representative consumer when $d(p) = ap^b$ and $b < -1$.

      iii. **(4 points)** Given past price and demand data $(D_1, p_1), (D_2, p_2), ..., (D_m, p_m)$ under m different price experiments (for example, under price p_1, the demand is D_1), the company would like to estimate the values of the parameters a and b using MLE. More specifically, it is assumed that the actual demand under posted price p is normally distributed with mean ap^b and known variance σ^2 (σ is small enough so that the chance of demand being negative can be neglected). Write down the loglikelihood function to estimate the parameters a and b. (You don't need to solve it.)

(f) **(4 points)** Suppose the airline company can identify all the arrival and purchase (or not) histories with customer ID in the past year. The number of arrivals A_i and number of purchases B_i of each customer $i = 1, 2, ..., I$ are recorded as follows:

$$(A_1, B_1), (A_2, B_2), ..., (A_I, B_I).$$

Describe how we can use K-means clustering approach to cluster the I customers into K segments.

Solution

(a) The sample space is $\{\text{buy}, \text{not buy}\}$. The probabilities are $P(\text{not buy}) = P(X \leq p) = F(p)$ and $P(\text{buy}) = P(X > p) = 1 - F(p)$.

(b) D is $\text{Binomial}(n, 1-F(p))$, with mean $n(1-F(p))$ and variance $n(1-F(p))F(p)$.

(c) The decision variable is the price p. The objective function is the expected profit $n(1 - F(p))(p - c)$. The constraint is $p \geq 1$.

(d) The second order derivative of the objective function is $-nF''(p)(p-c) - 2nF'(p)$. To ensure concavity of the objective function, we need $-nF''(p)(p-c) - 2nF'(p) \leq 0$.

(e) i. From part b we know that $d(p) = n(1 - F(p))$, hence $n(1 - F(p)) = ap^b$, i.e., $F(p) = 1 - \frac{ap^b}{n}$. Since $F(1) = 0$, we have $1 - a/n = 0$, i.e., $n = a$. Therefore $F(p) = 1 - p^b$.

ii. The expected consumer surplus of one representative consumer is

$$\begin{aligned}\mathbb{E}[\max\{X - p, 0\}] &= \int_p^\infty (x - p)F'(x)dx \\ &= \int_p^\infty -(x - p)bx^{b-1}dx \\ &= -\frac{1}{b+1}p^{b+1}.\end{aligned}$$

iii. Under price p_i, the demand distribution's pdf is

$$f(D_i|p_i) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{D_i - ap_i^b}{\sigma}\right)^2}.$$

Therefore, the likelihood function is

$$\begin{aligned} L(a,b) &= \Pi_{i=1}^m f(D_i|p_i) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^m} e^{-\frac{1}{2}\sum_{i=1}^m \left(\frac{D_i - ap_i^b}{\sigma}\right)^2}, \end{aligned}$$

and the loglikelihood function is

$$l(a,b) = -m\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^m \left(\frac{D_i - ap_i^b}{\sigma}\right)^2.$$

(f) Choose a distance function, say, Euclidean distance $d(x,y)$. Initialize the K cluster centers, say, $c_j^{(0)} = (A_j, B_j)$ for $j = 1, ..., K$. Then for $t = 0, 1, 2, ...,$ do

i. Cluster assignment: for each i, assign $\pi(i) = \arg\max_j d((A_i, B_i), c_j^{(t)})$.

ii. Cluster center update: for each j, let $c_j^{(t+1)} = \frac{1}{|\{i:\pi(i)=j\}|}\sum_{i:\pi(i)=j}(A_i, B_i)$.

Stop until no assignment and center update.

7. **(16 points) The Newsvendor's Problem: MLE + Convex Optimization**
A newspaper vendor faces the task of determining the optimal quantity of newspapers to stock, given uncertain demand D. The cost structure for ordering newspapers is as follows: the first twelve copies incur a cost of \$8 per copy, while any additional copies beyond twelve incur a cost of \$5 per copy. Each sold copy yields a revenue of \$10 for the vendor, while each unsold copy has a salvage value of \$2. Remark: for ease of analysis, we allow the vendor to order a non-integer number of copies.

(a) **(4 points)** Given that the demand D follows a uniform distribution with a support of $[\theta, 2\theta]$, where θ is unknown, the objective is to estimate the maximum likelihood estimator (MLE) of θ. Five samples of D are provided: $\{12, 15, 13, 20, 14\}$. Calculate the MLE of θ. Explain the result in detail.

(b) **(8 points)** Assuming that D follows a uniform distribution with a support of $[\theta, 2\theta]$, where θ is the MLE obtained in (a), calculate the vendor's optimal order decision. Explain the result in detail.

(c) **(4 points)** For (b), what's the vendor's optimal order decision, assuming that the vendor can only order an integer number of copies? Explain the result in detail.

Solution

(a) The likelihood is $\frac{1}{\theta^5}$ if $\theta \leq 12$ and $2\theta \geq 20$; otherwise, the likelihood is 0. To maximize the likelihood, we need to set $\theta = 10$.

(b) The demand $D \sim \mathcal{U}[10, 20]$. As the minimum of concave functions is concave and the non-negative weighted sum of concave functions is concave, $\mathbb{E}[\min(Q, D)]$ is concave in Q.

i. If $Q \leq 12$, the revenue is $10\mathbb{E}[\min(Q, D)]$, the salvage value is $2(Q - \mathbb{E}[\min(Q, D)])$, the cost is $8Q$. Thus, the profit is as follows:

$$8\mathbb{E}[\min(Q, D)] - 6Q$$

which is concave. Thus, if the optimal solution satisfies $Q^* < 12$, by the first-order optimality condition, the optimal decision shall be

$$\frac{20 - Q^*}{10} = \frac{6}{8},$$

namely $Q^* = 12.5$. However, $12.5 \geq 12$. Thus, the optimal decision shall be $Q^* = 12$.

ii. If $Q \geq 12$, the revenue is $10\mathbb{E}[\min(Q, D)]$, the salvage value is $2(Q - \mathbb{E}[\min(Q, D)])$, the cost is $5Q + 36$. Thus, the profit is as follows:

$$8\mathbb{E}[\min(Q, D)] - 3Q - 36$$

which is concave. Thus, if the optimal solution satisfies $Q^* > 12$, by the first-order optimality condition, the optimal decision shall be

$$\frac{20 - Q^*}{10} = \frac{3}{8},$$

namely $Q = 16.25$. As $16.25 \geq 12$, the optimal decision shall be $Q^* = 16.25$.

iii. As the cost is continuous in Q, the profit is continuous in Q. (i) implies that $\pi(Q) \leq \pi(12)$ for $Q \leq 12$ while (ii) implies that $\pi(65/4) \geq \pi(Q)$ for $Q \geq 12$. Thus the optimal Q shall be $65/4$.

(c) The optimal order quantity shall be either 16 or 17. $\pi(16) = 148/5$ and $\pi(17) = 147/5$. Thus the optimal order decision is 16 copies.