1. **(1 × 3=3 pts) Data Mining Applications:** Suppose you are working as a data mining consultant for an Internet Search Engine Company. Briefly describe (1 short paragraph each) how data mining can help the company by giving one example each of an application for which the techniques for (i) regression (ii) classification, (iii) anomaly detection can be used.

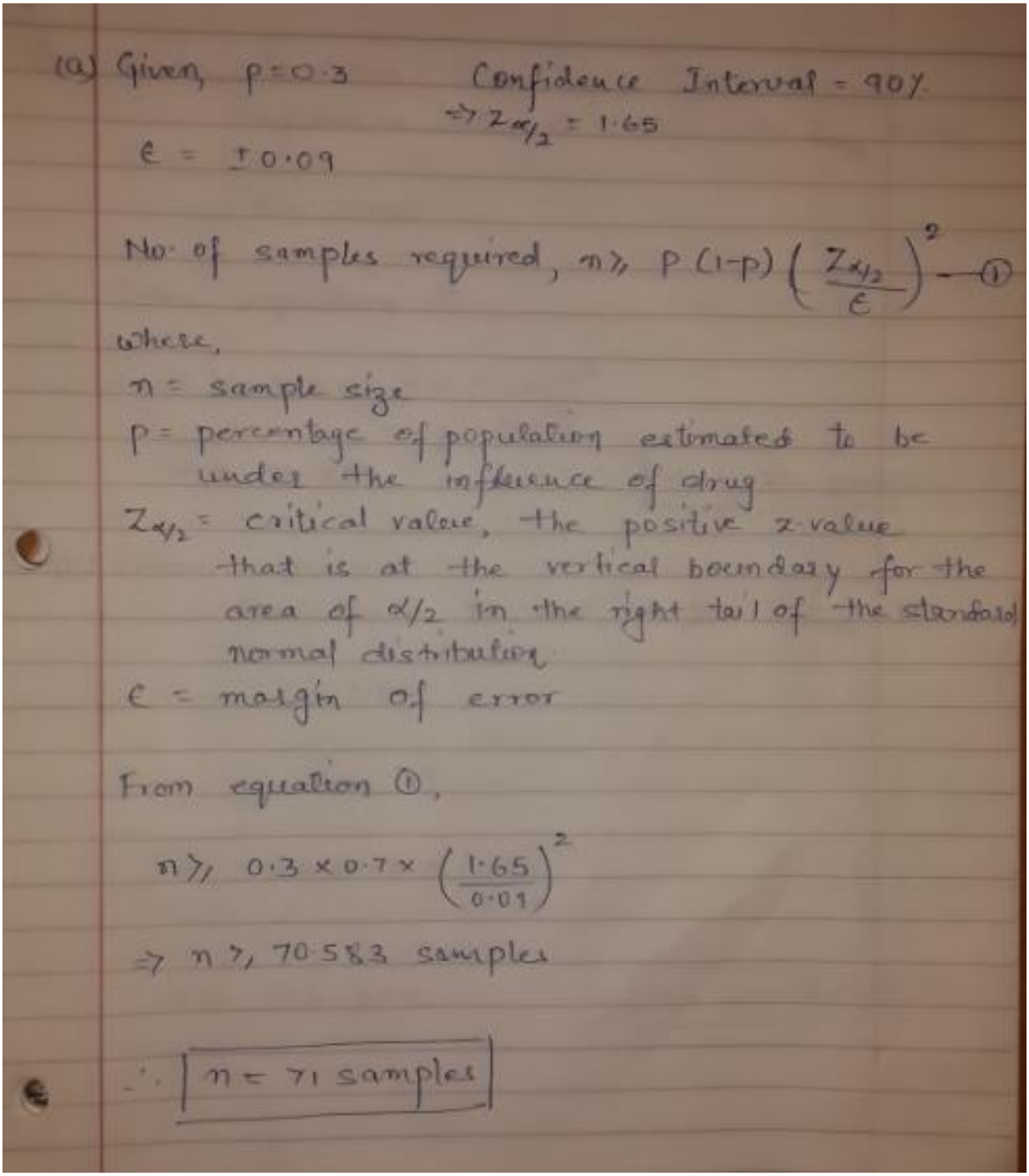
Answer:

- **Regression:** Search engines can use regression to rank the web pages that are to be displayed for a particular query. The training dataset would contain features like various queries, the web pages displayed, number of times the search keyword appears, the number of times users clicked on each of those web pages, etc. The number of clicks could be used as the dependent variable. This model could be used to determine the relevance of various pages for a new query and display the result according to this measure.
- **Classification:** Internet Search Engine Company can use unsupervised classification techniques like clustering for grouping different types of news online. They can use words (features) in various news articles to determine their categories. They can use a number of news articles whose category is known for the purpose of training the model. This would help search engine to show most relevant news related web pages for a particular searched keyword. This would also improve the efficiency of search engine, because the number of pages to be ranked would significantly decrease.
- **Anomaly Detection:** Search engine mostly uses number of clicks as a measure for the relevance or importance of a website/product/ads. But the implicit relation between websites might not be linear. For instance, number of searches/clicks for a mobile phone would be tremendously high compared to a real estate search. This does not mean real estates are trivial. In fact, they are more important. This kind of behavior can be understood using anomaly detection techniques. Search engines can use this information for appropriately pricing ads and ranking websites.

2. **(2 × 3=6 pts) Sampling:** A recent survey estimated that 30% of all Europeans aged 20 to 22 have driven under the influence of drugs or alcohol, based on a simple "Yes or No" question. A similar survey is being planned for Americans. The survey designers want the 90% confidence interval to have a margin of error of at most ± 0.09.
    a) Find the necessary sample size needed to conduct this survey assuming that the expected percentage of "yes" answers will be very close to that obtained from the European survey?
    b) Suppose the tolerance level was kept the same but the confidence level needs to increase to 95%. What is the required sample size for this new specification?

c) If one does not know where the true "p" may lie, one can conservatively conduct a survey assuming the worst case (in terms of required minimum sample size) scenario of p = 0.5. Redo part (b) for this "worst case" scenario.

Answer:



(a) Given, $p = 0.3$ Confidence Interval = 90%

$\Rightarrow Z_{\alpha/2} = 1.65$

$\epsilon = \pm 0.09$

No. of samples required, $n \geq p(1-p)\left(\frac{Z_{\alpha/2}}{\epsilon}\right)^2$ — ①

where,

n = sample size

p = percentage of population estimated to be under the influence of drug

$Z_{\alpha/2}$ = critical value, the positive z-value that is at the vertical boundary for the area of $\alpha/2$ in the right tail of the standard normal distribution

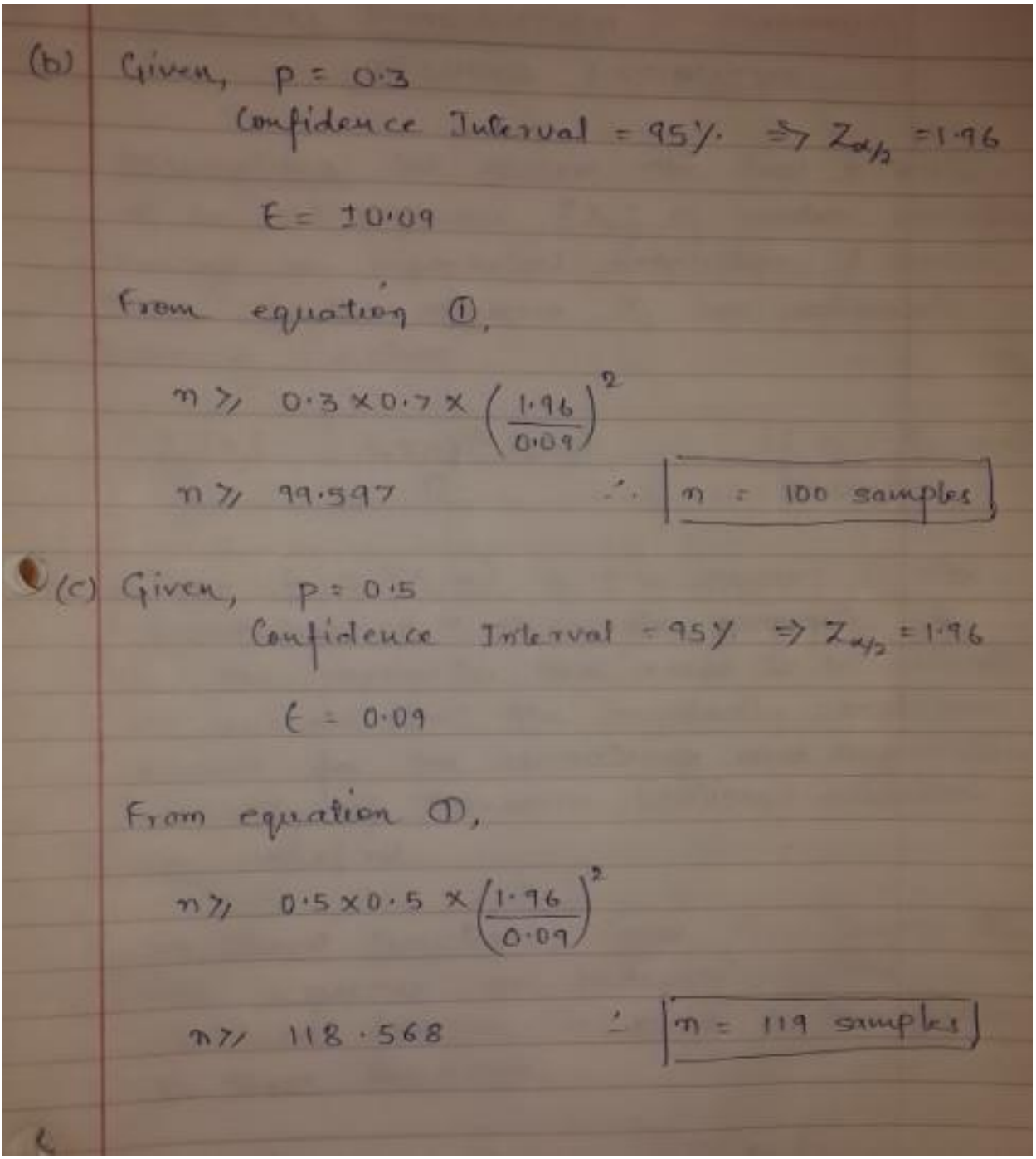
ϵ = margin of error

From equation ①,

$n \geq 0.3 \times 0.7 \times \left(\frac{1.65}{0.09}\right)^2$

$\Rightarrow n \geq 70.583$ samples

$\therefore$ **$n = 71$ samples**



3. **(5 pts) Maximum Likelihood Estimation:** Wishing to estimate the average time it takes to load Canvas on her tablet, Alice does the following study: She records the time xi that Canvas takes to load (in milliseconds), e i = 1,...,N, at N randomly selected time-points during one day. Suppose that the time it takes to load the webpage can be well represented by an exponential distribution with (unknown) parameter, λ. Derive the maximum likelihood estimate for λ from first principles. (i.e. do no just write down the answer).

Answer:

EXPONENTIAL DISTRIBUTION – MAXIMUM LIKELIHOOD ESTIMATION

Assumptions: We observe the first n terms of an IID sequence $\{X_n\}$ of random variable having an exponential distribution. A generic term of the sequence X_j has probability density function

$$f_X(x_j) = \begin{cases} \lambda_0 \exp(-\lambda_0 x_j) & \text{if } x_j \in R_X \\ 0 & \text{otherwise} \end{cases} \quad -(1)$$

where $R_X = [0, \infty)$ is the support of the distribution and the rate parameter λ_0 is the parameter that needs to be estimated. We assume that the regularity conditions needed for the consistency and asymptotic normality of maximum likelihood estimators are satisfied.

Likelihood Function: Since the terms of the sequence are independent, the likelihood function is equal to the product of their densities:

$$L(\lambda; x_1, \ldots, x_n) = \prod_{j=1}^{n} f_X(x_j; \lambda)$$

Because the observed values $x_1, \ldots, x_n$ can only belong to the support of the distribution, we can write

$$L(\lambda; x_1, \ldots, x_n) = \prod_{j=1}^{n} f_X(x_j; \lambda)$$

$$= \prod_{j=1}^{n} \lambda \exp(-\lambda x_j)$$

$$= \lambda^n \exp\left(-\lambda \sum_{j=1}^{n} x_j\right) \quad \text{—— (2)}$$

Log-likelihood function: This is obtained by taking the natural logarithm of the likelihood function:

$$l(\lambda; x_1, \ldots, x_n) = \ln(L(\lambda; x_1, \ldots, x_n))$$

$$= \ln\left(\lambda^n \exp\left(-\lambda \sum_{j=1}^{n} x_j\right)\right)$$

$$= \ln(\lambda^n) + \ln\left(\exp\left(-\lambda \sum_{j=1}^{n} x_j\right)\right)$$

$$= n \ln(\lambda) - \lambda \sum_{j=1}^{n} x_j \quad \text{—— (3)}$$

Maximum Likelihood Estimator: The estimator is obtained as a solution of the maximization problem

$$\hat{\lambda} = \arg\max_{\lambda} l(\lambda; x_1, \ldots, x_n)$$

The first order condition for a maximum is

$$\frac{d}{d\lambda} l(\lambda; x_1, \ldots, x_n) = 0$$

$$\Rightarrow \frac{d}{d\lambda}\left(n \ln(\lambda) - \lambda \sum_{j=1}^{n} x_j\right) = 0$$

$$\Rightarrow \frac{n}{\lambda} - \sum_{j=1}^{n} x_j = 0$$

$$\Rightarrow \lambda = \frac{n}{\sum_{j=1}^{n} x_j} \quad \text{——— (A)}$$

Therefore, the estimator $\hat{\lambda}_n$ is just the reciprocal of the sample mean

$$\boxed{\frac{\sum_{j=1}^{n} x_j}{n}}$$

Answer

4. **(6 pts) Illustration of "curse of dimensionality":** Consider the set of 2^{10000} binary vectors of length 10000 (i.e. 00...0 through 111...1). The Hamming distance between any two vectors (number of bits they differ by) from this will be a number between 0 and 10000. What is the probability that the Hamming distance of an arbitrary chosen pair of vectors is between 4950 and 5050? (Hint: distribution of distances is binomial; can be approximated by a Normal distribution as n >> 30. You can get a table for a Standard Normal distribution from many sources, e.g., http://www.stat.ucla.edu/y˜wu/teaching/normal.pdf)
This problem shows that if data is randomly distributed at the corners of a very high-dimensional hypercube, then for any point, most of the other points are bunched at about the same distance, so there is little scope of focusing on points with known "y" in a nearby neighborhood to do any prediction.

Answer:

Given,
n = 10000
p = 0.5

Since n >> 30, binomial distribution can be approximated as Normal distribution with

$$\text{Mean, } \mu = np = 10000*0.5 = 5000$$
$$\text{Standard Deviation, } \sigma = \text{sqrt}(np(1-p)) = \text{sqrt}(10000*0.5*0.5) = 50$$

We know that,

$$Z = (x - \mu)/\sigma$$

Therefore,

$$Z_{4950} = (4950 - 5000)/50 = -1$$
$$Z_{5050} = (5050 - 5000)/50 = 1$$

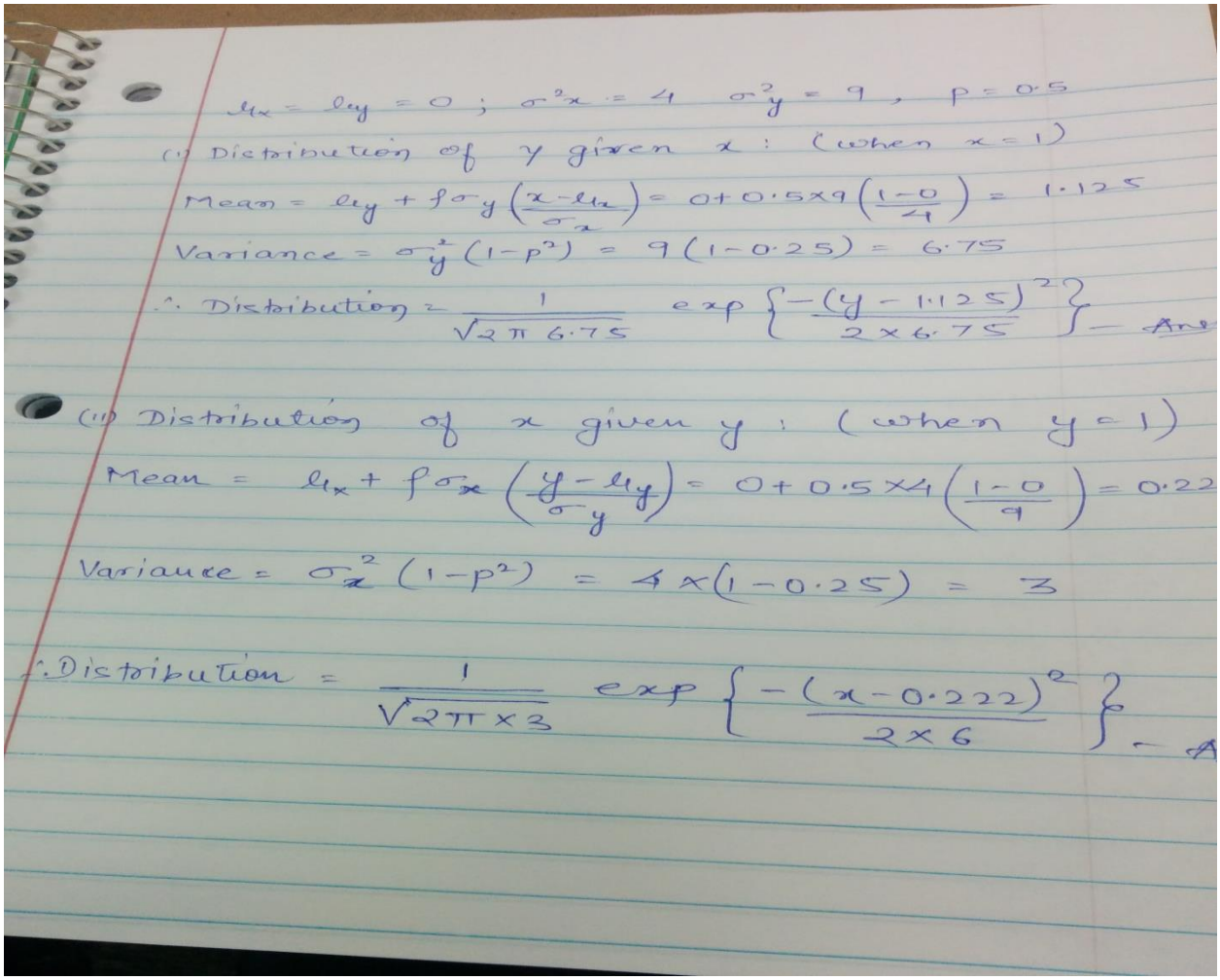
Since the normal distribution is symmetric and z-score is 1, the area under the curve would be approximately 68%. This implies that there is 68% probability that the Hamming distance of an arbitrary chosen pair of vectors is between 4950 and 5050.

5. **(3+4+3=6 pts). Bivariate Visualization and Mathematical Form:** Suppose X and Y are two random variables whose joint distribution is Normal (Gaussian), centered at (0,0) and with correlation ρ. (See "Bivariate Case" in the Wikipedia entry for "Multivariate Normal Distribution" for the equation, or use just use the vector form given in the class notes, with σ12 = σ21 = ρσxσy). Consider 2 cases i) σ2 x = 4;σ2 y = 9;ρ = 0 ii)σ2 x = 4;σ2 y = 9;ρ = 0.5

a) Obtain contour plots for each of the two distributions using the contour function in R to get the plots.
b) View 3-D plots for the two distributions from atleast two different viewing perspectives each. The function Persp allows you to specify viewing perspective by setting "theta" and "phi".
c) Consider the bivariate Normal Distribution given in part (ii). Reading the "Bivariate Case" under "Conditional distributions" in the Wikipedia entry will help you answer this problem; alternatively you can consult any undergraduate text on probability/statistics. What is the mathematical form of the conditional distribution that is obtained when (a) x is set to 1, and (b) when y is set to 1? (no need to actually derive the formulae from first principles; rather just obtain the result by substitution in the formula for a bivariate Gaussian.

Answer:

c) Mathematical Form of the conditional distributions:



Correction: In the last expression, 6 in the denominator should be replaced by 3.

6. **(2×5 =10 pts) Exploratory Data Analysis using R:** The "student" data set, part of the LearnBayes package, records properties of 657 students. For a description of the data, see http://cran.r-project.org/web/packages/LearnBayes/LearnBayes.pdf.
    a) Construct a histogram of the variable Shoes. (Set the value of the parameter breaks to 20)
    b) Use data visualization to check if the variable Dvds (approximately) follows a log-normal distribution.
    c) Summarize the variable Haircut using the summary command. Also, report the 2.5th and 97.5th percentiles.
    d) Construct a barplot of the individual values of Drink that were observed. Also, highlight the distribution of the variable Drink between the two genders on the same barplot.
    e) Construct a scatter plot of the variables ToSleep and WakeUp. Do you observe a positive correlation between the two variables?

    Note: Omit missing values if any are present.