DS5220 Homework 2 (written solution)

Aditya Singh

TOTAL POINTS

98 / 100

QUESTION 1

Problem 1 10 pts

1.1 2 / 2

✓ **- 0 pts Correct**

- 0.5 pts missing some information

1

1.2 1 / 1

✓ **- 0 pts Correct**

1.3 1 / 1

✓ **- 0 pts Correct**

- 0.5 pts not well explained

1.4 2 / 2

✓ **- 0 pts Correct**

- 1 pts unclear explaination

- 0.5 pts missing concept/understanding

1.5 2 / 2

✓ **- 0 pts Correct**

- 0.5 pts missing a few steps

- 1 pts improper explanation and steps

1.6 2 / 2

✓ **- 0 pts Correct**

- 0.5 pts incorrect/missing - when KNN is a better approach

- 2 pts incorrect

- 1 pts missing - how KNN works

- 0 pts Click here to replace this description.

QUESTION 2

Problem 2 30 pts

2.1 3 / 3

✓ **- 0 pts Correct**

- 2 pts Coding notebook (.ipynb) file not provided.

- 0 pts Click here to replace this description.

2.2 6 / 6

✓ **- 0 pts Correct**

2.3 3 / 3

✓ **- 0 pts Correct**

2.4 6 / 6

✓ **- 0 pts Correct**

2.5 6 / 6

✓ **- 0 pts Correct**

2.6 6 / 6

✓ **- 0 pts Correct**

- 2 pts Best number of neighbors not provided.

- 2 pts Best error not reported.

QUESTION 3

Problem 3 20 pts

3.1 4 / 4

✓ **- 0 pts Correct**

- 2 pts Part c and Part d missing

- 0.5 pts Partially correct. Wrong output probability

- 1 pts Explanation missing for all the parts.

- 0.25 pts Final simplified term not arrived in part d.

- 0.5 pts Explanation missing for some parts.

Hi. Could you submit your all the solutions for problem 3 on Gradescope as a pdf rather than .ipynb? Thanks.

3.2 2 / 4

- **0 pts** Correct
- **1 pts** Explanation not provided.
- ✓ **- 2 pts** **Wrong probability though there is good explanation.**

3.3 6 / 6

- ✓ **- 0 pts** **Correct**
- **2 pts** For n = 100, probability is not calculated.
- **1 pts** Formula in terms of 'n' not provided.
- **1 pts** Final answer not derived.
- **1 pts** Explanation not provided.
- **2 pts** No explanation provided.
- **6 pts** No answer.

Hi Aditya. Could you attach problem 3 c) bit question on gradescope. Thanks.

3.4 6 / 6

- ✓ **- 0 pts** **Correct**
- **1 pts** Explanation not provided.
- **6 pts** No answer.

Hi. Could you please attach problem 3 d) bit on gradescope. Thanks.

QUESTION 4

Problem 4 40 pts

4.1 2 / 2

- ✓ **- 0 pts** **Correct**
- **2 pts** Not Done Part (a): provide an estimate for the population mean of the Price
- **1 pts** Part (a): Incorrect estimatation for the population mean of the Price

4.2 3 / 3

- ✓ **- 0 pts** **Correct**
- **3 pts** Not Done Part (b): Provide an estimate of the standard error of $\hat{\mu}$
- **1.5 pts** Part (b): Incorrect/Incomplete estimate of the standard error of $\hat{\mu}$

4.3 10 / 10

- ✓ **- 0 pts** **Correct**
- **10 pts** Not Done Part (c): Estimate the standard error of $\hat{\mu}$ using 1,000 bootstraps. Comment on your findings
- **5 pts** Part (c): Incorrect/Incomplete standard error of $\hat{\mu}$ using 1,000 bootstraps.
- **2 pts** Comment on findings not done or incorrect.

4.4 3 / 3

- ✓ **- 0 pts** **Correct**
- **3 pts** Not Done Part (d): Provide a 95% confidence interval for the mean of the Price. Comment on your findings
- **2 pts** Part (d): Incorrect 95% confidence interval for the mean of the Price.
- **1 pts** Comment on findings not done or incorrect

4.5 2 / 2

- ✓ **- 0 pts** **Correct**
- **2 pts** Not Done Part (e): Estimate for the first 25% quantile of Price
- **1 pts** Part (e): Incorrect estimation for the first 25% quantile of Price

4.6 10 / 10

- ✓ **- 0 pts** **Correct**
- **10 pts** Not Done Part (f): Estimate the standard error of $\hat{\mu}0.25$. Comment on your findings
- **5 pts** Part (f): Incorrect estimation of the standard error of $\hat{\mu}0.25$.
- **2 pts** Comment on findings not done or incorrect.

4.7 10 / 10

- ✓ **- 0 pts** **Correct**
- **10 pts** Not Done Part (g): Train a linear regression model to predict Price using Rating. Compute estimates for the standard errors of the intercept $\beta 0$ and coefficient $\beta 1$
of Rating in two different ways: (1) using the bootstrap, and (2) using the standard errors. Comment on your findings

- **4 pts** Part (g): Incorrect computation of estimates for the standard errors of the intercept β0 and coefficient β1
of Rating using bootstrap.
- **4 pts** Part (g): Incorrect computation of estimates for the standard errors of the intercept β0 and coefficient β1
of Rating using the standard errors.
- **2 pts** Comment on findings not done or incorrect

Question 1

Aditya Singh

October 2022

1 Lasso Regression and Ridge Regression and its use cases

1. Ridge Regression

   Suppose we have a problem in which two features are correlated i.e if one feature increases the other is bound to increase in this case the value of coefficients beta0 ,beta1 will have high variance. therefore, to avoid this we make use of Ridge Regression we put one constraint on the beta values like $\beta_0 + \beta_1 + .. <= C$ where C is any arbitrary number. This is called as ridge penalty added on the OLS equation. λ is the parameter we have to hypertune larger the value of λ lesser would be the slope after gradient descent. If there are more parameters than data points ridge is again good choice to go with eg. In case of any space model where we need to do some kind of regression with stars and their location in uses cases like these the features are more then that of data points.

2. Lasso Regression

   In case of Lasso regression instead of taking the $slope^2$ times λ (Ridge penalty) we take the absolute value of slopes. The difference between Lasso and Ridge is that the value of slopes can be reduced asymptotically close to 0 in Ridge while in Lasso it can be reduced to absolute 0. Also Lasso uses something called as (stop gradient) instead of regular Gradient Descent to avoid calculation of mod at 0.

3. Ridge v/s Lasso

   When we a lot of parameters and all of those have good correlation with the target variable then we should chose 'Ridge Regression' and if there are a lot of parameters and many of them are redundant then we should use 'Lasso Regression' as it will bring the slope of those parameters to 0.

1.1 **2 / 2**

✓ **- 0 pts** Correct

- 0.5 pts missing some information

1

2 Explain in words the meaning of P-value and confidence interval.

1. P-value

   Probability of a event of interest and event equal or rarer is called as P-value. It consist of 3 part, the probability that random chance is generated or something equal or rarer. P-value helps us decide if we should reject the null-hypothesis or not.

2. Confidence Interval The confidence interval (CI) is a range of values that's likely to include a population value with a certain degree of confidence. It is often expressed as a % whereby a population mean lies between an upper and lower interval.

3 Maximum Likelihood in Logistic Regression

Maximum Likelihood estimation is a method of estimating the parameter of logistic regression model. It selects sets of parameter that maximize the value of model parameter in the likelihood function.

The Likelihood function is the probability that the observed values of the dependent variable may be predicted from observed value of independent variables. The likelihood varies from 0 to 1.

$\log L(\Theta) = \sum_{n=1}^{m} y_i * log(\sigma(\Theta^t * x_i)) + (1 - y_i) * log(1 - \sigma(\Theta^t * x_i))$

This is the log likelihood function we have to maximize using gradient ascent.

4 Define and explain the difference between variance, covariance matrix, and correlation coefficient.

1. Correlation coefficient

   In statistics, correlation is a measure that determines the degree to which two or more random variables move in sequence. When an equivalent movement of another variable reciprocates the movement of one variable in some way or another during the study of two variables, the variables are said to be correlated. The formula for correlation is:

   $\rho_{xy} = Correlation(x, y) = \frac{cov(x,y)}{\sqrt{var(x)} * \sqrt{var(y)}}$

2. Variance

   Variance tells us how much a quantity varies w.r.t. its mean. Its the spread of data around the mean value. You only know the magnitude here, as in how much the data is spread. A random variable is compared against itself.

1.2 1 / 1

✓ **- 0 pts** Correct

2 Explain in words the meaning of P-value and confidence interval.

1. P-value

   Probability of a event of interest and event equal or rarer is called as P-value. It consist of 3 part, the probability that random chance is generated or something equal or rarer. P-value helps us decide if we should reject the null-hypothesis or not.

2. Confidence Interval The confidence interval (CI) is a range of values that's likely to include a population value with a certain degree of confidence. It is often expressed as a % whereby a population mean lies between an upper and lower interval.

3 Maximum Likelihood in Logistic Regression

Maximum Likelihood estimation is a method of estimating the parameter of logistic regression model. It selects sets of parameter that maximize the value of model parameter in the likelihood function.

The Likelihood function is the probability that the observed values of the dependent variable may be predicted from observed value of independent variables. The likelihood varies from 0 to 1.

$\log L(\Theta) = \sum_{n=1}^{m} y_i * log(\sigma(\Theta^t * x_i)) + (1 - y_i) * log(1 - \sigma(\Theta^t * x_i))$

This is the log likelihood function we have to maximize using gradient ascent.

4 Define and explain the difference between variance, covariance matrix, and correlation coefficient.

1. Correlation coefficient

   In statistics, correlation is a measure that determines the degree to which two or more random variables move in sequence. When an equivalent movement of another variable reciprocates the movement of one variable in some way or another during the study of two variables, the variables are said to be correlated. The formula for correlation is:

   $\rho_{xy} = Correlation(x, y) = \frac{cov(x,y)}{\sqrt{var(x)} * \sqrt{var(y)}}$

2. Variance

   Variance tells us how much a quantity varies w.r.t. its mean. Its the spread of data around the mean value. You only know the magnitude here, as in how much the data is spread. A random variable is compared against itself.

1.3 **1 / 1**

✓ **- 0 pts** **Correct**

- 0.5 pts not well explained

2 Explain in words the meaning of P-value and confidence interval.

1. P-value

   Probability of a event of interest and event equal or rarer is called as P-value. It consist of 3 part, the probability that random chance is generated or something equal or rarer. P-value helps us decide if we should reject the null-hypothesis or not.

2. Confidence Interval The confidence interval (CI) is a range of values that's likely to include a population value with a certain degree of confidence. It is often expressed as a % whereby a population mean lies between an upper and lower interval.

3 Maximum Likelihood in Logistic Regression

Maximum Likelihood estimation is a method of estimating the parameter of logistic regression model. It selects sets of parameter that maximize the value of model parameter in the likelihood function.

The Likelihood function is the probability that the observed values of the dependent variable may be predicted from observed value of independent variables. The likelihood varies from 0 to 1.

$\log L(\Theta) = \sum_{n=1}^{m} y_i * log(\sigma(\Theta^t * x_i)) + (1 - y_i) * log(1 - \sigma(\Theta^t * x_i))$

This is the log likelihood function we have to maximize using gradient ascent.

4 Define and explain the difference between variance, covariance matrix, and correlation coefficient.

1. Correlation coefficient

   In statistics, correlation is a measure that determines the degree to which two or more random variables move in sequence. When an equivalent movement of another variable reciprocates the movement of one variable in some way or another during the study of two variables, the variables are said to be correlated. The formula for correlation is:

   $\rho_{xy} = Correlation(x, y) = \frac{cov(x,y)}{\sqrt{var(x)} * \sqrt{var(y)}}$

2. Variance

   Variance tells us how much a quantity varies w.r.t. its mean. Its the spread of data around the mean value. You only know the magnitude here, as in how much the data is spread. A random variable is compared against itself.

$Var(X) = E(X.X) - E(X) * E(X)$

3. Covariance Matrix

   covariance matrix is defined as a square matrix where the diagonal elements represent the variance and the off-diagonal elements represent the covariance. The covariance between two variables can be positive, negative, and zero. A positive covariance indicates that the two variables have a positive relationship whereas negative covariance shows that they have a negative relationship. If two elements do not vary together then they will display a zero covariance.

4. Difference between covariance matrix, Variance, Correlation Coefficient.

   Difference between Covariance matrix, correlation coefficient and Variance Variance tells us how much a quantity varies w.r.t. its mean. Its the spread of data around the mean value. You only know the magnitude here, as in how much the data is spread.

   Covariance tells us direction in which two quantities vary with each other.

   Correlation shows us both, the direction and magnitude of how two quantities vary with each other.

5 Steps needed for performing the linear discriminant analysis

Listed below are the 5 general steps for performing a linear discriminant analysis.

1. Compute the d-dimensional mean vectors for the different classes from the dataset.

2. Compute the scatter matrices (in-between-class and within-class scatter matrix).

3. Compute the eigenvectors ($ee_1, ee_2, ..., ee_d$) and corresponding eigenvalues ($\lambda\lambda_1, \lambda\lambda_2, ..., \lambda\lambda_d$) for the scatter matrices.

4. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d×k dimensional matrix WW (where every column represents an eigenvector).

5. Use this d×k eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: YY=XX×WW (where XX is a n×d-dimensional matrix representing the n samples, and yy are the transformed n×k-dimensional samples in the new subspace).

   These are the five steps taken to perform Linear Discriminant Analysis (LDA)

1.4 2 / 2

✓ **- 0 pts** **Correct**

- 1 pts unclear explaination

- 0.5 pts missing concept/understanding

$Var(X) = E(X.X) - E(X) * E(X)$

3. Covariance Matrix

   covariance matrix is defined as a square matrix where the diagonal elements represent the variance and the off-diagonal elements represent the covariance. The covariance between two variables can be positive, negative, and zero. A positive covariance indicates that the two variables have a positive relationship whereas negative covariance shows that they have a negative relationship. If two elements do not vary together then they will display a zero covariance.

4. Difference between covariance matrix, Variance, Correlation Coefficient.

   Difference between Covariance matrix, correlation coefficient and Variance Variance tells us how much a quantity varies w.r.t. its mean. Its the spread of data around the mean value. You only know the magnitude here, as in how much the data is spread.

   Covariance tells us direction in which two quantities vary with each other.

   Correlation shows us both, the direction and magnitude of how two quantities vary with each other.

5 Steps needed for performing the linear discriminant analysis

Listed below are the 5 general steps for performing a linear discriminant analysis.

1. Compute the d-dimensional mean vectors for the different classes from the dataset.

2. Compute the scatter matrices (in-between-class and within-class scatter matrix).

3. Compute the eigenvectors ($ee_1, ee_2, ..., ee_d$) and corresponding eigenvalues ($\lambda\lambda_1, \lambda\lambda_2, ..., \lambda\lambda_d$) for the scatter matrices.

4. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d×k dimensional matrix WW (where every column represents an eigenvector).

5. Use this d×k eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: YY=XX×WW (where XX is a n×d-dimensional matrix representing the n samples, and yy are the transformed n×k-dimensional samples in the new subspace).

   These are the five steps taken to perform Linear Discriminant Analysis (LDA)

1.5 **2 / 2**

✓ **- 0 pts** **Correct**

- 0.5 pts missing a few steps

- 1 pts improper explanation and steps

6 K Nearest Neighbors (KNN)

K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbours are classified.

In general, practice, choosing the value of k is k = sqrt(N) where N stands for the number of samples in your training dataset.

In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given "unseen" observation. Similarity is defined according to a distance metric between two data points. A popular one is the Euclidean distance method we can also use something like Manhattan, Minkowski, and Hamming distance methods.

When should we choose KNN over other Models:

1. We have properly labeled data. For example, if we are predicting someone is having diabetes or not the final label can be 1 or 0. It cannot be NaN or -1.

2. Data is noise-free. For the diabetes data set we cannot have a Glucose level as 0 or 10000. It's practically impossible.

3. Small dataset.

1.6 2 / 2

✓ **- 0 pts** Correct

- 0.5 pts incorrect/missing - when KNN is a better approach

- 2 pts incorrect

- 1 pts missing - how KNN works

- 0 pts Click here to replace this description.

2.1 3 / 3

✓ **- 0 pts** Correct

- 2 pts Coding notebook (.ipynb) file not provided.

- 0 pts Click here to replace this description.

2.2 6 / 6

✓ **- 0 pts** Correct

2.3 3 / 3

✓ **- 0 pts** Correct

2.4 6 / 6

✓ **- 0 pts** Correct

2.5 6 / 6

✓ **- 0 pts** Correct

2.6 6 / 6

✓ **- 0 pts** Correct

- 2 pts Best number of neighbors not provided.

- 2 pts Best error not reported.

Question 3

Aditya Singh

October 2022

1 What is the probability that z1 = x1?1

The Probability that our current bootstrap sample Z1, will be from the original sample is $\frac{1}{n}$ (n being number of sample). Therefore, The probability of the sample not being from that set is.

$$(1 - \frac{1}{n})$$

2 What is the probability that z2 = x1

We Know that the probability of sample not being in original sample is independent. Therefore P(Z1 != X1)*P(Z2 != X1) i.e

$$(1 - \frac{1}{n})^2$$

.

PART C and D are uploaded as .ipynb

3.1 4 / 4

✓ **- 0 pts** **Correct**

- 2 pts Part c and Part d missing

- 0.5 pts Partially correct. Wrong output probability

- 1 pts Explanation missing for all the parts.

- 0.25 pts Final simplified term not arrived in part d.

- 0.5 pts Explanation missing for some parts.

Hi. Could you submit your all the solutions for problem 3 on Gradescope as a pdf rather than .ipynb? Thanks.

Question 3

Aditya Singh

October 2022

1 What is the probability that z1 = x1?1

The Probability that our current bootstrap sample Z1, will be from the original sample is $\frac{1}{n}$ (n being number of sample). Therefore, The probability of the sample not being from that set is.

$$(1 - \frac{1}{n})$$

2 What is the probability that z2 = x1

We Know that the probability of sample not being in original sample is independent. Therefore P(Z1 != X1)*P(Z2 != X1) i.e

$$(1 - \frac{1}{n})^2$$

.

PART C and D are uploaded as .ipynb

3.2 **2 / 4**

- **0 pts** Correct
- **1 pts** Explanation not provided.

✓ **- 2 pts** **Wrong probability though there is good explanation.**

3.3 6 / 6

✓ **- 0 pts** **Correct**

- 2 pts For n = 100, probability is not calculated.

- 1 pts Formula in terms of 'n' not provided.

- 1 pts Final answer not derived.

- 1 pts Explanation not provided.

- 2 pts No explanation provided.

- 6 pts No answer.

Hi Aditya. Could you attach problem 3 c) bit question on gradescope. Thanks.

3.4 6 / 6

✓ **- 0 pts** **Correct**

- 1 pts Explanation not provided.

- 6 pts No answer.

Hi. Could you please attach problem 3 d) bit on gradescope. Thanks.

4.1 2 / 2

✓ **- 0 pts** Correct

- 2 pts Not Done Part (a): provide an estimate for the population mean of the Price

- 1 pts Part (a): Incorrect estimatation for the population mean of the Price

4.2 3 / 3

✓ **- 0 pts** **Correct**

- 3 pts Not Done Part (b): Provide an estimate of the standard error of $\hat{\mu}$

- 1.5 pts Part (b): Incorrect/Incomplete estimate of the standard error of $\hat{\mu}$

4.3 10 / 10

✓ **- 0 pts** **Correct**

- 10 pts Not Done Part (c): Estimate the standard error of $\hat{\mu}$ using 1,000 bootstraps. Comment on your findings

- 5 pts Part (c): Incorrect/Incomplete standard error of $\hat{\mu}$ using 1,000 bootstraps.

- 2 pts Comment on findings not done or incorrect.

4.4 3 / 3

✓ **- 0 pts** **Correct**

- 3 pts Not Done Part (d): Provide a 95% confidence interval for the mean of the Price. Comment on your findings

- 2 pts Part (d): Incorrect 95% confidence interval for the mean of the Price.

- 1 pts Comment on findings not done or incorrect

4.5 2 / 2

✓ **- 0 pts** Correct

- 2 pts Not Done Part (e): Estimate for the first 25% quantile of Price

- 1 pts Part (e): Incorrect estimation for the first 25% quantile of Price

4.6 10 / 10

✓ **- 0 pts** **Correct**

- 10 pts Not Done Part (f): Estimate the standard error of $\hat{\mu}0.25$. Comment on your findings

- 5 pts Part (f): Incorrect estimation of the standard error of $\hat{\mu}0.25$.

- 2 pts Comment on findings not done or incorrect.

4.7 10 / 10

✓ **- 0 pts** **Correct**

- 10 pts Not Done Part (g): Train a linear regression model to predict Price using Rating. Compute estimates for the standard errors of the intercept β0 and coefficient β1
of Rating in two different ways: (1) using the bootstrap, and (2) using the standard errors. Comment on your findings

- 4 pts Part (g): Incorrect computation of estimates for the standard errors of the intercept β0 and coefficient β1
of Rating using bootstrap.

- 4 pts Part (g): Incorrect computation of estimates for the standard errors of the intercept β0 and coefficient β1
of Rating using the standard errors.

- 2 pts Comment on findings not done or incorrect