FOUNDATIONS OF MACHINE LEARNING

M.SC. IN DATA SCIENCES AND BUSINESS ANALYTICS CENTRALESUPE´LEC

Assignment 1

Instructor: Fragkiskos Malliaros

TA: Benjamin Maheu and Lily Monnier Due: **November 9, 2022 at 23:00**

**How to submit:** Please complete the first assignment **individually**. *Typeset* all your answers (**PDF** file only). Submissions should be made on **gradescope** (Assignment 1; Entry Code: ZZPBXP). You have already received an email on your cs email account from *gradescope* (if not, please contact me). Make sure that the answer to each question is on a **separate page** (questions 1-10).

# General Questions

## Question 1 [10 points]

True/False questions, *with justification*. *[Keep your answer short]*

* 1. Cross-validation for the tuning of a model guarantees that the model will not overfit.

TRUE: Data trained and tested in chunks does lead to ensuring no overfitting.

* 1. Both PCA and linear regression can be thought of as algorithms for minimizing a sum of squared errors.

TRUE: PCA choses eigen vectors such that data loss is lowest. RMS of this new space will be a good error estimator

* 1. Let’s assume that we perform Singular Value Decomposition on a dataset and observe that the singular values that are returned are all approximately equal. Then, we should expect that the variance explained is approximately linear to the number of principal components used for Prin- cipal Component Analysis (PCA).

TRUE: Principal components explain the variance in each axis. Since all eigen values of SVD are equal, the variance explained will be approximately linear to the number of principal components

* 1. Let *yi* = *xα*1 *eα*2 + *ϵi* be a model, where *ϵi* ∼ N (0*, σ*2) corresponds to Gaussian noise. Then, the maximum likelihood parameters of the model (*α*) can be learned using linear regression.

FALSE: The error term makes it impossible to apply MLE

* 1. The eigenvectors of **AA***T* and **A***T***A** are the same.

FALSE: While the eigen values will be equal, the eigen vectors will be different

# Dimensionality Reduction

## Question 2 [15 points]

Let **M***m×n* be a data matrix (*m* observations (i.e., data points), *n* dimensions (i.e., features)).

* 1. [2 p] Are the matrices **MM***T* and **M***T***M** symmetric, square and real? Justify your answer.

TRUE: A picture containing text, whiteboard

Description automatically generated

* 1. [4 p] Show that the eigenvalues of **MM***T* are the same as the ones of **M***T***M**. Are their eigenvectors the same too? Justify your answer.

EIGEN VALUES ARE SAME, EIGEN VECTORS DIFFERENT: Text, letter

Description automatically generated

* 1. [4 p] SVD decomposes the matrix **M** into the product **UΣV***T*, where **U** and **V** are orthonormal and **Σ** is a diagonal matrix. Given that **M** = **UΣV***T*, write a simplified expression of **M***T***M** in terms of **V**, **V***T* and **Σ**. Can we find an analogous expression for **MM***T*?

M=A, PFB Answer

Text, letter

Description automatically generated

* 1. [5 p] What is the relationship (if any) between the eigenvalues of **M***T***M** and the singular values of **M**? Justify your answer.

The values of M.T \* M are squares of M. The above answer shows us the same

## Question 3 [10 points]

PCA transforms (zero-mean) data points **x** ∈ R*d* into low-dimensional reconstructions that lie in the span of the top *k* eigenvectors of the sample covariance matrix. Let **U***k* denote the *d* × *k* matrix of the top *k* eigenvectors of the covariance matrix (**U***k* is a truncated version of **U**, which is the matrix of eigenvectors of the covariance matrix).

There are two approaches to computing the low-dimensional reconstruction **w** ∈ R*k* of a data point

**x** ∈ R*d*:

1. Solve a least squares problem to minimize the reconstruction error.
2. Project **x** onto the span of the columns of **U***k*.

In this question, you will show that these approaches are equivalent.

1. [5 p] Formulate the least squares problem in terms of **U***k*, **x** and **w**. (Hint: the optimization problem should resemble linear regression.)

A piece of paper with writing on it

Description automatically generated

1. [5 p] Show that the solution of the least squares problem is equal to **U***Tk* **x**, which is the projection of **x** onto the span of the columns of **U***k*. (Hint: use the closed-form solution of the least-squares problem).

A piece of paper with writing on it

Description automatically generated

# Model Evaluation

## Question 4 [4 points]

We evaluated two algorithms on a task consisting in classifying mushrooms between poisonous and edible based on some descriptors. We obtained the two following confusion matrices:

|  |  |  |
| --- | --- | --- |
|  | Labeled Edible | Labeled Poisonous |
| Edible | 100 | 0 |
| Poisonous | 3 | 97 |

**Table 1:** Confusion table of Algorithm 1.

|  |  |  |
| --- | --- | --- |
|  | Labeled Edible | Labeled Poisonous |
| Edible | 96 | 4 |
| Poisonous | 0 | 100 |

**Table 2:** Confusion table of Algorithm 2.

1. [2 p] Compute the accuracies of Algorithm 1 and Algorithm 2.

Accuracy of Algo1: 197/200 = 0.985

Accuracy of Algo2: 196/200 = 0.980

1. [2 p] For the task of identifying poisonous mushrooms, which algorithm is better? Explain your answer.

Algo 2 is better because it has not classified any poisonous mushroom as edible. This is contrary to algo1 which has higher accuracy but huge risk.

# Regression and MLE

## Question 5 [3 points]

*Multiple choice questions,* ***with short justiflcation*** *(max 5 lines). Indicate* ***all the correct choices****; there might be more than one correct choice per question.* ***No partial credit will be given.*** *All the correct answers should be selected.*

Your role as a machine learning engineer in a consulting firm is to use social media data of 100 million

(108) users to train a classification model to predict the binary election vote of each person, represented by *y* = ±1. In your solution, you decide to use regularized logistic regression with the following loss function:

108

min 1 Σ log 1 + exp −*y* w*T*X + *λ*ǁwǁ2*.*

w

108

*i*

*i*

2

*i*=1

Using cross-validation, you find the best regularization hyperparameter *λ*1. Later, you are informed that only 10 million of these voters consented to this experiment. Considering the ethical concerns raised, you decide to re-train your model using only 10 million people, and discard the rest. That way, following a similar methodology, you find the best hyperparameter *λ*2. Which of the following statements are true?

1. *λ*2 is expected to be greater than *λ*1.
2. *λ*2 is expected to be smaller than *λ*1.

3. *λ*2 ≈ *λ*1.

4. 10 × *λ*2 ≈ *λ*1.

5. None of the above.

## Question 6 [10 points]

Let {*yi, Xi*}*m* denotes a set of *m* observations, where each *Xi* is an *n*-dimensional vector. In *Ridge*

*i*=1

*Regression*, a regularization term is added in the linear regression model in order to penalize the model complexity, leading to the following optimization problem:

arg min ǁ**y** − **X***θ*ǁ2 + *λ*ǁ*θ*ǁ2*,*

2 2

*θ*

where *λ >* 0 is a regularization parameter.

1. [8 p] Find the closed form solution of the ridge regression problem. Text

   Description automatically generated
2. [2 p] Explain briefly why the ridge regression estimator is more robust to overfitting compared to the least-squares regression.

It is robust to overfitting because it penalizes large weights. This ensures the best fit with small weights

## Question 7 [13 points]

Let

*α −α*

tanh(*α*) = *eα* + *e−α* be the hyperbolic tangent function.

*e* − *e*

1. [3 p] Show that the tanh(·) function and the *logistic sigmoid function σ*(·) are related by:

tanh(*α*) = 2*σ*(2*α*) − 1*.*

Text, letter

Description automatically generated

1. [10 p] Show that a general (*M* -th order polynomial) linear combination of logistic sigmoid func- tions of the form

*M*

*y*(*x,* ***θ***) = *θ*0 +

Σ*j*=1

*θj σ*

*x* − *µj*

*s*

is equivalent to a linear combination of tanh(·) functions of the form

*y*(*x,* **u**) = *u*0 +

*M*

Σ*j*=1

*uj* tanh

*x* − *µj*

2*s*

and find expressions to relate the new parameters {*u*0*, . . . , uM* } to the original parameters {*θ*0*, . . . , θM* }.

A whiteboard with writing on it

Description automatically generated with low confidence

## Question 8 [10 points]

Consider the following probability distribution *Pθ*(*x*) = 2*θxe−θx* , where *θ* is a parameter and *x* is a positive real number. Suppose you get *m i.i.d.* samples *xi* drawn from this distribution. Show how one can compute the *maximum likelihood estimator* for *θ* based on these samples.

2

Text, letter

Description automatically generated

# Naive Bayes

## Question 9 [5 points]

Let’s consider a binary classification problem using the Naive Bayes classifier. You are given data points representated as two dimensional features (*X*1*, X*2). The categorical class conditional distributions *P* (*X*1 = *x*1|*Ci*) and *P* (*X*2 = *x*2|*Ci*) for *X*1 and *X*2 respectively, are given in the tables below. The two classes are equally likely.

|  |  |  |
| --- | --- | --- |
| *X*1 | *C*1 | *C*2 |
| -1 | 0.2 | 0.3 |
| 0 | 0.4 | 0.6 |
| 1 | 0.4 | 0.1 |

|  |  |  |
| --- | --- | --- |
| *X*2 | *C*1 | *C*2 |
| -1 | 0.4 | 0.1 |
| 0 | 0.5 | 0.3 |
| 1 | 0.1 | 0.6 |

Given a new data point (−1*,* 1), calculate the following posterior probabilities:

* *P* (*C*1|*X*1 = −1*, X*2 = 1)

A piece of paper with writing on it next to a keyboard

Description automatically generated with medium confidence

0.1

* *P* (*C*2|*X*1 = −1*, X*2 = 1)

0.9

# Regression in Practice

## Question 10 [20 points]

In this exercise you will need to use the *Amazon Gift Card* dataset ([https://s3.amazonaws.com/](https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Gift_Card_v1_00.tsv.gz) [amazon-reviews-pds/tsv/amazon\_reviews\_us\_Gift\_Card\_v1\_00.tsv.gz](https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Gift_Card_v1_00.tsv.gz)). The above is a tab-separated values dataset, which includes reviews from Amazon products. You can import the data using a .csv reader of Python. Using the dataset, you will need to answer the following questions. You can use the scikit-learn[1](#_bookmark0) library for your models. **Include only the basic parts of your code in the report – Python scripts will not be submitted**.

1. [2 p] What is the distribution of ratings in the dataset (e.g., number of 1-star, 2-star, 3-star (etc.) reviews)? Your answer can either be a table or a plot showing the distribution.

Text

Description automatically generated with medium confidence

1. [6 p] Now, we will train a simple *linear regression* model to predict the star rating of each review using just two features of the dataset:

star rating ' *θ*0 + *θ*1 × verified purchase + *θ*2 × length of review*,*

where the ’length of review’ is the number of words in the review (excluding ’,’ and ’.’). Report the values of *θ*0, *θ*1, and *θ*2 and briefly provide an interpretation of these values (i.e., what do they represent). Explain these in terms of the features and labels, e.g., if the coefficient of ’length of review’ is negative, what would that say about positive versus negative reviews?

Graphical user interface, text, application, email

Description automatically generated

Verified purchases is positively related to the ratings, while the length of the reviews is negatively linked to the ratings

1. [6 p] Split the dataset into two fractions: the first 90% of the dataset will be used for training, while the remaining 10% for testing (do not shuffle the dataset; use the order of the instances as they appear in the file). Now, train the same model as in question (b) on the training set, and report the model’s mean-square error on both the training and test sets.

Text

Description automatically generated

1. [6 p] Consider a similar prediction problem as above, where you can use all four numeric features of the dataset (’verified purchase’, ’length of review’, ’helpful votes’, ’total votes’). Try to obtain a more accurate model using *polynomial features*, as we have examined in the class[2](#_bookmark1). Give the expression of the feature vector you have designed and report the training and test accuracies.

Text

Description automatically generated Table

Description automatically generated

Poly 2 has better performance than 1st degree fit. The test error drops sharply. Poly 2 is the best fit for the data.

Please find file below, double click to open:

1[https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)

[html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)

2[https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html)

[html](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html)