FOUNDATIONS OF MACHINE LEARNING
MASTER IN DATA SCIENCES AND BUSINESS ANALYTICS
CENTRALESUPÉLEC

Assignment 1

Instructor: Fragkiskos Malliaros
TA: Hakim Benkirane and Konstantinos Florakis

Due: **November 12, 2023 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: PWR777). 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 1-8).

I. General Questions

Question 1 [10 points]

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

(a) Stochastic gradient descent performs less computation per update than gradient descent.

(b) Both PCA and linear regression can be thought of as algorithms for minimizing a sum of squared errors.

(c) Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be the matrix representation of our data. Let's assume that we project our data on the k-dimensional space using Principal Component Analysis, where k equals the rank of $\mathbf{A}$. Then, no loss is incurred in the reconstruction of the data.

(d) Let $y_i = \log(x^{\alpha_1} e^{\alpha_2}) + \epsilon_i$ be a model, where $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ corresponds to Gaussian noise. Then, the maximum likelihood parameters of the model (α) can be learned using linear regression.

(e) The eigenvectors of $\mathbf{A}\mathbf{A}^\top$ and $\mathbf{A}^\top\mathbf{A}$ are the same.

II. Dimensionality Reduction

Question 2 [10 points]

Let $\mathbf{M}_{m \times n}$ be a data matrix (m observations (i.e., data points), n dimensions (i.e., features)).

(a) [2 p] Are the matrices $\mathbf{M}\mathbf{M}^\top$ and $\mathbf{M}^\top\mathbf{M}$ symmetric, square and real? Justify your answer.

(b) [2 p] Show that the eigenvalues of $\mathbf{M}\mathbf{M}^\top$ are the same as the ones of $\mathbf{M}^\top\mathbf{M}$. Are their eigenvectors the same too? Justify your answer.

(c) [3 p] SVD decomposes the matrix $\mathbf{M}$ into the product $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and $\boldsymbol{\Sigma}$ is a diagonal matrix. Given that $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, write a simplified expression of $\mathbf{M}^\top\mathbf{M}$ in terms of $\mathbf{V}$, $\mathbf{V}^\top$ and $\boldsymbol{\Sigma}$. Can we find an analogous expression for $\mathbf{M}\mathbf{M}^\top$?

(d) [3 p] What is the relationship (if any) between the eigenvalues of $\mathbf{M}^\top\mathbf{M}$ and the singular values of $\mathbf{M}$? Justify your answer.

Question 3 [10 points]

As we have seen in the course, PCA projects data points from $\mathbf{x} \in \mathbb{R}^d$ to low-dimensional space defined by the k eigenvectors of the covariance matrix that correspond to the largest eigenvalues. Let $\mathbf{U}_k$ denote the $d \times k$ matrix of the top k eigenvectors of the covariance matrix ($\mathbf{U}_k$ is a truncated version of $\mathbf{U}$, which is the matrix of eigenvectors of the covariance matrix).
We have two ways to find the low-dimensional representation $\mathbf{w} \in \mathbb{R}^k$ of a data point $\mathbf{x} \in \mathbb{R}^d$:

1. Solve a least squares problem to minimize the reconstruction error.
2. Project $\mathbf{x}$ onto the span of the columns of $\mathbf{U}_k$.

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

(a) [5 p] Formulate the least squares problem in terms of $\mathbf{U}_k$, $\mathbf{x}$ and $\mathbf{w}$. (Hint: the optimization problem should resemble linear regression.)

(b) [5 p] Show that the solution of the least squares problem is equal to $\mathbf{U}_k^\top \mathbf{x}$, which is the projection of $\mathbf{x}$ onto the span of the columns of $\mathbf{U}_k$. (Hint: use the closed-form solution of the least-squares problem).

III. Model Evaluation, Regression, and MLE

Question 4 [15 points]

*Multiple choice questions, **with short justification** (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.*

(a) [5 p] Your role as a machine learning engineer in a consulting firm is to use social media data of 100 million (10^8) users to train a classification model to predict the binary election vote of each person, represented by $y = \pm 1$. In your solution, you decide to use regularized logistic regression with the following loss function:

$$\min_{\mathrm{w}} \frac{1}{10^8} \sum_{i=1}^{10^8} \log\left(1 + \exp\left(-y_i \mathrm{w}^\top \mathrm{X}_i\right)\right) + \lambda \|\mathrm{w}\|_2^2.$$

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. $\lambda_2 \approx \lambda_1$.
4. $10 \times \lambda_2 \approx \lambda_1$.
5. None of the above.

(b) [5 p] Consider a least-squares linear regression model. Which of the following will never negatively impact the training error (mean squared error)?

1. Using polynomial features
2. Using Ridge to reduce the model complexity by coefficient shrinkage
3. Using Lasso to encourage sparse coefficients
4. Normalizing the data points

(c) [5 p] Given a data matrix $X \in \mathbb{R}^{n \times d}$, labels Y, and $\lambda > 0$, we find the weighted vector $\mathbf{w}^*$ that minimizes $\|Y - X\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|^2$. Let's assume that $\mathbf{w}^* \neq 0$. Choose the correct answer(s).

1. The variance of the method decreases if λ increases enough
2. There might be multiple solutions for $\mathbf{w}^*$
3. The bias of the method decreases if λ increases enough
4. $\mathbf{w}^* = X^+ Y$, where X^+ is the pseudoinverse of X

Question 5 [10 points]

Let $\{y_i, X_i\}_{i=1}^m$ denotes a set of m observations, where each X_i is an n-dimensional vector. In *Ridge Regression*, a regularization term is added in the linear regression model in order to penalize the model complexity, leading to the following optimization problem:

$$\arg\min_{\theta} \|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \lambda\|\theta\|_2^2,$$

where $\lambda > 0$ is a regularization parameter.

(a) [8 p] Find the closed form solution of the ridge regression problem.

(b) [2 p] Explain briefly why the ridge regression estimator is more robust to overfitting compared to the least-squares regression.

Question 6 [15 points]

Let's consider n random variables $x_i, i \in [1, n]$ drawn independenty from a Bernoulli distribution with mean θ. Reminder: in a Bernoulli distribution $X \in \{0, 1\}$ and $p(X\theta) = \theta^x(1-\theta)^{1-x}$.

(a) [3 p] Express the likelihood function $L(\theta; X_1, \ldots, X_n)$.

(b) [5 p] Find the expression of the log-likelihood (show the steps of your solution in detail).

(c) [7 p] Prove that the expression of the Maximim Likelihood Estimate is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$

V. Naive Bayes

Question 7 [10 points]

We consider a problem where have data points, as shown in the matrix below, composed of four feature $X = (x_1, x_2, x_3, x_4)$ and three labels $y = \{+1, 0, -1\}$. Now, let's assume that $p(X, y)$ and $p(y)$ are both Bernoulli distributions.

x_1	x_2	x_3	x_4	y
1	1	0	1	$+1$
0	1	1	0	$+1$
1	0	1	1	0
0	1	1	1	0
0	1	0	0	-1
1	0	0	1	-1
0	0	1	1	-1

(a) [3 p] Fill in the table below with the MLE for $p(x_i = 1|y)$ for all different values of i and y.

	$y = +1$	$y = 0$	$y = -1$
$x_1 = 1$			
$x_2 = 1$			
$x_3 = 1$			
$x_4 = 1$			

(b) [1 p] Compute the MLE for $p(y = +1)$, $p(y = 0)$, and $p(y = -1)$.

(c) [6 p] Based on the values computed in the previous two sub-questions, clasify a new data point with feature values $(x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1)$ to one of the three classes.

VI. Regression in Practice

Question 8 [20 points]

In this exercise you will need to use the *GoodReads* dataset provided in the assignment. The above is a *.json* file, which includes reviews of fantasy novels from *Goodreads*. You can import the data using the following code or any other reader of Python.

```
path = dataDir + "fantasy_100.json"
f = open(path)
data = []
for l in f:
d = json.loads(l)
data.append(d)

f.close()
```

Using the dataset, you will need to answer the following questions. You can use the `scikit-learn`[1] library for your models. **Include only the basic parts of your code in the report – Python scripts will not be submitted**.

(a) [2 p] What is the distribution of ratings in this 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.

(b) [6 p] Now, we will train a simple *linear regression* model to predict the star rating of each review using only the review length:

$$\text{star rating} \simeq \theta_0 + \theta_1 \times (\text{review length in characters}),$$

where the 'review lenght in characters' is the number of characters in the review. Report the values of θ_0 and θ_1, and briefly provide an interpretation of these values (i.e., what do they represent). Also, compute the Mean Squared Error of your predictions.

(c) [6 p] Now, build a new model including a second feature based on the number of comments, i.e.,

$$\text{star rating} \simeq \theta_0 + \theta_1 \times (\text{review length}) + \theta_2 \times (\text{number of comments}).$$

Compute the new coefficients and the Mean Squared Error. Explain your observations (why θ_1 is different from the one of sub-question (b)).

(d) [6 p] Finally, try to obtain a more powerful model using *polynomial features*, as we have examined in the class[2]. Give the expression of the feature vector you have designed and the Mean Squared Error. Please explain your observations.

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

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