Assignment 2

Data Mining and Machine Learning (2IIG0)

Ghada Sokar

This is your second assignment of the Data Mining and Machine learning course. The assignment is composed of six questions, the last one is open-ended. Multiple-choice questions have four possible choices; among these, only one answer is correct. The number of points for each question is written at the beginning of it. The maximum achievable score is 100 points. In all the questions, you have to implement the algorithm yourself required in the question using Python. You are only allowed to use scikit-learn for some steps where it is explicitly stated in the questions.

Submit the answers for the first five questions in the **HW2** quiz in Canvas. You have to submit this quiz individually. That is, every group member has to fill out his/her own quiz, even when you are encouraged to work on the exercises together. You can submit the quiz multiple times, the latest one will be graded. Please fill this quiz out carefully.

The open question can be submitted on a group level by means of the **Open Question HW2** assignment in Canvas. That is, one of your group members should upload:

- A PDF document containing the answer to the open question. Include a paragraph about the workload distribution as well. We assume that you can decide yourself how to organize the work in your group. But if someone didn't contribute at all, or only marginally, then you should note this in the PDF.
- A Python file/ notebook containing the implementations you used to answer the implementation-based questions. This notebook is your insurance, it shows what you did and which results you got. If you want to argue for a higher grade/get partial points then you can back it up with the results stated in the uploaded notebook. However, if we see that you couldn't provide a working implementation, but just guessed the correct answer in the MC format, then we will also subtract the corresponding points from your grade.

In the first two questions, we will work with NO2 dataset, collected by the Norwegian Public Roads Administration and is available at StatLib. The dataset was originated in a study where air pollution at a road is related to traffic volume and meteorological variables. In the Assignment .zip file, you can find a modified version of this dataset, named *data/no2_dataset.csv*, in which we consider three attributes from the original ones. The three considered attributes are cars per hour, wind speed, and wind direction. In the following two exercises, we will explore different regression models to fit this data.

1. (20 points) In this exercise, we would like to build a regression model to predict the NO2 concentration. We will fit the regression model with an affine function. Therefore, you are asked to implement a function that returns the regression parameter β which minimizes the RSS. You can assume here that the matrix $X^T X$ has an inverse. To this end:

   1. Import the data in Python.
   2. Explore the data. You can visualize the relationship between each attribute and NO2 concentration. Therefore, make three plots; each has the data points represented by one attribute (feature) and the target y (NO2).
   3. Split a separate test set using *train_test_split* function in the *model_selection* module of scikit-learn, with $random\_state = 10$ and $test\_size = 0.2$.
   4. Construct the design matrix X using ϕ_{aff} as explained in the lecture.
   5. Implement a function named *regression* that takes the design matrix X and target vector y (NO2) and returns the regression parameter β which minimizes the RSS.
   6. Analyze how well the model fit the data. Therefore, repeat step 2 (visualization), but now plot the relationship between each attribute and the corresponding predictions $\hat{y}$ from the regression model according to the regression parameter β.
   7. Calculate the MSE on the training and test splits.

   Which of the following statements is **correct**?

   A. The prediction f(cars_per_hours, wind_speed, wind_direction) is most susceptible to noise in the feature wind_direction. 如何判断？

   B. The predicted model suffers from overfitting.

   C. The MSE on the test set is 0.327 (rounded).

   D. The value of the (rounded) intercept β_0 is 0.368.

2. (10 points) Now, we would like to fit the NO2 dataset using a regression model with polynomial function. Therefore, you have to transform the data using ϕ_{pk} to construct the design matrix X as discussed in the lecture. **For this step only**, you can use the *PolynomialFeatures* function in the *preprocessing* module of scikit-learn to construct the design matrix X. The function takes the degree of the polynomial k.

   Use the *regression* function you implemented in the previous question to get the regression parameters β that minimizes the RSS, given the design matrix X and the target y.

   Try different values for the degree of the polynomial k and analyze how the model behaves. Visualize the relationship between the predicted output $\hat{y}$ and each feature.

   Which of the following statements is **correct**?

   degree 2应该是最理想的回归拟合，那么degree = 3肯定不是underfitting

   overfitting和underfitting只能看图得到吗

   A. The polynomial regression model of degree 3 suffers from underfitting.

   B. The polynomial regression model of degree 4 is a good fit for the data.

   0.29573719515366503

   C. The MSE on the test set by polynomial regression of degree 2 ($k = 2$) is less than the MSE on the test set by polynomial regression of degree 3. 0.3515826550364887

   D. The predicted NO2 by the polynomial regression model of degree 2 is 3.171 for the following attributes: cars_per_hour = 6, wind_speed = 3, and wind_direction = 100. 3.186

   Random因子一样，那么划分的训练集也是一样的？

看Training Curve 以及 Validation Curve，在其他条件理想的情况下，如果Training Accuracy 高， Validation Accuracy 低，也就是过拟合 了，可以尝试去减少层数或者参数。如果Training Accuracy 低，说明模型学的不好，可以尝试增加参数或者层数。至于是增加长度和宽度，这个又要根据实际情况来考虑了。

PPT: Lecture 5, p42–47

3. (10 points) Imagine you train three regression models $f(x) = \beta_1 x + \beta_0$ on three independently sampled data sets $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$:

所以说，三个数据集是从同一个数据集中独立取样获得的？

$$f_{\mathcal{D}_1}(x) = -2x + 0.6$$
$$f_{\mathcal{D}_2}(x) = -1.4x + 0.4$$
$$f_{\mathcal{D}_3}(x) = -2x + 0.8$$

Use the three regression models to compute a sample mean estimate of the expected value $\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)]$. Assuming that $f^*(x) = -x + 1$, report an estimate of the:

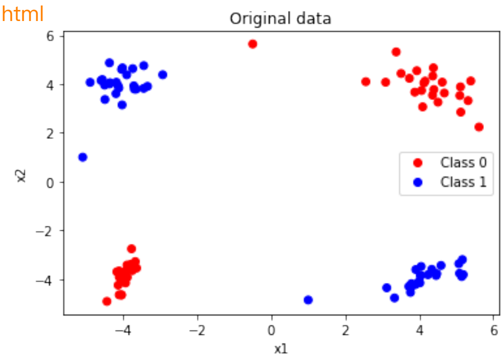
f* 完美拟合方程

1. $bias^2(x) = (f^*(x) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)])^2$
2. $variance(x) = \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)] - f_{\mathcal{D}}(x))^2]$

for $x = 2$. Round the result to one decimal after the point (if necessary).

4. (10 points) This exercise is about the theoretical understanding of multilayer perceptron network. Consider the two-classes dataset shown in the Figure below. The data has two features (x_1, x_2). Suppose that we want to fit this data using multi-layer feed-forward neural networks. Which of the following statements is **correct**?

https://www.cnblogs.com/fanghao/p/7533385.html
https://www.cnblogs.com/weijiazheng/p/10910139.html



层数越多，过拟合可能性越大
Neuron越多，过拟合可能性越大

线性激活只能得到线性组合的结果，那么多层的效果其实用一层就可以实现，例如第一层是output = a*input, 第二层是output = b*input。那么第一层实际上就可以实现第二层的功能，即output = a*b*input

尽量达到最大的overfitting

This dataset can be modeled with zero training error using:

线性不可分割

~~A.~~ A neural network with a single hidden layer and linear activation.

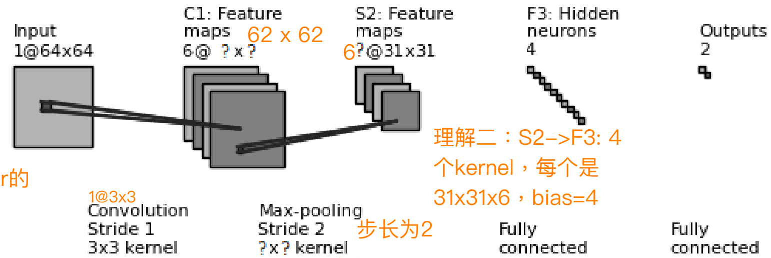
B. A neural network with a single hidden layer and non-linear activation.

~~C.~~ A neural network with at least two hidden layers and linear activation.

D. The data could not be modeled with zero training error using multilayer perception network.

5. (10 points) This exercise is about the theoretical understanding of convolutional neural networks. Consider the figure below which represents a convolutional neural network that converts a 64x64 image into 2 output values. The network consists of the following layers from input to output: convolution, max pooling, and two fully-connected layers. The notation 1@64x64 means 1 channel of size 64x64. No padding is used.

理解一：将S2中的数据作为input，平铺到4个神经元中，因为是fully connected, 每个neuron获得所有数据，并且每个neuron还有一个bias



理解二：S2–>F3: 4个kernel，每个是31x31x6，bias=4

从F3到output:
四个neuron到2个neuron，全连接
4 * 2 + 2 = 10

kernel的channel数 = 当前layer的channel数，对于input layer，channel = 1

pooling不会导致channel数量的变化
Pooling也不需要parameter，因为它不是用w和b来计算，而是直接滑动筛选

Which of the following statements is **correct**?

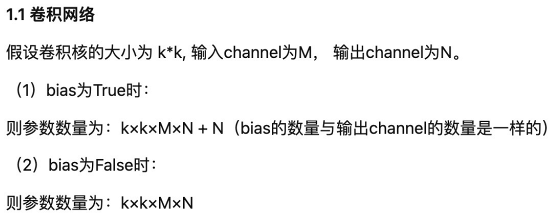
A. The width of the kernel in layer S2 is ~~3~~. 2 62 / 31 = 2

~~B.~~ The number of the feature maps in layer S2 is ~~3.~~ 6

C. The number of parameters (weights, bias) in layer F3 is 23068.

6x31x31x4 + 4 = 23068, 每一个neuron一个bias，所以bias = 4

D. The total number of network parameters (weights, bias) is ~~23136.~~



1.1 卷积网络

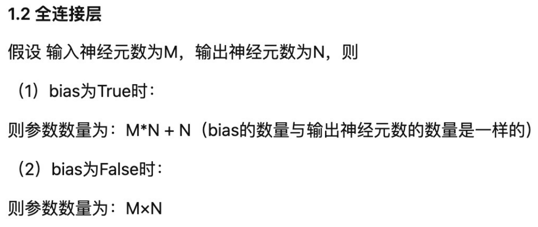
假设卷积核的大小为 k*k, 输入channel为M， 输出channel为N。

（1）bias为True时：

则参数数量为：k×k×M×N + N（bias的数量与输出channel的数量是一样的）

（2）bias为False时：

则参数数量为：k×k×M×N



1.2 全连接层

假设 输入神经元数为M，输出神经元数为N，则

（1）bias为True时：

则参数数量为：M*N + N（bias的数量与输出神经元数的数量是一样的）

（2）bias为False时：

则参数数量为：M×N

6 x (1 x 3 x 3 + 1)
+ 0
+ 4 x (6 x 31 x 31 + 1)
+ 2 x (4 + 1)
= 23138

Input –> C1
C1 –> S2 pooling 不需要parameter
S2 –> F3
F3 –> output

6. (40 points) **Open Question** In this exercise, you are asked to implement a Multi Layer Perceptrons (MLPs) network **from scratch**. In the assignment .zip file, you can find Python notebook (MLP.ipynb) that contains a skeleton for parts of the code. You may add as many input parameters as you need to any of the existing function prototypes. You can also find a small two-classes dataset divided into two *.csv* files: *data/train_data.csv* and *data/validate_data.csv*. Each row represents a data point. The first two columns (x_1, x_2) represent the two input features. The third column (y) represents the class label (0 or 1).

We would like to fit this data using MLP network. The network consists of 2 hidden layers. Each layer consists of 10 neurons. The network should be trained using backpropagation and mini-batches Stochastic Gradient Descent (SGD) as explained in the lecture. To this end:

(a) Import the data in Python and explore it a bit.

(b) (5 points) Perform data normalization. Choose one of the data normalization methods discussed in the lecture (e.g. Zero-mean-unit-variance, Min-Max normalization, etc.) and apply it on the data. Report the formula that you used in the report and discuss your choice. Include in your report the first 4 samples of the training data after transformation.

size of input layer
= #(features) = 2

output: for 2 variable classification, 1 output is enough

(c) (1 points) Determine the size of the input and output layers. Report your answer. ppt7, p22

loss function + activation function:
https://cloud.tencent.com/developer/article/1518599

(d) (2 points) Choose a suitable loss function and activation function for the first and second hidden layers ϕ_1, ϕ_2. Report and justify your choice.

如果使用RE，那么output必须先用softmax处理。
如果使用MSE，那么可以不加softmax，也可以加softmax

(e) (2 points) Choose an initialization of the parameter values. Report and discuss your choice.

(f) (10 points) Implement the MLP network from scratch with the specification stated above and the backpropagation algorithm to train the network. We will run and check the uploaded Python code. To obtain the points for this subproblem: 1) the Python code has to run with **no errors** and 2) the MLP model and the Backpropagation algorithm have to be implemented completely from scratch. **Note** that you are not allowed to use any library which implements MLP models, but you are allowed to use auxiliary libraries (e.g. Numpy, Matplotlib, Pandas).

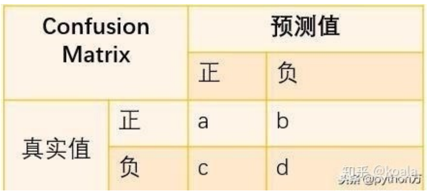
(g) (10 points) Tune the hyperparameters: learning rate, batch size, number of epochs. Try to achieve the best validation accuracy you can. In this subproblem, part of your grade depends on how well your model performs. You should at least get 70% on the validation set to obtain points greater than 0. You can get the full points when you are close to the maximal performance, ~ 97%, on the validation set (i.e. $\geq$ 90% should be fine to obtain the full points). Report the best values for each hyperparameter.

maybe: validation set的cross entropy最小的时候
作CE–epoch关系图，找到CE最小的epoch作为stopping criteria
或者是validation accuracy–epoch关系图，在accuracy最大时的epoch作为stopping criteria

(h) (5 points) Plot the loss computed over the training set and over the validation set. In addition, plot the classification accuracy computed over the two sets. Choose a stopping criteria for your training. State and justify your stopping criteria. Include your answer and the plots in your report.

也可以用梯度下降的方法来判断终止，当梯度变化量小于某个阈值，就停止

(i) (5 points) Report the final accuracy obtained and the confusion matrix on the training and validation sets.



Confusion Matrix		预测值	
		正	负
真实值	正	a	b
	负	c	d