ITU Computer and Informatics Faculty
BLG 454E Learning From Data, Spring 2018
Homework #2
Due April 10, 2018 11pm

1. **[Multivariate Analysis]**

   You will use the dataset *dataTrain.csv* and *dataTest.csv* given for question 1. The last column of the file represents the class label (class 0 or class 1 or class 2):

   dataTrain.csv: Training data.

   dataTest.csv: Test data.

   Suppose that each class i's ($i = 0\ ,\ 1,\ 2$) inputs, x_i, is distributed according to normal distribution $N(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$ in dataset that is given to you. Class conditional density $p(x|C_i)$ is $N(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$:

$$p(x|C_i) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_i|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i{}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)\right) \tag{1}$$

   You are going to introduce the **discriminant functions** $g_i(x) = P(x|C_i).P(C_i)$ for each class according to the case below:
   General case $\Sigma_i = \Sigma$ are arbitrary. Meaning that features x_1, x_2 are not necessarily independent. Thus the discriminant function and decision boundaries are quadratic. Therefore you should find a seperate Σ_i covariance matrix for each class.

   (a) (20 pts) Formulate and implement g(x) discriminant function clearly (add comments) in your code and write its formula into the report.
   Hint: You are required to calculate the mean and covariance matrix for each class.

   (b) (20 pts) Draw the decision boundaries for each classifier in part(a) for training set is similar to Figure 1 and report it.
   Hint: Line equations can be found by $g_i(x) = g_j(x)$ for $i, j = 1, 2, 3..$ You can use the following example in order to plot decision boundary:
   `http://scikit-learn.org/stable/auto_examples/linear_model/plot_iris_logistic.html`

   (c) (10 pts) Calculate test accuracy and write it into the report.
   Hint : Do not recompute the discriminant functions for *testData.csv set*, just reuse the ones you computed for the *trainData.csv* set.

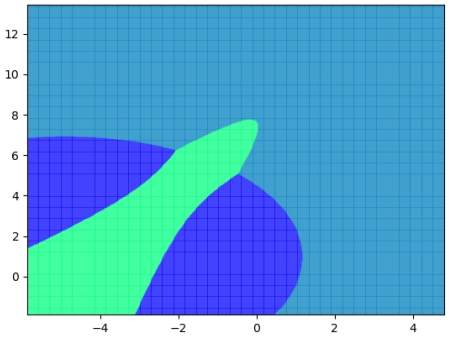

Figure 1: Decision boundaries

Purpose : We want you to show that decision boundaries are nonlinear for arbitrary Σ_i.

2. **[Logistic Regression]**

   You will use the following dataset(iris.data) for question 2. You use the whole dataset therefore you have only training set in this question. Please check the website carefully:

   https://archive.ics.uci.edu/ml/datasets/iris

   Logistic regression produces the sigmoidal function that best describes the given data. Implement logistic regression classifier and write its code clearly(add comments). Classify given dataset using logistic regression.

   Hint: You may use pseudo codes in Ethem Alpaydins's book(Figure 2) for logistic regression classifier. You are not allowed use logistic regression built-in function.

   (a) (25 pts) Calculate accuracy and confusion matrix using 10 fold cross validation and write them into the report. Which classes are most confused with each other?

   (b) (25 pts) Analyze the effect of learning rate(η). Use 10 fold cross validation and compare your results with different learning rates are 10, 1, 0.1, 0.01 based on the number of iterations and classification accuracy.

For $i = 1, \ldots, K$, For $j = 0, \ldots, d$, $w_{ij} \leftarrow \text{rand}(-0.01, 0.01)$
Repeat
  For $i = 1, \ldots, K$, For $j = 0, \ldots, d$, $\Delta w_{ij} \leftarrow 0$
  For $t = 1, \ldots, N$
    For $i = 1, \ldots, K$
      $o_i \leftarrow 0$
      For $j = 0, \ldots, d$
        $o_i \leftarrow o_i + w_{ij} x_j^t$
    For $i = 1, \ldots, K$
      $y_i \leftarrow \exp(o_i) / \sum_k \exp(o_k)$
    For $i = 1, \ldots, K$
      For $j = 0, \ldots, d$
        $\Delta w_{ij} \leftarrow \Delta w_{ij} + (r_i^t - y_i) x_j^t$
  For $i = 1, \ldots, K$
    For $j = 0, \ldots, d$
      $w_{ij} \leftarrow w_{ij} + \eta \Delta w_{ij}$
Until convergence

Figure 2: Logistic discrimination algorithm implementing gradient descent for the case with $K > 2$ classes. For generality, we take $x_0^t = 1, \forall t.$. (Alpaydin, E., 2014. Introduction to machine learning. MIT press.)

Submission Policy

- Prepare the report and code. Only electronic submissions through Ninova will be accepted no later than April, 10 at 11pm.
- You may discuss the problems at an abstract level with your classmates, but you should not **share or copy code** from your classmates or from the Internet. You should submit your **own, individual** homework.
- Note that your codes and reports will be checked with the plagiarism tools **including previous years submissions**!
- Academic dishonesty, including cheating, plagiarism, and direct copying, is unacceptable.
- If a question is not clear, please let the teaching assistants know by email (**cebeci16@itu.edu.tr**).

Bonus marks (10pts)

- Clarity and nicely described report
- Using Latex template for the report

Deductions (-10pts)

- Spelling errors.
- Messiness
- Lack of content.
- Irrelevant / mistaken content.