მანქანური სწავლების მონაცემთა დამუშავების ალგორითმი

გურამი ასანიშვილი

თ.ს.უ ზუსტ და საბუნებისმეტყველო მეცნიერებათა

ფაკულტეტის დოქტორანტი

ხელმძღვანელი: მანანა ხაჩიძე

თბილისი 2018

აბსტრაქტი

მანქანური სწავლების დანიშნულებაა რაიმე სისტემის შემავალი და გამომავალი სიდედეების, მდგომარეობების დაკვირვებათა ნაკრებიდან შედგეს ამ სისტემის მოდელი, რომლის საშუალებითაც შესაძლებელი იქნება სხვდასხვა ამოცანების გადაწყვეტა. ინტელექტუალური სისტემები ეფუძვნება მანქანურ სწავლებას აქედან გამომდინარე უდიდესი მნიშვნელობა აქვს შესაბამისი მოდელის აგებას. ასეთი ტიპის სისტემები შეიძლება გათვლილი იყოს მხოლოდ კონკრეტული ტიპის ინფორმაციაზე. არსებობს უამრავი სხვადასხვა ტიპის ალგორითმი, რომლებიც გამოიყენება ასეთ სისტემებში, მაგ: კლასიფიკაციის, კლასტერიზაციის ტიპის ალგორითმები. როგორც წესი მათი ცალ ცალკე გამოყენება ვერ იძლევა ეფექტურ შედეგს, აქედან გამომდინარე საჭირო ხდება ასეთი ტიპის ალგორითმების დაჯგუფება, ერთმანეთში კომბინაცია.ინტელექტუალური სისტემა შეიძლება მრავალნაირი იყოს, აქედან გამომდინარე სრულიად შესაძლებელია შემავალი ინფორმაცია იყოს ნებისმიერი სახით წარმოდგენილი. ინფორმაციის ტიპები მრავალნაირია, შესაბამისად თავდაპირველად მოწოდებული შემავალი ინფორმაცია შეიძლება იყოს სხვადასხვა ტიპის: ტექსტური , რიცხვითი და ა.შ. თუმცა ინტელექტუალური ტიპის სისტემა ძირითადად ემყარება ორი ტიპის ინფორმაციას, რომელთაგან პირველია რაოდენობრივი ხოლო მეორე ხარისხობრივი, ხოლო ყოველი ინფორმაცია საჭიროებს დამუშავებას და შედეგის მიღებას. სწორედ ამიტომ, მანქანურ სწავლებაში გამოიყენება სხვადასხვა დანიშნულების ალგორითმები, რომელიც საშუალებას გვაძლევენ ეფექტურად დავამუშავოთ ინფორმაცია, რათა მოხდეს მათი სწორედ წარმოდგენა და გაანალიზება.რეფერატში განხილულია ერთ-ერთი ესეთი მანქნური სწავლების მონაცემთა დამუშავების ალგორითმი.

Abstract

The main purpose of machine learning is to create a model from input and output quantities, state of training set which will be used to solve some kind of problems Intelligence systems are based on machine learning Therefore it's important to create an adequate model.Such kind of systems might be based on some kind of information. There exists a lot of different algorithms which are used for such systems, for example: Classification, Clustering. As usual if they are used separately, it does not affect high performance. It appears from this we have to combine/merge such kind of algorithms In any case it's required to define these important questions: Input and output information, the main processes and principles of whole system. It's important to define input information adequately for every used algorithm. There exists many kind of intelligence system, therefore input information might be defined in any format. There exists many kind of information, therefore it may be: String based, Number based and etc. As usual intelligence systems are based on two kinds of information: Quantitative and Qualitative. for this machine learning to use several types of algorithm that help us process data and then use them to more

quickly analyze them. in this lecture one of the algorithms of machine learning which helps us to process the data will be discussed.

გასაღები სიტყვები: მანქანური სწავლება, მონაცემთა დამუშავება, ალგორითმები, მონაცემთა დამუშავების ალგორითმები.

Keywords: Machine learning, Data processing, Algorithms , Data processing algorithms.

შესავალი

რა არის მანქანური სწავლება?

მანქანური სწავლება ეს არის კომპიუტერულ მეცნიერებათა დარგი რომელიც შეისწავლის თუ როგორ უნდა შეისწავლოს მანქანამ რაიმე, მონაცემებზე დაყრდნობით. უფრო ზუსტად რომ ვთქვათ მანქანამ პროგრამისტის ჩარევის გარეშე მასში შემავალი მონაცემების გაანალიზების შემდეგად უნდა მოგვცეს შედეგი და გარკვეული ინფორმაცია. მაგალითად თანამედროვე სისტემებში სპამის ფილრაციისთვის გამოიყენემა მანქანური სწავლება, როდესაც მომხმარებელი ამა თუ იმ შეტყობინებას ან ელ.ფოსტას ნიშნავს როგორც სპამს, პროგრამა იმახსოვრებს და ინიშნავს ყველა იმ თავისებურებას რომელიც შეტყობინებაში იყო, რომ შემდგომ იგივე ან მსგავსი სიტუაციის დროს მოიქცეს შესაბამისად: აღიქვას შეტყობინება როგორც შესაბამისი მაილი თუ აღნიშნოს იგი როგორც სპამი და არ მიიტანოს ადრესატამდე.

არსებული მდგომარეობა

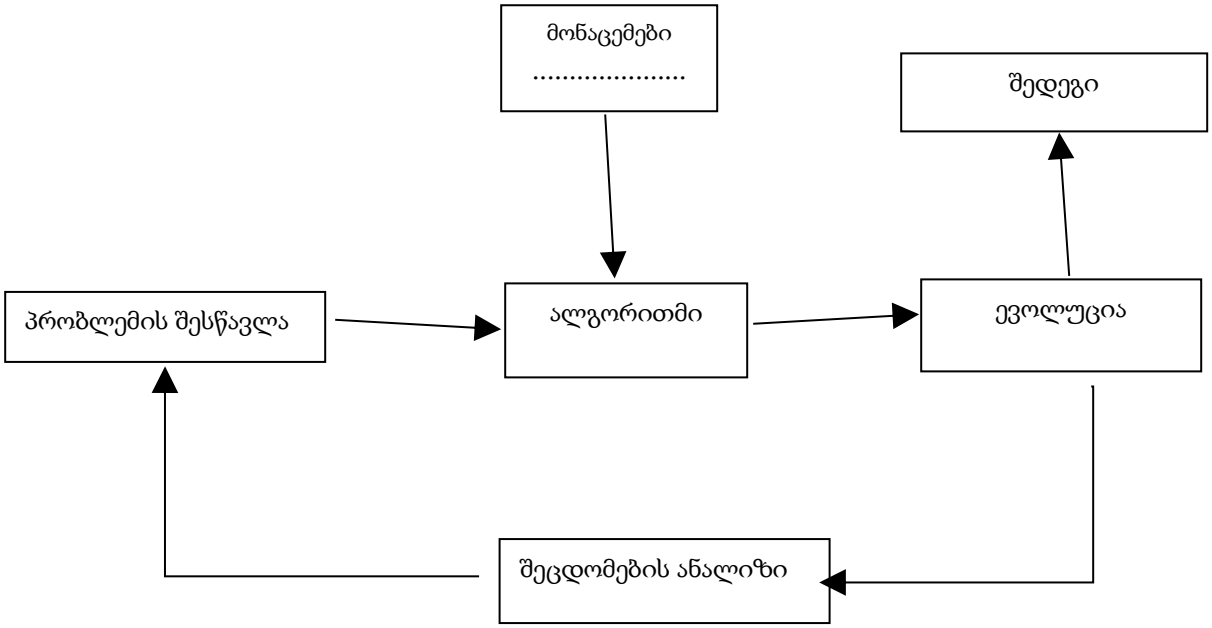
დღესდღეისობით მაქანური სწავლება , მისი მიდგომები და ალგორითმები ფართოდ გამოიყენება თითქმის ყველა სფეროში. ყველაზე ხშირად კი:

- მონაცემთა უსაფრთხოება: მონაცემთა უსაფრთხოება ერთ-ერთ ყველაზე დიდი გამოწვევაა თანამედროვე სისტემეში. ბოლო ხანებში კარსპერსკის ლაბორატორიე აცხადებს რომ ყოველდღიურად დაახლოებით 300 000 მავნე პროგრამას ან პროგრამულ კოდს. შესაბამისად ამ რაოდენობის ფაილების დამუშევება დიდ შრომას მოითხოვს. პრობლემის გადაწყვეტა მოიძებნა მანქანურ სწავლებაში რომლის ალგორითმი 2-10% შუალედში პოულობს და ადგენს ფაილის საფრთხის დონეს, რაც კარგი შედეგია.
- პერსონალური უსაფრთხოება: მანქანური სწავლება პერსონალურ უსაფრთხოებაში გამოიყენბა ისეთი სისტემეში რომლებიც ახდენენ ადამიანისთვის საშიშ საფრთხის

პრევენციას და დეტექციას.. მაგალითად: ტრანსპორტის მართვის დეტექტორები, მასობრივი თავშეყრის ადგილების უსაფრთხოება და ა.შ

- საფინანსო სისტემები: მრავალი ადამიანი ცდილობს მოახერხოს საფონდო ბირჟაზე არსებული სიტუაცია. მანქანური სწავლების გამოყენებით საფონდო ბირჟებზე მოვაჭრე კომპანიები ცდილობენ მაქსიმალურად გაზარდონ მოგების მოცულობა მინიმალური რისკის და დანახარჯების ქვეშ.
- ჯანდაცვაში: ჯანდაცვის სფეროში მომუშავე ალგორითმებს შესწევთ შესაძლებლობა დიდი სიზუსტით მოახდინონ დაავადებების დიაგნოსტიკა და შესაბამისი მკურნალობის დანიშვნა. ასევე მანქანური სწავლების ალგორითმების გამოყენებით შესაძლებელია ეპიდემიების რისკი შემცირება მათი დროულად დეტექციის ხარჯზე.
- ონლაინ ძებნა: ყველაზე მაშტაბური გამოყენაბა მანქანური სწავლებს ალგორითმებს ალბათ ონლაინ ძებნის სისტემებში აქვთ, ისეთი კომპანიები როგორიცაა Google,Yahoo და სხვა კომპანიები აქტიურად ნერგავენ ამ ალგორითმებს რათა მომხმარებლებისთვის ონლაინში ინფორმაციის ძებნის პროცესი უფრო მოსახერხებელი და ზუსტი გახადონ.
- თაღლითობების დეტექციაში: მანქანური სწავლება ასევე აქტიურად გამოყენება თაღლითობების აღმოჩენაში.მაგალითად: კომპანია PayPal იყენებს მანქანური სწავლების ალგორითმს ფულის გათეთრებისგან თავის დაზღვევითვის.
- მარკეტინგში: რაც უფრო კარგად იცნობთ თქვენს კლიენტს მით უფრო უკეთესეად შეძლებთ მის მომსახურეობას. ეს მარკეტიგის ძირითადი მიდგომაა. მანქანური სწავლების ალგორითმები მარკეტინგში გამოიყენება ისეთი ფაქტორების შესაწავლად როგორიცაა: მომხმარებლების სურვილების უპირატესობები, იმ პროდუქტების განსაზღვრისთვის რომლებიც ხშირად იყიდება და ა.შ

მანქანური სწავლების ალგორითმები ზოგადად მოქმედებენ შემდეგი სქემით:



სისტემური მოდელი და ამოცანის დასმა

იქამდე სანამ გამოჩნდებოდა "ინტელექტუალური" პროგრამები, მრავალი სისტემა მონაცემთა დამუშავებისთვის ან ინფორმაციის კორექტირებისთვის გამოიყენებდა მკაცრ „თუ-მაშინ" წესს,რაც არც თუ ისე მოსახერხებელია. ამასთან ისიც გასათვალისწინებელია რომ მსგავს მკაცრ მიდომას აქვს თავისი უარყოფითი მხარეები:

- ლოგიკა,რომელიც აუცილებელია გადაწყვეტილების მიღებისთვის მიეკუთვნება მხოლოდ ერთ კონკრეტულ არეს და ამოცანას. სისტემაში მცირედმა ცვლილემაც კი შესაიძლება გამოიწვიოს სისტემის მთლიად გადაწერა.
- წესების შემუშავება საჭიროებს პროცესების ღრმა ცოდნას.

ერთ-ერთი მაგალითი რომლის მიხედვითაც მკაცრი მიდგომა ყოვეთვის განიცდის კრახს , ეს არის სახის ამოცნობს სისტემა. დღესდღეისობით თითქმის ყველა სმატფონს შეუძლია სახის ამოცნობა სურათის მიხედვით.ძირითადი პრობლემა მდგომაობს იმაში , რომ პიქსელები რომლითაც ფორმირდება სურათი ძალიან განსხვავდება ადამიანის აღქმისგან. ეს წესი ადამიანს არ აძლევს საშუალებას მოარგოს შესაბამისი წესების მიმდევრობა ციფრულ სურათს. თუმცა მანქანური სწავლების მეთოდების და მიდგომების გამოყენებით ეს მაინც შესაძლებელი გახდა, შესაბამისი ალგორითმი ადვილად უმკლავდება ამ პრობლემას.

მანქანურ სწავლებაში შედარებით წარმატებულ შეიძლება ჩაითვალოს ისეთი ალგორითმები რომლეიც ახდნენ პროცესის ავტომატიზაციას და ყოველი ახლის შესასწავლად იყენებენ ძველ გამოცდილებას. ეს ალგორითმები ცნობილია როგორც მასწავლებლით შესწავლადი ან კონტროლირებადი ალგორითმები. მათ შესწევთ უნარი მოგვცეს პასუხი ობიეტის შესახებ მაშინაც კი როდესაც ობიექტი მისთვის უცნობია.

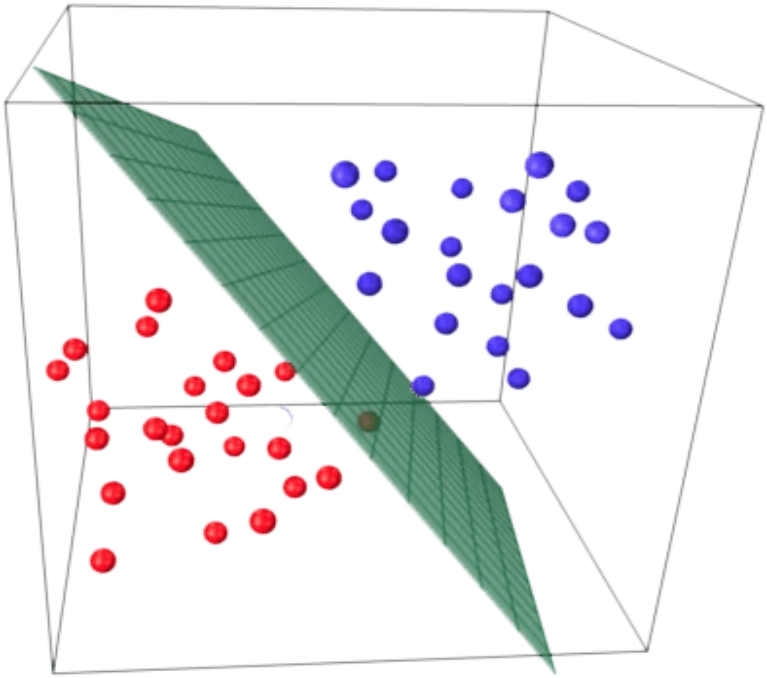
მასწავლებლით შესწავლადი ალგორითმები ძირითადად იყოფა ორ ტიპად: კლასიფიკაცია და რეგრესია. კლასიფიკაციის მიზანს წარმოადგენს მოახდინოს კლასის ნიშნის პროგნოზირება, რომელიც თავის თავში მოიცავს წინასწარ განსაზღვრულ შესაძლო ვარიანტების სიას, ხოლო რეგრესის მიზანს წარმოადგენს იპროგნოზიროს უწყვეტი რიცხვი ან რიცხვი მცურავი წერტილებით. მაგალითად, წინასწარ განისაზღვროს პიროვნების წლიური შემოსავალი მისი განათლების, ადგლიმდებარეობის და ასაკის მიხედვით. მსაგავსი მეთოდი შეიძლება გამოვიყენოთ წლიური მოსავლის პროგნოზირებისთვია და ა.შ. ყველაზე ადვილი განისაზღვროს ამოცანა კლასიფიკაციის არის თუ რეგრესიის, უნდა დაისვას კითხვა: თუ პასუხის მიღებაში შეიმჩნევა გარკვეული უწყვეტობა მაშინ ამოცანა მიეკუთვნება რეგრესიას.

ანალიზი

ლოგიკური რეგრესია

მანქანურ სწავლებაში რეგრესიების ამოცანებში ფართოდ გამოიყენება „ლოგიკური რეგრესიის“ ალგორითმი. ალგორითმის ძირითადი აზრი მდგომარეობს იწინასწარმეტყველოს გარკვეული მოვლენები. ჩვეულებრივი რეგრესიისგან განსხვავებით „ლოგიკური რეგრესიის“ მეთოდი არ ახდენს პროგნოზირებას გავალი მონაცემების შერჩევის საფუძველზე, არამედ გამომავალი მონაცემის შედეგი დამოკიდებულია იმაზე რომ ისითი მიეკუთვნებიან გარკვეულ კლასს. მაგალითად, დავიშვათ რომ გვაქვს ორი კლასი და ალბათობა რომელიც უნდა განისაზღვროს. P_+ ალბათობა რომ ზოგიერთი მნიშვნელობები ეკუთვნის „+“ კლასს და რა თქმაუნდა $P_- = 1 - P_+$, ამის შედეგად ლოგიკური რეგრესიის შედეგი ყოველთვის იქნება [0,1] იტერვალში.

ლოგიკური რეგრესიის ძირითადი იდეა მდგომარეობს იმაში ,რომ გამომავალი მონაცემების სივრცე შეიძლება გაოყოს წრფივი საზღვრით ორი სხვადასხვა კლასის შესაბამისად. რა იგულისხმება წრფივი საზღვრის ქვეშ? ორ განზომილებიანი



დავუშვათ გვაქვს მონაცემები y_i, x_i , სადაც $x_i \in R^n$ ანუ იგი წარმოადგენს n განზომილებიან ვექტორს სივრცეში, y_i=0,1. აუცილებელია შეიქმნას მოდელი y მნიშვნელობის განსაზღვრება R^n სივრცეში, სხვა სიტყვებით რომ ვთქვათ უნდა აღდგეს გამომავლი მონაცემებით აღვადგინოთ ალბათური განაწილება P_y(x) , თუ ცლობილია რომ $P_y = 0$(x) , მაშინ $P_{y-1} = 1$-$P_y = 0$(x), ასეთ შემთხვევაში y მიიღებს მნიშვნელობას 0 ან 1 . შემდეგ ნაბიჯის წარმოადგენს $P_y = 0$(x) -თვის გაყოფის აღდგენის მოდელის შერჩევაში. ლოგიკური ზრდა f(x) = $\frac{1}{1+e^{-x}}$

თუმცა საჭიროა ფუნქციისთვის მოქნილობის დამატება, რომელიც ფუნქციაც ცოტათი ფორმას შეუცვლის, ამიტომაც სრულიად ბუნებრივია მოდელში პარამეტრების დამეტება (w_1,… w_n, c) მოდელისთვის $P_y = 0$(x)= f(x) = $\frac{1}{1+e^{w^t x+c}}$, P_{y-1} =1- f(x) , სადაც $w^t + c = w_1 * x_1 + w_2 * x_2 + \dots w_n * x_n$+C მონაცემებზე დაყრდნობიდან გამომდინარე. ასეთ ამოცანებში პარამეტრების შეფასების დროს გამოყენება მეთოდი რომლის დროსაც იარსებებეს მაქსიმალური ალბათობა, ასეთ შემთხვევაში მოვლენები რომლებიც მოხდება ამოცანის დასრულების შემდეგ უფრო მეტად სამართლიანი და სწორი იქნება. არათანაბარი განზომილებების შემთხვავში პირობა უნდა ვრცელდებოდეს ყველა მონაცემზე.

ლოგიკური რეგრესია იძლევა საშუალებას შეფასდეს ალბათობა მონაცემების ორ კლასად გამიჯვნის შემთხვევაში თუ რომელ კლასს მიაკუთვნება ესა თუ ის მონაცემი. მსგავსი შეფასებებზე სქემის აგება არ წარმოადგენს სირთულეს:

1. შეირჩეს ის კლასი სადაც სადაც ალბათობა იქნება 0,5-ზე დიდი
2. ლოგიკური რეგრესიის დახმარებით შეფასდეს მონაცემების კონკრეტულ კლასზე მიკუთვნების ალბათობა.

სიმულაცია და ექსპერიმენტაცია

განვიხილოთ მაგალითი ფაქტორიალური სივრცისორი კლასის წერტილების კუთვნილების ალბათობა სპეციალურად გენერირებული მონაცემებით.

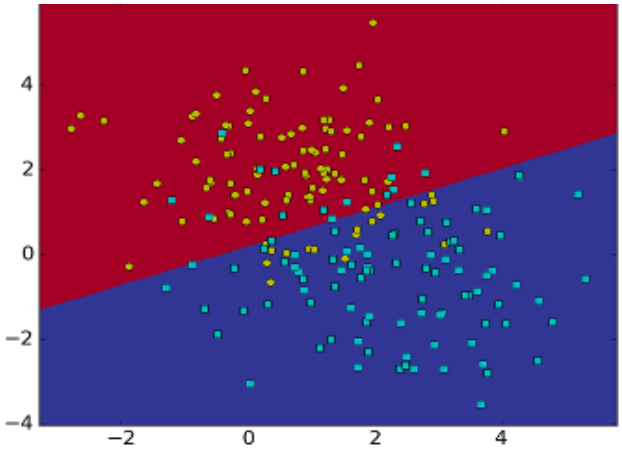
პროგრამული კოდი დაწერილია, პროგრამულ ენა Python

```
import numpy as np   #კომპიუტერ მეცნიერებათა ფუნდამეტური პაკეტი
from pylab import *
from sklearn import linear_model  #მანქანური სწავლების ბიბლიოთეკა
#რეზულტატის მისაღებად,მოვახდინოთ შემთხვევითი რიცხვების გენერირება
np.random.seed(1000)
# მოდელური მონაცემების შექმნა
mean1 = [0.5, 2]
cov1 = [[1, 1.1], [-1.1, 1]]
mean2 = [2.3, -0.5]
cov2 = [[1.3, -1.5], [1.5, 1.6]]
# მონაცემები პირველი კლასისთვის
data1 = np.random.multivariate_normal(mean1, cov1, 100)
# მონაცემები მეორე კლასისთვის
data2 = np.random.multivariate_normal(mean2, cov2, 100)
# ----------------------------------------------------------------
```

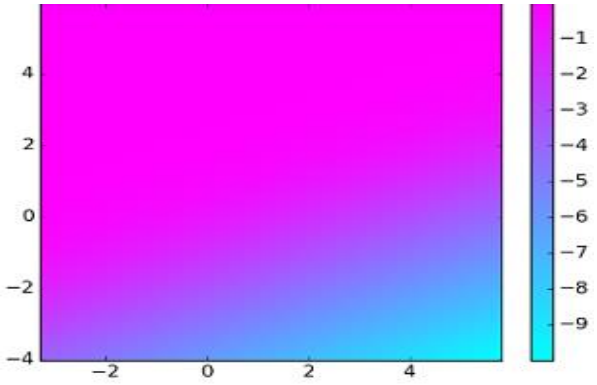
```
# შეწავლადი ფორმის ნიმუშის შექმანა
X = np.vstack([data1, data2])
# ცვლადების დაჯგუფება
Y = [0]*len(data1) + [1]*len(data2)
# ლოგუკური რეგრესიის მოდელის შექმნა
logreg = linear_model.LogisticRegression(C=1e5)
logreg.fit(X, Y)
# წერტილების მასივი შესაბამის კლასებში კრასიფიკაციისთვის
x_min, x_max = X[:, 0].min() - .5, X[:, 0].max() + .5
y_min, y_max = X[:, 1].min() - .5, X[:, 1].max() + .5
xx, yy = np.meshgrid(np.arange(x_min, x_max, 0.02), np.arange(y_min, y_max, 0.02))

# ყოველი წერტილის კლასიფიკაციის შესრულება
Z = logreg.predict(np.c_[xx.ravel(), yy.ravel()])

# კლასების მიკუთვლების ალბათობის ლოგარითმი
probabilities = logreg.predict_log_proba(np.c_[xx.ravel(), yy.ravel()])

# pcolormesh მოთხოვნის საფუძველზე ფორმატის სახეცვლილება
Z = Z.reshape(xx.shape)
p1 = probabilities[:,0].reshape(xx.shape)
p2 = probabilities[:,1].reshape(xx.shape)
title(u'result’)
plot(data1[:, 0], data1[:, 1], 'oy')
plot(data2[:, 0], data2[:, 1], 'sc')
pcolormesh(xx, yy, Z, cmap='RdYlBu')
gca().set_xlim([x_min, x_max])
gca().set_ylim([y_min, y_max])
```

```
figure()
title(u'Log-Class Probabilities 1')
pcolormesh(xx, yy, p2, cmap='cool')
colorbar()
gca().set_xlim([x_min, x_max])
gca().set_ylim([y_min, y_max])
show()
```

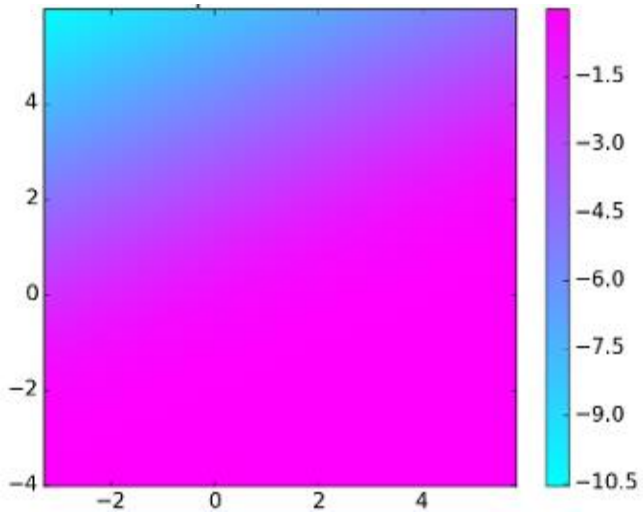
კლასიფიკაციის შედეგი



პირველი კლასის ლოგი



მეორე კლასის ლოგი



დასკვნა

ორ განზომილებიანი ფაქტორული სივრციდან ყოველი წერტილის კლასიფიკაციამ, საშუალება მოგვცა ვიზუალურად დაგვენახა გადაწყვეტილების მიღების არე. მოდელური მონაცემების გენერერიბისთვის გამოვიყენეთ ორი ორ განზომილებიანი ნორმალური განაწილება საშუალო მონაცემებით, რამაც გამოიწვია წერტილების გადაადგილება სასწავლო გარემოში მათი მიკუთვნების მიხედვით.

ლიტერატურა

- Machine Learning by Tom M Mitchell
- The Elements of Statistical Learning
- Pattern Recognition and Machine Learning
- Learning from Data