VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
KOMPIUTERIJOS KATEDRA

Baigiamasis bakalauro darbas

Duomenų dimensiškumo mažinimas ir klasifikavimas

Atliko:
Donatas Kučinskas parašas

Vadovas:
Vytautas Valaitis

Vilnius
2015

Turinys

Sutartinis terminų žodynas

Pateikiamas terminų sąrašas (jei reikia)

Santrauka

Santraukos tekstas rašto darbo kalba...

Summary

Darbo pavadinimas kita kalba

This is a summary in English...

Įvadas

Klasifikavimas - tai dažnai sutinkama užduotis, turintį įvairių sprendimo būdų. Šios uždavinio tikslas - identifikuoti, kuriai grupei priklauso tiriamas objektas. Tiriamieji objektai dažniausiai būna vienos rūšies, aprašomi tam tikrais parametrais, o grupės, kuriems jie yra priskiriami - iš anksto žinomos. Pavyzdžiui, galima klasifikuoti gyvūnus pagal tam tikras jų fizines savybes - kojų ilgį, storį, kitas kūno apimtis, kailio ilgį ir pan. Natūralu, kad kiekvienas net ir tos pačios rūšies gyvūnas turės šiek tiek kitokius parametrus, tačiau šie parametrai dažniausiai turi įvairius proporcingumus, pagal kuriuos galima bandyti atspėti, kuriai rūšiai tam tikras gyvūnas priklauso.

Norint išspręsti konkretų klasifikavimo uždavinį, paprasčiausias sprendimas atrodo galėtų būti šių grupių parametrų ištyrimas - pavyzdžiui, norint mokėti atskirti triušius nuo liūtų turint jų ilgius nėra sunki užduotis. Tačiau problema kyla, kai atskiriamos klasės yra labai panašios viena į kitą - tokiu atveju pastebėti tam tikrus dėsningumus ir juos sumodeliuoti bei realizuoti ir kur kas sunkiau. Be to, sprendžiant konkretų klasifikavimo uždavinį, tektų gilintis į klasifikuojamus objektus - pavyzdžiui, norint sukurti tam tikrų kiškių rūšių klasifikavimą, gilios žinios apie šias kiškių rūšių savybes būtų privalomos.

1. Dirbtinių neuronų tinklas

Dirbtinis neuronų tinklas - tai tarpusavyje susijungusių dirbtinių neuronų tinklas, kurio užduotis yra spręsti tam tikrą užduotį arba užduotis. Dirbtinis neuronų tinklas gavęs pradinius užduoties duomenis, juos apdoroja ir taip gaunamas tam tikras atsakymas. Šis atsakymas nebūtinai yra teisingas - neuronų tinklai suprojektuoti taip, kad galėtų būti mokomi kai gauna neteisingą atsakymą.

1.1. Dirbtinis neuronas

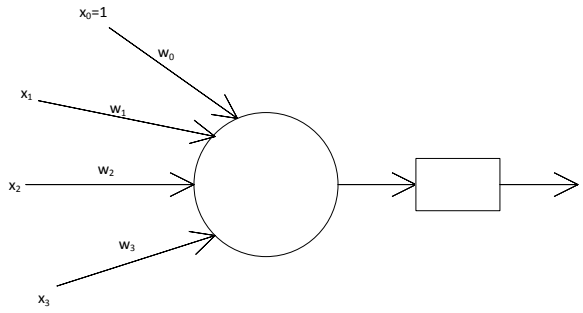
Dirbinių neuronų tinklas sudarytas iš daugybės dirbtinių neuronų, todėl norint suprasti tinklą, reikia pradėti nuo vieno dirbtinio neurono. Žmogaus smegenys sudarytos iš daugybės neuronų. Dirbtinis neuronas - tai supaprastintas šių biologinių neuronų modelis. Jo modelis pavaizduotas 1 paveiksliuke. Dirbtinio neurono veikimo principas gan paprastas - per kairėje esančias jungtis dirbtinis neuronas gauna signalus iš kitų dirbtinių neuronų - iš k-tosios jungties gaunamas x_k dydžio signalas. Šiuos signalus neuronas apjungia ir pertvarko, ir taip sugeneruojamas dirbtinio neurono išeinamasis signalas. Šis išeinamasis signalas gali būti siunčiamas daugybei kitų neuronų - dešinėje esančios jungtys yra neurono išeinamojo signalo jungtys, kuriomis ir yra siunčiamas išeinamasis signalas.

Dirbtinis neuronas generuoja išeinamąjį signalą pagal tam tikrą modelį. Pirmiausia, kiekviena įeinančioji jungtis k turi savo svorį w_k - šis svoris yra padauginamas iš įeinančio signalo dydžio x_k. Tada visos šios signalų dydžių ir svorių sandaugos yra susumuojamos - taip gaunamas skaičius a (1.1 formulė). Tada šis skaičius a yra paduodamas kaip argumentas tam tikrai funkcijai f ir gaunamas neurono išeities signalas $y = f(a)$. Šį funkcija f yra vadinama aktyvacijos funkcija - ją galima keisti pagal tai, kokio tikslo siekiama iš šio dirbtinio neurono. Populiariausios aktyvacijos funkcijos - slenkstinė, tiesinė, hiperbolinis tangentas bei sigmoidinė (1.2 formulė). Iš esmės akvyvacijos funkcija gali būti bet kokia funkcija, tačiau vėliau norint apmokyti dirbtinį neuronų tinklą, reikia rasti šios funkcijos išvestinę. Dėl šios priežasties dažniausiai pasirenkamos tokios aktyvacijos funkcijos, kurios ne tik tinkamai pertvarko signalą išvedimui, tačiau ir kurios išvestinės yra paprastos.

Įeinamosios neurono jungtys numeruojamos nuo 1 iki k. Norint a reikšmę padaryti tinkamesnę neuroninio tinklo funkcijoms, dažniausiai įvedama papildoma 0-inė jungtis su svoriu w_0 ir signalo stiprumu $x_0 = 1$. Tokiu būdu prie a (formulė 1.1) reikšmės papildomai pridedama $w_0 * x_0 = w_0$ reikšmė.

$$a = \sum_{k=1}^{N} w_k x_k \tag{1.1}$$

$$f(a) = \frac{1}{1 + e^{-a}} \tag{1.2}$$



1 pav. Dirbtinis neuronas TODO: add functions

1.2. Dirbtiniai neuronai/tinklas?TODO

Visi

[TODO: dirbtinio neurono paveiksliukas]

[TODO: citata?]

[TODO: 110 iš knygos]

2. Dimensiškumo mažinimas

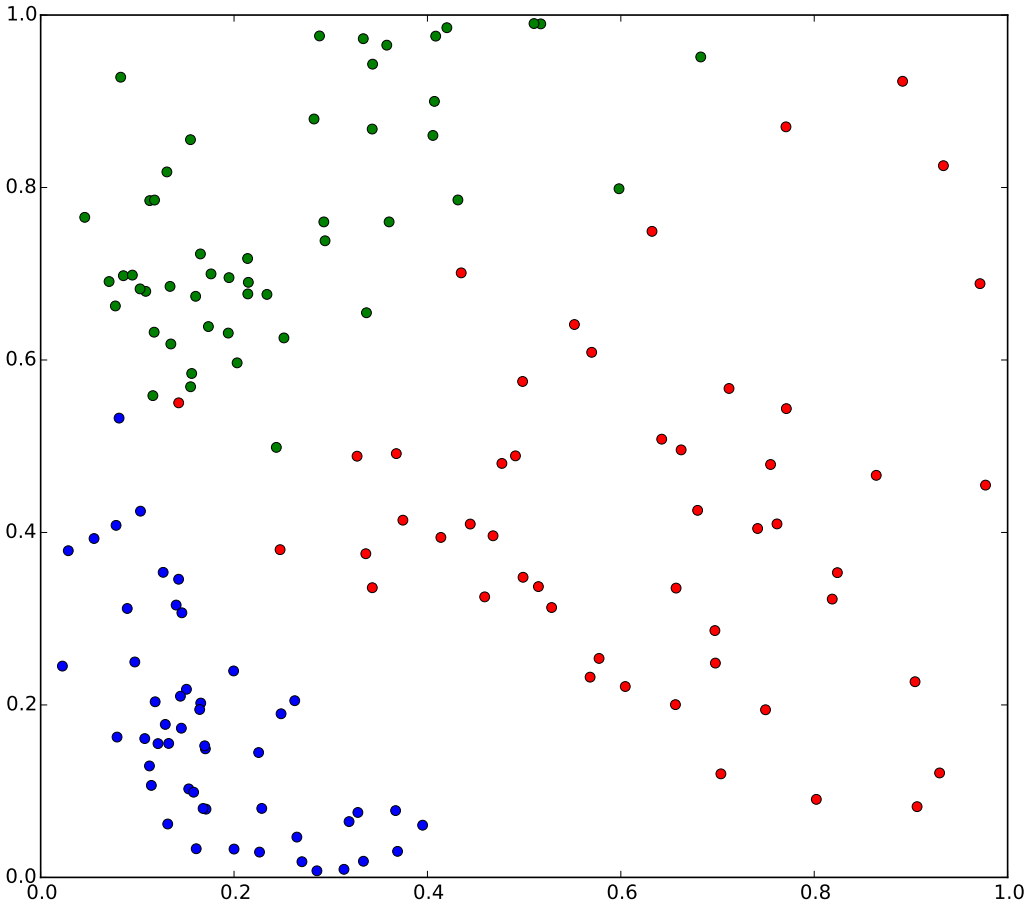
Klasifikavimo problema

Galimi sprendimai:

* statistinis sprendimas * neuroniniai tinklai * veikimas * apmokymas * validavimas? * Klasifikavimas požymių išskyrimui * Dimensiškumo mažinimas -> Klasifikavimas

2.1. Statistinis sprendimas

Vienas iš galimų dimensiškumo mažinimo sprendimo būdų - tiesinė diskriminantinė analizė (angl. *Linear discriminant analysis*).



3. Vilkdagių duomenys

Programuojant neuroninius tinklus, testavimui buvo panaudoti vilkdagių (angl. *Iris flower*) duomenys. Tai plačiai taikomi ir viešai pasiekiami duomenys, aprašantys 3 rūšių vilkdagius. Aprašyta po 50 kiekvienos rūšies vilkdagių. Kiekvienas vilkdagis aprašomas pateikiant 4 dydžius: taurėlapio ilgis, taurėlapio plotis, vainiklapio ilgis bei vainiklapio plotis. Šiuos vilkdagių duomenis sudaro 150 gėlių, kurių kiekviena aprašyta 4 parametrais bei priskirta vienai iš 3 vilkdagių grupių.

Šie duomenys puikiai tinka klasifikavimo tinklo apmokymui - tinklo tikslas yra kuo mažiau klystant pasakyti, kuriai iš 3 vilkdagių rūšių tam tikra gėlė su tam tikrais parametrais priklauso. Be to, yra pakankamai duomenų, kad būtų galima dalį jų panaudoti tinklo apmokymui, o kitą dalį - testavimui. Tokiu būdu bus užtikrinama, kad tinklas teisingai išmoko atskirti vilkdagių rūšis pagal parametrus, o ne tiesiog prisitaikė prie mokymui panaudotų duomenų.

TODO: nuoroda į Vilkdagių duomenis?

4. Dimensiškumo mažinimas neuroniniu tinklu

TODO: įžanga

Dimensiškumo mažinimui taip pat buvo panaudotas neuroninis tinklas. Turint N dimensijų ir norint jas sumažinti iki M, kai $M < N$, tai buvo atliekama sukūrus neuroninį tiklą, kurio pirmąjame ir paskutinimae sluoksniuose yra po N neuronų, o viename iš vidinių sluoksnių - M (šį vidinį sluoksnį vadinkime kompresijos sluoksniu). Tokio neuroninio tinklo užduotis nėra tiesiog sumažinti dimensijų skaičių - tai daroma netiesiogiai. Šiam tinklui perduodant tam tikrus M dimensijų turinčius duomenis, iš jo tikimasi, kad išeities neuronuose susiformuos rezultatas, lygus pradiniams duomenims - tai yra neuronų tinklas šių duomenų nepakeis. Ši užduotis paprastai nebūtų sunki, jeigu visi vidiniai turėtų bent N dimensijų - tada pateikiami duomenys galėtų būti tiesiog perkeliami iš vieno neuronų sluoksnio į kitą nepakeisti. Tačiau kompresijos sluoksnis turi tik M neuronų - vadinasi, duomenis reikės tam tikru būdu pertvarkyti, kad jie galėtų būti perduodami per šį sluoksnį prarandant kuo mažiau savybių. Būtent čia ir įvyksta dimensiškumo mažinimas - neuroninis tinklas yra apmokomas pateikti kuo panašesnius duomenis į pradinius, ko pasekoje kompresijos sluoksnyje su M neuronų yra gaunami duomenys, turintys mažiau dimensijų. Norint sumažinti tam tikro duomens dimensijas, užtenka šį duomenį paduoti apmokytui neuroniniui tinklui ir pažiūrėti, kokie duomenys susidarė kompresijos sluoksnyje. Nuskaičius šių neuronų reikšmes ir bus gaunamas duomuo, turintis mažiau dimensijų.

Tokiame apmokytame neuroniniame tinkle visi sluoksniai, esantys kairėje nuo kompresijos sluoksnio, yra naudojami dimensiškumo mažinimui. Būtent per šiuos sluoksnius einant signalams ir yra sudaromas mažiau dimensijų turintis duomuo. Kadangi šio neuroninio tinklo tikslas yra pateikti rezultatą, kuris būtų kuo panašesnis į pateiktus duomenis, todėl galima teikti, kad sluoksniai, esantys dešinėje nuo kompresijos sluoksnio, yra naudojami pradinių duomenų atstatymui.

TODO: diagrama su dimensijų mažinimo neuroniniu tinku (kompresijos, dekompresijos pusės; N, M neuronų; rezultatas toks pat kaip duomenys)

5. NOTES

Šaltinis [**?**].

1 lentelė. Lentelė ...

test	test
test	test

Išvados ir rekomendacijos

Išvados bei rekomendacijos.

Ateities tyrimų planas

Pristatomi ateities darbai ir/ar jų planas, gairės tolimesniems darbams....

Priedai

Dokumentą sudaro du priedai: A priede ....

A. Pirmojo priedo pavadinimas

Pirmojo priedo tekstas ...

B. Šaltiniai

1. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=298007