Turinys

1 Teorinis darbo pagrindas

Šiame skyriuje aprašysiu teorinį darbo pagrindą.

1.1 Mokymasis su ir be mokytojo

Šiame skyriuje stengsiuosi atsakyti į klausimą kuo skiriasi mokymas su mokytoju (angl. supervised learning) nuo mokymo be mokytojo (angl. unsupervised learning). Mokymasis, duomenų klasifikavimo kontekste, reiškia modelių (pvz. klasifikatorių) kūrimo metodus (algoritmus), kurie naudoja mokymosi duomenis[1], kitaip tariant, tai mokymasis iš pavyzdžių.

1.1.1 Mokymas su mokytoju

Mokymas su mokytoju tai toks mokymas, kai turime mokymo duomenis, kuriems jau yra priskirtos tam tikras teisingas atsakymas. Kitaip tariant, mes sprendžiame uždavinį, kuriam atsakymą galime pasitikrinti. Mokymas su mokytoju yra skirstomas į dvi rūšis:

1. Klasifikavimas (angl. classification) - pagal nepriklausomus kintamuosius bandome nuspėti kokybinius (kategorinius) priklausomus kintamuosius.

2. Regresija (angl. regression) - pagal nepriklausomus kintamuosius bandome nuspėti kiekybinius priklausomus kintamuosius.

1.1.2 Klasifikavimo uždavinio pavyzdys

Klasifikavimo uždavinių aktualumą galima pagrįsti paprastu pavyzdžiu.

Uždavinys: Pašto skyriuose laiškai siunčiami įvairiomis kryptimis pagal gavėjo adresą ir (arba) pašto kodą. Mes norime automatizuoti laiškų rūšiavimą pagal siuntimo kryptį. Tam, kad galėtume laiškų rūšiavimą pagal kryptį automatizuoti, mums reikėtų galimybės nuo voko nuskaityti pašto kodą.

Sprendimas: Šią problemą mums padėtų išspręsti skeneris ir programine įranga, kuri sugebėtų ranka rašytus skaitmenis atpažinti ir konvertuoti į skaitmeninį formatą. Tų skaitmenų atpažinimui ir konvertavimui į skaitmeninį formatą, tikėtina, kad mes naudosime klasifikavimo algortimus, nes uždavinys pasižymi visomis klasifikavimui būdingomis savybėmis: turime aibę duomenų (vaizdinė informacija su ranka rašytais skaitmenimis), turime teisingus atsakymus (žmogus pažiūrėjęs į ranka rašytą skaitmenį gali pasakyti programai, koks ten yra skaitmuo), bei galimų sprendimai yra kategorinio tipo (dešimt skaitmenų nuo 0 iki 9).

Įgyvendinę aukščiau aprašyto uždavinio sprendimą, pašto skyrių vadybininkai galėtume atlaisvinti žmones nuo iš esmės mechaninio darbo - rūšiuoti laiškus. Tokiu būdu būtų optimizuotas pašto skyrių efektyvumas.

[1] Mokymosi duomenys (angl. sample data)- duomenys, kurie yra paruošti darbui programų, kurios kurs modelius (pvz. klasifikatorius).

1.1.3 Regresijos uždavinio payzdys

Abiejų mokymo su mokytoju rūšių tikslas yra pagal mokymosi duomenis sukurti modelį, kuriuo remiantis būtų galima identifikuoti naujų objektų savybes.[Hal99] Šiame darbe negrinėsime klasifikavimo problemą.

2 Teorija

2.1 Bayesian decision theory

2.2 Klasifikavimas “artimiausio kaimyno” metodu

2.3 Klasifikavimas “mažiausių kvadratų metotu”

2.4 Naive Bayesian classifier

2.5 Bias and variance tradeoff

2.6 Klasifikavimo metodo įvertinimas

2.6.1 Klasifikavimo metodo įvertinimas “cross-validation” metodu

2.6.2 Klasifikavimo metodo įvertinimas “bootstrapping” metodu

2.7 Atraminių vektorių metodai

Atraminių vektorių klasifikatorius[Vap00] (angl. support vector machines) - tai mašininio mokymosi (angl. machine learning) algoritmas išvestas iš statistinio mokymosi. Jis priskiriamas mokymuisi su mokytoju. Metodas taikomas ir klasifikavime, ir regresinėje analizėje.

Naudojant atraminių vektorių klasifikatorių, yra sukuriama hiperplokštuma, atskirianti duomenis į dvi klases. Hiperplokštuma parenkama tokia, kad atstumas tarp skirtingų klasių artimiausių elementų ir hiperplokštumos būtų didžiausias.

Konstruojant hiperplokštumą yra spendžiamas optimizavimo su ribojimais algoritmas.

Gali būti ir taip, kad ieškoma hiperplokštuma gali ir neegzistuoti pavyzdžiui, kai klasės stipriai persidengia. Tada įvedamas parametras ir pasikeičia optimizavimo uždavinys.

Viena iš atraminių vektorių metodų klasifikavimo ypatybių yra gebėjimas mokytis iš labai mažos mokymosi duomenų aibės.

2.8 Random forests

2.9 Kuo ypatingas daugiamačių duomenų klasifikavimas

2.9.1 the curse of dimensionality

3 Susiję darbai

4 Klasifikavimo metodų palyginimo karkasas

5 Klasifikavimo metodų palyginimo rezultatai

SĄVOKŲ APIBRĖŽIMAI

Prižiūrimas mokymasis (angl. supervised learning) -

Neprižiūrimas mokymasis (angl. unsupervised learning) -

Mašininis[Mam08] (kompiuterinis, sistemos[Mar08]) mokymasis (angl. machine learning) - tai mokslas siekiantis priversti kompiuterius atlikti tam tikrą darbą be išreikšto programavimo.

Hiperplokštuma (angl. hyperplane) - plokštumos generalizacija daugiadimansėje erdvėje.

Atraminių vektorių klasifikatoriai (angl. support vector machines, SVM) - yra klasifikavimo su mokymu metodas, taikomas ir klasifikavime, ir regresinei analizei.[Ber08]

Regrèsija [lot. regressio – grįžimas, traukimasis]: tikimybių teorijoje ir mat. statistikoje – atsitiktinio dydžio vidurkio priklausomybės nuo kt. dydžio (kelių dydžių) išraiška;[tzz10]

Literatūra

[Ber08] Jolita Bernatavičienė. *Vizualios žinių gavybos metodologija ir jos tyrimas.* PhD thesis, Vilniaus Gedimino technikos universitetas, 2008. Prieiga internetu: `http://vddb.laba.lt/fedora/get/LT-eLABa-0001:E.02~2008~D_20080930_090520-93322/DS.005.0.02.ETD`.

[Hal99] Mark A. Hall. *Correlation-based Feature Selection for Machine Learning.* PhD thesis, University of Waikato, 1999. Prieiga internetu: `http://www.cs.waikato.ac.nz/~ml/publications/1999/99MH-Thesis.ps`.

[Mam08] Jelena Mamčenko. *Duomenų gavybos technologijų taikymas išskirstytų serverių darbui gerinti.* PhD thesis, Vilniaus Gedimino technikos universitetas, 2008. Prieiga internetu: `http://vddb.laba.lt/fedora/get/LT-eLABa-0001:E.02~2008~D_20090105_150124-79076/DS.005.0.02.ETD`.

[Mar08] Dalia Martišiūtė. Vaizdų klasterizavimas. Master's thesis, Vilniaus universitetas, 2008. Prieiga internetu: `http://vddb.laba.lt/fedora/get/LT-eLABa-0001:E.02~2008~D_20090908_201754-37094/DS.005.1.01.ETD`.

[tzz10] *Tarptautinių žodžių žodynas.* Vyriausioji enciklopedijų redakcija, 2010.

[Vap00] V.N. Vapnik. *The nature of statistical learning theory.* Statistics for engineering and information science. Springer, 2000.