Turinys

1 Teorinis darbo pagrindas

Šiame skyriuje aprašysiu teorinį darbo pagrindą.

1.1 Prižiūrimas ir neprižiūrimas mokymasis

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

1.1.1 Prižiūrimas mokymasis

Prižiūrimas mokymasis tai toks mokymasis, kai turime iš anksto nustatytas klases bei mokymosi duomenis, kuriems jau yra priskirtos tam tikros teisingos klasės. Tikslas yra pagal mokymosi duomenis sukurti klasifikatorių, kuriuo remiantis būtų galima identifikuoti naujų objektų priklausomybę vienai iš žinomų klasių.[Hal99]

Prižiūrimo mokymosi metodų pagrindinė prielaida yra ta, kad kontekstas suteikia pakankamai informacijos. Kitaip tariant - jei žinai pakankamai daug objektų priklausančių kažkokioms tai klasėms, tai naujiems objektams pakankamai tiksliai gali priskirti tas klases.

Prižiūrėtasis mokymasis turi du pagrindinius būdus:

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.

Išskiriami trys pagrindiniai klasifikavimo etapai:

1. diskriminavimo (atskiriančiųjų) kintamųjų parinkimas,

2. klasifikavimo taisyklių sudarymas,

3. klasifikavimo kokybės įvertinimas.

1.1.2 Neprižiūrimas mokymasis

Neprižiūrimas mokymasis dar vadinamas klasterizavimu (angl. clustering) arba mokymusi be mokytojo. Patogumo dėlei, toliau naudosiu klasterizavimo sąvoką kaip ekvivalentą neprižiūrimojo makymosi sąvokai.

Klasterizavimas (angl. clustering) - tai viena iš duomenų gavybos sričių. Klasterizavimo algoritmo užduotis – objektų suskirstymas į prasmingas grupes

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

– klasterius, kai jokia papildoma informacija apie tas grupes (jų dydį, kiekį, grupavimo požymius) nėra iš anksto žinoma. Klasterizavimo algoritmas pats, pagal pasirinktus algoritmo parametrus, turi nurodyti, kokioms grupėms priklauso atitinkami įvesties duomenys.[Mar08]

Klasterizavimo algoritmų pagrindinis privalumas – gebėjimas atpažinti grupavimo struktūrą be jokios išankstinės informacijos.

Klasterizavimo principas - maksimizuoti objektų, esančių vienoje grupėje, tarpusavio panašumą ir minimizuoti tarpgrupinį objektų panašumą.

1.1.3 Prižiūrimojo ir neprižiūrimojo mokymosi skirtumai

Pagrindinis skirtumas tarp prižiūrimojo ir neprižiūrimojo mokymosi slypi mokymosi duomenyse: prižiūrimojo mokymosi algoritmų įeities duomenyse yra išreikštinai pasakyta, kokio rezultato mes laukiame, o neprižiūrimojo mokymosi duomenyse tokios papildomos informacijos nėra. Aptarkime pavyzdį: mums reikia sukurti klasifikatorių, kuris pasakytų, ar nuotraukoje yra žmogaus veidas.

Prižiūrimojo mokymosi programai kaip įeities duomenis paduotume keletą nuotraukų su žymėmis pasakančiomis, ar nuotraukoje yra žmogaus veidas ar jo ten nėra, kitaip tariant, duotume keletą pavyzdžių su teisingais atsakymais. Programa peržvelgs visas nuotraukas ir susikurs klasifikatorių (modelį), kuris kažkokiu tikslumu galės atskirti nuotraukas su žmogaus veidu. Tokiu būdu mūsų prižiūrimojo mokymosi programa “išmoks”, kas yra veidas.

Neprižiūrimojo mokymosi programai kaip įeities duomenis paduotume keletą nuotraukų be jokių papildomų žymių. Žinoma, mūsų programa pati nesugebės “išrasti”, kas yra žmogaus veidas, tačiau ji tikriausiai sugrupuos nuotraukas su žmonių veidais ir tarkim peizažais į skirtingas grupes. Kitaip tariant, nuotraukos su žmonių veidais mūsų neprižiūrimo mokymosi programai bus nepanašios į nuotraukas su peizažais, todėl ji į vieną klasterį susidės nuotraukas, kurios jai atrodo tarpusavyje panašiausios: viename klasteryje nuotraukos su žmonių veidais, o kitoje su gamtos peizažais.

Apibendrinant galime pasakyti, kad abi mokymosi rūšys siekia to paties tikslo, tik skitingomis priemonėmis. Pvz. atskirti nuotraukas su žmonių veidais nuo kitų nuotraukų su ar be teisingos žymės apie konkrečią nuotrauką. Bendras bruožas yra tai, kad jos mokymosi procese naudoja pavyzdžius, tik tie pavyzdžiai skiriasi programai suteikiama informacija.

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[Š06] (angl. support vector machines) - tai mašininio mokymosi (angl. machine learning) algoritmas išvestas iš statistinio mokymosi. Jis priskiriamas prižiūrimajam mokymuisi. 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]

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`.

[Š06] Simonas Šimkevičius. Klasifikavimo su mokytoju metodų lyginamoji analizė. Master's thesis, Kauno technologijos universitetas, 2006. Prieiga internetu: http://vddb.laba.lt/fedora/get/LT-eLABa-0001:E.02~2006~D_20060605_092832-93331/DS.005.0.02.ETD.