1. Domača naloga: Linearna regresija

Luka Krsnik (63110179)

11. marec 2015

1 Uvod

Cilj te naloge je bil, da iz testnih podatkov o intenziteti vonja v povezavi z lastnostmi molekul, ki jih vonjamo, zgeneriram neko "funkcijo", s pomočjo katere bom lahko čim bolje napovedal, kako intenzivna bo molekula z določenimi lastnostmi.

2 Metode

Pearson Pearsonova metoda pomaga pri izbiri značilk. Sicer deluje tako, da izračuna korelacijo med dvema značilkama. Če je vrednost 1 pomeni da sta vektorja v pozitivni korelaciji, če je -1 sta v negativni korelaciji, pri 0 pa nista v korelaciji. Implementirana je glede na velikost korelacije po absolutni vrednosti, večja kot je, večji pomen ima določen stolpec.

Regresija Regresija je postopek, s pomočjo katerega je na podlagi učnih podatkov zgrajena "funkcija", ki iz značilk čim bolje predvidi tisto kar je naučena (v konkretnem primeru vonj). Osnovni princip njenega izračuna je tak, da je izražena funkcija za izračun vsote kvadratov napak, v katero so vstavljeni takšni parametri, da bo njen rezultat čim manjši. V programu je vgrajena preko funkcije L-bfgs in njenih parametrov.

Regularizacija Regularizacija sama po sebi ni metoda, pač pa le izboljšana regresija. Izboljšana je tako, da na funkcijo delno vpliva tudi vrednost thete. Kakšen vpliv bo imela theta je nadzorovano z velikostjo lambde.

Metoda končnih razlik Ta metoda je uporabljena pri preverjanju pravilnosti zastavljenih enačb. Uporablja se tako, da se primerja rezultat odvoda (v gradientu) z vrednostmi, ki se pridobijo kot rezultat iz metode končnih diferenc. Če so skoraj enake, potem so zastavljene enačbe pravilne, sicer ne.

3 Rezultati

Rezultati so bili v primerjavi z drugimi boljši od pričakovanj, ko pa pomislim na to, da so le malo manjši od povprečja (mean), vidim da nisem naredil zelo veliko. Mogoče bi na tem mestu moral povedati, da sem pri Pearsonu (čeprav bi v teoriji morali biti boljši rezultati, če bi vzel

največje absolutne korelacije), vzel le največje pozitivne korelacije, saj so bili rezultati tako precej boljši - vzel sem 19 največjih vrednosti (Tabela: 1).

Tabela 1: Tabela rezultatov.

ime metode	oddaja	ocena strežnika
regresija	predzadnja	17.71364
regresija z regularizacijo	zadnja	17.71460

Rezultati brez uporabe regularizacije so skoraj enaki tistim z regularizacijo, celo nekoliko slabši. Po mojem mneju je to posledica tega, da uporabljamo relativno malo različnih značilk, saj zato spreminjanje thete ne izboljša rezultatov, temveč jih morda celo nekoliko poslabša. Druga možnost za takšne rezultate je ta, da sem izbral napačno lambdo (mogoče je bila premajhna - 0.00003).

4 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.