Uvod v odkrivanje znanj iz podatkov
2. velika domača naloga
Bicikelj

Gregor Kovač

Maj 2023

1 Podatki

Dani so podatki o tem, koliko je koles na posamezni postaji Bicikelj ob določenem času. Želeli smo zgraditi model, ki bo čim bolj točno napovedal število koles ob nekem novem času za posamezno postajo.

2 Obdelava podatkov in ustvarjanje novih značilk

Najprej sem podatke za čas (značilka *timestamp*) pretvoril v objekt Pandas Datetime za lažjo ekstrakcijo podatkov. Nato sem iz tega ustvaril novi značilki *weekday* (dan v tednu) in *hour* (ura), staro značilko *timestamp* pa sem zavrgel. Obe novi značilki sem transformiral z metodo *one-hot-encoding* in nato dobil še več novih atributov, kot so *weekday_1, weekday_2, ..., weekday_6*. Vsi te atributi so binarni. Potrebujemo eno manj novih značilk kot je možnih vrednosti stare značilke, saj lahko s samimi ničlami "zakodiramo" še eno vrednost. Prednost tega napran eni sami značilki je, da ima lahko naprimer ponedeljek večjo utež v končnem modelu kot sobota, kar je smiselno, ker bo ob ponedeljkih poraba koles verjetno drugačna kot ob vikendih.

Nato sem ustaviril še 3 nove atribute, ki predstavljajo količino koles na postaji pred 60, 90 in 120 minutami. Podatke o tem sem pridobil tako, da sem od trenutnega časa odštel N minut ($N \in \{60, 90, 120\}$) in pogledal, kateri čas v učni množici je najbližji. Ker se lahko zgodi, da čas pred N minutami ne obstaja, najbližji pa je precej oddaljen (po pregledu podatkov sem mejo nastavil na 15 minut), sem iz učne množice odstranil vse primere, kjer se to pojavi. V testni množici se za veliko primerov zgodi, da čas pred eno uro ne obstaja in je najbližji oddaljen za okoli 60 minut, zato sem vsem testnim primerom dodal zastavico *is_two_hour_break* in jo za take primere nastavil na 1, za ostale pa na 0. Za primere z vrednostjo 1 sem zgradil ločen model, ki ne vsebuje značilk za število koles pred 60 in 90 minutami.

3 Modeli

Nad podatki sem preizkusil več različnih modelov, najbolje pa se je obnesla linearna regresija. Model sem ustvaril za vsako postajo posebej, torej sem za ciljno spremenljivko izbral število koles na postaji, za vhodne spremenljivke pa sem izbral zgoraj omenjene značilke. Uporabil sem dve verziji modelov, kot sem že omenil zgoraj. Pred treniranjem modela pa sem še nadomestil čase za števila koles pred 60, 90 in 120 minutami z dejanskimi števili koles iz učne množice. Ko sem dobil končne rezultate, sem vse negativne vrednosti nastavil na 0, saj ni smiselno, da je na neki postaji negativno število koles.