Uvod. Pri peti domači nalogi smo napovedovali prihode avtobusov na zadnjo postajo z linearno regresijo. Za predtekmovanje smo napovedovali prihod samo za eno linijo, za tekmovanje pa za vse linije.

Učna množica je vsebovala LPP podatke za 11 mesecev (januar - november), testno množico pa je predstavljal mesec december za katerega so bili časi prihodov neznani - le za tega smo svoje napovedi testirali na strežniku, ki nam je nato sporočil absolutno povprečno napako. Lokalno smo rezultate testiranja pridobili s prečnim preverjanjem na učni množici: izmenično smo učili model na desetih mesecih ter preverjali za povprečno absolutno napako na tistem, ki ga nismo uporabili za učenje.

Predtekmovanje Lokalno in na strežniku (lestvici) smo preizkusili več načinov določevanja značilk. Spodaj so opisani najboljši pristopi, rezultati napovedi pa so prikazani v tebeli 1. Pri vseh pristopih smo uporabili regularizacijo z vrednostjo $lambda = 0.4$.

predtekmovanje160	predtekmovanje156	predtekmovanje150.4	predtekmovanje150.3 *
160.11469	156.00678	150.38881	150.28703 *

Tabela 1: Absolutna povprečna napaka napovedi za predtekmovanje na lestvici (v sekundah). Z * je označena končna rešitev.

predtekmovanje160 Iz podatkov o času in datumu odhoda smo določili značilke:

- ali je mesec poletnih počitnic (julij, junij, avgust)
- ali je dan v tednu delavnik, sobota ali nedelja
- ura odhoda spada v jutro (7, 8, 9), dopoldne (10, 11, 12), popoldne (13, 14, 15, 16, 17, 18), večer (19, 20, 21, 22) ali noč (23)

Vseh 9 značilk ima vrednost 0 ali 1 zato smo uporabili kar obicajno linearno regresijo.

predtekmovanje156 Iz podatkov o času in datumu odhoda smo določili značilke:

- ali je mesec poletnih počitnic (julij, junij, avgust)
- ali je dan v tednu delavnik, sobota ali nedelja
- normalizirana ura odhoda ter njena druge ter tretja potenca

Značilke počitnice in dnevi so imajo vrednost 0 ali 1, ure pa vrednosti v intervalu [0,1] zato smo za njih uporabili polinomsko razširitev regresije. Z razširitvijo se lahko regresija bolj natančno prileže podatkom.

predtekmovanje150.4 Iz podatkov o času in datumu odhoda smo določili značilke:

- ali je mesec poletnih počitnic (julij, junij, avgust)
- ali je dan v tednu delavnik, sobota ali nedelja
- ali je ura 0, 1, 2, ..., 23

Vseh 28 značilk ima vrednost 0 ali 1 zato smo uporabili kar obicajno linearno regresijo.

predtekmovanje150.3 Iz podatkov o času in datumu odhoda smo določili značilke:

- ali je mesec poletnih počitnic (julij, junij, avgust)
- ali je dan ponedeljek, torek, sreda, četrtek, petek, sobota ali nedelja
- ali je ura 0, 1, 2, ..., 23

Vseh 31 značilk ima vrednost 0 ali 1 zato smo uporabili kar obicajno linearno regresijo. Pristop se je izkazal za najbolj uspešnega, saj smo lahko najnatančnejše prilagodili podatkom.

Tekmovanje Tako kot pri predtekmovanju smo preizukusili delovanje z različnimi značilkami in opisali tri najbolj zanimive pristope. Rezultati testiranja so vidni v tabeli 2. Pri vseh pristopih smo uporabili regularizacijo z vrednostjo $lambda = 0.5$.

tekmovanjeX	tekmovanje214	tekmovanje207 *
/	214.99199	207.43110 *

Tabela 2: Absolutna povprečna napaka napovedi za tekmovanje na lestvici (v sekundah). Z * je označena končna rešitev.

tekmovanjeX Najprej smo za vse linije konstruirali en model. Linija je določena z ID, ki je sestavljen iz vseh atributov, ki opisujejo linijo, npr. '1VIŽMARJE - MESTNI LOG MESTNI LOG; sejemŠentvidMESTNI LOG'. Torej tudi različna smer je določena kot nova linija. Značilke so:

- za vsak ID še ali je mesec poletnih počitnic (julij, junij, avgust)
- za vsak ID še ali je dan ponedeljek, torek, sreda, četrtek, petek, sobota ali nedelja
- za vsak ID še ali je datum državni praznik
- za vsak ID še ali je ura 0, 1, 2, ..., 23

Zaradi velikega števila značilk ima takšen model 4484 dimenzij in ne uspe napovedati prihodov v razumljivem času. Vse značilke imajo vrednost 0 ali 1.

tekmovanje214 Podobno smo naredili še en enotni model. Linije so določene z ID enakim kot v pristopu tekmovanje(X). V želji, da bi zmanjšali število dimenzij smo značilke določili kot:

- za vsak ID še ali je mesec poletnih počitnic (julij, junij, avgust)
- za vsak ID še ali je dan v tednu delavnik, sobota
- za vsak ID še ali je datum državni praznik ali nedelja
- za vsak ID še normalizirana ura odhoda ter njena druge ter tretja potenca

Tako smo dobili 951 dimenzionalni model, ki je napovedal prihode v manj kot 5 minut. Potence ur imajo vrednosti v intervalu [0,1] zato smo za njih uporabili polinomsko razširitev regresije. Ostale značilke imajo vrednost 0 ali 1.

tekmovanje207 Najboljše je deloval pristop, kjer smo za vsako linijo definirali svoj model z značilkami:

- ali je mesec poletnih počitnic (julij, junij, avgust)
- ali je dan v tednu delavnik, sobota
- ali je datum državni praznik ali nedelja
- ali je ura 0, 1, 2, ..., 23

Linije so določene z ID enakim kot v pristopu tekmovanje(X). Pristop se je obnesel boljše kot enotni model za vse linije.