1. domača naloga: Linearna regresija

Jernej Henigman (63100259) Kaggle: JernejHenigman

25. april 2016

1 Uvod

Cilj domače naloga je bil izboljšati točnost napovedi na podatkovni množici "Digit-Recognizer"iz portala Kaggle, z uporabo metode skladanja (štacking")

2 Metode

Uporabili smo že prej pripravljeno predlogo, v kateri smo morali za pravilno delovanje štackinga"implementirati metodo za učenje (fit_storage) ter metodo za napovedovanje (predict_storage).

Uporabljali smo Orangove knjižnice za delo s podatki.
V fit_storage smo zgradili novo učno množico, ki je zgrajena iz napovedi posameznih učnih modelov. Te napovedi zložimo skupaj (od tod ime skladanje). Vektor, ki vsebuje pravilen razred posameznega učnega primera se ohrani. Vse napovedi posameznega učnega modela pridobimo s pomočjo prečnega preverjanja. Nato zgradimo "meta_learner"model (v našem primeru logistična regresija), ki agregira vmesne rezultate v končne napovedi.

3 Rezultati (AUC)

	Kaggle	digits-358	digits-79
Logistična regresija	0.8812	0.8727	0.9239
Softmax	0.8911	0.8902	0.9078
Stacking	0.9498	0.9338	0.9500

Digits-358 in digits-79 se dela na vseh podatkih s prečnim preverjanjem (500 primerov). Notranji k = 5, zunanji k = 5. Podatke iz portala kaggle se preverja na celotni tesni množici (28k primerov), uči pa na podmnožici učne množice (12k primerov). Pri rezultatih v stolpcu kaggle se dejansko gleda accuracy (oz. mero, ki je pač uporabljena na kagglu).
Pri metodi stacking uporabimo naslednje algoritme: KNNLearner, RandomForestLearner, LogisticRegressionLearner, SoftmaxRegressionLearner, LinearSVMLearner. Edina optimizacija, ki je bila narejena, je bila ta, da smo odstranili konstantne stolpce iz podatkov. (Pomankanje časa =)).