Sedma domača naloga

Anže Pečar (63060257)

21. maj 2012

1 Uvod

Cilj domače naloge je bil preizkusiti delovanje logistične regresije z regularizacijo na podatkih za tekmovanje iz področja kemoinfromatike.

2 Metode

2.1 Logistična regresija

Logistična regresija se uporablja pri binarnih klasifikacijskih problemih. Za te probleme sicer lahko uporabimo tudi linearno regresijo, vendar je, kot smo videli na predavanjih, trivialno sestaviti primer, kjer se linearna regresija ne izkaže. Intuitivno tudi nima nobenega smisla, da nam hipoteza $h_\Theta(x)$ vrača vrednosti manjša od 0 in večja od 1, če pa vemo, da je $y \in \{0, 1\}$.

V našo hipotezo vstavimo logistično funkcijo, da dobimo

$$h_\Theta(x) = \frac{1}{1 + e^{-\Theta^T x}}.$$

2.2 Cenovna funkcija

Cenovna funkcija, ki jo želimo žmanjšati je tako imenovana *log loss* funkcija. Definirana je na naslednji način

$$logloss = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{x_i}) (1 - y_i) \log(1 - \hat{x_i}).$$

Za vrednost $\hat{x_i}$ smo uporabili logistično funkcijo po naslednji enačbi

$$\hat{y_i} = h_\Theta(x_i).$$

2.3 Gradient

Da bi maksimizirali verjetje (*liklehood*) $L(\Theta) = p(\vec{y}|X; \Theta)$, lahko uporabimo metodo najhitrejšega sestopa (*gradient ascent*). V vektorski notaciji ga zapišemo kot $\Theta = \Theta + \alpha * \Delta_\Theta l(\Theta)$, kjer je $l(\Theta) = logL(\Theta)$. Iz vektorskega zapisa lahko izpeljemo stohastično formulo, ki je zelo podobna

tisti, ki smo jo uporabili pri linearni regresiji

$$\Theta_j = \Theta_j + \alpha(y^{(i)} - h_\Theta(x^{(i)}))x_j^{(i)}.$$

Razlika je seveda v tem, da je tukaj $h_\Theta(x^{(i)}))$ logistična funkcija, kot smo jo definirali zgoraj.

3 Podatki in opis problemske domene

Podatki so iz tekmovanja iz področja kemoinformatike. Problemska domena ima 1776 atributov in 3751 primerov. Vrednosti atributov so med 0 in 1, kar precej pa jih ima samo dve različni vrednosti.

4 Rezultati

Rezultati 5-kratnega prečnega preverjanja so zbrani v tabeli 1. Kot je razvidno iz tabele smo najboljši rezultat dobili pri lambda_ 0.01. Nismo pa uspeli izboljšati rezultata prejšne domače naloge, kjer smo z RF algoritmom dobili logloss 0.4643.

Tabela 1: Rezultati

lambda_	log_loss
0.1	0.569092454919
0.01	0.506691088418
0.001	0.557776033395
0.0001	0.745162922321
0.0	2.86933161772

5 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.

Literatura

[1] Ian H. Witten & Eibe Frank, *Data Mining Practical Machine Learning Tools and Techniques, Second Edition* Morgan Kaufmann Publishers, 2005.