Predikcia filmových hodnotení

Peter Bašista, Radoslav Krivák

Neurónové siete
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze
2010

Popis dát pre úlohu

- Pôvod dát: Filmová databáza Netflix
- Použité dáta:

  http://archive.ics.uci.edu/ml/datasets/Netflix+Prize

  boli určené pre súťaž Netflix Prize

  http://www.netflixprize.com/
- Pôvodných dát je veľa (665MB vo formáte gzip)
- Vybrali sme si iba veľmi malú podmnožinu (2MB bez kompresie)
- Ale vyberali sme ju dôkladne...

Vybrané dáta

- Pôvodná databáza obsahovala:
  - Názvy filmov (odstránili sme)
  - Dátumy jednotlivých hodnotení (nebrali sme ich do úvahy)
  - Veľa filmov (viac ako 17000)
  - A veľmi veľa hodnotení od užívateľov (viac ako 10^8)
- My sme si vybrali:
  - 1000 filmov, ktoré mali v databáze najviac hodnotení
  - 1000 užívateľov, ktorí najviac hodnotili nami vybrané filmy

Dostali sme:

- Maticu `M` typu 1000 x 1000, ktorá je ”hustá” (81,68% dát)
- `M[i,j]` = *hodnotenie i. - teho filmu j. - tym užívateľom*

Postup

- Predspracovanie dát:
  - Nahradenie chýbajúcich hodnotení (priemerom)
  - Naškálovanie hodnotení na interval [-2, 2]
- Výber vstupných dát:
  - Hodnotenia prvých 200 najviac hodnotených filmov
- Výber výstupných dát:
  - Hodnotenie j. - teho filmu (j > 200)
- Trénovanie siete:
  - Metódou Scaled conjugate gradient
- Skript a dáta sú dostupné na:
  - http://basip6am.matfyz.cz/nnproject/

Výsledky 1

Architektúra	Filmy	mae(train)	mae(valid.)	mae(test)	úspešnosť
200-1-1	201-250	0.6227	0.6516	0.6733	46.77%
200-1-1	251-300	0.6274	0.6649	0.6562	46.8%
201-1-1	301-350	0.6232	0.6554	0.6612	47%
200-5-1	201-250	0.5987	0.6757	0.6665	49.52%
200-5-1	251-300	0.5824	0.6623	0.6673	50.57%
200-5-1	301-350	0.5735	0.6465	0.6561	50.67%
200-10-1	201-250	0.5580	0.6778	0.6778	52.69%
200-10-1	251-300	0.5645	0.6834	0.6839	51.85%
200-10-1	301-350	0.5569	0.6866	0.6987	51.47%
200-1-1	951-1000	0.6441	0.6754	0.6697	44.53%
200-5-1	951-1000	0.5684	0.6666	0.6733	49.81%
200-10-1	951-1000	0.5406	0.6735	0.6946	52.68%

Výsledky 2

Architektúra	Filmy	mae(train)	mae(valid.)	mae(test)	úspešnosť
200-15-1	201-250	0.5692	0.7378	0.7326	51.68%
200-15-1	251-300	0.5291	0.6968	0.7068	54.6%
200-15-1	301-350	0.5660	0.7414	0.7331	51.58%
200-20-1	201-250	0.5769	0.8158	0.7929	51.57%
200-20-1	251-300	0.5761	0.7991	0.8100	51.03%
200-20-1	301-350	0.5488	0.7811	0.7874	52.01%
200-20-20-1	201-250	0.5756	0.7116	0.7053	51.38%
200-20-10-1	201-250	0.5728	0.6653	0.6687	51.43%
200-10-5-1	201-250	0.5734	0.6489	0.6438	50.99%
200-5-3-1	201-250	0.5797	0.6397	0.6487	50.71%
200-30-1	201-250	0.6640	0.9342	0.9623	48.17%
200-40-1	201-250	0.7871	1.1365	1.1419	44.31%

Výsledky 3

Architektúra	Filmy	mae(train)	mae(valid.)	mae(test)	úspešnosť
200-20-10-5-1	201-250	0.5689	0.6463	0.6245	51.29%
200-20-15-10-1	201-250	0.5756	0.6525	0.6615	51.26%
200-15-10-5-1	201-250	0.5886	0.6471	0.6477	49.86%
200-20-5-1	201-250	0.5782	0.6426	0.6494	50.27%
200-14-7-1	201-250	0.5612	0.6397	0.6382	51.81%
200-15-5-1	201-250	0.5664	0.6443	0.6549	50.84%
200-7-5-1	201-250	0.5900	0.6594	0.6498	49.67%
200-10-3-1	201-250	0.5918	0.6405	0.6411	49.4%
200-15-3-1	201-250	0.5837	0.6481	0.6464	49.37%
200-10-7-1	201-250	0.5700	0.6523	0.6533	51.71%
200-5-2-1	201-250	0.6160	0.6663	0.6638	47.64%
200-5-10-1	201-250	0.6311	0.6838	0.6953	47.05%

Záver

Z výsledkov vyplýva, že:

- Priemerná úspešnosť sa takmer nemení
- Nalepší počet neurónov v 1. skrytej vrstve je medzi 10 a 20
- Priemerná absolútna chyba na testovacích dátach nikdy neklesla pod 0.6