Projekt zaliczeniowy

Techniki Optymalizacji

11.12.2018

Projekt dotyczy klasyfikacji za pomocą regresji logistycznej. Wykonywany będzie w zespołach **dwuosobowych**. Każdy zespół otrzyma własne zbiory danych (treningowy i testowy) w sposób następujący:

1. Osoby w zespole sumują swoje numery indeksów, i biorą resztę z dzielenia przez 40. Jest to numer przyporządkowany zespołowi. *Przykład:* w zespole są dwie osoby o numerach 143511 i 112351, po zsumowaniu 255862, modulo 40 daje to numer 22. *Uwaga:* jeśli ktoś wykonuje zadanie sam, należy po prostu wziąć resztę z dzielenia przez 40 ze swojego indeksu.

2. Pliki treningowy i testowy są dostępne pod adresem:

   ```
   http://www.cs.put.poznan.pl/wkotlowski/to/project/fileX_train.txt
   http://www.cs.put.poznan.pl/wkotlowski/to/project/fileX_test.txt
   ```

   gdzie `X` jest przydzielonym zespołowi numerem między 0 a 39 (np. dla numeru 9 będzie to nazwa `file9_train.txt`, itp.). Zbiór treningowy zawiera 20 000 obserwacji, 500 cech wejściowych o wartościach zero-jedynkowych i jedną cechę wyjściową (klasę) z wartościami 1 lub -1. Zbiór testowy ma podobną strukturę, i również zawiera 20 000 obserwacji (różnych od tych ze zbioru treningowego). **Uwaga: wszelka optymalizacja ma mieć miejsce wyłącznie na zbiorze treningowym! Zbiór testowy służy wyłącznie do przetestowania regresji.**

Waszym zadaniem będzie:

1. Zaimplementowanie algorytmu regresji logistycznej metodą Newtona-Raphsona (stała λ powinna być równa zero, jeśli hesjan jest nieosobliwy). **Uwaga: proszę pamiętać o uwzględnieniu wyrazu wolnego!** W sprawozdaniu należy zamieścić:

   (a) Liczbę iteracji i orientacyjny czas obliczeń (może być przybliżony).

   (b) Średnią wartość błędu logistycznego na zbiorze treningowym i testowym.

   (c) Średnią wartość błędu zero-jedynkowego na zbiorze treningowym i testowym.

   (d) Histogram (rozkład) wartości wag (parametrów). Można użyć np. polecenia `hist` z programu Octave.

2. Zaimplementowanie algorytmu regresji logistycznej metodą stochastycznego spadku wzdłuż gradientu. Należy dobrać długość kroku metodą prób i błędów, kierując się wskazówkami z wykładu. **Uwaga: w implementacji stochastycznego gradientu nie trzeba normalizować/standaryzować danych, ponieważ wszystkie zmienne są zero-jedynkowe. Należy natomiast dodać cechę związaną z wyrazem wolnym (same jedynki)**. W sprawozdaniu należy umieścić:

   (a) Wybraną długość kroku (z komentarzem, dlaczego taka została wybrana).

   (b) Liczbę epok i orientacyjny czas obliczeń (może być przybliżony).

   (c) Średnią wartość błędu logistycznego na zbiorze treningowym i testowym dla końcowych wartości parametrów.

   (d) Średnią wartość błędu zero-jedynkowego na zbiorze treningowym i testowym dla końcowych wartości parametrów.

(e) Wykres, który pokazuje zbieżność algorytmu. W tym celu należy po każdej epoce (nie iteracji!) obliczyć średni błąd logistyczny na zbiorze treningowym dla aktualnych wartości wag i zamieścić wykres błędu w funkcji liczby epok (końcowy błąd powinien być równy błędowi na zbiorze treningowym raportowanemu w punkcie (c)). Na wykresie należy również zaznaczyć optymalną wartość błędu treningowego uzyskanego metodą Newtona-Raphsona (np. poziomą linią przerywaną). Należy zamieścić komentarz do wykresu.

(f) Podobny wykres dla błędu zero-jedynkowego.

(g) Histogram (rozkład) wartości końcowych wag (parametrów).

3. Podanie krótkich wniosków z eksperymentu. Powinny się tu znaleźć następujące wnioski:

   (a) Która metoda optymalizacji wydaje się lepsza dla tego zbioru danych?

   (b) Jak mają się do siebie wartości błędów na zbiorze testowym i treningowym?

   (c) Jak mają się do siebie zbieżność w sensie błędu logistycznego i zero-jedynkowego (wnioski należy wyciągnąć z wykresów).

   (d) Jak mają się do siebie wagi zwracane przez obie metody? Czy są podobne? Można policzyć jakąś miarę odległości między wektorami wag (np. odległość euklidesową).

4. **Dla chętnych**: Zależność błędu logistycznego i zero-jedynkowego na zbiorze treningowym i testowym w metodzie Newtona-Raphsona dla niezerowych wartości stałej regularyzacyjnej λ. Proponuję zmieniać λ w zakresie od 2^{-10} do 2^{10} (tzn. $\lambda = 2^{-10}, 2^{-9}, \ldots, 2^{9}, 2^{10}$). Należy przedstawić wyniki na dwóch wykresach (błędów treningowych i testowych w funkcji λ, ze skalą logarytmiczną na λ oczywiście; osobny wykres dla błędu logistycznego i osobny dla błędu zero-jedynkowego) i podać wnioski. To zadanie nie jest obowiązkowe, ale jego poprawne wykonanie może podwyższyć wynik o dodatkowe 3 punkty.

Język programowania jest dowolny, ale użycie języka Octave/Matlab ułatwi zadanie z uwagi na to, że spora część kodu została już napisana w trakcie zajęć. Sprawozdania nie trzeba drukować, tylko przesłać e-mailem na adres `wkotlowski@cs.put.poznan.pl` o tytule „`[TO] Zadanie X`", gdzie `X` to numer zadania. W treści proszę podać imiona, nazwiska i numery indeksu obu studentów. E-mail powinien posiadać następujące załączniki:

1. Główny plik ze sprawozdaniem w formacie PDF. **Proszę nie przesyłać formatu .doc, .odt lub innego, tylko zamienić go na PDF**.

2. Kod źródłowy (spakowany do archiwum) lub link do repozytorium z którego mogę taki kod pobrać. Uwaga: proszę nie umieszczać kodu źródłowego w dokumencie PDF sprawozdania; proszę też nie wysyłać plików wykonywalnych (filtr antyspamowy je zwykle odrzuca).

3. Listing wag zwróconych przez obie metody, umieszczony w zwykłych plikach tekstowych (mogą być spakowane). Proszę przypadkiem nie umieszczać tych list 501 wag w sprawozdaniu.

Proszę nie załączać w e-mailu zbiorów danych (są bardzo duże)! Długość sprawozdania nie jest kryterium podlegającym ocenie, stąd proszę być zwięzłym. Nie trzeba pisać żadnych wstępów, opisu danych, itp., tylko odpowiedzieć na pytania i zamieścić odpowiednie wnioski.

Maksymalna liczba punktów do zdobycia: 16 (+3 dodatkowe)
Termin nadesłania sprawozdania: 11.01.2019. Nadesłanie po terminie będzie skutkować ujemnymi punktami.