Standaryzacja danych

18 października 2016

1 Tematy prezentacji na wykładzie

Trzy wykłady są przeznaczone na prezentacje studentów na wybrane tematy związane z przedmiotem. Terminy wykładów:

- 2 grudnia 2016,
- 14 grudnia 2016,
- 23 grudnia 2016 (lub 13 stycznia 2017 w przypadku ewentualnych godzin rektorskich)

Prezentacje można przygotowywać w grupach konkursowych, do 3 osób.

Lista proponowanych tematów znajduje sie na stronie http://www.cs.put.poznan.pl/kdembczynski/lectures/mmds-challenge/talks/list_of_topics.html. Można zaproponować własny temat.

Proszę stworzyć listy preferencji: trzy tematy prezentacji oraz data (rozumiana jako ńie wcześniej niż") prezentacji. Na ich podstawie zostanie utworzony plan prezentacji. W przypadku konfliktów decydować będzie data wystąpienia (im wcześniej tym lepiej) lub losowanie.

2 Omówienie poprzedniego zadania

3 Standaryzacja danych ⋆

Treść

Przeprowadź standaryzację danych (X_train.csv) z poprzednich zajęć na dwa sposoby:

- tak, aby średnia wartość każdej cechy wynosiła 0, a odchylenie standardowe 1,
- tak, aby dla każdej cechy minimum wynosiło 0, a maksimum 1.

Sprawdź w przypadku surowych danych oraz dla obu transformacji danych jakie są:

- minimum i maksimum każdej cechy (wystarczy jak wyswietlisz wartości dla pierwszych 20 cech),
- średnią i odchylenie standardowe dla każdej cechy,
- minimalną i maksymalną wartość spośród wszystkich cech.

4 Zaokrąglanie a średnia ⋆

Treść

Podczas wykonywania poprzednich czynności wielokrotnie używana była wartość średnia cechy. Jak na wartość średniej wpływa precyzja z jaką przechowujesz dane?

Przeprowadź eksperyment na używanym zbiorze danych. Sprawdź jaka jest średnia każdej cechy używając oryginalnych wartości ze zbioru danych oraz wartości całkowitych - zaokrąglonych do najbliższej liczby całkowitej. Czy te wartości są sobie równe (bardzo zbliżone) czy znacznie odmienne? Z czego wynika to zjawisko? Jak można rozwiązać ten problem?