Inteligencja obliczeniowa

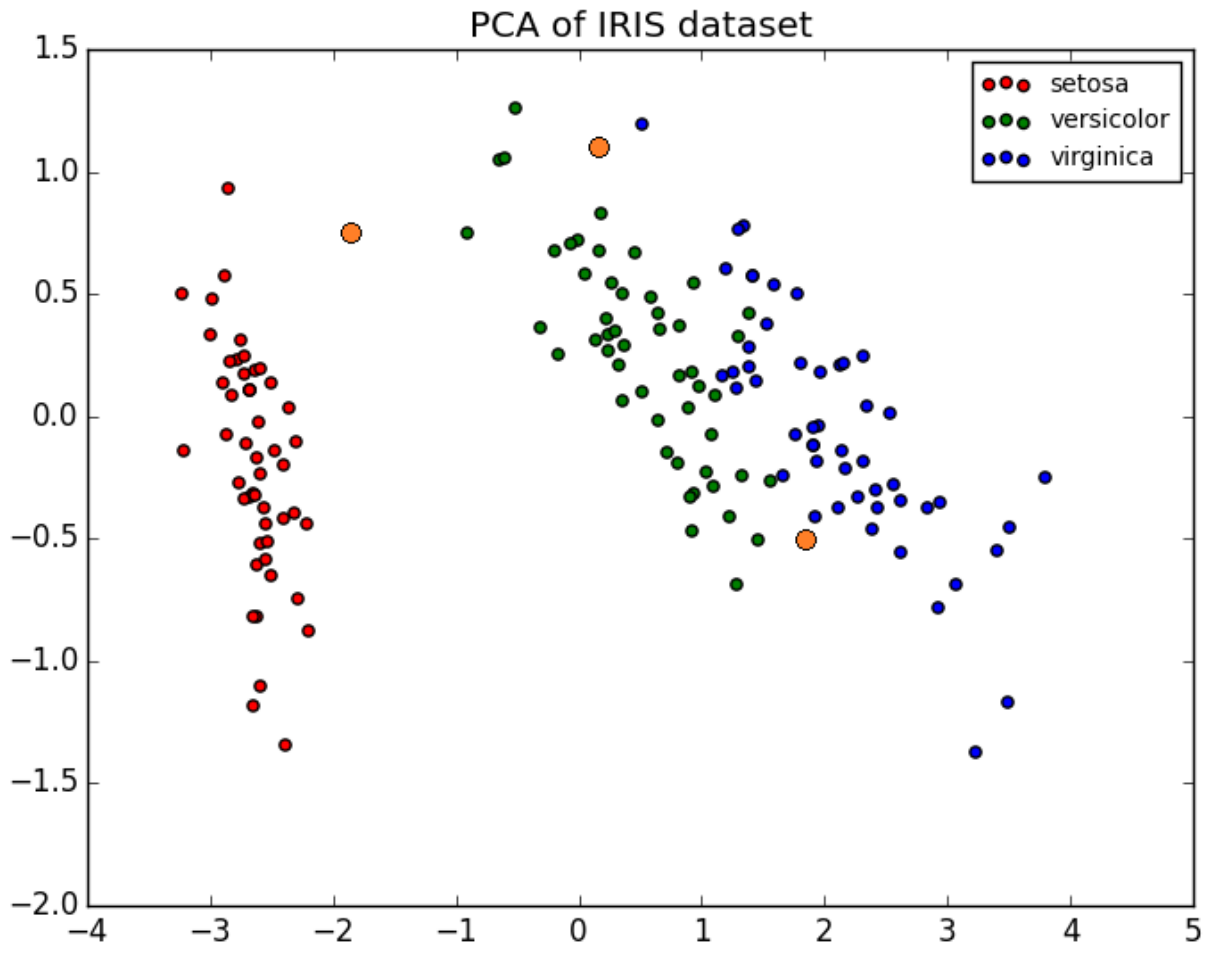
Laboratorium 5: Zadanie klasyfikacji – Algorytm k-Najbliższych Sąsiadów

Przy pomnijmy sobie bazę danych z irysami. Na poprzednich laboratoriach zmniejszlaiśmy wymiarowość bazy danych, by lepiej zobrazować przynależność do gatunków na wykresie. Jednak nie napisaliśmy programu, który rozstrzyga czy irys z takimi a nie innymi parametrami należy do tego gatunku czy innego.

Zadanie rozstrzygania do jakiej klasy należy dana instancja (rekord) zwane jest zadaniem klasyfikacji. Istnieje wiele algorytmów, które klasyfikują rekordy. Na dzisiejszych zajęciach poznamy klasyfikator, który przydziela irysy do odpowiedniego gatunku (klasy) szukając n (na przykład 3) najbardziej podobnych irysów do niego i przydzielając mu gatunek co większość z tych irysów.

Spójrz na rysunek i spróbuj powiedzieć, jaki gatunek będą miały nowe irysy (3 pomarańczowe kropki) jeśli klasyfikacji dokonamy algorytmem:

- 1-najbliższych sąsiadów
- 3-najbliższych sąsiadów



Zadanie 1
Korzystając ze wskazówek w poradniku http://blog.datacamp.com/machine-learning-in-r/ i poniższych informacji dokonaj klasyfikacji irysów metodą k-najbliższych sąsiadów.

Krok. 1. (przygotowanie i podział danych na zbiór testowy i treningowy)
Żeby algorytm nauczył się rozpoznawać gatunki irysów, trzeba dać mu zbiór treningowy, na którym dokona „nauki". Potrzebny jest też zbiór testowy, na którym przetestuje czy dobrze działa (czyli czy jego odpowiedzi, pokrywają się z klasą w tabeli).

a) wczytaj bazę danych irysów i znormalizuj dane liczbowe

b) podziel na zbiór treningowy i testowy

Krok.2 (klasyfikacja)

c) Uruchamiamy nasz algorytm, który pracuje na zbiorze treningowym. Skorzystaj z komendy knn by dokonać klasyfikacji.

d) Patrzymy teraz jakie gatunki przyporzadkowane będą irysom ze zbioru testowego. Można to rozumieć jak wstawianie pomarańczowych kropek na wykres składający się tylko z pokolorowanych kropek treningowych.

Krok. 3 (ewaluacja)
Irysy ze zbioru testowego zostały sklasyfikowane – dostałe etykietki przewidywane (predicted) od algorytmu. Czy faktycznie algorytm dobrze działał? Można go zewaluować go porównując klasę przewidzianą z realna (tą z tabeli).

e) Dokonaj ewaluacji. Sklej przewidziane gatunki z prawdziwymi ze zbioru testowego. Wyświetl jaki prcocent rekordów ma te same etykiety (liczbowo i w procentach). Do tego celu możesz napisać prosta funkcję.

f) Wyświetl macierz błędu (confusion matrix), tak jak zrobiono to w tutorialu. Sprawdź jakie gatunki pomylono.

Zadanie 2
Użyj algorytmu k-najblizszych sąsiadów do wykrywania cukrzycy u kobiet z USA. Skorzystaj z załączonej tabeli diabetes.csv. Sprawdź dla wariantu k=1, k=3, k=5, k=7. Który z nich klasyfikuje instancje najdokładniej?

Wzoruj się na krokach z zadania 1.