

Tutorial 3: PCA

1. Zbiór danych: własny (2 klasowy), MNIST (10 klasowy), kolorowy obrazek 200x200 pikseli.
2. Problem 1. Zbiór uczący: 5% , 25% , 75% (po 10 instancji dla każdego przypadku) oraz one-to-all. Reszta zbioru to zbiór testowy (założmy, że o M elementach).
3. Dla zbiorów danych wielowymiarowych zastosować klasyfikator k-NN dla każdych 10 instancji, policzyć średni błąd klasyfikacji oraz jego odchylenie standardowe, dla $k=1,3,5,100$. Błąd klasyfikacji dla elementów x_i zbioru testowego liczyć jako średnią (suma przez M) stosunku ilości k najbliższych punktów uczących należących do tej samej klasy co x_i , dzielone przez k . Odchylenie standardowe liczyć od średnich dla wszystkich 10 instancji. (W przypadku one-to-all mamy tylko średnią!). Dla jakiego k otrzymujemy najmniejszy błąd klasyfikacji oraz najmniejsze odchylenie standardowe (wariancję) dla poszczególnych przypadków uczenia i zbiorów danych? Dlaczego?
4. Zwizualizować przy pomocy PCA wyniki dla zbioru własnego, zakładając że zbiór uczący to 5% oraz 75% (dwa rysunki) oraz najlepsze k wybrane w punkcie 3. Z tym, że przynależność punktów zbioru treningowego określać na podstawie większości k najbliższych sąsiadów. Porównać do PCA w którym znana jest przynależność do klas 100% punktów. Przedyskutować wynik.
5. Dla obrazka wygenerować losowo zbiór pikseli uczących w ilości 10%, 50%, 75%. Zrobić to 5- krotnie (5 instancji dla każdego zbioru uczącego). Kolory reszty z pikseli policzyć jako średnią z $k=1, 3, 10$ najbliższych sąsiadów ze zbioru uczącego. Jak wyglądają wygenerowane obrazki?
6. Znaleźć błąd dla każdego wygenerowanego obrazka w porównaniu z oryginałem.
7. Dla każdego przypadku uczącego (10%, 50%, 75%.) policzyć średnie wartości pikseli testowych (nieznanych) z wygenerowanych 5 obrazków. Jak wygląda tak wygenerowany obrazek. Czy błąd się zmniejszył???