Autor: Szymon Tokarz

Data: 19.11.2024 r. Godz. 8.00

Ćwiczenie: Uczenie maszynowe

Rezultaty

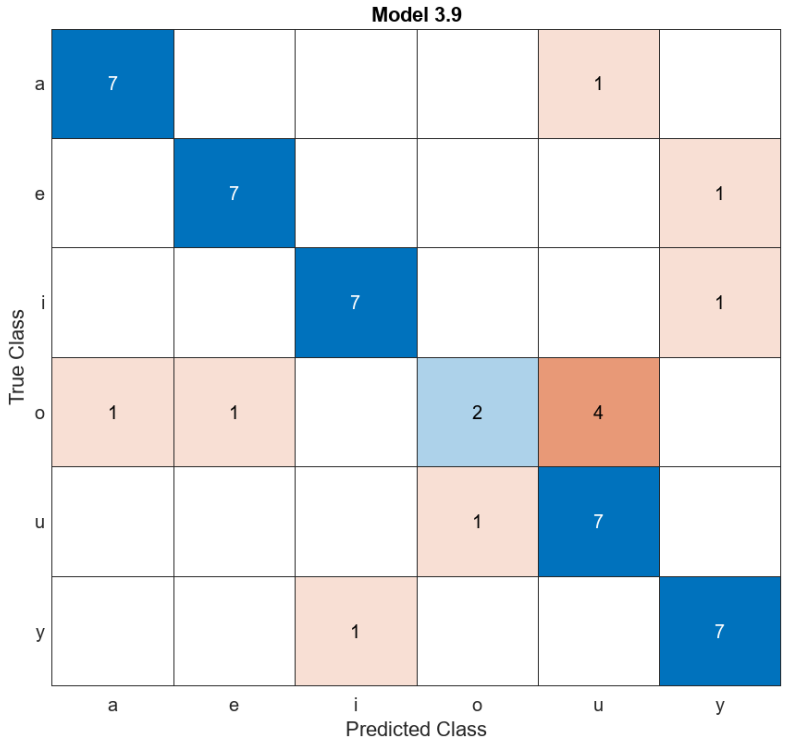
Część 1

Uzupełniony kod:

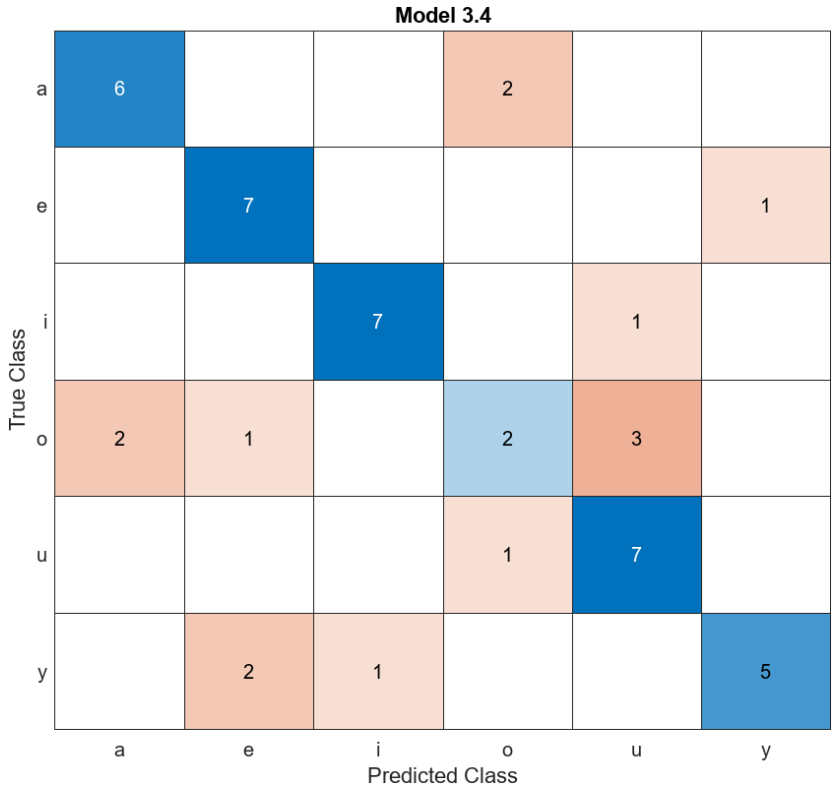
```
for i=1:nrOfVowels                          % iteracja po kolejnych samogło
    select1 = strcmp(vowelClass,uWovels(i));          % wybór element
    % UZUPELNIJ_1 (wykorzystaj sumowanie)
    nrOfexamples(i) = length(find(cell2mat(vowelClass)==uWovels{i}));
```

Część 2

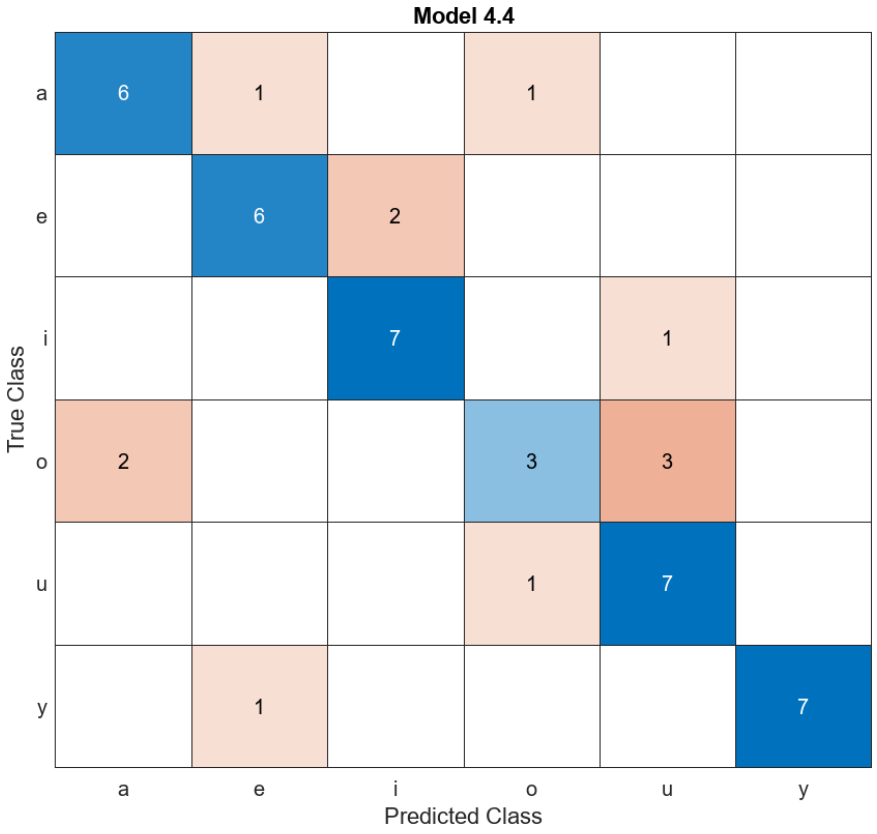
Należało zamieścić macierze pomyłek dla najlepszych dokładności klasyfikacji bez włączonej opcji PCA.



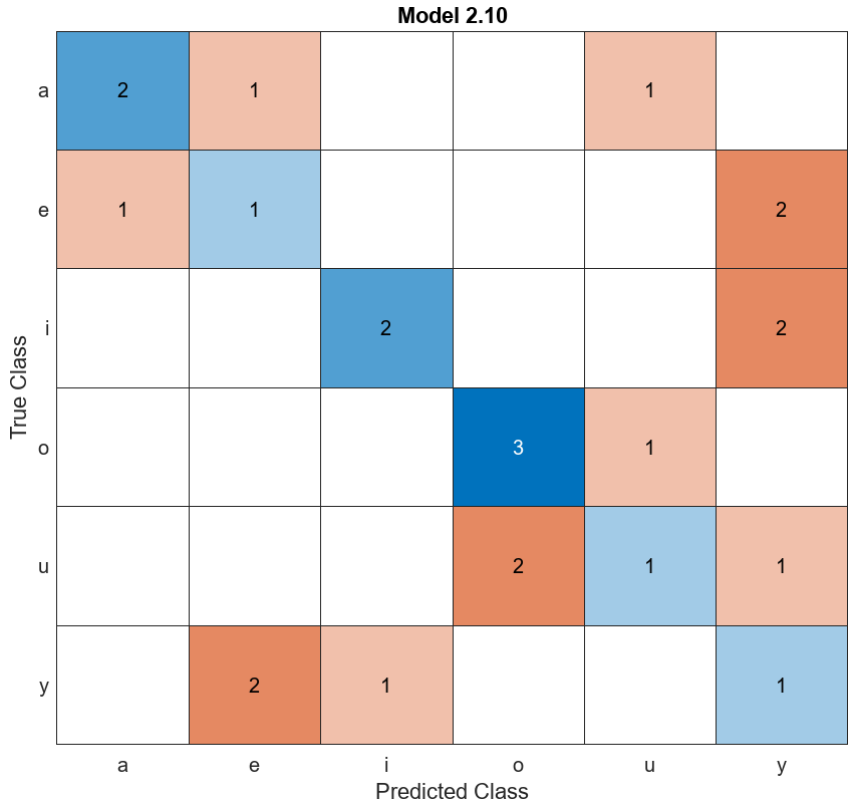
Rys.1 Macierz pomyłek dla dokładności 77.1%



Rys. 2 Macierz pomyłek dla dokładności 70.8%



Rys. 3 Macierz pomyłek dla dokładności 75% z PCA



Rys. 4 Macierz pomyłek dla dokładności 41.7% dla formantsTableTest

Analiza i wnioski

Liczebność dla wszystkich samogłosek wynosi 12. Przy podziale 70% i 30% wszystkie samogłoski mają 8 przykładów w zbiorze treningowym i 4 w zbiorze walidacyjnym. Przy podziale walidacyjnym niektóre w zbiorze treningowym mają 8 albo 9 przykładów, lecz jest to wyrównane przez liczbę przykładów w z biorze walidacyjnym. Ważne powodem, dla którego w uczeniu maszynowym klasy powinny mieć podobną liczebność jest chęć uniknięcia zbytniego skupienia się modelu na elementach neutralnych.

Z macierzy można zauważyć, żę najczęśćiej mylonymi literą jest „o”, która jest mylona z „u” i „a”. Jest to spowodowane podobnymi kształtami tych znaków.

Użycie PCA nie poprawiło dokładności klasyfikacji. Najlepszym klasyfikatorem jest model 3.9 z dokładnością 77.1% nie używający PCA.

Dla klasyfikatora wygenerowanego dla danych treningowych wartości validationAccuracy = 0.7917 i TrainAccuracy = 1 różnią się. Jest to spowodowane tym, że zbiór walidacyjny jest inny niż treningowy, przez co klasyfikacja może się różnić.

Pytania

Czym różni się prosta walidacja (holdout validation) od ręcznego podziału na zbiór uczący i testowy?

Losowo przypisujemy punkty danych do dwóch zestawów d0 i d1, zwykle nazywanych odpowiednio zestawem uczącym i testowym. Rozmiar każdego z zestawów jest dowolny, chociaż zazwyczaj zestaw testowy jest mniejszy niż zestaw treningowy. Następnie trenujemy (budujemy model) na d0 i testujemy (oceniamy jego wydajność) na d1.

Wyjaśnij na czym polega walidacja krzyżowa (k-fold cross-validation) i jaki jest jej cel. Jakie są inne sposoby podziału zbioru danych na zbiór uczący i testowy? Jaką walidację powinno się stosować w przypadku gdy zbiór danych jest niewielki?

W k-fold cross validation oryginalna próbka jest losowo dzielona na k równej wielkości podpróbek, często określanych jako „fałdy”. Spośród k podpróbek, pojedyncza podpróbka jest zachowywana jako dane walidacyjne do testowania modelu, a pozostałe k - 1 podpróbek jest wykorzystywanych jako dane treningowe. Proces walidacji krzyżowej jest następnie powtarzany k razy, przy czym każda z k podpróbek jest używana dokładnie raz jako dane walidacyjne.

Inne metody podziału zbioru:

- Leave-p-out cross-validation,

- Monte Carlo cross-validation.

W przypadku małego zbioru danych powinno wykorzystać walidacje

Wyjaśnij na czym polega specyficzność oraz precyzja klasyfikatora.

Specyficzność to miara wskazująca w jakim procencie klasa faktycznie negatywna została pokryta przewidywaniem negatywnym.

Precyzja miara wskazująca z jaką pewnością można ufać wskazaniom klasyfikatora.

Jak jest rola PCA (ang. Principal Component Analysis) w klasyfikacji?

Usunięcie wielowymiarowości przy zachowaniu najważniejszych cech.

Dlaczego stosujemy 3 zbiory: uczący, walidacyjny oraz testowy?

W celu uniknięcia nadmiernego dopasowania do zbioru uczącego.