Konkurs

Opis problemu

Do dyspozycji masz zbiór danych składający się z około dwóch tysięcy maili przerobionych na reprezentację, w której każdy email zapisany jest w postaci częstości występowania poszczególnych słów, a następnie wartość ta jest znormalizowana do zakresu [0,1] tak, iż im częściej dany wyraz występuje tym ma wartość bliższą 1.

Maile podzielone są na dwie grupy SPAM (wart. 1) i Normalny Mail (wart. 0). W zbiorze danych stosunek Spam / Nie Spam = 65%

Zadanie polega więc na wykorzystaniu dotychczasowej wiedzy do zbudowania jak najlepszego modelu predykcyjnego, który potrafiłby przewidzieć czy dany wektor odpowiada mailowi będącemu Spamem czy tez jest on Nie Spamem.

Kryteria oceny

Do oceny uzyskanych wyników zostanie wykorzystana funkcja dokładności zdefiniowana jako

$$acc = 0.5\left(\frac{\sum_{i=1}^{n} y_i{=}d_i{=}1}{\sum_{i=1}^{n} d_i{=}1} + \frac{\sum_{i=1}^{n} y_i{=}d_i{=}0}{\sum_{i=1}^{n} d_i{=}0}\right)$$

Gdzie

- n – liczba wektorów w zbiorze danych
- d – rzeczywisty wynik, który powinniśmy uzyskać
- y – wynik przewidywania sieci neuronowej
- zapis $\sum_{i=1}^{n} d_i = 1$ oznacza liczbę maili będących spamem, czyli przyjmujących wartość 1
- zapis $\sum_{i=1}^{n} d_i = 0$ oznacza liczbę maili nie będących spamem, czyli przyjmujących wartość 0
- zapis $\sum_{i=1}^{n} y_i = d_i = 1$ oznacza liczbę maili, które system przewidział jako spam i które w rzeczywistości były spamem
- zapis $\sum_{i=1}^{n} y_i = d_i = 0$ oznacza liczbę maili, które system przewidział jako nie spam (normalny email) i które w rzeczywistości były normalnymi mailami

Funkcja ta zostanie wykorzystana na etapie oceny wyników, gdzie prowadzący zajęcia dostarczy każdemu zespołowi zestaw wektorów, dla których należy dokonać przewidywania, czy dany email jest spamem czy tez zwykłym mailem.

Wyniki przewidywania w postaci jednej kolumny liczb należy przesłać prowadzącemu mailem podając w nazwie zbioru danych nazwiska osób w grupie.

Ponieważ prowadzący zna poprawne etykiety dla każdego z emaili, dokona on oceny jakości uzyskanych wyników za pomocą opisanej powyżej funkcji.

Grupa studentów która uzyska największa wartość dokładności wygrywa i jest zwolniona z zaliczenia.