13522103 - Steven Tjhia

Logistic Regression

1. Jelaskan cara kerja dari algoritma tersebut! (boleh dalam bentuk *pseudocode* ataupun narasi)

   ⇒ Pertama, model menyimpan data latihan dan melakukan inisiasi *weight* dan *bias*. Kemudian memperbarui nilai *weights* dan *bias* dengan *gradient descent* sampai didapatkan *weight* dan *bias* yang paling bagus.
   Proses pembaruan nilai *weight* dan *bias* dimulai dengan menghitung prediksi linear dengan rumus:

   z = sum(x1*w1+x2*w2+...+xm*wm+b)
   w : *weight*
   b : *bias*

   Kemudian diterapkan fungsi *sigmoid* pada hasil tersebut.

   sigmoid(z) = 1/(1+e^(-z))

   Setelah itu dihitung selisih antara hasil tersebut dengan hasil yang seharusnya yang selanjutnya akan digunakan untuk melakukan penyesuaian nilai *weight* dan *bias*. Fungsi dibawah ini merupakan hasil dari turunan fungsi *cross-entropy loss function*.

   dw = (1/n_samples) * (transpose(X).dot(predictions - y))
   db = (1/n_samples) * sum(predictions - y)
   n_samples : jumlah total sampel
   X : seluruh data tiap fitur tanpa fitur target
   y : fitur target
   predictions : hasil dari persamaan linear yang sudah diterapkan fungsi sigmoid

   Penyesuaian nilai *weight* dan *bias* tersebut dilakukan sebanyak jumlah iterasi yang diberikan. Setelah iterasi berakhir, model akan menyimpan *weight* dan *bias* yang terakhir dihasilkan. Selanjutnya nilai *weight* dan *bias* tersebut akan digunakan untuk melakukan prediksi. Hasil dari fungsi sigmoid memiliki rentang nilai antara 0 dan 1, sedangkan pada klasifikasi yang akan dilakukan hanya

terdapat 2 klasifikasi, yaitu 0 atau 1 sehingga hasil dari fungsi sigmoid perlu disesuaikan dengan mengubah hasil yang bernilai 0.5 atau lebih besar menjadi 1 dan hasil yang bernilai lebih kecil dari 0.5 akan diubah menjadi 0.

2. Bandingkan hasil evaluasi model from scratch dan *library*, bagaimana hasil perbandingannya? Jika ada perbedaan, jelaskan alasannya!

Akurasi [holdout 80-20]	
Scratch	***Library***
0.677	0.675
Akurasi [k-fold 5]	
Scratch	***Library***
0.668	0.670
Waktu Eksekusi	
Scratch	***Library***
0.015 detik	0.019 detik

Berdasarkan data hasil percobaan, hasil antara model *Logistic Regression* dengan implementasi sendiri dan dengan menggunakan *library* tidak memiliki perbedaan yang signifikan.

3. Jelaskan *improvement* apa saja yang bisa Anda lakukan untuk mencapai hasil yang lebih baik dibandingkan dengan hasil yang Anda punya saat ini! *Improvement* yang dimaksud tidak terbatas pada bagaimana algoritma diimplementasikan, namun juga mencakup tahap sebelum *modeling and validation*.

   ⇒ Feature selection : fitur yang tidak relevan akan sangat mengganggu kinerja model sehingga diperlukan proses *feature selection* agar hasil yang didapatkan bisa lebih baik.

   ⇒ Melakukan normalisasi data misalnya dengan menggunakan StandardScaler ataupun RobustScaler. Untuk data yang tidak terdistribusi normal lebih baik

menggunakan RobustScaler karena RobustScaler menggunakan *interquartile range* sedangkan StandardScaler menggunakan rata-rata dan standar deviasi.