Projek KB

“Memprediksi berapa kemungkinan terbanyak bet yang dimenangkan dari 10 kali bet”

1.Neural Network

Adalah bentuk pola yang mengadopsi layaknya kerja otak manusia dimana terdapat input data masuk,yang kemudian diproses,dan menghasilkan output.

Kenapa menggunakan Neural Network?

Neural network memprediksi lebih baik, karena menggunakan hidden layer,

dalam menguji dataset yang kita miliki. Dibanding menggunakan regresi linear,

hanya menggunakan node input dan output dalam prediksi tanpa mengolah data lebih dalam.

Supervised atau unsupervised?

Menggunakan Supervised karena tujuan dari adanya training data adalah untuk menghasilkan hanya output yang kita kehendaki.

Dalam penggunaan data ,data dikelompokan berdasar suatu nama kategori tertentu,misal horse age,horse country, horse type, horse rating,actual weight.

Kalau menggunakan unsupervised,output yang dihasilkan masih tidak diketahui polanya,abstak,jadi bisa tidak sesuai harapan, dan dapat membuang waktu.

2.Data Preprocessing

Data yang digunakan berasal dari 2 file csv,yaitu races.csv dan runs.csv

2 tabel ini saling berkaitan pada kolom race\_id dan diperlukan join dalam memproses data

1. Read races.csv dulu

Line 1-6: Untuk import package yang dibutuhkan selama project seperti pandas,tensorflow,numpy

Line 8-9: Read csv dari race,dengan bantuan pandas

Pilih 6 column yang dianggap penting

(venue,config,surface,distance,going,race\_class)

Line 11-1 :Mengecek data valid / tidak ,misal kolom kosong akan dihilangkan/drop

Line 15-23: Encode data,karena data yang ada di kolom beberapa adalah string,jadi harus diubah ke int dulu

Ada 2 tipe yaitu nominal encode dan ordinal encode

1.nonimal label encoder : buat data 1d(n\_sample)

2.ordinal label encoder : buat data 2d(n\_sample,n\_features)

\*\*Nominal Label encoder : Data non-numerik tanpa urutan yang melekat/urutan tertentu

\*\*Ordinal Label Encoder : Data non-numerik dengan urutan status, misal survey kepuasan

2. Process Runs.csv

Line 1-4:memasukan data/read(sama seperti races.csv)

Line 6-8 : cek kolom kosong(sama seperti races.csv)

Line 10-13:dalam 1 pertandingan terdapat 14 peserta,jadi kalau peserta yang lebih akan dieliminasi

Line 15-19:encode nominal column yaitu horse country,horse type

3.Memasukan data ke neural network

Yang dimasukan ke neural network

1 baris dari races.csv + semua dari runs.csv

Code yang digunakan berfungsi untuk merubah urutan data/reshape runs.csv

Supaya baris dari race yang sama bisa jadi 1 row

Pembahasan coding:

Line 1-5: function buat sorting column

Line 7 : proses reshape supaya dalam race yang sama jadi 1 baris

Co:awalnya

[(horse\_age1),(horse\_age2),(horese\_age3)]

Jadi

[(horse\_age1),(horse\_country1),(horse\_type1)...

..(horse\_age2),(horse\_country2),(horse\_type2)]

Jadi akan terbentuk pola

Horse 1 feature,horse 2 feature,horse 3 feature

Dalam prosesnya,menghindari loop agar lebih efisien

Line 8-10:sort column function dari line 1-5 dipanggil

Selain horse feature yang sudah dikategorikan,di akhir tiap feature

Ditambah kolom result,jadi kolom paling kanan polanya

[(result,1).....(result,11),(result,12),(result,13)]

3. Siapkan train dan dataset

Penjelasan code

Line 1:join data race dengan runs

Line 2:select semua kolom kecuali 14 kolom terakhir(dianggap sebagai input x)

Line 3-4:gunakan scaling technique

Standardisation (co:declared weight nominalnya bisa lebih dari 1000+)

Jika tidak network akan kesulitan mempelajari

Line 6:convert kolom result sebagai output y,jika 1 artinya horse menang,0 kalah

Line 8-9:print untuk melihat bentuk input dan outputnya

Hasilnya:

(6048,104)

(6348,14)

Artinya ada 6348 race,104 input feature,dan 14 output

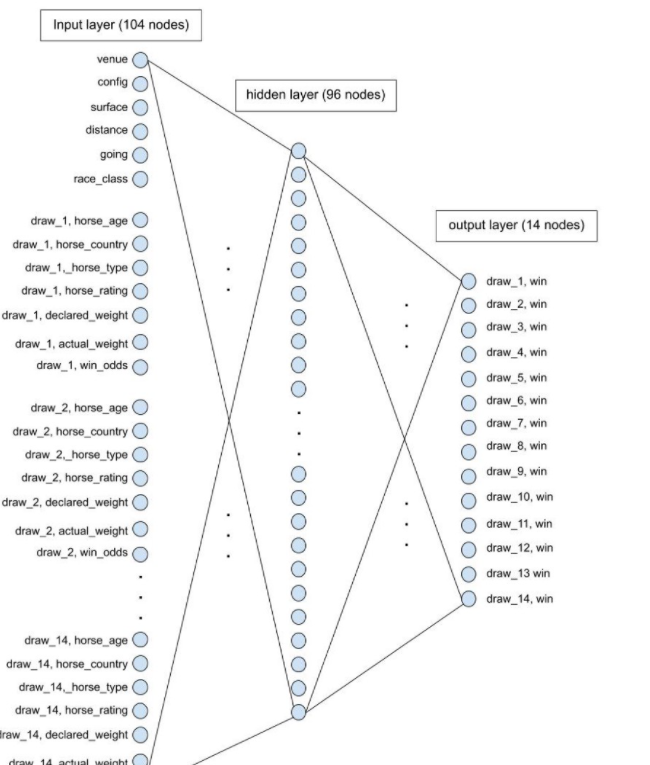
Train\_test\_split untuk split data

* 80% for training
* 20% for testing(validation)

Line 11-12:pakai handy util method dari sk learn untuk split data jadi train dan test

Validation set

4.Membuat Model Neural Network



Ada 3 layer:

1. Input layer

Ada 104 node

6 dari race.csv (cuma 1 match saja jadi 6\*1)

7 \*14 peserta ->karena dari 14 peserta masing-masing punya 7 data jadi 7\*14

Total 104 node

1. Hidden Layer

Jumlah node pada hidden layer tidak ditentukan,jumlahnya bebas seusai kebutuhan

Tapi ada beberapa kriteria

1.Harus di antara input dan output,misal diantara 14-104

2.Untuk menghitung estimasi node yang dibutuhkan menggunakan rumus

sqrt(input node \*output node )

1. Output Layer

Jumlah output layer yang ada sebanyak 14 node,karena terdapat 14 peserta ,

Dimana tiap node mengindikasikan apakah seekor kuda memenangkan

Pertandingan

Penjelasan code

Bisa disimpulkan banyak parameter yang bisa di train adalah 11,438

4.Train the network

Line 1-4:mengacak data,dari 11,438 diambil sebanyak 500 data

Acak dilakukan 2 kali untuk train dan test

Line 6-8:training dilakukan

Line 10-26:Menampilkan graph dengan plot

tf.data.Dataset.from\_tensor\_slices

Untuk mendapat irisan dari array data x\_train dan y\_train

dataset.shuffle(len(X\_train)).batch(500)

Mengambil 500 data dari 11.438 data secara acak

validation\_dataset=data untuk metest training data

Model.fit

Untuk call model agar mentrain data yang dimasukan dalam parameter

Hasil yang dihasilkan:

Sebuah nilai precision tertinggi dari berapa banyak epochs training yang dilakukan

Misal precision tertinggi adalah 0.3 maka jika memprediksi win sebanyak 10 kali,hanya 3 yang benar

precision = 0.3, means If we bet 'Win' 10 times based on the model's prediction, only 3 times is correct.

1. Penjelasan model.fit
2. Penggunaan sk learn buat split data
3. Dense dense?

Data-data yang digunakan:

Races.csv

Race\_id = Id untuk membedakan suatu pertandingan

Venue = Tempat pertandingan berlangsung , ST(Shatin), HV(Happy Valley)

Config = Konfigurasi lapangan(A , A+2, A+3, B, B+2) semakin data ke bawah, semakin kecil lebar lapangan

Surface = Permukaan lapangan (1 = Tanah, 0 = Tanah Berumput)

Distance = Jarak lomba (satuan meter)

Going = Kondisi track

Race class = (Rating kuda yang dipertandingkan)

Semakin kuat kuda class semakin tinggi,berhubungan dengan kolom rating

Runs.CSV

Race\_id=id pertandingan

draw=urutan kuda di starting gate

Horse age=karir kuda umur 2-10,peak 4-6

Horse country=asal kuda,menentukan umur kuda dari northern hemisphere(negara utara) ulang tahun tiap 1 januari ,dari southern hemisphere (negara selatan )ulang tahun tiap 1 agustus

Horse type=tipe kuda

Horse rating=rating kuda,menentukan class

Declared weight =berat yang didaftarkan termasuk kuda dan jockey dalam lbs

Actual weight = berat yang dibawa kuda dalam lbs

Wind odds =kemungkinan menang ,semakin besar semakin bagus

Result = hasil pertandingan sebelumnya