

- a. Apa itu supervised learning?
- b. Jelaskan bagaimana cara kerja dari algoritma yang anda implementasikan!
- c. Bandingkan ketiga algoritma tersebut, kemudian tuliskan kelebihan dan kelemahannya!
- d. Jelaskan penerapan dari algoritma supervised! (misal di bidang kesehatan atau industri)

Jawab

- a. Supervised learning merupakan algoritma machine learning yang membutuhkan “supervisor”. Supervisor yang dimaksud merupakan data berlabel yang digunakan untuk membangun model machine learning dengan tingkat akurasi yang dapat ditingkatkan seiring dengan banyaknya data yang diolah oleh model.
- b.
 - Id3
 1. Hitung entropi dari tiap atribut, dan dari entropi yang didapat hitung information gain untuk tiap atribut
 2. Pilih atribut dengan information gain tertinggi sebagai node
 3. Lakukan partisi pada atribut pada no 2 berdasarkan nilai unik pada atribut tersebut
 4. Partisi pada no 3, selanjutnya, akan menjadi cabang dari node pada no 2
 5. Atribut pada no 2, selanjutnya, tidak akan dihitung lagi information gainnya
 6. Ulangi langkah 1-5 hingga:
 - a. Setiap data pada partisi merupakan bagian dari kelas yang sama
 - b. Tidak ada lagi atribut yang tersisa
 - c. Tidak ada data pada partisi
 - Knn
 1. Tentukan nilai parameter K, yaitu jumlah tetangga terdekat
 2. Hitung jarak untuk setiap data yang akan diklasifikasikan (data tes) dengan setiap data latih
 3. Untuk tiap data tes, cari K data latih terdekat berdasarkan hasil dari no 2
 4. Hasil klasifikasi akan berupa kelas yang paling sering muncul pada no 3
 - Logistic regression
 1. Logistic regression menggunakan fungsi sigmoid untuk melakukan prediksi dengan masukan berupa vector weight (w) • vector dari fitur (x) + bias (b)

2. $\text{sigmoid}(z) = 1/(1+e^{-z})$, dengan $z = wx + b$
3. Lakukan iterasi sebanyak epochs, dengan setiap iterasi:
 - a. Hitung sigmoid dari data latih dengan w awal bernilai 0 seluruhnya, dan b juga bernilai 0
 - b. Hasil kembalian dari poin a akan disebut sebagai \hat{y}
 - c. Untuk setiap \hat{y} yang didapat, hitung gradient dari w (dw) dan b (db)
 - d. $dw = (\text{transpose}(x) \cdot (\hat{y} - y)) / m$, dengan y adalah vector label dan m banyak data latih
 - e. $db = \text{sum}(\hat{y} - y)/m$
 - f. Dengan tiap dw dan db yang didapat, hitung w dan b baru
 - g. $w_{\text{baru}} = w_{\text{lama}} - \text{learning_rate} \times dw$
 - h. $b_{\text{baru}} = b_{\text{lama}} - \text{learning_rate} \times db$
 - i.
4. Prediksi akan dihitung menggunakan fungsi sigmoid dengan w dan b baru yang didapat pada no 3, dan x berupa data tes yang akan diprediksi

c.

- Id3
 - i. Kelebihan
 1. Tidak membutuhkan normalisasi atau standardisasi
 2. Pra-proses lebih sedikit dibandingkan algoritma lain
 3. Tidak dipengaruhi oleh missing value
 - ii. Kekurangan
 1. Proses pelatihan relatif mahal seiring kompleksitas data bertambah
 2. Tidak dapat digunakan pada data kontinu
 3. Perubahan kecil pada data dapat memicu perubahan yang besar pada struktur pohon yang akan dibuat
- Knn
 - i. Kelebihan
 1. Implementasi relatif lebih mudah

2. Tidak ada proses pelatihan
 3. Penambahan data tidak merusak model, sebab tidak ada proses pelatihan
- ii. Kekurangan
 1. Seiring bertambahnya ukuran dan dimensi data, proses penghitungan dapat menjadi sangat lama
 2. Membutuhkan standardisasi atau normalisasi
 3. Sensitif terhadap outlier dan missing value
- Logistic regression
 - i. Kelebihan
 1. Proses implementasi cukup mudah
 2. Proses pengklasifikasian relatif cepat
 3. Tidak melakukan asumsi mengenai distribusi data
 - ii. Kekurangan
 1. Tidak dapat diaplikasikan pada masalah-masalah non-linear
 2. Akan terjadi overfit jika jumlah data kurang dari jumlah atribut
 3. Melakukan asumsi mengenai linearitas antara variabel dependen dan variabel independen
- d. Algoritma supervised learning dapat diterapkan pada berbagai bidang, salah satunya kesehatan. Salah satu penerapan supervised learning pada bidang kesehatan adalah GeNose, yaitu sebuah alat yang dibangun untuk memprediksi apakah seseorang terjangkit covid-19 atau tidak berdasarkan nafas seseorang.