GAIB Easy Supervised Learning - Michael Jonathan Halim - 13521124

1. Jelaskan yang dimaksud dengan *supervised learning* dan cakupannya!

   Supervised learning merupakan salah satu pendekatan dalam machine learning dimana algoritma trainingnya menggunakan data yang sudah diberi label untuk membuat prediksi. Algoritma menerima data yang berisi fitur-fitur yang diperlukan untuk melakukan klasifikasi ataupun regresi terhadap label target. Klasifikasi merupakan salah satu cakupan dalam supervised learning dimana algoritma akan mengklasifikasikan data ke dalam kategori atau kelas sesuai dengan data yang sudah di-train. Contohnya adalah deteksi email spam, apakah suatu email spam atau tidak. Regresi adalah cakupan dalam supervised learning juga dimana algoritma akan memprediksi suatu nilai numerik yang kontinu.

2. Jelaskan cara kerja algoritma yang telah diimplementasikan!

   Untuk algoritma yang diimplementasikan, mulai dari KNN. Untuk KNN sendiri, dibuat kelas KNN yang memiliki konstruktor untuk set nilai k. Lalu, terdapat method euclidean distance untuk menghitung euclidean distance antara satu baris predictor dan baris lainnya. Terdapat method untuk mengubah nilai k juga, dan juga terdapat method untuk mendapatkan tetangga terdekat dari data yang ingin diprediksi. Method fit digunakan untuk menyimpan data training. Terakhir, method predict digunakan untuk prediksi dengan mencari k nearest neighbours dan mencari label dengan count terbanyak/mode sebagai output label prediksi.

   Untuk Logistic Regression, terdapat kelas LogisticRegression yang memiliki konstruktor untuk set batch_size, epochs, dan learning rate. Metode sigmoid untuk menghasilkan probabilitas kelas. Metode loss untuk menghitung loss function dari logistic regression. Metode gradients untuk menghitung gradien dari loss yang digunakan untuk memperbarui weight dan bias nantinya. Metode fit untuk men-train model dengan menghitung bobot dan bias yang optimal dengan memanfaatkan metode-metode yang tadi sudah disebutkan. Dalam setiap iterasi batch, dihitung y_hat dengan metode sigmoid untuk perhitungan gradien, lalu dihitung gradien (dw = gradien terhadap bobot dan db = gradien terhadap bias) untuk memperbarui weight dan bias (weight dikurangi oleh learning rate * dw dan bias dikurangi oleh learning rate * db). Metode predict digunakan

untuk prediksi dengan menghitung probabilitas kelas dan mengklasifikasikannya berdasarkan threshold 0.5 (binary classification).

Untuk ID3, terdapat kelas ID3 yang memiliki metode entropy untuk menghitung entropy dari daftar label, metode information_gain, metode find_best_split untuk mencari atribut terbaik untuk membagi data, metode create_tree untuk membuat decision tree secara rekursif, metode fit untuk membuat tree berdasarkan X_train dan y_train, metode predict_row untuk prediksi label dari suatu baris data, dan metode predict untuk prediksi seluruh X_test. Proses training terjadi dengan alur yaitu mendapatkan daftar fitur dan target label untuk membuat tree. Proses pembuatan tree dilakukan dengan mengambil label dari data, jika semua baris memiliki label yang sama maka pohon selesai. Jika tidak ada atribut yang tersedia untuk dipisah, maka simpul terakhir adalah label yang paling umum. Jika masih ada atribut yang bisa dipisah, kita cari atribut terbaik berdasarkan gain information tertinggi. Cabang dari node sekarang akan terbentuk berdasarkan nilai-nilai unik dari atribut terbaik sekarang, pembuatan cabang tree akan dipanggil secara rekursif dengan fitur-fitur tersisa. Proses prediksi dilakukan dengan mengiterasi decision tree secara rekursif. Yang dilakukan adalah memeriksa node dari akar dengan memeriksa apakah terdapat data prediksi pada cabang. Jika tidak maka nilai default dari cabang tersebut akan dikembalikan. Jika ada maka akan dioperasikan subtree tersebut secara rekursif hingga didapatkan label prediksi.

3. Bandingkan ketiga algoritma tersebut, lalu tuliskan kelebihan dan kelemahannya!
    a. KNN:
        i. Kelebihan: Mudah untuk diimplementasikan, tidak memerlukan suatu periode untuk proses training karena data yang di-fit ke model akan digunakan untuk perhitungan saat prediksi, dan data baru lebih cepat untuk di-fit ke model karena tidak memerlukan periode training.
        ii. Kekurangan: Tidak bagus untuk big data karena akan memakan waktu lama untuk melakukan perhitungan euclidean distance dan tidak bagus juga untuk data dengan dimensi yang besar.
    b. Logistic Regression:
        i. Kelebihan: Lebih akurat pada data dengan dimensi rendah, sangat efisien untuk dataset yang memiliki fitur-fitur yang terpisah secara linear

ii. Kekurangan: Akan overfit pada data dengan dimensi tinggi, tidak bisa handle data non-linear

c. ID3:

i. Kelebihan: Sangat cepat dalam prediksi klasifikasi data baru, dapat mengatasi atribut yang tidak relevan dengan mudah

ii. Kekurangan: Tidak memperhitungkan interaksi antar atribut, setiap batas keputusan melibatkan satu atribut saja

4. Jelaskan penerapan dari algoritma *supervised* di berbagai bidang (misalnya industri atau kesehatan)!

Algoritma supervised learning sudah digunakan di berbagai bidang, seperti

a. Industri: Dalam bidang industri, supervised learning digunakan untuk deteksi kecacatan dalam suatu barang, optimisasi proses produksi, dan prediksi kegagalan mesin untuk membantu proses maintenance.

b. Kesehatan: Dalam bidang kesehatan, supervised learning digunakan untuk diagnosis penyakit berdasarkan gejala dan tes yang dilakukan, misalkan deteksi kanker payudara dan deteksi diabetes pada seseorang.

c. Finance: Dalam bidang finance, supervised learning digunakan untuk analisis risiko dan manajemen portofolio, fraud detection dalam transaksi keuangan, dan menilai skor kredit seseorang berdasarkan data historikal.