Soal 1

Supervised learning merupakan salah satu sub-bidang dalam machine learning yang membahas tentang pembelajaran pada data terhadap suatu label target tertentu. Hasil pembelajaran digunakan untuk memprediksi label target pada data lain. Beberapa contoh algoritma yang tercakup didalamnya adalah KNN, Logistic Regression, Decision Tree, dan ANN.

Soal 2

KNN (Classification)

1. Untuk sebuah data baru X, urutkan seluruh data yang ada berdasarkan suatu metrik jarak tertentu (pada program ini yang digunakan adalah Euclidean Distance) terhadap X.
2. Ambil sejumlah k data terdekat dengan X, kemudian beri label kepada X sesuai dengan mayoritas label pada k data yang dipilih
3. Ulangi poin 1 dan 2 sampai seluruh data test sudah dilabeli

Logistic Regression

Secara general, logistic regression akan mempelajari suatu parameter w dan b sehingga fungsi loss (loss function) dapat diminimalisasi. Berikut penjelasan secara lebih detail:

1. Inisialisasi parameter w dan b dengan nilai 0
2. Lakukan proses forward propagation dengan urutan kalkulasi sebagai berikut:

(1) $z^{(j)} = w^{(j)} \cdot x^{(j)} + b^{(j)}$

(2) $a^{(j)} = \sigma\left(z^{(j)}\right)$ dengan $\sigma(x) = \frac{1}{1-e^{-x}}$

$\sigma(x)$ disebut juga sebagai sigmoid function.

3. Hitung loss dari hasil forward propagation (loss function yang digunakan di program ini adalah cross-entropy loss).

(3) $L_{CE} = -\sum_{i=1}^{m} y_{true}^{(i)} * \log a^{(i)}$ $atau$ (3′) $L_{CE} = -\left(\sum_{i=1}^{n} y_{true}^{(i)} * \log a^{(i)}\right) + \frac{\lambda}{2m}\sum_{i=1}^{n} {w^{(j)}}^2$

3 -> Tanpa l2 regularization

3’ -> Dengan l2 regularization

4. Lakukan proses backpropagation dengan urutan kalkulasi sebagai berikut:

(4) $dw = \frac{1}{m}\sum_{i=1}^{m}\left(a^{(i)} - y_{true}^{(i)}\right) \cdot x^{(j)}$ $atau$ (4′) $dw = \left(\frac{1}{m}\sum_{i=1}^{m}\left(a^{(i)} - y_{true}^{(i)}\right) \cdot x^{(j)}\right) + \frac{\lambda w^{(j)}}{m}$

4 -> Tanpa l2 regularization

4’ -> Dengan l2 regularization

(5) $db = \frac{1}{m}\sum_{i=1}^{n}\left(a^{(i)} - y_{true}^{(i)}\right)$

Persamaan pada tahap backpropagation didapatkan dengan cara menghitung turunan parsial setiap parameter terhadap loss function menggunakan chain rule.

5. Perbarui parameter w dan b dengan cara sebagai berikut:

(6) $w := w - \alpha * dw$

(7) $b := b - \alpha * db$

6. Ulangi poin 2 sampai 5 sebanyak n kali (ditentukan oleh pengguna)

ID3

1. Pilih suatu fitur X dari data yang memiliki nilai gain tertinggi
2. Buatlah cabang pohon yang dapat memisahkan setiap data berdasarkan fitur X
3. Jika setelah pemisahan data, data yang diperolah pada suatu cabang sudah murni (hanya ada data dengan suatu label target tertentu saja), maka cabang tersebut sudah mencapai posisi paling dalam (leaf node). Jika data yang diperoleh belum murni, maka pilih kembali fitur lain selain X yang memiliki gain paling tinggi terhadap data yang ada pada cabang tersebut.
4. Ulangi poin 1 sampai 3 hingga seluruh leaf node menghasilkan data yang murni.

Soal 3

KNN

Kelebihan:

- Salah satu algoritma yang paling mudah dipahami dan mudah diimplementasikan
- Tidak ada fase training, dapat langsung memprediksi data baru
- Tidak ada asumsi spesifik terkait distribusi data train, sehingga algoritma dapat memprediksi secara fleksibel

Kekurangan:

- Berat secara komputasional
- Sensitif terhadap skala dari tiap fitur
- Performa yang kurang bagus pada data dengan fitur yang banyak

Logistic Regression

Kelebihan:

- Cepat dan efisien secara komputasional
- Memiliki performa yang baik terhadap data yang dapat dipisahkan secara linear

Kekurangan:

- Karena diasumsikan data dapat dipisah secara linear, algoritma ini tidak dapat bekerja dengan baik terhada data yang tidak dapat dipisah secara linear.
- Mudah overfit

ID3

Kelebihan:

- Hasil training model sangat mudah untuk dipahami dan intuitif
- Dapat menghandle fitur kategorikal
- Seleksi fitur dilakukan secara otomatis

Kekurangan:

- Mudah overfit
- Solusi yang ditemukan belum tentu optimal karena menggunakan strategi greedy untuk setiap proses splitting
- Pohon yang dibentuk belum tentu seimbang.

Soal 4

- Diagnosis penyakit, contohnya prediksi tumor berbahaya.
- Fraud detection, biasanya digunakan untuk memprediksi kejanggalan-kejanggalan transaksi berdasarkan aktivitas transaksi sebelumnya.
- Klasifikasi spesies makhluk hidup
- Natural language processing (NLP)
- Speech recognition
- Prediksi cuaca
- Filter spam email