GEDE PRASIDHA BHAWARNAWA

IF 2020 - 13520004

SUPERVISED LEARNING

Jelaskan yang dimaksud dengan supervised learning dan cakupannya!

1. Supervised learning adalah sebuah algoritma pembelajaran mesin atau machine learning yang pada proses pembelajarannya memiliki sebuah parameter tertentu yang dijadikan sebuah pembanding dan basis utama pembelajaran (umumnya dilabeli sebagai result/target). Secara umum, supervised learning dapat mengolah dua jenis kasus, regresi dan klasifikasi. Pada kasus regresi, hasil dari supervised learning adalah sebuah fungsi linear yang dengan sejumlah variable input, koefisien per variable, dan konstanta. Fungsi linear ini dapat digunakan untuk memprediksi data yang baru namun masih satu tipe dengan learning dataset yang digunakan. Pada kasus klasifikasi, seperti pada tugas ini, hasil dari supervised learning digunakan untuk mengklasifikasikan sebuah data baru ke dalam sebuah atau beberapa kelas. Cakupan dari algoritma supervised learning termasuk neural network, naïve bayes classification algorithm, linear and logistic regression, support vector machine (SVM), k-nearest neighbors (KNN), random tree, dan random forest. Cakupan jenis data yang dapat diproses oleh algoritma supervised learning berupa data numerik dan kategorikal.

Jelaskan cara kerja algoritma yang telah diimplementasikan!

1. K-Nearest Neighbors adalah sebuah algoritma yang didasarkan pada asumsi bahwa data-data yang tergolong pada suatu kelompok akan berposisi berdekat-dekatan pada sebuah bidang-(N-1) dengan asumsi data tersebut memiliki N fitur terdiri atas 1 fitur klasifikasi dan N-1 fitur informasi atau data tambahan. Maka dari itu, algoritma ini mengukur jarak antara data yang diuji dengan dataset benchmark. Setiap data pada benchmark dataset diukur jaraknya terhadap data yang diuji lalu diurutkan di dalam sebuah priority queue dari data dengan jarak terkecil ke data dengan jarak terbesar. Lalu, diambillah sejumlah K data dengan jarak terdekat dari priority queue tersebut. Karena pada tugas ini hanya diminta untuk melakukan klasifikasi dan bukan regresi, maka dari K data tersebut cukup dicari modus label klasifikasi (0/1, True/False). Perhitungan jarak dapat dilakukan dengan berbagai metode, namun untuk implementasi di tugas ini, digunakan Euclidean distance atau jarak mutlak antara fitur dengan rumus: (masukin rumus). Alasan pemilihan rumus tersebut adalah karena data yang digunakan semuanya bertipe numerik dan nilai jaraknya absolut (tanpa ada penggunaan asumsi atau heuristic)
2. Logistic Regression adalah sebuah metode yang mirip dengan metode klasifikasi regresi linear, namun menggunakan fungsi sigmoid sebagai ganti dari fungsi linear. Fungsi sigmoid didasarkan pada fungsi dan penurunan rumus matematika berikut. Pada metode ini, fungsi sigmoid menyatakan nilai dari probabilitas suatu kejadian memiliki sebuah klasifikasi tertentu. Pada umumnya, pembatas dari nilai probabilitas ini adalah 0.5 namun dapat disesuaikan oleh pengguna. Fungsi sigmoid dapat diperoleh dengan menggunakan parameter perkalin dot product antara matriks fitur dari dataset yang digunakan untuk pembelajaran dengan theta atau sebuah variabel bebas yang pada awalnya diinisiasikan dengan nol. Lalu, terdapat elemen learning rate dan epochs. Epochs adalah jumlah pengulangan yang dilakukan untuk memperbarui nilai dari theta dan learning rate adalah faktor yang menentukan seberapa besar atau kecil perubahan terhadap theta pada setiap

pengulangan. Suatu indikator apakah nilai theta sudah cukup mewakili model dataset adalah loss rate. Loss rate yang terlalu besar berarti model tersebut tidak dapat mewakili dataset tersebut dan berlaku juga sebaliknya. Suatu cara untuk mengurangi loss rate adalah dengan mengubah nilai epoch dan learning rate. Setelah model siap, maka model tersebut dapat di-feed dengan data uji untuk dilakukan proses klasifikasi terhadap sebuah threshold atau nilai batasan probabilitas tertentu.

3. Algoritma ID3 atau Iterative Dichotomiser 3 adalah sebuah algoritma yang menggunakan metode greedy best first search dan breadth first tree-search untuk mengklasifikasikan data. Dari namanya sendiri "iteratif" dan "dichotomiser" atau membagi menjadi dua kelas atau kategori, kita dapat menarik kesimpulan bahwa algoritma ini akan mengiterasi setiap fitur dan membagi data menjadi dua bagian hingga model tree yang dibentuk sudah cukup mewakili dataset yang ada tanpa ada cabang pohon yang redundan dan minim fungsi. Dari dataset awal, algoritma ID3 akan mencari fitur

Bandingkan ketiga algoritma tersebut, lalu tuliskan kelebihan dan kelemahannya!

1. Algoritma KNN
    a. Kelebihan dari algoritma KNN adalah tidak adanya kebutuhan untuk melakukan training. Alasannya adalah KNN tergolong algoritma Lazy Learner atau Instance Based Learning. KNN hanya menyimpan data training dan baru akan melakukan prediksi ketika diperlukan tanpa perlu ada pembentukan model terlebih dahulu. Maka, proses prediksi dapat dilakukan secara real time. Selain itu, karena algoritma KNN tidak memiliki proses training, maka penambahan data baru ke dalam training set tidak akan mempengaruhi akurasi dari algoritma. Kelebihan lainnya dari algoritma KNN adalah kemudahan untuk mengimplementasikannya. Hanya diperlukan dua parameter untuk menggunakan algoritma ini, yaitu nilai K untuk dicari modus untuk kasus klasifikasi atau mean untuk kasus regresi serta rumus jarak yang dapat berupa Minkwoski, Euclidean, dan Manhattan.
    b. Kelemahan dari algoritma KNN adalah waktunya lama pemrosesan untuk data dengan jumlah yang banyak dan/atau dengan jumlah fitur atau atribut yang banyak. Jika pada training set terdapat m buah data dengan n buah fitur, maka jumlah kompleksitas proses algoritmanya menjadi O(m * n * 1) untuk penghitungan jarak terhadap test data, O(m * log(m)) untuk mengurutkan nilai semua data berdasarkan jarak dari terkecil ke terbesar menggunakan quicksort, dan O(k) untuk mengiterasikan sejumlah k buah data untuk dicari mean atau modusnya dengan k < m. Dengan ini, maka total kompleksitas waktu algoritma KNN adalah O(m * n). Selain itu, algoritma ini juga sangat sensitif terhadap data yang atributnya ada yang NULL atau jika terdapat outlier. Ini dapat dicegah dengan cara melakukan pembersihan data sebelum diinput ke dalam training set.
2. Algoritma Logistic Regression
    a. Kelebihan dari algoritma logistic regression adalah dapat diimplementasikan dengan mudah karena termasuk salah satu model pembelajaran mesin paling sederhana dan juga tidak membutuhkan tenaga komputasi yang besar. Algoritma ini juga tepat digunakan untuk dataset dengan jumlah fitur yang tidak sedikit karena kecil kemungkinan adanya kasus overfitting. Algoritma ini juga umum dimanfaatkan sebagai benchmark untuk membandingkan performa dengan model-model pembelajaran mesin lainnya. Lalu, algoritma ini juga mudah diterapkan untuk model-model yang terpisah secara linear.

b. Kekurangan dari algoritma ini adalah adanya kemungkinan tinggi terjadi overfitting untuk dataset yang memiliki banyak fitur. Lalu, algoritma ini juga tidak dapat menyelesaikan masalah yang non-linear dan sulit menyelesaikan masalah yang bersifat kompleks. Terakhir, model ini rentan terpengaruhi oleh data duplikat, data outlier, dan data dengan fitur yang tidak relevan.

3. Algoritma ID3
    a. Kelebihan dari algoritma ID3 adalah karena didasarkan pada algoritma Decision Tree, maka dapat menghasilkan aturan klasifikasi yang mudah dipahami oleh manusia sehingga memudahkan proses debugging dan testing. Selain itu dia dapat membuat pohon klasifikasi dalam waktu yang relatif singkat untuk kedalaman yang relatif kecil. Karena menggunakan nilai dari entropi dan information gain untuk melakukan klasifikasi, maka tidak perlu dilakukan testing terhadap semua atribut hingga mencapai pure class karena terdapat kemungkinan adanya atribut yang nilai informationnya kecil sehingga relatif dapat diabaikan. Selain itu, dengan asumsi bahwa data berjumlah n dan pohon selalu membagi menjadi dua kelas baru, maka algoritma ini memproses sejumlah 2n data dari hasil deret geometri tak hingga berrasio 0.5.
    b. Kekurangan dari algoritma ID3 adalah adanya kemungkinan terbentuk model yang underfitting atau overclassified. Underfitting adalah kasus dimana model gagal mendeteksi fitur-fitur utama dari sebuah kelompok atau kelas sehingga mengakibatkan error rate yang tinggi pada training dan testing. Ini dapat dicegah dengan menambah jumlah data ke dalam proses training dan kompleksitas pada data dengan catatan tidak boleh terlalu kompleks atau dapat mengakibatkan overfitting. Kelemahan kedua adalah algoritma ini dapat berjalan waktu yang lama untuk membuat model dari data dengan jumlah fitur atau atribut yang banyak karena algoritma ini hanya dapat memproses satu atribut dalam satu waktu atau satu kali pengulangan.

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

1. Salah satu contoh yang paling mudah diamati dari penggunaan algoritma supervised learning, spesifiknya algoritma yang dibangun dengan Decision Tree sebagai fundamentalnya, adalah pada aplikasi kesehatan seperti Ada Health di Android dan iOS. Aplikasi ini dapat dibuat dengan mengimplementasikan decision tree atau ID3. Dari database penyakit yang ada, akan dicari terlebih dahulu gejala-gejala yang paling sering ditemukan sebagai fitur atau atribut dengan nilai information gain tertinggi. Lalu, semakin pengguna berinteraksi dengan aplikasi tersebut dan memasukkan lebih banyak gejala dari penyakit mereka, maka program akan dapat semakin akurat dengan klasifikasi penyakit. Tentu saja, ada opsi “Mungkin” dan “Saya tidak tahu/tidak yakin” yang dapat digunakan oleh program untuk mengiterasi model tree untuk kemungkinan iya dan tidak.