HOMEWORK

DS – 4 #W13

1. How does K-Nearest Neighbors algorithm work? Please answer with detailed  
   explanation more than what I had told in the class  
     
   Answer :

Algoritma K-Nearest Neighbor adalah algoritma supervised learning dimana hasil dari instance yang baru di klasifikasikan berdasarkan mayoritas dari kategori - tetangga terdekat.

Tujuan dari Algoritma ini adalah utuk mengklasifikasikan obyek baru berdasarkan atribut dan sample-sample dari training data.

Secara umum, cara kerja algoritma KNN adalah sebagai berikut:

- Tentukan jumlah tetangga (K) yang akan digunakan untuk pertimbangan penentuan kelas.

- Hitung jarak dari data baru ke masing-masing data point di dataset.

- Ambil sejumlah K data dengan jarak terdekat, kemudian tentukan kelas dari data baru tersebut.

Untuk menghitung jarak antara dua titik pada algoritma KNN di gunakan metode Euclidean Distance yang dapat digunakan pada 1-dimensional space, 2-dimensional space, atau multi-dimensional space.

1-dimensional space berarti perhitungan jarak hanya menggunakan satu variable bebas(Independent variable), 2-dimensional space berarti ada dua variable bebas, dan multi-dimensional space berarti ada lebih dari dua variable.

1. What are the advantages and disadvantages of K-Nearest Neighbors  
   algoritm?

Answer :

Kelebihan :

- Sangat Nonlinear = KNN merupakan salah satu algoritma (model) pembelajaran mesin yang bersifat nonparametric. Model nonparametric adalah model yang tidak mengasumsikan apa-apa mengenai distribusi instance di dalam dataset.

- Mudah di pahami dan di implementasikan = Untuk mengklasifikasi instance x menggunakan KNN, kita cukup mendefinisikan fungsi untuk menghitung jarak antar-instance.

Kekurangan :

- Perlu menunjukkan parameter K (Jumlah tetangga terdekat).

- Tidak menangani nilai hilang (Missing Value) secara implisit.

- Sensitif terhadap data pencilan (outlier)

- Rentan terhadap variable yang non-informatif

- Rentan terhadap dimensiionalitas yang tinggi

- Rentan terhadap perbedaan rentang variable

- Nilai komputasi yang tinggi.

1. What are the advantages and disadvantages of Logistic Regression and  
   when does it applied?

Answer :

Kelebihan :

* Logistic Regression tidak memiliki asumsi normalitas dan heteroskedastisitas atas variable bebas yang digunakan dalam model sehingga tidak diperlukan uji asumsi klasik walaupun variable independent berjumlah lebih dari satu.
* Variabel independent dalam Logistic Regression bisa campuran dari variable kontinu, diskrit, dan dikotomis.
* Logistic Regression tidak membutuhkan keterbatasan dari variable independennya.
* Logistic Regression tidak mengharuskan variable bebasnya dalam bentuk interval.

Kekurangan :

* Rentan terhadap Underfitting pada dataset yang kelasnya tidak seimbang.

1. Explain the differences between Decision Tree and Random Forest  
   classifier!  
     
   Answer :

Random Forest

Random Forest adalah salah satu metode berbasis klasifikasi dan regresi dimana terdapat proses agregasi decision tree. Oleh karena itu, prinsip dasar Random Forest mirip dengan Decision Tree. Masing-masing decision tree akan menghasilkan output yang bisa saja berbeda-beda. Random Forest akan memberikan output berupa mayoritas hasil dari semua decision tree. Algoritma ini memberikan akurasi yang bagus dalam klasifikasi, dapat menangani data training yang jumlahnya besar, dan juga efektif untuk mengatasi data yang tidak lengkap.

Decision Tree

Decision tree merupakan salah satu algoritma yang sangat mudah dipahami dalam klasifikasi objek. Algoritma ini adalah salah satu algoritma Supervised Learning. Decision Tree membagi data menjadi himpunan bagian berdasarkan variable inputnya. Algoritma ini merupakan jenis diagram alir yang membantu dalam proses pengambilan keputusan.

1. List points from every aspect of differences among ensemble methods!

Answer :

* Bagging : Merupakan metode yang dapat memperbaiki hasil dari algoritma klasifikasi machine learning dengan menggabungkan klasifikasi prediksi dari beberapa model. Bagging paling cocok untuk masalah dengan dataset pelatihan yang relatif kecil.
* Bossting : Merupakan cara utuk menghasilkan beberapa model atau penggolongan untuk prediksi atau klasifikasi, dan juga menggabungkan prediksi dari berbagai model ke dalam prediksi tunggal. Boosting di rancang untuk masalah biner, menciptakan kumpulan dari tiga klasifikasi yang lemah pada satu waktu.
* Stacking : Merupakan cara untuk mengkombinasi beberapa model, dengan konsep meta learner. Dipakai setelah Bagging dan Bossting.

Keuntungan

* Ensembling adalah metode yang terbukti untuk meningkatkan akurasi model dan bekerja di sebagian besar kasus.
* Ini adalah bahan utama untuk memenangkan hampir semua hackathon pembelajaran mesin.
* Ensembling membuat model lebih kuat dan stabil sehingga memastikan kinerja yang layak pada uji kasus di sebagian besar skenario.
* Untuk menangkap hubungan kompleks linier dan sederhana serta non-linear dalam data. Ini dapat dilakukan dengan menggunakan dua model yang berbeda dan membentuk ensemble dua.

Kekurangan

* Ensembling mengurangi interpretability model dan membuatnya sangat sulit untuk menarik wawasan bisnis penting di akhir.
* Memakan waktu dan dengan demikian mungkin bukan ide terbaik untuk aplikasi real-time.
* Pemilihan model untuk menciptakan ensemble adalah seni yang benar-benar sulit untuk dikuasai.

1. Does Support Vector Machine need feature scaling before the data is  
   trained? Why? Explain it!

Answer :

Standarisasi atau penskalaan biasanya bertujuan untuk mengubah skala data menjadi rata-rata 0 dan deviasi standar 1. Penskalaan berfokus pada perubahan rentang nilai kumpulan data.