NAIVE BAYESIAN

KONSEP DASAR BAYESIAN

Bayesian adalah algoritma yang digunakan untuk klasifikasi (class) secara statistik. Algoritma ini biasanya diimplementasikan untuk melakukan prediksi kemungkinan data dalam suatu kelas berdasarkan sampel atau contoh data yang sudah ada.

Algoritma Bayesian disebut juga klasifikasi Bayesian (Naive Bayesian Classifiers). Dengan rujukan algoritma ini pada model statistik, maka klasifikasi Bayesian tidak terlepas dari teorema atau rumusan Bayes. Rumusan ini memang diimplementasikan sebagai komparasi dengan algoritma lain seperti Decision Tree (termasuk algoritma C-4.5) dan Neural Network (misalnya jaringan syaraf tiruan).

Dalam berbagai analisis perbandingan, algoritma ini lebih tinggi akurasi hasil dan kecepatan, khususnya jika diimplementasikan dalam database berukuran besar. Berikut adalah konsep Teorema Bayes berdasarkan konsep statistik.

Diketahui X adalah sampel data yang klasifikasinya belum diketahui, sedangkan H adalah hipotesis dari sampel data X yang masuk ke dalam klasifikasi C. Untuk persoalan klasifikasi, maka perlu mendefinisikan P(H|X), dimana probabilitas dari hipotesis H dapat diberikan berdasarkan observasi dari sampel data X. P(H|X) disebut sebagai *posterior probability* dari hipotesis H terhadap kondisi data X.

Sebagai analogi dari konsep ini, bayangkan bahwa terdapat sekumpulan data yang terdiri atas buah-buahan. Data tersebut dideskripsikan dari parameter warna buah dan bentuknya. Jika X adalah berwarna merah dan bulat, dan hipotesis dari H bahwa X adalah buah apel. Maka P(H|X) adalah refleksi dari keyakinan bahwa X adalah apel

berdasarkan warna merah dan berbentuk bulat. Sebaliknya P(H) adalah prior probability dari H; pada contoh analogi buah apel, data sampel yang diberikan adalah buah apel, berdasarkan pengetahuan sebelumnya bagaimana buah apel tersebut dideskripsikan. Posterior probability, P(H|X) didasarkan pada informasi dari prior probability, P(H) yang secara independen dari X.

Dengan konsep yang sama, P(X|H) adalah posterior probability dari X terhadap kondisi data H. Sehingga, probabilitas bahwa X adalah berwarna merah dan bulat sudah diketahui benar bahwa X adalah buah apel. P(X) adalah prior probability dari X. Berdasarkan contoh analogi buah apel, tentunya dapat dikatakan bahwa contoh data yang digunakan dari buah-buahan adalah berwarna merah dan berbentuk bulat.

Berdasarkan konsep di atas, maka muncul pertanyaan mendasar, bagaimana probabilitas dapat dihitung? P(X), P(H) dan P(X | H) dapat dihitung berdasarkan data yang diberikan. Teorema Bayes digunakan untuk mengkalkulasikan posterior probability, P(H|X), dari P(H), P(X) dan P(X|H). Rumusan yang digunakan adalah sebagai berikut:

$$\boldsymbol{P(H|X)} = \frac{\boldsymbol{P(X|H)P(H)}}{\boldsymbol{P(X)}}$$

Langkah-langkah penyelesaian klasifikasi data dengan Naive Bayes adalah sebagai berikut:

1. Setiap sampel data direpresentasikan dalam bentuk vektor n dimensional, X=(X_1, X_2, X_3, …, X_n) yang menggambarkan n sampel dari atribut A_1, A_2, A_3, …, A_n)
2. Jika terdapat m kelas yaitu C_1, C_2, C_3, …, C_m dan diberikan data sampel yang tidak diketahui yaitu X, maka dapat dilakukan prediksi terhadap klasifikasi data X ke dalam kelas dengan nilai *posterior probability* yang

terbesar berdasarkan kondisi data X. Bayesian akan memasukan data X ke dalam kelas C_i, jika dan hanya jika $P(C_i|X)>P(C_j|X)$ untuk $1\leq j \leq m$, $j\neq1$. Dengan demikian nilai $P(C_i|X)$ dapat dimaksimalkan menjadi *maximum posterior hypothesis* menggunakan prinsip dasar Bayesian.

$$\boldsymbol{P(C_i|H)} = \frac{\boldsymbol{P(X|C_i)P(C_i)}}{\boldsymbol{P(X)}}$$

3. Jika P(X) adalah konstan untuk semua kelas data, maka hanya $P(C_i|X)P(C_i)$ yang butuh dimaksimalkan nilainya. Jika kelas *prior probabilities* tidak diketahui, maka dapat diasumsikan bahwa kelas data tersebut adalah sama, $P(C_1)=P(C_2)= \ldots = P(C_m)$ dan nilai $P(X|C_i)$ dapat dimaksimalkan. Sebaliknya jika tidak dimaksimalkan, maka nilai $P(X|C_i)P(C_i)$ yang dapat dimaksimalkan. Sebagai catatan bahwa kelas *prior probabilities* dapat diestimasi dengan cara $\boldsymbol{P(C_i)} = \frac{\boldsymbol{S_i}}{\boldsymbol{s}}$ dimana S_i adalah jumlah data sampel yang digunakan dari kelas C_i dan S adalah jumlah total data sampe keseluruhan.
4. Untuk melakukan prediksi data X yang tidak diketahui dari data sampel X, $P(X|C_i)P(C_i)$ dapat dihitung dari setiap kelas C_i. Sampel data X menjadi bagian dari kelas C_i, jika dan hanya jika memenuhi $P(X|C_i)P(C_i)>P(X|C_j)P(C_j)$ untuk $1\leq j \leq m$, $j\neq1$ atau dengan kata lain data sampel X masuk dalam kelas C_i ketika$P(X|C_i)P(C_i)$ adalah maksimum.

Konsep Bayesian di atas, dapat diimplementasikan dengan lebih jelas pada latihan soal di bawah ini.

LATIHAN SOAL: TRANSAKSI ELEKTRONIK PEMBELIAN KOMPUTER

Data berikut adalah data transaksi elektronik pembelian komputer oleh pelanggan. Terdapat 2 kelas dengan atribut buys_computer yaitu kelas yang bernilai yes

dan kelas yang bernilai no. Keputusan yes atau no, bergantung pada parameter umur (age), penghasilan (income), status mahasiswa (student) dan tingkat nilai credit (credit rating)

Tabel 25 Data Transaksi Elektronik Pelanggan Komputer

No	AGE	INCOME	STUDENT	CREDIT RATING	BUYS COMPUTER?
1	<=30	HIGH	NO	FAIR	NO
2	<=30	HIGH	NO	EXCELLENT	NO
3	31..40	HIGH	NO	FAIR	YES
4	>40	MEDIUM	NO	FAIR	YES
5	>40	LOW	YES	FAIR	YES
6	>40	LOW	YES	EXCELLENT	NO
7	31..40	LOW	YES	EXCELLENT	YES
8	<=30	MEDIUM	NO	FAIR	NO
9	<=30	LOW	YES	FAIR	YES
10	>40	MEDIUM	YES	FAIR	YES
11	<=30	MEDIUM	YES	EXCELLENT	YES
12	31..40	MEDIUM	NO	EXCELLENT	YES
13	31..40	HIGH	YES	FAIR	YES
14	>40	MEDIUM	NO	EXCELLENT	NO

Jika terdapat data X dengan spesifikasi age = "<=30", income = "medium", student = "yes", credit rating = "fair", maka lakukan prediksi menggunakan Bayesian untuk keputusan Buys Computer? Tentukan klasifikasi dari data X!