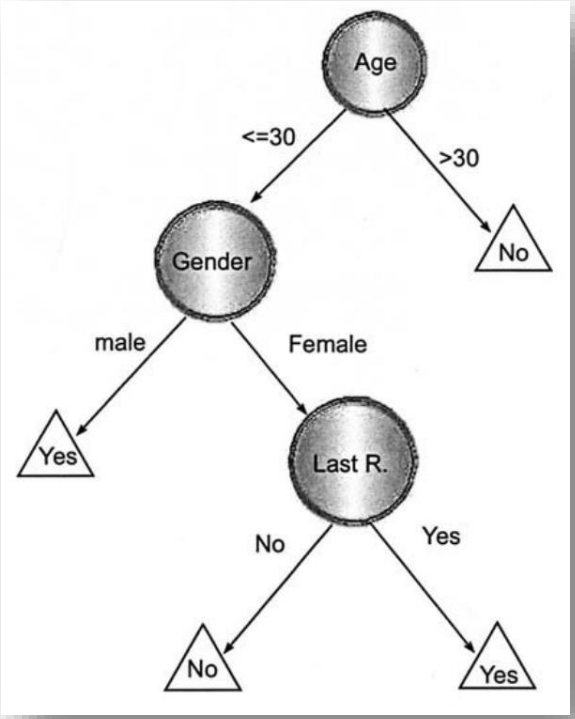
Nama : Akmal Zuhdy Prasetya
NIM : H071191035
Kelas : Machine Learning

Resume Pertemuan 4

A. Apa itu Decision Tree?

Decision Tree adalah jenis pengklasifikasi yang diekspresikan sebagai partisi rekursif dari ruang instance. Pohon keputusan terdiri dari simpul-simpul yang membentuk pohon berakar, artinya merupakan pohon berarah dengan simpul yang disebut "root" yang tidak memiliki tepi masuk. Semua node lain memiliki tepat satu edge yang masuk. Sebuah node dengan tepi keluar disebut internal atau test node. Semua node lain disebut daun (juga dikenal sebagai terminal atau node keputusan). Dalam pohon keputusan, setiap simpul internal membagi ruang contoh menjadi dua atau lebih sub-ruang sesuai dengan fungsi diskrit tertentu dari nilai atribut input. Dalam kasus yang paling sederhana dan paling sering, setiap pengujian mempertimbangkan satu atribut, sehingga ruang instance dipartisi sesuai dengan nilai atribut. Dalam kasus atribut numerik, kondisi mengacu pada rentang.

Setiap daun ditugaskan ke satu kelas yang mewakili nilai target yang paling tepat. Sebagai alternatif, daun dapat menyimpan vektor probabilitas yang menunjukkan probabilitas atribut target yang memiliki nilai tertentu. Instance diklasifikasikan dengan menavigasinya dari akar pohon ke daun, menurut hasil pengujian di sepanjang jalur. Gambar 1.1 di bawah menjelaskan pohon keputusan yang menjadi alasan apakah calon pelanggan akan menanggapi surat langsung atau tidak. Node internal direpresentasikan sebagai lingkaran, sedangkan daun dilambangkan sebagai trangles. Perhatikan bahwa pohon keputusan ini menggabungkan atribut nominal dan numerik. Dengan pengklasifikasi ini, analis dapat memprediksi respons pelanggan potensial (dengan menyortirnya ke bawah pohon), dan memahami karakteristik perilaku seluruh populasi pelanggan potensial terkait pengiriman langsung. Setiap node diberi label dengan atribut yang diujinya, dan cabangnya diberi label dengan nilai yang sesuai.



Gambar 1.1. Pohon Keputusan Menyajikan Respon untuk Direct Mailing.

Dalam hal atribut numerik, pohon keputusan dapat ditafsirkan secara geometris sebagai kumpulan hyperplanes, masing-masing ortogonal ke salah satu sumbu. Secara alami, pengambil keputusan lebih menyukai pohon keputusan yang tidak terlalu rumit, karena dapat dianggap lebih mudah dipahami. Selanjutnya kompleksitas pohon memiliki efek penting pada akurasinya. Kompleksitas pohon secara eksplisit dikendalikan oleh kriteria penghentian yang digunakan dan metode pemangkasan yang digunakan. Biasanya kompleksitas pohon diukur dengan salah satu metrik berikut: jumlah total node, jumlah daun, kedalaman pohon dan jumlah atribut yang digunakan. Induksi pohon keputusan berkaitan erat dengan induksi aturan. Setiap jalur dari akar pohon keputusan ke salah satu daunnya dapat diubah menjadi aturan hanya dengan menggabungkan tes di sepanjang jalur untuk membentuk bagian antecedent, dan mengambil prediksi kelas daun sebagai nilai kelas. Misalnya, salah satu jalur pada Gambar 1.1 dapat diubah menjadi aturan: "Jika usia pelanggan kurang dari atau sama dengan atau sama dengan 30, dan jenis kelamin pelanggan adalah "Pria", maka pelanggan akan merespons surat". Kumpulan aturan yang dihasilkan kemudian dapat disederhanakan untuk meningkatkan pemahamannya kepada pengguna manusia, dan mungkin akurasinya.

B. Kriteria Pemisahan Univariat

Dalam kebanyakan kasus, fungsi pemisahan diskrit adalah univariat. Univariat berarti bahwa simpul internal dibagi sesuai dengan nilai atribut tunggal. Akibatnya, penginduksi mencari atribut terbaik untuk dipisah. Ada berbagai kriteria univariat. Kriteria ini dapat dicirikan dengan cara yang berbeda, seperti:

- Menurut asal ukuran: teori informasi, ketergantungan, dan jarak.
- Menurut struktur ukuran: kriteria berbasis ketidakmurnian, kriteria berbasis ketidakmurnian yang dinormalisasi dan kriteria biner.

Berikut merupakan penjelasan tentang beberapa kriteria yang paling umum kemunculannya.

1. Impurity Based Criteria (Kriteria Berbasis Ketidakmurnian)

   Diberikan variabel acak x dengan k nilai diskrit, didistribusikan menurut $P = (p_1, p_2, .., p_k)$, ukuran ketidakmurnian adalah fungsi $\phi: [0,1]^k \to R$ yang memenuhi kondisi berikut:

   - $\phi(P) \geq 0$
   - $\phi(P)$ itu minimum jika $\exists i$ sehingga komponen p_i
   - $\phi(P)$ itu maksimum jika $\forall i, 1 \leq i \leq k, p_i = 1/k$
   - $\phi(P)$ simetris terhadap komponen P
   - $\phi(P)$ itu halus (dibedakan di mana-mana) dalam jangkauannya

   Perhatikan bahwa jika vektor probabilitas memiliki komponen 1 (variabel x hanya mendapat satu nilai), maka variabel tersebut didefinisikan sebagai murni. Di sisi lain, jika semua komponen sama, tingkat ketidakmurnian akan mencapai maksimum.

   Diberikan dataset training S, vektor probabilitas dari atribut target y didefinisikan sebagai:

$$P_y(S) = \left(\frac{\left|\sigma_{y=c_1} S\right|}{|S|}, \dots, \frac{\left|\sigma_{y=c_{|dom(y)|}} S\right|}{|S|} \right)$$

   Kebaikan pemisahan karena atribut diskrit a_i didefinisikan sebagai pengurangan ketidakmurnian dari atribut target setelah mempartisi S menurut nilai $v_{i,j} \in dom(a_i)$:

$$\Delta\Phi(a_i, S) = \phi\left(P_y(S)\right) - \sum_{j=1}^{|dom(a_i)|} \frac{\left|\sigma_{a_i=v_{i,j}} S\right|}{|S|} \cdot \phi\left(P_y\left(\sigma_{a_i=v_{i,j}} S\right)\right)$$

2. Information Gain

Information Gain atau Perolehan Informasi adalah kriteria berbasis ketidakmurnian yang menggunakan ukuran Entropi sebagai ukuran ketidakmurnian. Cara menghitungnya adalah sebagai berikut:

$$InformationGain(a_i, S) = Entropy(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{\left|\sigma_{a_i=v_{i,j}} S\right|}{|S|} \cdot Entropy\left(y, \sigma_{a_i=v_{i,j}} S\right)$$

dimana:

$$Entropy(y, S) = \sum_{c_j \in dom(y)} -\frac{\left|\sigma_{y=c_j} S\right|}{|S|} \cdot \log_2 \frac{\left|\sigma_{y=c_j} S\right|}{|S|}$$

3. Gini Index

Indeks Gini adalah kriteria berbasis ketidakmurnian yang mengukur perbedaan antara distribusi probabilitas dari nilai atribut target. Indeks Gini telah digunakan dalam berbagai karya seperti dan dan didefinisikan sebagai:

$$Gini(y, S) = 1 - \sum_{c_j \in dom(y)} \left(\frac{\left|\sigma_{y=c_j} S\right|}{|S|}\right)^2$$

Akibatnya kriteria evaluasi untuk memilih atribut a_i didefinisikan sebagai:

$$GiniGain(a_i, S) = Gini(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{\left|\sigma_{a_i=v_{i,j}} S\right|}{|S|} \cdot Gini\left(y, \sigma_{a_i=v_{i,j}} S\right)$$

4. Likelihood-Ratio Chi-Squared Statistics

Likelihood-Ratio didefinisikan sebagai:

$$G^2(a_i, S) = 2 \cdot \ln(2) \cdot |S| \cdot InformationGain(a_i, S)$$

Rasio ini berguna untuk mengukur signifikansi statistik dari kriteria perolehan informasi. Hipotesis nol (H_0) adalah bahwa atribut input dan atribut target ini

merupakan atribut independen bersyarat. Jika H_0 berlaku, statistik uji terdistribusi sebagai χ^2 dengan derajat kebebasan sama dengan: $(dom(a_i) - 1) \cdot (dom(y) - 1)$.