Decision Tree

Bahan Kuliah SD3104 Machine Learning

Sevi **Nurafni**

Fakultas Sains dan Teknologi
Universitas Koperasi Indonesia 2024

Aproksimasi Fungsi

Problem setting

- Himpunan yang mungkin terjadi X
- Himpunan yang mungkin label Y
- Fungsi target yang tidak diketahui $f : X \rightarrow Y$
- Himpunan hipotesis fungsi $H = \{h \mid h : X \rightarrow Y\}$

Input: Contoh Training dari fungsi target f yang tidak diketahui

$$\{(x_i, y_i)\}_{i=1}^{n} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Output: Hipotesis $h \in H$ yang paling mendekati f

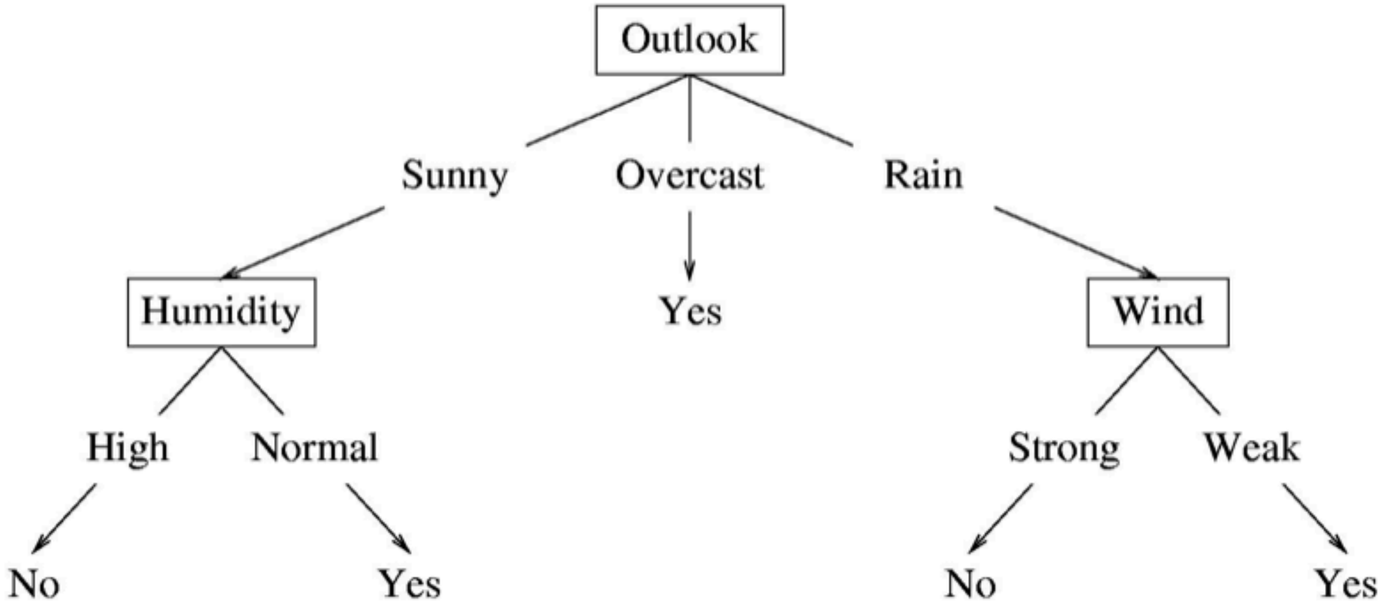
Sample Dataset

- Kolom menunjukkan fitur X_i
- Baris menunjukkan contoh yang diberi label (x_i, y_i)
- Label kelas menunjukkan apakah permainan tenis telah dimainkan $\langle \boldsymbol{x}_i, y_i \rangle$

Predictors				Response
Outlook	**Temperature**	**Humidity**	**Wind**	**Class**
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Decision Tree

- Sebuah pohon keputusan yang mungkin untuk data:

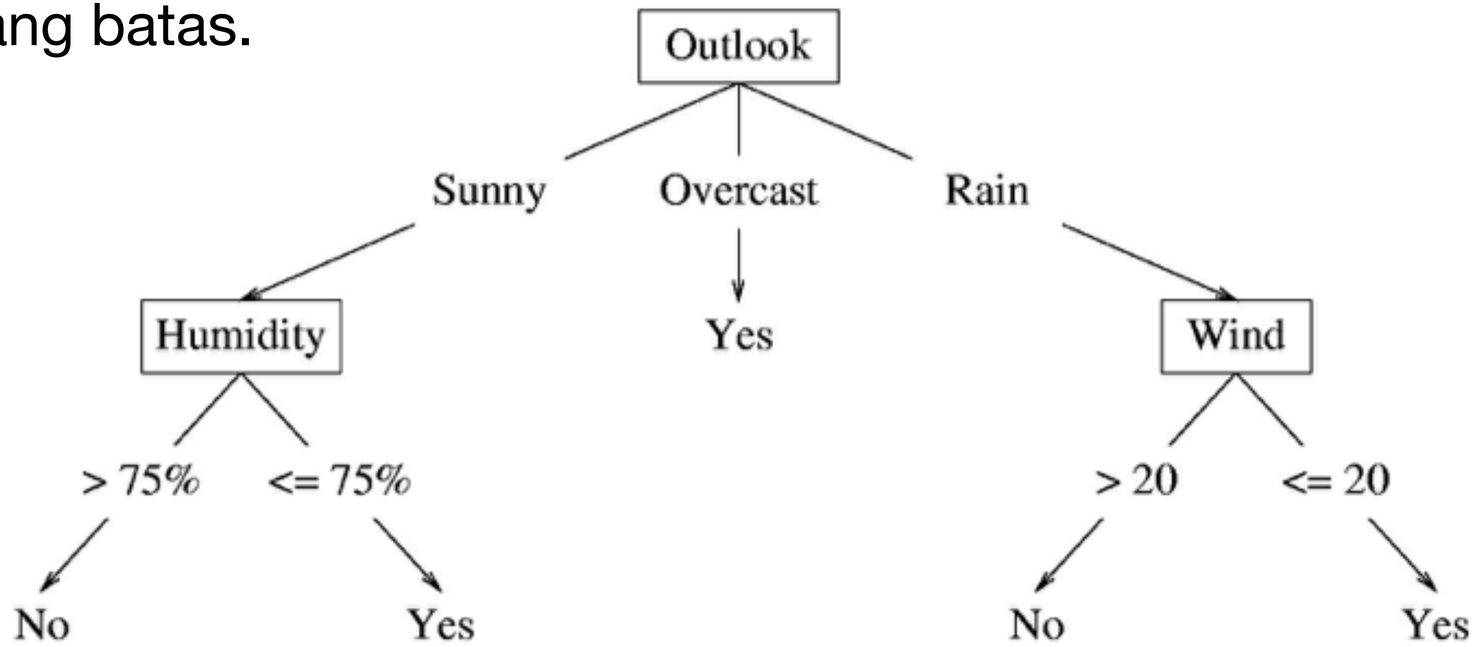


- node internal menguji nilai fitur tertentu x_i dan cabang sesuai dengan hasil pengujian
- simpul daun menentukan kelas $h(x)$

Misalkan fiturnya adalah Outlook (x_1) Temperature (x_2), Humidity (x_3), dan Wind (x_4). Kemudian fitur dari vector x = (Sunny, Hot, High, Strong) akan diklasifikasikan sebagai No. Fitur Temparature tidak relevan.

Decision Tree

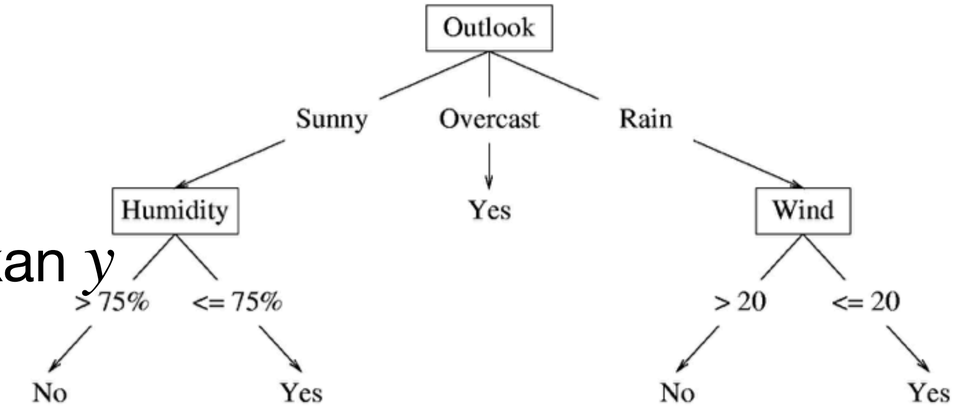
- Jika fitur bersifat kontinu, node internal dapat menguji nilai fitur tersebut terhadap sebuah ambang batas.



Decision Tree Learning

Problem setting

- Himpunan yang mungkin terjadi X
  - <Humidity=low, Wind=weak, Outlook=rain, Temp=hot>
- Himpunan yang mungkin label Y
  - Y bernilai diskrit
- Himpunan hipotesis fungsi $H = \{h \mid h : X \rightarrow Y\}$
  - Setiap hipotesis h adalah decision tree
  - Pohon mengurutkan x ke daun, yang memberikan y



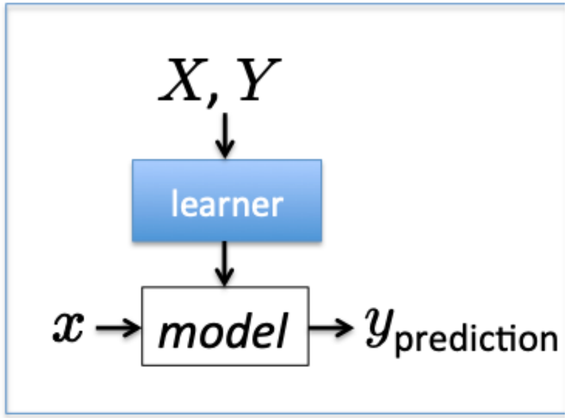
Stages of (Batch) Machine Learning

Given: data training berlabel $X, Y = \{(x_i, y_i)\}_{i=1}^{n}$

Asumsikan setiap $x_i \sim D(X)$ dengan $y_i = f_{\text{target}}(x_i)$

Melatih model:

Model ← classifier.train (X, Y)



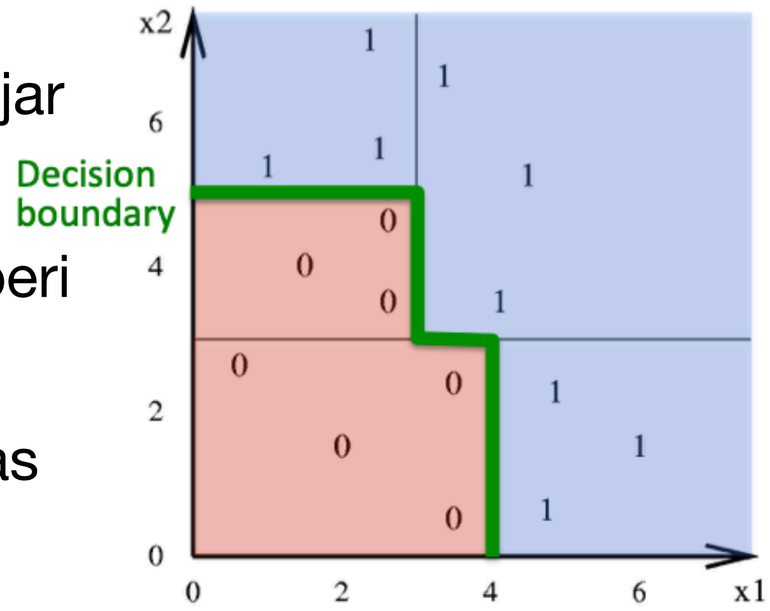
Menerapkan model ke data baru:

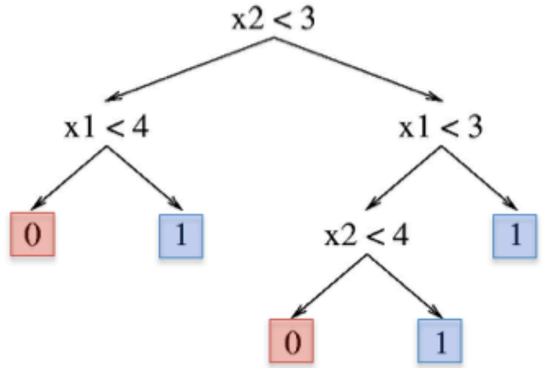
Given: contoh baru yang tidak berlabel $x \sim D(X)$

$y_{\text{prediction}}$ ← model.predict(x)

Decision Tree Untuk Menentukan Batasan

- Decision Tree membagi ruang fitur menjadi persegi panjang yang sejajar dengan sumbu
- Setiap wilayah persegi panjang diberi label dengan satu label
    - atau distribusi probabilitas atas label





Decision Trees (ID3, C4.5 by Quinlan)

- Node = root dari decision tree
- Main loop:

1. A ← atribut “terbaik” untuk node berikutnya
2. Tetapkan A sebagai atribut keputusan untuk node.
3. Untuk setiap nilai A, buatlah turunan baru dari node.
4. Mengurutkan contoh training dengan menggunakan node daun.
5. Jika contoh-contoh training diklasifikasikan dengan sempurna, STOP. Jika tidak, ulangi lagi pada simpul daun yang baru.

Memilih Atribut terbaik

Masalah utama: memilih atribut mana yang akan dibagi dari sekumpulan contoh yang diberikan

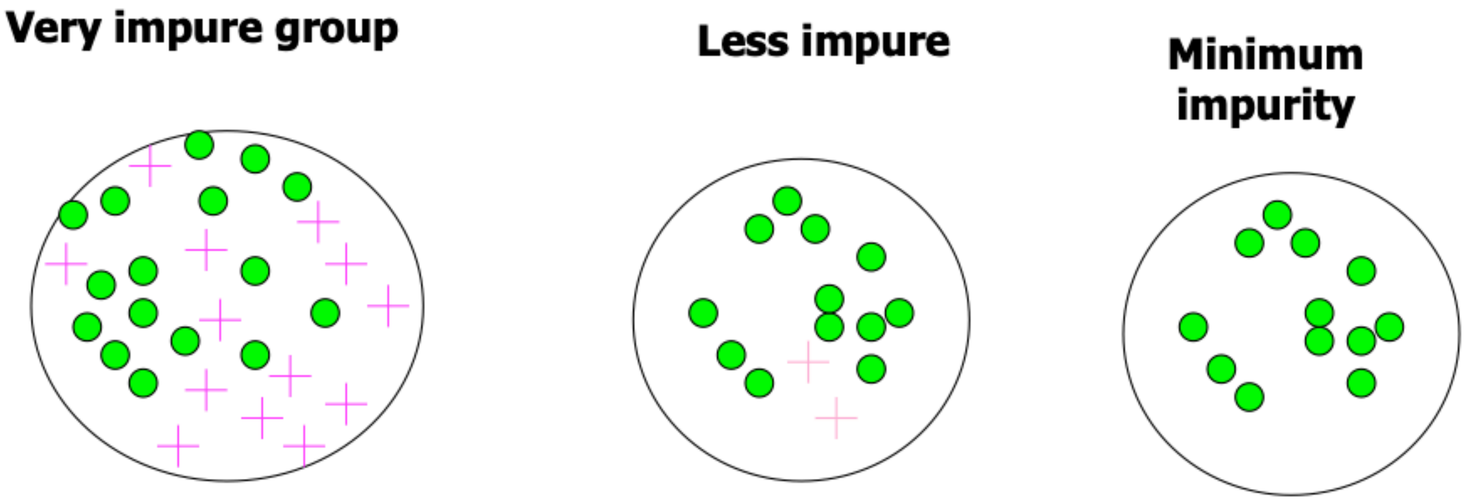
- Beberapa kemungkinannya adalah:
  - Random: Pilih atribut apa pun secara acak
  - Least-Values: Pilih atribut dengan jumlah nilai terkecil yang mungkin
  - Most-Value: Pilih atribut dengan jumlah nilai terbesar yang mungkin
  - Max-Gain: Pilih atribut yang memiliki perolehan informasi terbesar yang diharapkan

    yaitu, atribut yang menghasilkan ukuran terkecil yang diharapkan dari sub-pohon yang berakar pada anak-anaknya
- Algoritma ID3 menggunakan metode Max-Gain untuk memilih atribut terbaik

Information Gain

Impurity/Entropy (informal)

- Mengukur tingkat ketidakmurnian dalam suatu kelompok



Information Gain

Information Gain adalah informasi timbal balik antara atribut input A dan variabel target Y
Information Gain adalah pengurangan entropi yang diharapkan dari variabel target Y untuk sampel data S, karena pemilahan pada variabel A

$$\text{Gain}(S, A) = \text{Entropy}_S(Y) - \text{Entropy}_S(Y \mid A)$$

Entropy

Entropy(S) adalah jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sample S.

Besarnya Entropy pada ruang sample S ditentukan oleh

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

S adalah ruang (data) sample yang digunakan untuk training.
P+ adalah jumlah yang bersolusi positif (mendukung) pada data sample untuk kriteria tertentu.
P- adalah jumlah yang bersolusi negatif (tidak mendukung) pada data sample untuk kriteria tertentu.

SELAMAT BELAJAR