Data Mining

Pertemuan 6: Classification
(lanjutan)

Review Kuliah Sebelumnya

- Algoritma Klasifikasi
  - Induksi Pohon Keputusan dengan ID3
    - Gunakan sampel untuk membuat training set
    - Tentukan class dari tiap sampel
    - Pilih salah satu atribut (atribut terbaik, yang memiliki nilai entropi terkecil) sebagai root
    - Bagi training set sesuai dengan nilai atribut terbaik yang sama
    - Jika nilai class dapat terpartisi semua, proses selesai. Jika tidak, pilih atribut baru dan bagi kembali training set (dari yang tidak memiliki nilai class yang sama) berdasarkan nilai atribut baru tersebut

Algoritma Klasifikasi

- Induksi pohon keputusan (*decision tree*)
- Bayesian Classification
- Neural Network Classification
- Support Vector Machine Classification

BEBERAPA REVIEW DARI ALGORITMA KLASIFIKASI LAINNYA

1. Bayesian Classification

- Menggunakan pendekatan statistik, dengan memprediksi probabilitas sebuah sampel termasuk dalam class yang mana
- Menggunakan Teorema Bayes:

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

- Dimana: P(H|X)= Peluang hipotesis H jika diketahui data X
  P(X|H)= Peluang data X jika diketahui hipotesis H
  P(H) = Peluang hipotesis H
  P(X) = Peluang data X

1. Bayesian Classification (lanjutan)

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

- Contoh:
  - Misal dataset adalah data customer yang dideskripsikan dengan atribut *usia* dan *gaji* dan misal X sebuah data customer dengan usia 35 dan gaji $40,000.
  - Misal hipotesis H=hipotesis customer akan membeli komputer. Maka P(H|X) = peluang customer X membeli komputer jika diketahui usia dan gajinya
  - P(H) = *prior probability dari H*, yakni peluang sembarang customer membeli komputer (tanpa peduli usia dan gajinya) → independen
  - P(X|H) = *posterior probability*, yakni peluang seorang customer X, dengan usia 35 dan gaji $40,000, membeli komputer
  - P(X) = *prior probability dari X*, yakni peluang seorang customer memiliki usia 35 dan gaji $40,000

1. Naïve Bayesian Classification (lanjutan): CONTOH

- Kita akan menggunakan dataset dari contoh ID3.
- Data tabel memiliki 3 atribut: Age, Competition, dan Type
- Label class, Profit, memiliki 2 nilai: {Down, Up}. Misal class C_1=DOWN dan class C_2=UP
- Misal kita ingin mengklasifikasikan sebuah sampel data:

```
X=(age=midlife,
competition=no,
type=hardware)
```

Profit	Age	Competition	Type
Down	Old	No	Software
Down	Midlife	Yes	Software
Up	Midlife	No	Hardware
Down	Old	No	Hardware
Up	New	No	Hardware
Up	New	No	Software
Up	Midlife	No	Software
Up	New	Yes	Software
Down	Midlife	Yes	Hardware
Down	Old	Yes	Software

1. Naïve Bayesian Classification (lanjutan): CONTOH

X=(age=midlife, competition=no, type=hardware)

- Kita memaksimalkan $P(X|C_i)P(C_i)$ untuk i=1, 2
- Prior probability untuk tiap class:
  - P(Profit=DOWN) = 5/10 = ½
  - P(Profit=UP) = 5/10 = ½
- Untuk menghitung $P(X|C_i)$ untuk i=1, 2, hitung probabilitas kondisional sebagai berikut:
  - P(age=midlife|profit=down) = 2/5
  - P(age=midlife|profit=up) = 2/5
  - P(competition=no|profit=down) = 2/5
  - P(competition=no|profit=up) = 4/5
  - P(type=hardware|profit=down) = 2/5
  - P(type=hardware|profit=up) = 2/5

Profit	Age	Competition	Type
Down	Old	No	Software
Down	Midlife	Yes	Software
Up	Midlife	No	Hardware
Down	Old	No	Hardware
Up	New	No	Hardware
Up	New	No	Software
Up	Midlife	No	Software
Up	New	Yes	Software
Down	Midlife	Yes	Hardware
Down	Old	Yes	Software

1. Naïve Bayesian Classification (lanjutan): CONTOH

X=(age=midlife, competition=no, type=hardware)

- Untuk class C1: $P(X|profit = DOWN) = P(age = midlife|profit = down) * P(competition = no|profit = down) * P(type = hardware|profit = down)$

  $P(X|profit = DOWN) = \frac{2}{5} * \frac{2}{5} * \frac{2}{5} = 0.064$

- Sedangkan untuk class C2:

  $P(X|profit = UP)$

  $= P(age = midlife|profit = up) * P(competition = no|profit = up) * P(type = hardware|profit = up)$

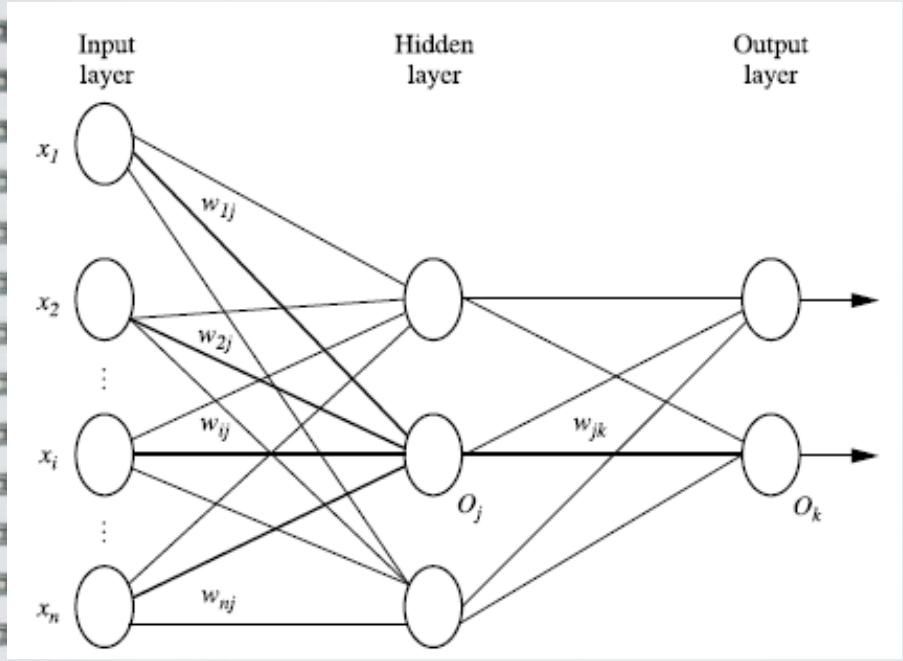
  $P(X|profit = UP) = \frac{2}{5} * \frac{4}{5} * \frac{2}{5} = 0.128$

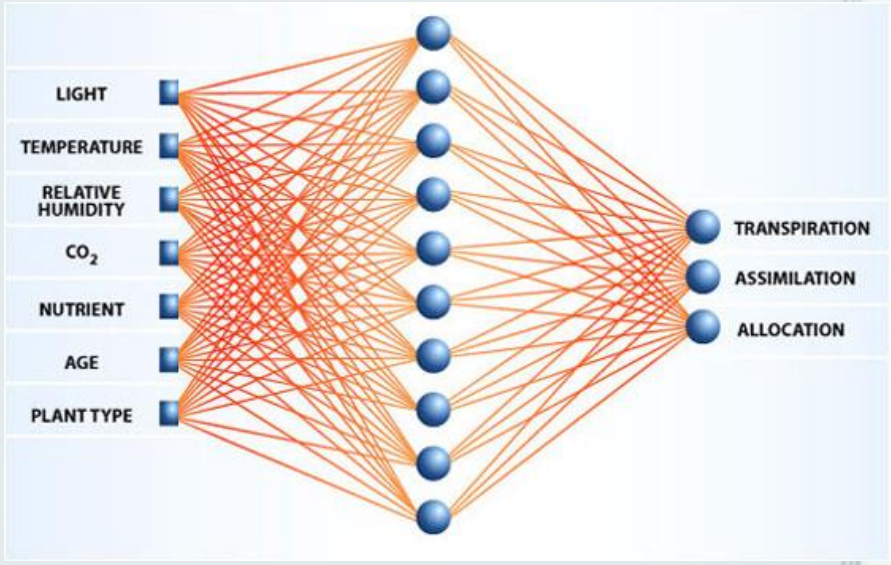
- Untuk mencari class C_i yang memaksimalkan $P(X|C_i)$, maka:
- $P(X|Profit = DOWN) * P(Profit = DOWN) = 0.064 * 0.5 = 0.032$
- $P(X|Profit = UP) * P(Profit = UP) = 0.128 * 0.5 = 0.064$ → **tertinggi**
- Maka hasil prediksi Naïve Bayes adalah Profit = UP untuk data X

2. Neural Network Classification

- Mengunakan jaringan saraf tiruan yang mengkoneksikan unit input dengan output beserta bobot dari tiap koneksi.
- Pada proses pembelajaran, bobot koneksi dapat disesuaikan.
- Digunakan untuk data atribut yang bertipe numerik

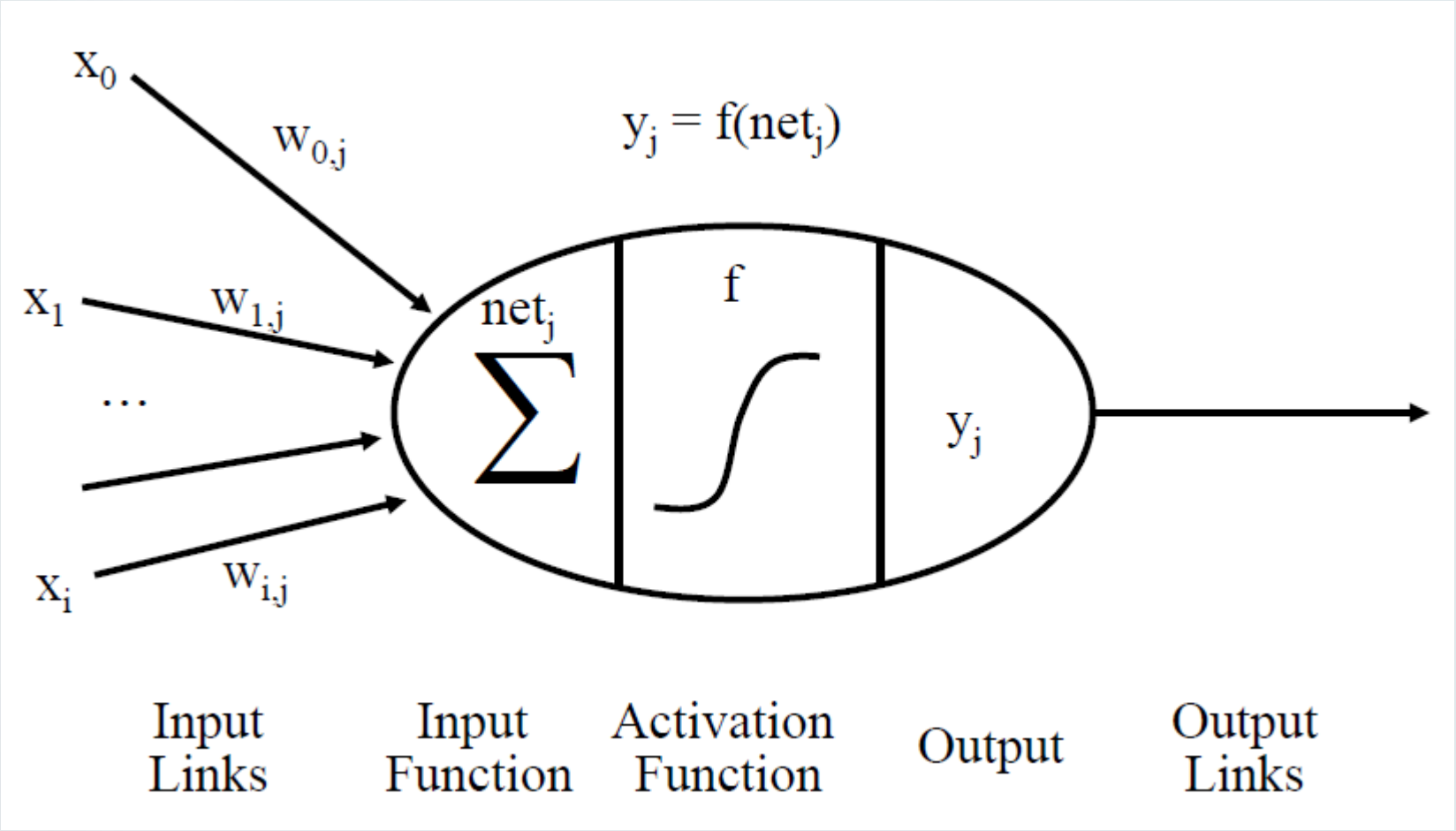
2. Neural Network Classification (lanjutan)





Multilayer neural network

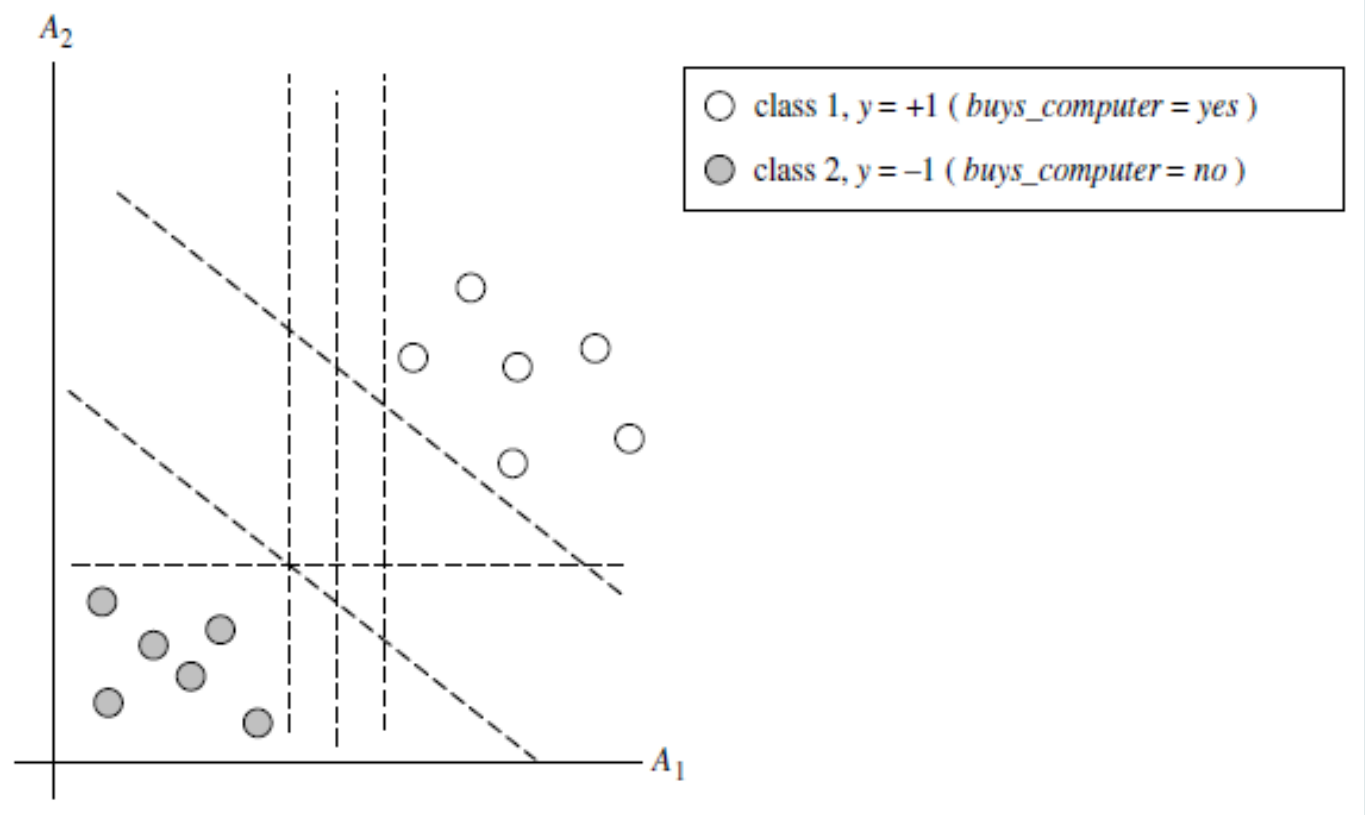
2. Neural Network Classification (lanjutan)



3. Support Vector Machine Classification

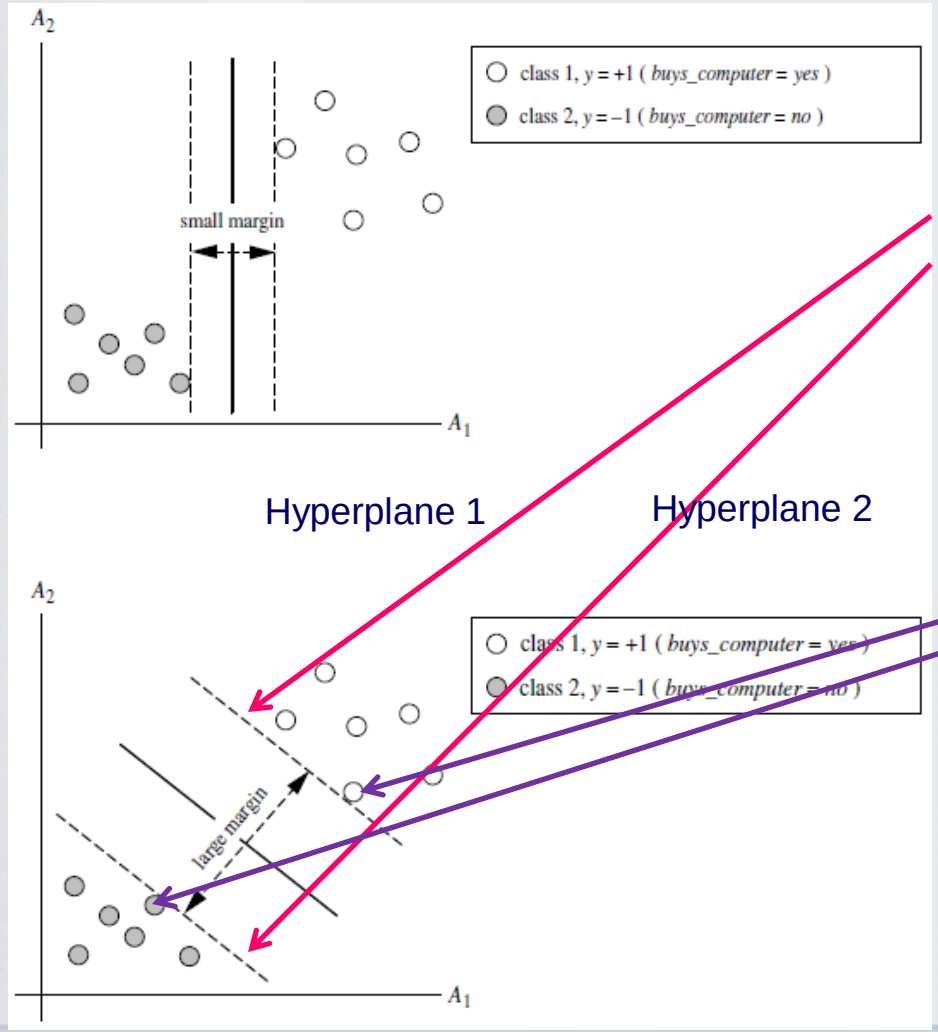
- Menggunakan mapping non-linier untuk mentransformasi data training ke dimensi yang lebih tinggi
- Di dimensi baru ini, SVM mencari “batas pemisah (*hyperplane*)” linier yang optimal yang memisahkan sebuah kelompok class dengan kelompok class yang lain

3. Support Vector Machine Classification (lanjutan)



Untuk training set dengan 2 atribut, A1 dan A2, dan 2 class, class 1 dan class 2, dapat dibuat garis pemisah. Garis yang mana yang paling optimal?

3. Support Vector Machine Classification (lanjutan)



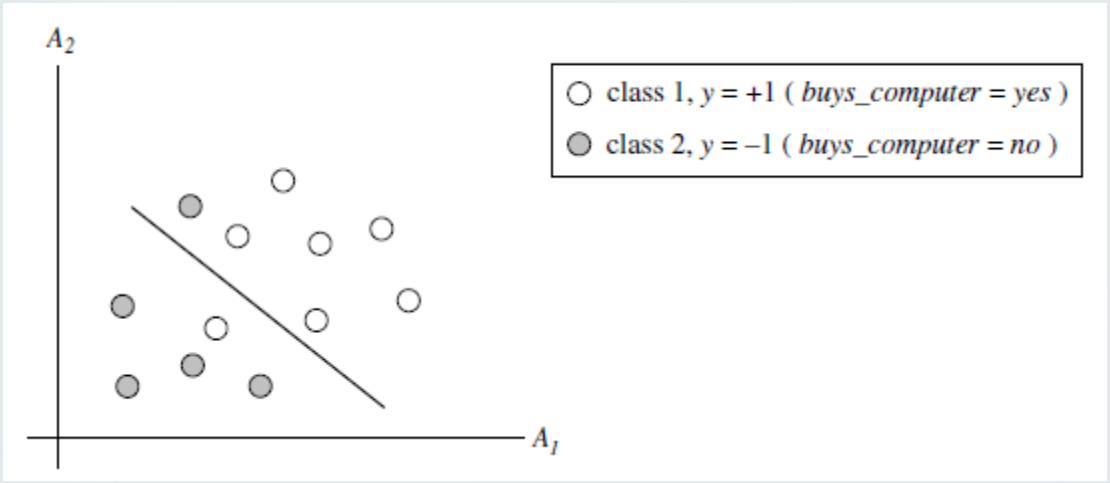
SVM memilih pemisah yang memiliki margin (jarak) lebih besar

H1: $w_0+w_1x_1+w_2x_2 \geq 1$ utk $y_i = +1$
H2: $w_0+w_1x_1+w_2x_2 \leq -1$ utk $y_i = -1$

Data training yang jatuh tepat pada garis *hyperplane* disebut sebagai *Support vectors*.

3. Support Vector Machine Classification (lanjutan)

Problem: Bagaimana jika data tidak dapat dipisahkan secara linier?



Solusi: SVM dapat diekstensi menjadi nonlinear SVM, dengan menggunakan fungsi kernel Polinomial derajat h, fungsi basis radial Gaussian, atau Sigmoid

TUGAS PENGGANTI KULIAH TANGGAL 15 MARET 2018

- Gunakan data seperti pada tugas sebelumnya (terlampir berikut ini) untuk melakukan klasifikasi pada data uji berikut:

  X=(Age=Senior, Student=No, Credit_Rating=Excellent)

- Tunjukkan langkah-langkah perhitungannya

Age	Student	Credit_Rating	Buys_Computer
Middle_Aged	No	Excellent	Yes
Youth	No	Fair	No
Senior	No	Fair	No
Youth	Yes	Excellent	Yes
Middle_Aged	No	Fair	No
Senior	Yes	Fair	Yes
Youth	No	Excellent	No
Senior	Yes	Excellent	Yes