BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

COMP8014

Knowledge Data Discovery

Classification: Basic Concepts

LEARNING OUTCOMES

LO4: Mahasiswa diharapkan mampu untuk menentukan metode data mining yang sesuai dengan sifat kealamiahan permasalahan yang dihadapi.

LO5: Mahasiswa diharapkan mampu untuk mengimplementasikan metode data mining.

OUTLINE MATERI :

1. Apakah klasifikasi itu?
2. Pendekatan umum?
3. Regresi dan Klasifikasi.
4. Evaluasi dan Seleksi Model
5. Meningkatkan akurasi klasifikasi
6. Metode Decision Tree.

ISI MATERI

Apakah Klasifikasi itu?

- Klasifikasi adalah tugas untuk menempatkan objek ke salah satu dari beberapa kelas (kategori) yang telah ditentukan berdasarkan kumpulan data pelatihan yang berisi pengamatan (atau contoh) yang keanggotaan kelasnya diketahui.
- Dalam terminologi pembelajaran mesin, klasifikasi mempertimbangkan contoh masalah belajar yang diawasi Kumpulan data pelatihan diberi label data
- Setiap record (dikenal sebagai instance atau example) dicirikan oleh tupel (x, y), di mana x adalah himpunan atribut dan y adalah atribut khusus, yang ditunjuk sebagai label kelas (juga dikenal sebagai kategori atau atribut target)

Tujuan dari Klasifikasi?

- Data baru yang harus dimasukkan ke kelas tertentu seakurat mungkin.
- Kumpulan data uji digunakan untuk menentukan keakuratan model. Biasanya, kumpulan data yang diberikan dibagi menjadi set pelatihan dan set tes, dengan set pelatihan digunakan untuk membangun model dan set tes yang digunakan untuk memvalidasinya.
- Jika set tes digunakan untuk memilih model, ini disebut validasi (uji) yang ditetapkan.

Note: Klasifikasi adalah supervised learning karena adanya label yang digunakan untuk mengawasi proses pembelajaran.

Supervised learning vs unsupervised learning?

Supervised learning (Pembelajaran yang diawasi) →untuk klasifikasi, regresi

- Data: data lebed (label kelas data pelatihan diketahui)
- Pengawasan: Data pelatihan (observasi, pengukuran, dll.) Disertai oleh label yang menunjukkan kelas pengamatan

- Data baru dikelompokkan berdasarkan set pelatihan

Unsupervised learning (Pembelajaran tanpa pengawasan) →untuk klustering.

- Data: unalebed data (label kelas data pelatihan tidak diketahui)
- Dengan seperangkat pengukuran, observasi, dll dengan tujuan membangun keberadaan kelas atau cluster dalam data.

Perbedaan Klasifikasi dengan Regresi?

- Klasifikasi:
  - Memprediksi label kelas kategoris
  - Mengklasifikasikan data (menyusun model) berdasarkan kumpulan pelatihan dan nilai (label kelas) dalam atribut klasifikasi dan menggunakannya dalam mengklasifikasikan data baru
- Regresi:
  - Model fungsi bernilai kontinu, yaitu, memprediksi nilai yang tidak diketahui atau yang hilang

Aplikasi dari Klasifikasi?

- Banking: Credit/loan approval
- Fraud detection: if a transaction is fraudulent
- Detecting spam email messages based upon the message header and content
- Medical diagnosis: Predicting tumor cells as benign or malignant based upon the results of MRI scans
- Biology: Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Web page categorization: Categorizing news stories as finance, weather, entertainment, sports, etc

2 langkah proses di klasifikasi.

1. Model Konstruksi: menggambarkan seperangkat kelas yang telah ditentukan Dengan seperangkat set data berlabel (sebagai rangkaian pelatihan) untuk konstruksi model. Model diwakili sebagai aturan klasifikasi, pohon keputusan, atau fungsi matematika (penggolong)

2. Penggunaan Model: untuk mengklasifikasikan objek yang akan datang atau yang tidak diketahui.

    a. Perkirakan akurasi model

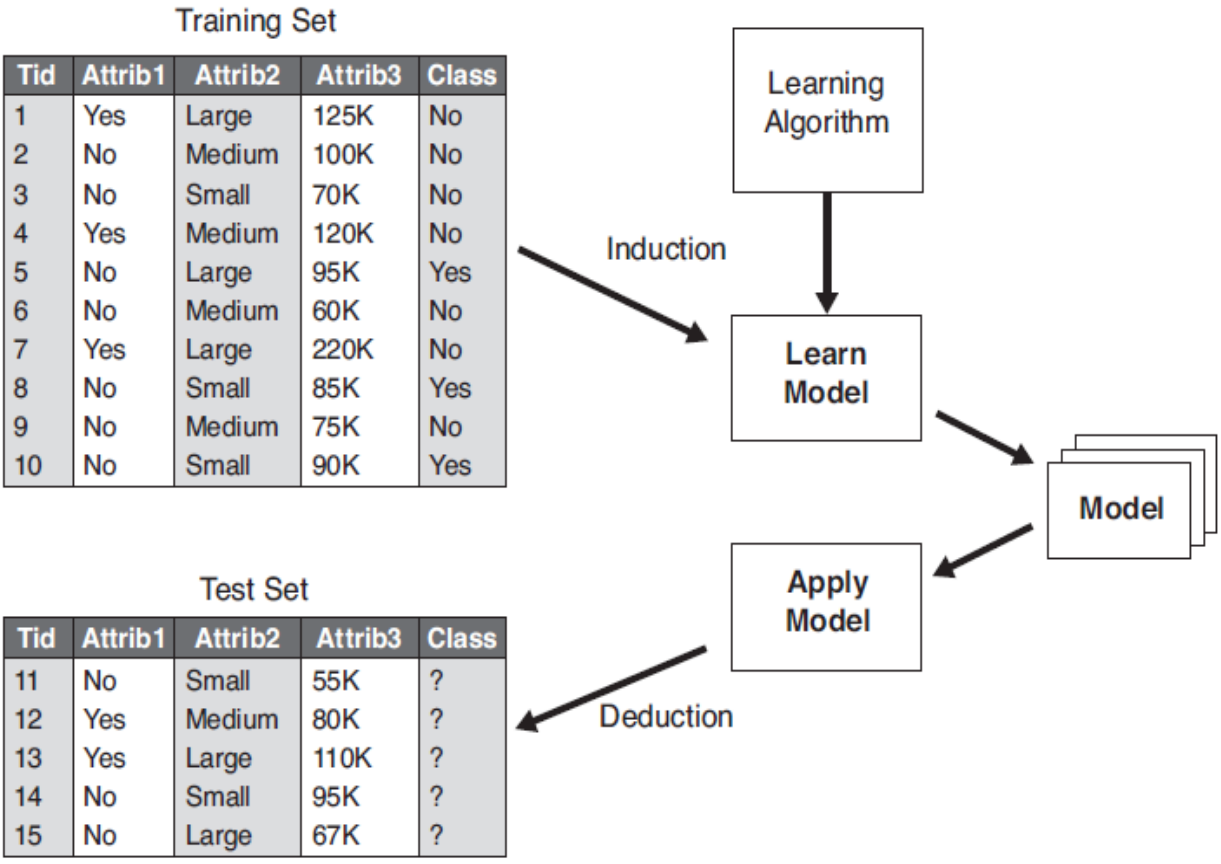
        i. Label sampel uji yang diketahui dibandingkan dengan hasil klasifikasi dari model

        ii. Tingkat akurasi adalah persentase sampel uji yang diklasifikasikan dengan benar oleh model

        iii. Test set independen dari set pelatihan (jika tidak overfitting) Jika ketepatannya bisa diterima, gunakan model untuk mengklasifikasikan data baru

        Catatan: Jika set tes digunakan untuk memilih model, ini disebut validasi (uji) yang ditetapkan

Ilustrasi klasifikasi:



Metode Klasifikasi?

- Decision Tree-based Methods → Lihat penjelasan di slide PPT.
- Rule-based Methods
- Naive Bayes Classifiers
- Bayesian Belief Networks
- Nearest-Neighbor Classifiers (KNN)
- Artificial Neural Networks (ANN)
- Support Vector Machines (SVM)

SIMPULAN

1. Klasifikasi adalah suatu bentuk analisis data yang mengekstrak model yang menggambarkan kelas data penting.
2. Metode yang efektif dan terukur telah dikembangkan untuk pengambilan keputusan pohon, klasifikasi Naive Bayesian, klasifikasi berbasis aturan, klasifikasi tetangga terdekat dan banyak metode klasifikasi lainnya.

DAFTAR PUSTAKA

1. Han, J., Kamber, M., & Pei, Y. (2006). “Data Mining: Concepts and Technique”. Edisi 3. Morgan Kaufman. San Francisco
2. Tan, P.N., Steinbach, M., & Kumar, V. (2006). “Introduction to Data Mining”. Addison-Wesley. Michigan
3. Witten, I. H., & Frank, E. (2005). “Data Mining : Practical Machine Learning Tools and Techniques”. Second edition. Morgan Kaufmann. San Francisco