

# LECTURE NOTES

**COMP8014**

**Knowledge Data Discovery**

**Classification: Advanced Methods 2**

## LEARNING OUTCOMES

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

LO5: Mahasiswa diharapkan mampu untuk mengimplementasikan metode data mining.

LO6: Mahasiswa mampu mengevaluasi aplikasi dan trend Data Mining.

### OUTLINE MATERI :

1. Support Vector Machines (SVM)
2. Nearest neighbor classifier
3. Topik Tambahan

## ISI MATERI

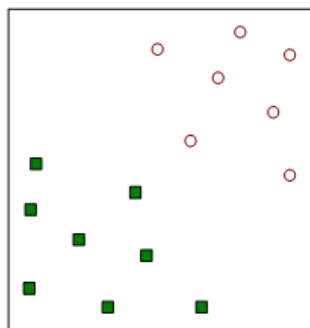
### Support Vector Machine?

- Metode klasifikasi yang relatif baru untuk data linier dan nonlinier
- Menggunakan pemetaan nonlinier untuk mengubah data pelatihan asli menjadi dimensi yang lebih tinggi
- Dengan dimensi baru, ia mencari linier pemisah linier yang linier (yaitu, "batas keputusan")
- Dengan pemetaan nonlinier yang sesuai dengan dimensi yang cukup tinggi, data dari dua kelas selalu dapat dipisahkan oleh hyperplane
- SVM menemukan hyperplane ini menggunakan vektor pendukung (tupel pelatihan "esensial") dan margin (ditentukan oleh vektor pendukung)
- SVM adalah klasifikasi biner.

### Contoh: Linear SVM



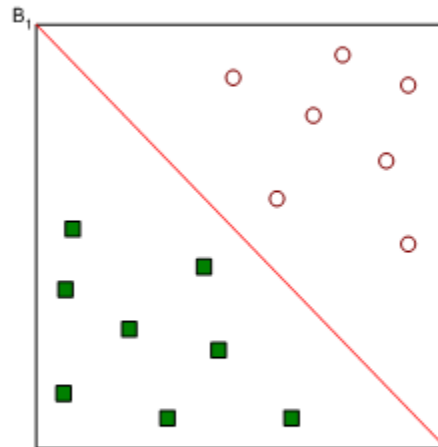
### Support Vector Machines



- Find a linear hyperplane (decision boundary) that will separate the data

Problemnya: mencari hyperplane linear terbaik yang memisahkan dua kelas dari dataset.

## Support Vector Machines

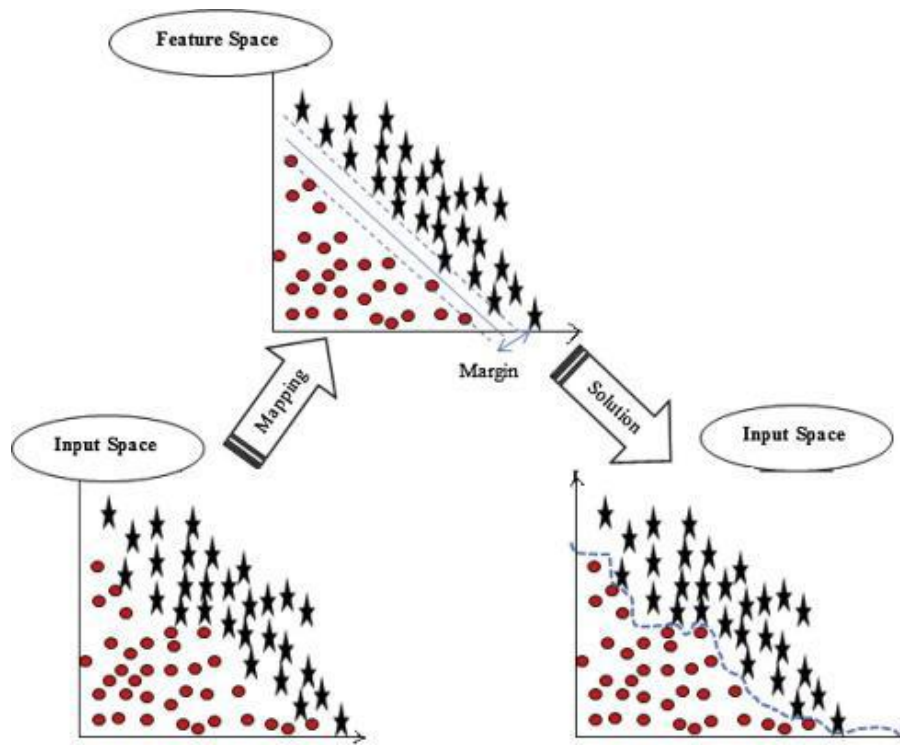


- One Possible Solution

Hyperplane terbaik adalah one possible solution.

### Bagaimana apabila karakter datanya tidak linear?

- Linear SVM tidak sesuai untuk memisahkan data yang mempunyai karakteristik nonlinear.
- Linear SVM perlu dimodifikasi dengan menggunakan suatu trik kernel berdasarkan teorema Mercer.
- Prosedur dari SVM dapat diilustrasikan pada gambar berikut:



- Adapun langkah-langkah yang dilakukan adalah sebagai berikut:
  - Data di original space (input space) ditransformasikan ke feature space dengan menggunakan suatu fungsi kernel.
  - Selanjutnya di feature space dibuat model linear SVM dan melakukan suatu trik kernel sehingga diperoleh suatu model optimisasi yang dapat diselesaikan.
  - Setelah dilakukan trik kernel, dilakukan perhitungan untuk mencari solusi dari model optimisasi linear kernel tersebut, dan akan diperoleh nonlinear SVM di original space.

**Bagaimana kalau masalah yang dihadapi mempunyai lebih dari 2 kelas?**

- Linear dan nonlinear SVM tidak dapat digunakan secara langsung.
- Dibuat mekanisme baru yang menggunakan linear/nonlinear SVM → Multiclass SVM.

## SIMPULAN

1. Tes signifikansi dan kurva ROC berguna untuk pemilihan model.
2. Ada banyak perbandingan metode klasifikasi yang berbeda; Masalahnya tetap menjadi topik penelitian
3. Tidak ada metode tunggal yang terbukti lebih unggul dari semua yang lain untuk semua kumpulan data
4. Masalah seperti akurasi, waktu pelatihan, ketahanan, skalabilitas, dan interpretasi harus dipertimbangkan dan dapat melibatkan trade-off, yang selanjutnya mempersulit pencarian metode unggul secara keseluruhan.
5. Topik Tambahan pada Klasifikasi: Klasifikasi multiclass, klasifikasi semi-supervisi, pembelajaran aktif, dll.

## 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