# **Data Mining**

Pertemuan 5 Classification

#### Klasifikasi – Definisi Umum

 Proses pengenalan ciri dari objek untuk dapat dibedakan, dan dipahami

#### Klasifikasi – definisi statistik

 Identifikasi penentuan set kategori (subpopulasi) yang menjadi observasi baru (test / evaluation), berdasarkan kumpulan data training yang berisi observasi yang keanggotaan kategorinya diketahui

#### Klasifikasi terdiri dari:

- Kategori / Label / Class
- Training Data → Label diketahui
  - Supervised Learning
- Test / Evaluation Data → Label belum diketahui

## Supervised Learning

 Pembelajaran mesin dengan mempelajari fungsi yang memetakan input ke output berdasarkan contoh/training pasangan inputoutput.

# Cara kerja supervised

 Menyimpulkan fungsi dari data pelatihan berlabel yang terdiri dari serangkaian contoh pelatihan.

# Input Vektor Supervised

- Setiap contoh adalah pasangan yang terdiri dari objek input (biasanya vektor) dan nilai output yang diinginkan (juga disebut supervised signal).
- Memerlukan dimensionality reduction!

# Analisis Data Supervised

- Menganalisis data pelatihan dan menghasilkan fungsi yang disimpulkan, yang dapat digunakan untuk memetakan data baru dengan klasifikasi yang sama.
- Skenario optimal akan memungkinkan algoritma menentukan label kelas dengan benar untuk instance yang tidak terlihat.

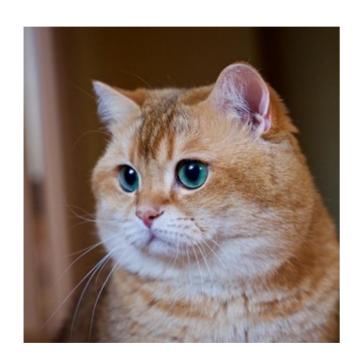
# Isu dalam Supervised Learning

- Imbalan Bias-varians
- Kompleksitas fungsi dan jumlah data pelatihan
- Dimensi ruang input
- Kebisingan dalam nilai output
- Faktor-faktor lain yang perlu dipertimbangkan (penting)

#### Trade-off Bias-varians

- Masalah pertama adalah pertukaran (Trade-off) antara bias dan varians.
- Bias → kesalahan pengenalan karena terlalu banyak referensi untuk satu kelas yang sama.
- Varians → kurangnya variasi antara dua atau lebih kelas yang berbeda.

### Kurang referensi



Training Data Kelas Kucing



Karena kurang varians Foto ini tidak dikenal sebagai kucing

### Terlalu banyak referensi



Terlalu banyak varians Training Data Kelas Kucing



Cheetah ter-bias sebagai kucing

### Kesimpulan bias dan varians

- Terlalu banyak varians untuk referensi akan menyebabkan bias terhadap satu kelas.
- Terlalu sedikit varians untuk referensi akan menyebabkan klasifikasi tidak fleksibel.
- Untuk mengurangi bias dan varians diperlukan pemilihan dan dimentionality reduction yang sangat ketat di tahap training.

#### Kompleksitas fungsi dan jumlah data pelatihan

 Masalah kedua adalah jumlah data pelatihan yang tersedia relatif terhadap kompleksitas fungsi "benar" (fungsi pengklasifikasi atau regresi).

### Fungsi Sederhana cukup sedikit data

 Jika fungsi sebenarnya sederhana, maka algoritma pembelajaran "tidak fleksibel" dengan bias tinggi dan varians rendah akan dapat mempelajarinya dari hanya sedikit data.

### Fungsi Kompleks harus banyak data

- Fungsi sangat kompleks karena melibatkan interaksi kompleks di antara banyak fitur input yang berbeda dan berperilaku berbeda di berbagai bagian ruang input
- Fungsi kompleks hanya akan dapat dipelajari dari sejumlah besar data pelatihan dan menggunakan algoritma pembelajaran "fleksibel" dengan bias rendah dan varian tinggi.