Pengantar Machine Learning

Data Science Camp

Bayu | github.com/pyk

Tujuan Materi

1. Mengerti konsep dasar Machine Learning.
2. Mengerti apa aja yang harus dipelajari jika ingin lebih mendalami Machine Learning.
3. Mengetahui bagaimana penerapan Machine Learning di industri.

Gambaran Umum Materi

1. Konsep Dasar Machine Learning
2. Penerapan Machine Learning
3. Studi kasus: Rojak

Konsep Dasar Machine Learning

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.”

Definisi Machine Learning oleh Tom Mitchell

Konsep Dasar Machine Learning

Task T

Masalah apa yang ingin diselesaikan atau **tujuan** apa yang ingin dicapai.

Experience E

Dari mana program tersebut bisa belajar. Sering disebut **data training**.

Metrics P

Metrik untuk mengukur performa program yang dibuat.

Task	Experience	Metrics	Algorithm
Analisa sentimen	- Ulasan Konsumen - Respon survey - Tweet - Status Facebook - Berita di media	- Akurasi - Presisi - Recall - False negative - False positive - True positive - True negative - F-measure - Jaccard Index	- Naive Bayes - Support Vector Machine - Decision Tree - Neural networks - Convolutional Neural network - Recurrent Neural network
Personalisasi Iklan	- Data pribadi pelanggan - Aktivitas pelanggan - History pelanggan		
Aplikasi kartu kredit	- Data pribadi pemohon - History transaksi pemohon		
Pendeteksi kanker	- Data pribadi pasien - Foto rontgen pasien		

Task T

Masalah apa yang ingin diselesaikan atau tujuan apa yang ingin dicapai menggunakan Machine Learning.

Jenis task/masalah yang paling sering muncul:

1. Klasifikasi
2. Clustering
3. Dimensionality Reduction
4. Regresi

Task Klasifikasi

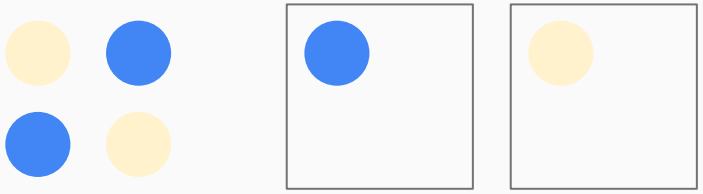
Memasukkan data kedalam kelas yang telah ditentukan.

Data training terdiri dari pasangan data dan kelas/label/kategori nya.

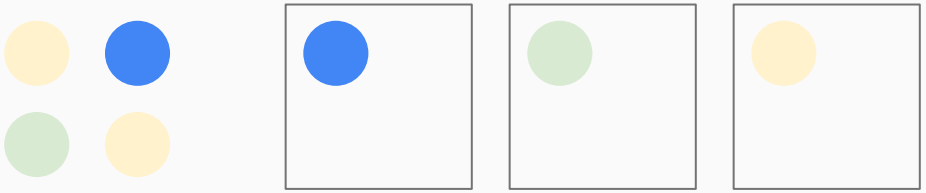
Metode belajarnya disebut **Supervised Learning** karena data training ada labelnya.

Algoritma Machine Learning yang dipakai disebut **Classifier**.

Binary



Multiclass



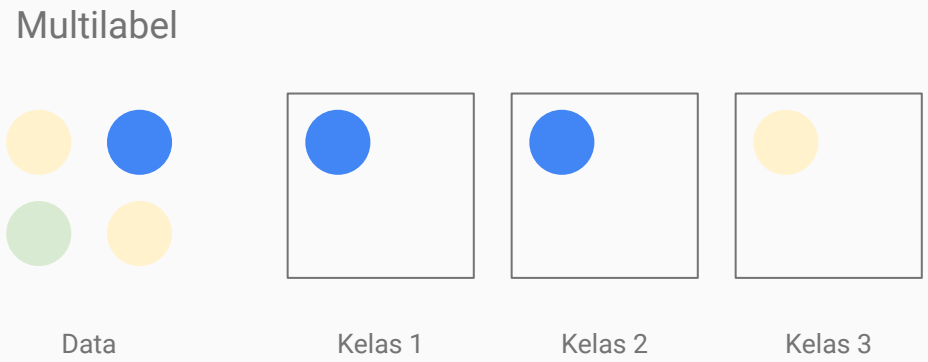
Task Klasifikasi

Memasukkan data kedalam kelas yang telah ditentukan.

Data training terdiri dari pasangan data dan kelas/label/kategori nya.

Metode belajarnya disebut **Supervised Learning** karena data training ada labelnya.

Algoritma Machine Learning yang dipakai disebut **Classifier**.



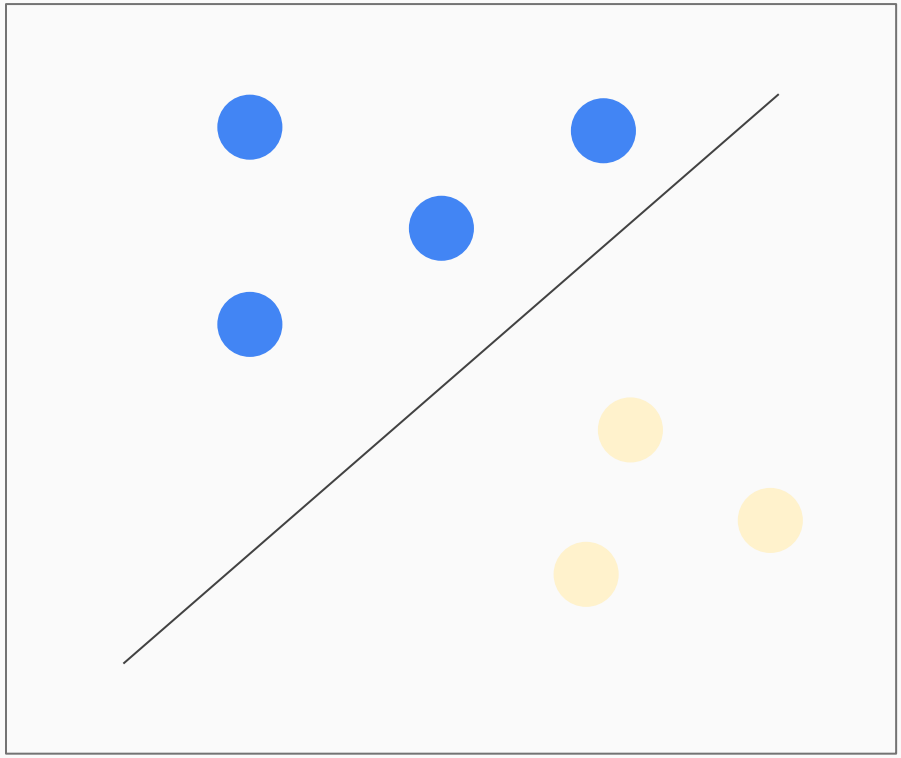
Task Klasifikasi

Jenis-jenis classifier dan beserta contohnya.

Linear	**Non-linear**
Logistic regression	Multi-layer perceptron
Perceptron	
Support vector machines	
Naive bayes	

Task Klasifikasi

Linear classifier hanya bisa memisahkan data secara linear.



Space

Task Klasifikasi

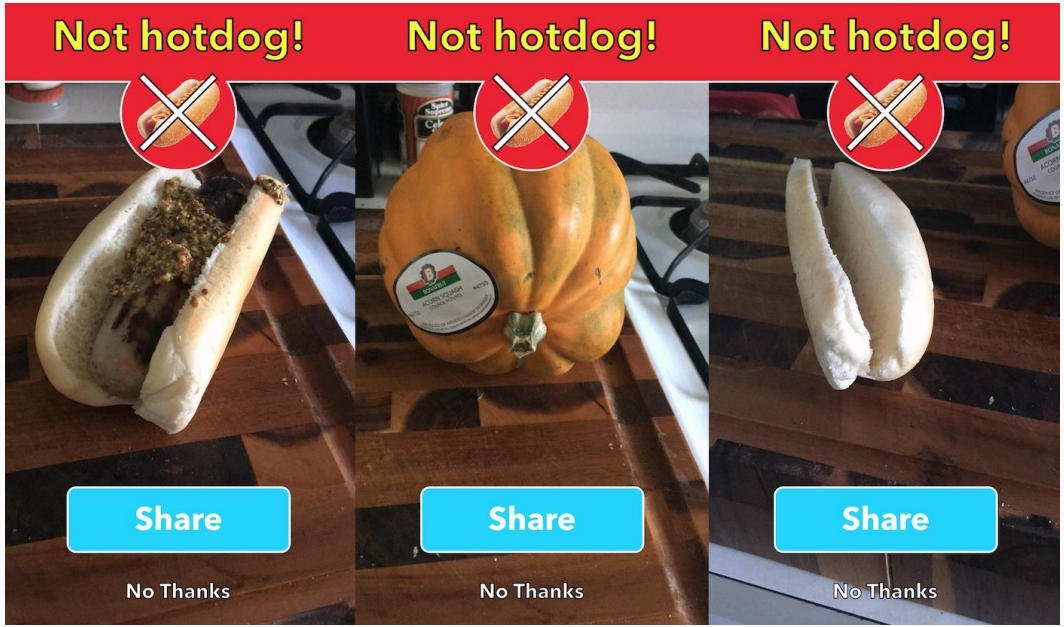
Teknik untuk klasifikasi Multiclass

One-vs-Rest (OvR)
Membuat satu classifier untuk tiap kelas/kategori.

One-vs-One (OvO)
Untuk sebanyak K kelas, kita membuat sebanyak (k(k-1))/2 binary classifier.

Task Klasifikasi

Contoh task klasifikasi.



Not Hotdog App

Task Klasifikasi

Contoh task klasifikasi.

Nama task	YouTube Faces DB
Tujuan	Pengenalan wajah
Data	Frame video dan informasi jumlah orang didalam video
URL	https://www.cs.tau.ac.il/~wolf/ytfaces/

Task Klasifikasi

Metrics untuk mengukur kinerja classifier.

Akurasi $$\frac{TP + TN}{TP + TN + FP + FN}$$

Presisi $$\frac{TP}{TP + FP}$$

Recall $$\frac{TP}{TP + FN}$$

Kondisi Sebenarnya

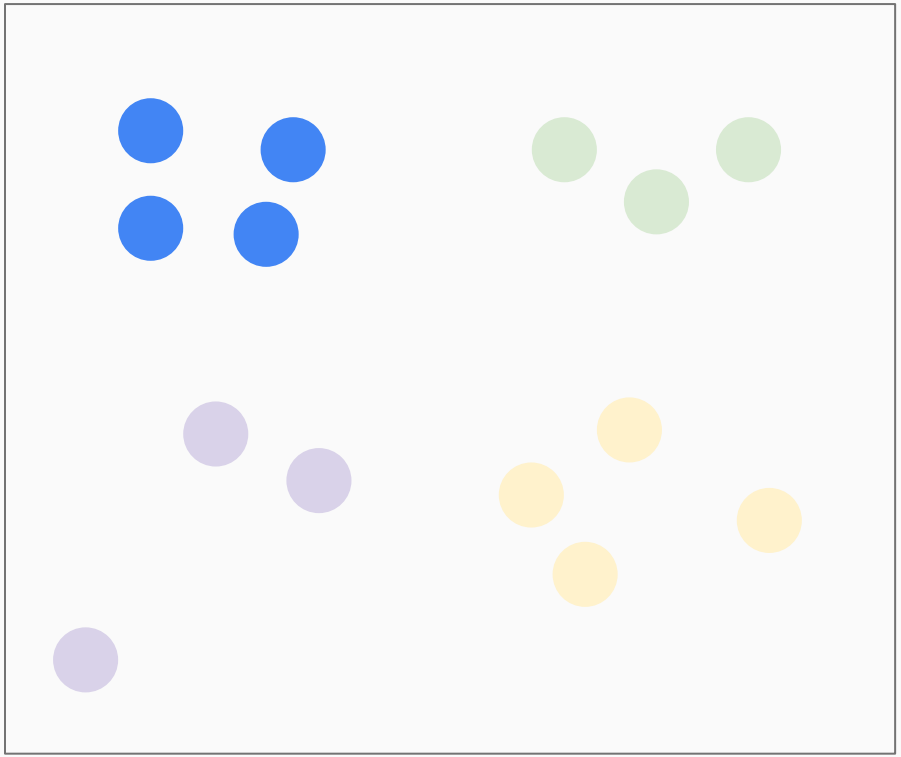
Hasil Prediksi	True	False
True	True Positive (TP)	False Positive (FP)
False	False Negative (FN)	True Negative (TN)

Task Clustering

Mengelompokkan data yang mempunyai karateristik sama kedalam satu kelompok.

Data training tidak ada label atau kelasnya.

Karena data training tidak ada label atau informasi kelasnya, maka cara belajarnya disebut **Unsupervised Learning**.



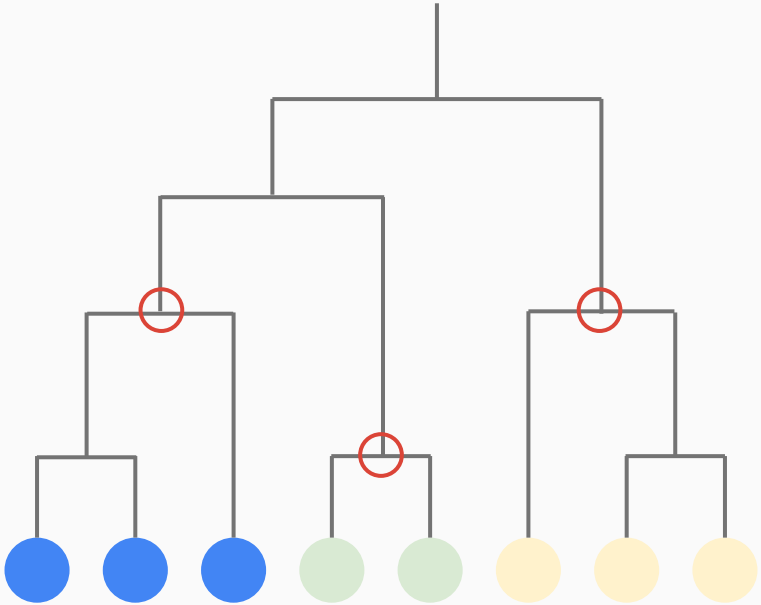
Space

Task Clustering

Jenis-jenis clustering.

Hierarchical Clustering

Tujuannya untuk membuat cluster hierarki.

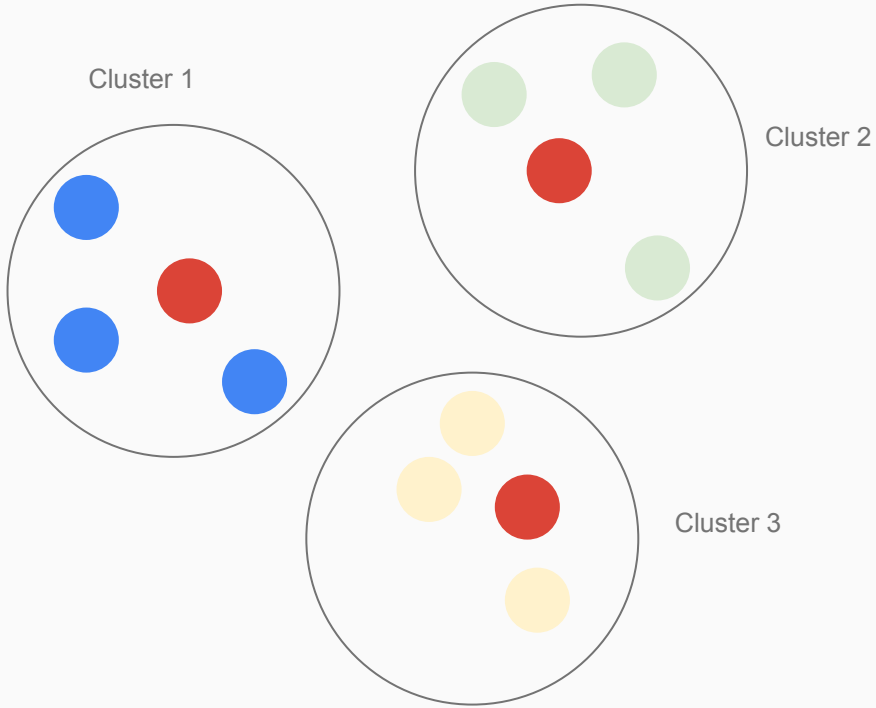


Task Clustering

Jenis-jenis clustering.

Centroid-based Clustering

Sebelumnya sudah ditentukan berapa centroid atau pusat clusternya.



Task Clustering

Metrics untuk mengukur kinerja algoritma clustering.

Rand-ms $$\frac{TP + TN}{TP + TN + FP + FN}$$

Jaccard $$\frac{TP}{TP + FP + FN}$$

Nama Metrics	Keterangan
Rand-measure	Mengukur seberapa mirip cluster yang dihasilkan
Jaccard Index	Mengukur kesamaan antara 2 dataset

Task Clustering

Contoh dataset clustering.

Nama task	FMA: A Dataset For Music Analysis Data Set
Tujuan	Analisis musik
Data	106,574 tracks dan termasuk title, album, artist, genres; play counts, favorites, comments; description, biography, tags.
URL	https://archive.ics.uci.edu/ml/datasets/FMA%3A+A+Dataset+For+Music+Analysis

Algoritma

Cara kerja algoritma Machine Learning tidak dibahas di materi ini.

Cara memilih algoritma yang tepat untuk menyelesaikan task:

1. Baca paper tentang task yang ingin dilakukan
2. Lihat hasil benchmark nya
3. Test dengan data training

Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	76.1	45.0	82.7	89.6	91.2	79.8	83.4
CNN-static	81.0	45.5	86.8	93.0	92.8	84.7	**89.6**
CNN-non-static	**81.5**	48.0	87.2	93.4	93.6	84.3	89.5
CNN-multichannel	81.1	47.4	**88.1**	93.2	92.2	**85.0**	89.4
RAE (Socher et al., 2011)	77.7	43.2	82.4	–	–	–	86.4
MV-RNN (Socher et al., 2012)	79.0	44.4	82.9	–	–	–	–
RNTN (Socher et al., 2013)	–	45.7	85.4	–	–	–	–
DCNN (Kalchbrenner et al., 2014)	–	48.5	86.8	–	93.0	–	–
Paragraph-Vec (Le and Mikolov, 2014)	–	**48.7**	87.8	–	–	–	–
CCAE (Hermann and Blunsom, 2013)	77.8	–	–	–	–	–	87.2
Sent-Parser (Dong et al., 2014)	79.5	–	–	–	–	–	86.3
NBSVM (Wang and Manning, 2012)	79.4	–	–	93.2	–	81.8	86.3
MNB (Wang and Manning, 2012)	79.0	–	–	**93.6**	–	80.0	86.3
G-Dropout (Wang and Manning, 2013)	79.0	–	–	93.4	–	82.1	86.1
F-Dropout (Wang and Manning, 2013)	79.1	–	–	**93.6**	–	81.9	86.3
Tree-CRF (Nakagawa et al., 2010)	77.3	–	–	–	–	81.4	86.1
CRF-PR (Yang and Cardie, 2014)	–	–	–	–	–	82.7	–
SVM_S (Silva et al., 2011)	–	–	–	–	**95.0**	–	–

Kim (2014)

Istilah-istilah

- **Features**: Sifat yang bisa diukur dari data
- **Samples**: Beberapa data poin
- **Feature vector**: Representasi features secara numerik
- **Feature extraction**: Fase untuk mengekstrak feature
- **Training set**: Kumpulan data untuk training
- **Development set**: Kumpulan data untuk men-tuning hyperparams.
- **Test set**: Kumpulan data untuk test

Penerapan Machine Learning

Untuk analisa teks