



DIGITAL
TALENT
SCHOLARSHIP

Lecture 29

Clustering dan Classification

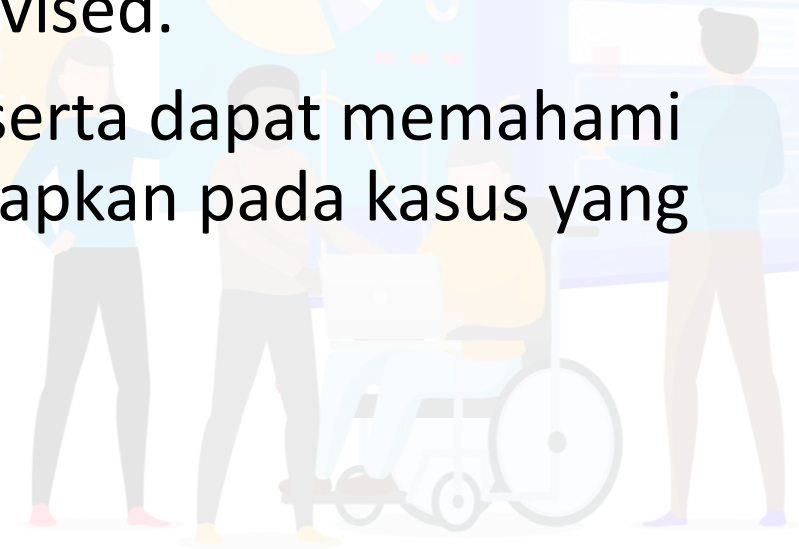


Lecture's Objective

- Mempelajari pengelompokan obyek atau beberapa variabel berdasarkan kecenderungan yang ada pada data secara supervised dan unsupervised.
- Setelah mengikuti sesi ini, peserta dapat memahami metode clustering dan menerapkan pada kasus yang tepat

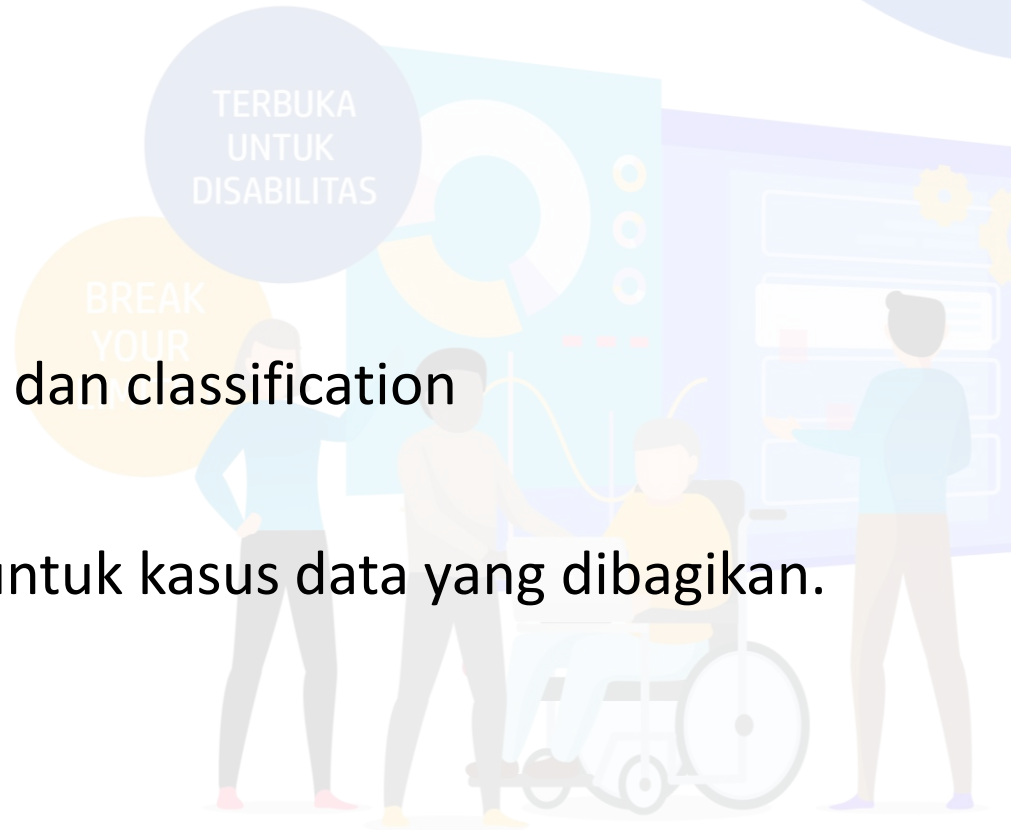
TERBUKA
DISABILITAS

YOUR



Outline

- 60 menit:
 - Clustering
 - Classification
 - Contoh kasus clustering dan classification
- 140 menit:
 - Mencoba source code untuk kasus data yang dibagikan.



Evaluasi Persiapan Kelas

- Apa perbedaan antara supervised dengan unsupervised Learning, dan clustering dengan classification?
- Mengapa clustering dinyatakan unsupervised sedangkan classification dinyatakan supervised?
- Apa itu Neural network? bagaimana cara neuron di neural network bekerja?
- Sebutkan 5 algoritma populer yang biasa digunakan untuk clustering!
- Sebutkan dan jelaskan secara singkat 2 algoritma yang biasa digunakan untuk classification!
- Sebutkan 3 contoh penggunaan clustering dan classification!

Clustering

- **Clustering** adalah teknik *machine learning* berupa algoritma pengelompokkan objek-objek data berjumlah N menjadi kelompok-kelompok data tertentu (*cluster*).
- Objek data yang berada dalam satu kelompok / *cluster* harus memiliki kemiripan.
- Semakin banyak data yang diperoleh = Semakin akurat hasil yang didapatkan.
- *Clustering* merupakan salah satu jenis dari algoritma **unsupervised learning**, algoritma yang bertujuan untuk mempelajari dan menemukan pola dari suatu input yang diberikan tanpa menggunakan label.
- Dengan penggunaan *supervised learning*, maka beberapa hal berikut ini dapat dilakukan:
 1. *Search*: Membandingkan antar dokumen, gambar atau suara untuk menampilkan *item* serupa.
 2. Deteksi anomali: Mendeteksi perilaku yang tidak biasa yang biasanya berhubungan dengan hal-hal yang ingin dicegah atau dideteksi, seperti contoh penipuan.

Algoritma *Clustering*

- K-Means Clustering
- Mean-Shift Clustering
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)
- Agglomerative Hierarchical Clustering

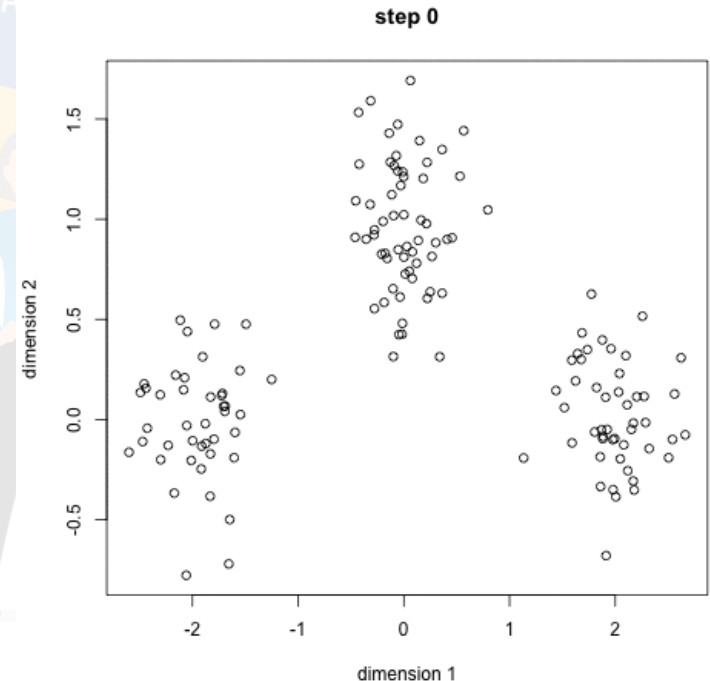


K-Means *Clustering*

- Tentukan jumlah cluster
- Alokasikan data ke dalam cluster secara random
- Hitung centroid/rata-rata dari data yang ada di masing-masing cluster
- Alokasikan masing-masing data ke centroid/rata-rata terdekat
- Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan di atas nilai threshold yang ditentukan

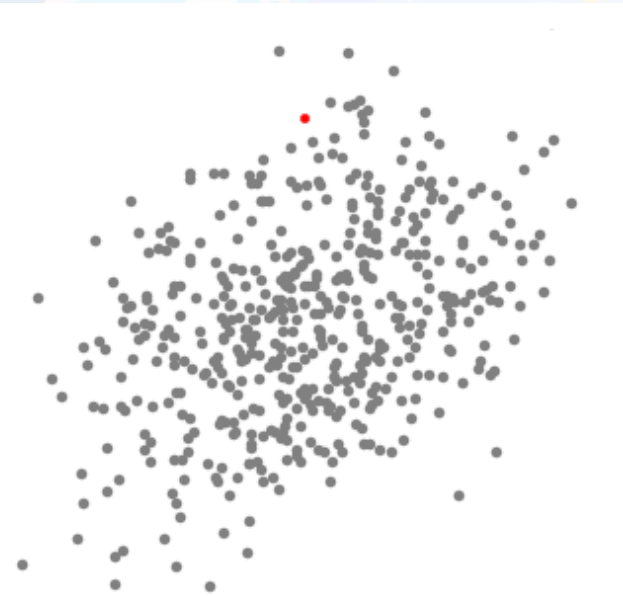
TERBUKA
UNTUK
DISA

AK
OUR
UNITS!

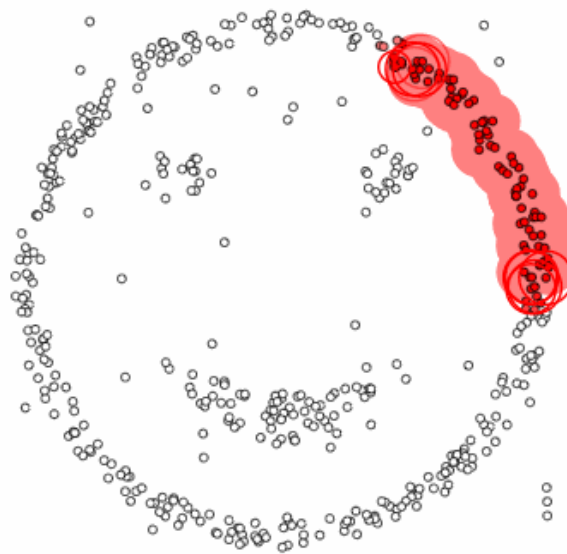


Mean-Shift Clustering

- Untuk menjelaskan mean-shift, kita ilustrasikan sekumpulan titik dalam ruang dua dimensi seperti ilustrasi di samping.
- Kita mulai dengan circular sliding window yang berpusat pada titik C (dipilih secara acak) dan memiliki jari-jari r sebagai kernel. Mean shift adalah algoritma yang melakukan pergeseran kernel ini secara iteratif ke daerah kepadatan yang lebih tinggi pada setiap langkah hingga konvergensi.
- Pada setiap iterasi, sliding window digeser ke arah daerah dengan kepadatan lebih tinggi dengan menggeser titik tengahnya. Kepadatan pada sliding window sebanding dengan jumlah titik di dalamnya.
- Terus melakukan pergeseran pada sliding window sesuai dengan rata-rata sampai tidak ada arah di mana pergeseran dapat mengakomodasi lebih banyak titik di dalam kernel (tidak lagi meningkatkan densitas / jumlah titik di window)
- Langkah 1 hingga 3 ini dilakukan dengan banyak sliding window sampai semua titik terletak di dalam window. Ketika beberapa window saling bertumpang tindih, window yang berisi titik terbanyak akan dipertahankan. Objek data kemudian dikelompokkan sesuai dengan sliding window tempat mereka berada.



Density-Based Spatial Clustering of Applications with Noise (DBSCAN)



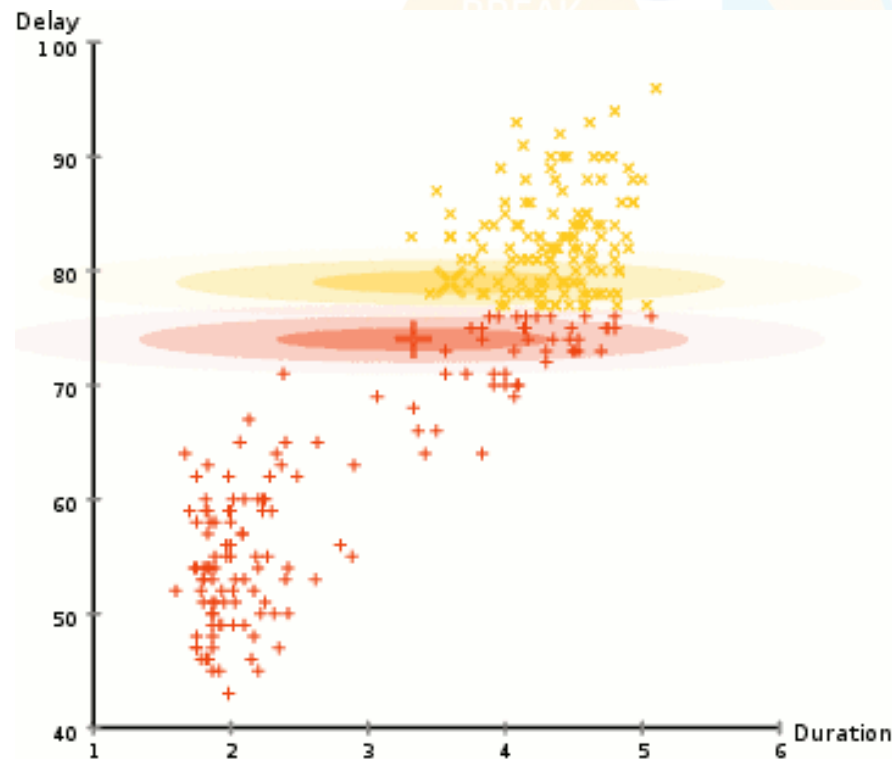
epsilon = 1.00
minPoints = 4

Restart

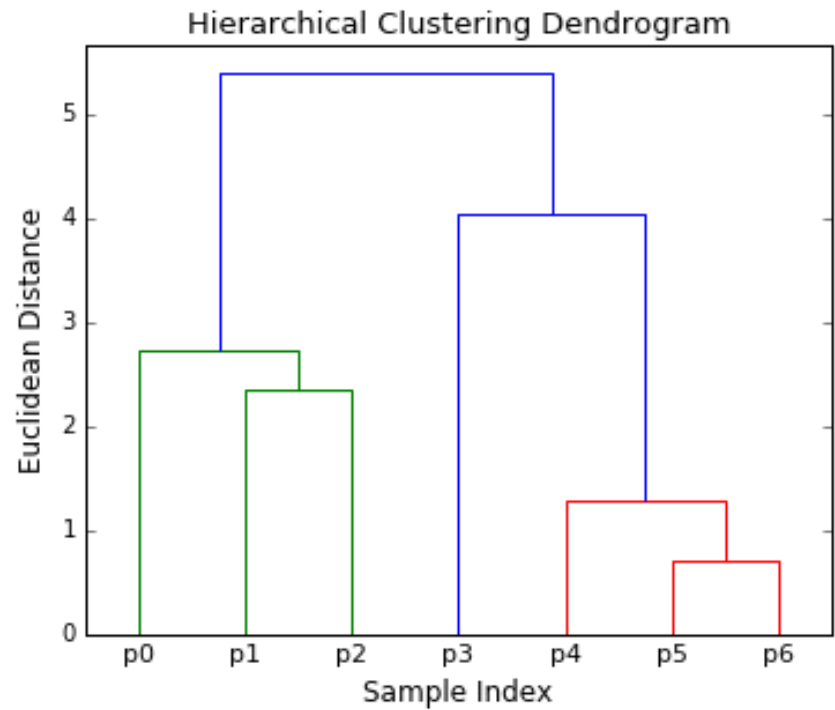
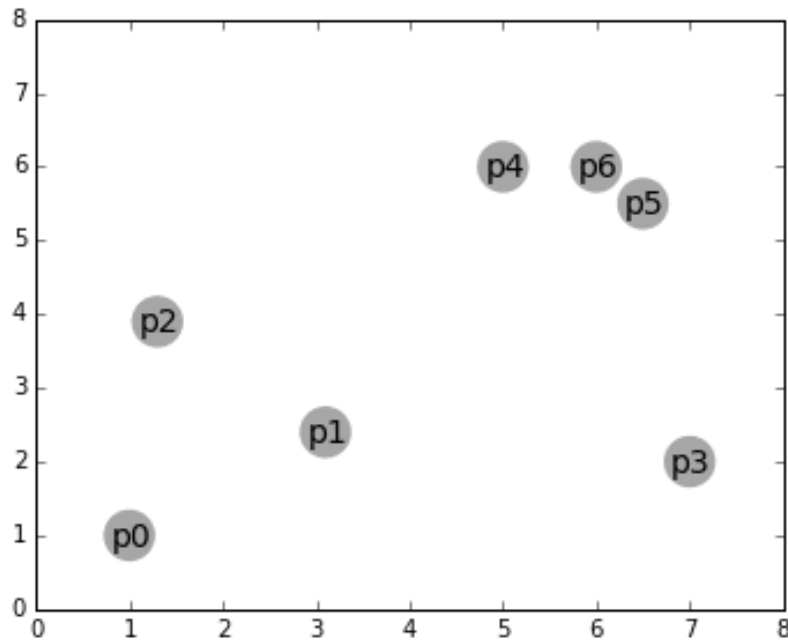


Pause

Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)



Agglomerative Hierarchical Clustering



Tugas 1: K-Means Clustering

Ikuti tutorial di:

<https://blog.floydhub.com/introduction-to-k-means-clustering-in-python-with-scikit-learn/>

TERBUKA
UNTUK
DISABILITAS

BREAK
POINTS



Classification

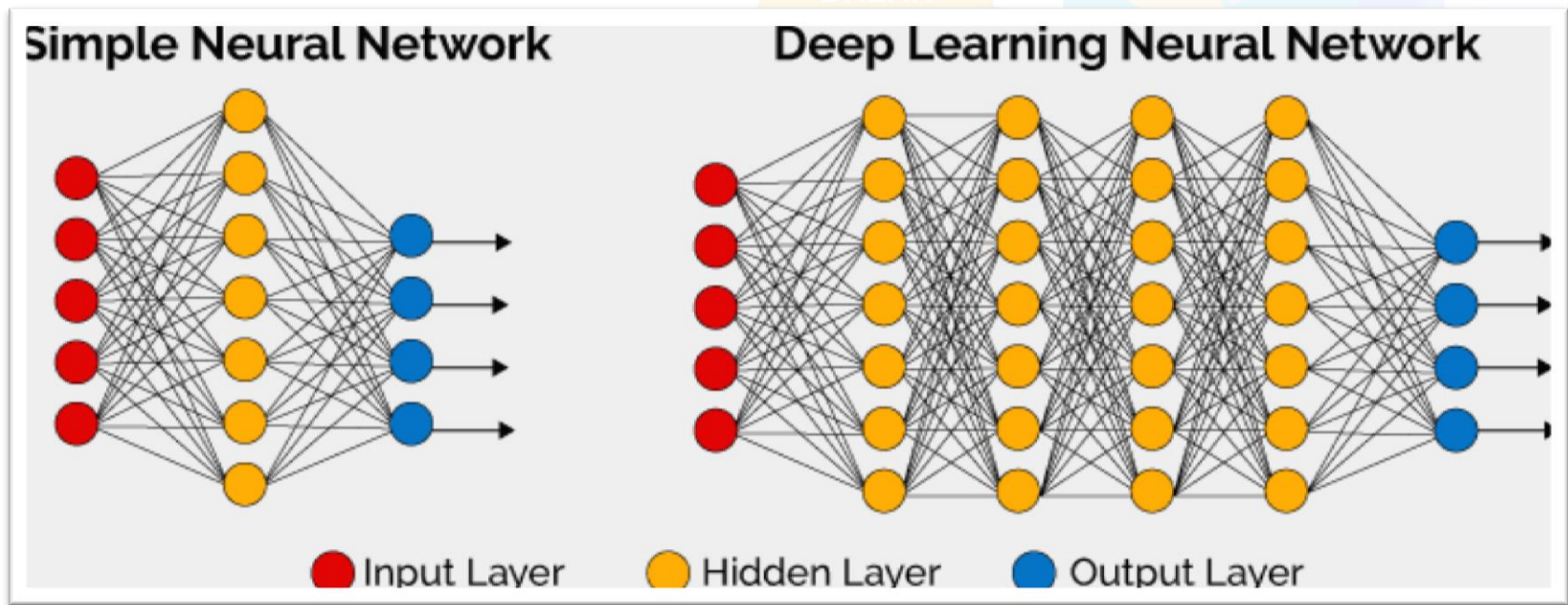
- Classification adalah teknik machine learning berupa algoritma pengklasifikasian objek-objek data ke dalam kelompok kelas yang telah ada.
- Pada classification, tidak akan ada pembentukan kelompok kelas baru.
- Classification merupakan salah satu jenis dari algoritma supervised learning, algoritma yang mempelajari korelasi antara sekumpulan input-output yang diinginkan dalam jumlah yang cukup besar menggunakan label.

Classification

- Dengan penggunaan supervised learning, maka beberapa hal berikut ini dapat dilakukan:
 1. Mendeteksi wajah, mendeteksi orang di dalam suatu gambar, mengenali ekspresi muka.
 2. Mengidentifikasi objek pada suatu gambar.
 3. Mengenali gestur pada suatu video.
 4. Mendeteksi suara, mendeteksi speaker, mentranskripsikan pidato menjadi teks, mengenali sentimen dalam suara.
 5. Mengklasifikasi teks sebagai spam (pada email) atau penipuan (pada asuransi), mengenali sentimen dari suatu teks.

Classification: Neural Networks

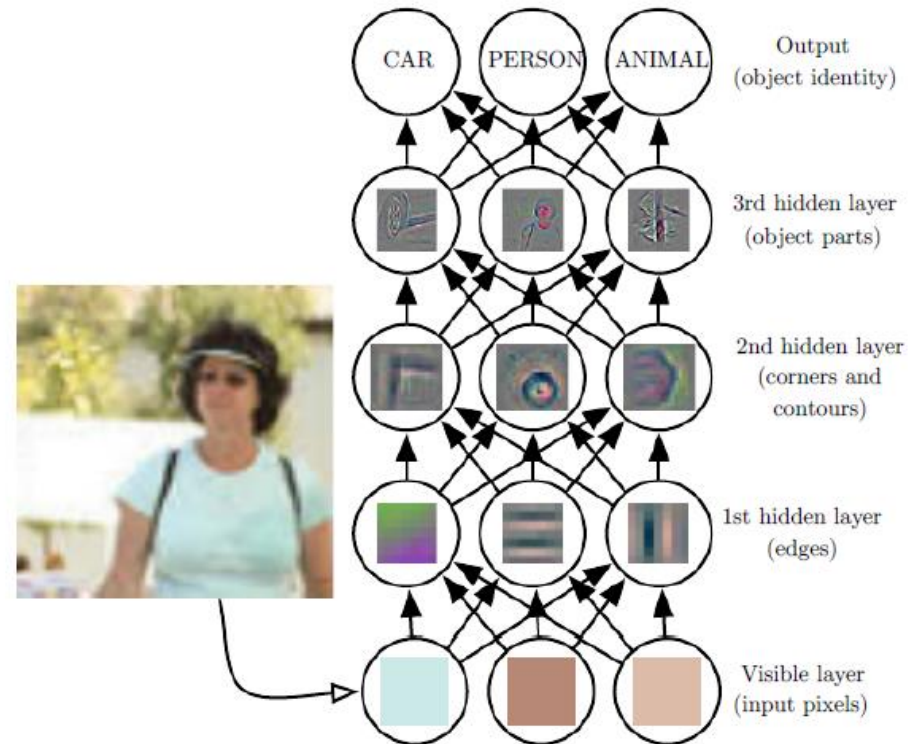
Salah satu pendekatan classification yaitu pendekatan neural network.



Classification: Neural Networks (2)

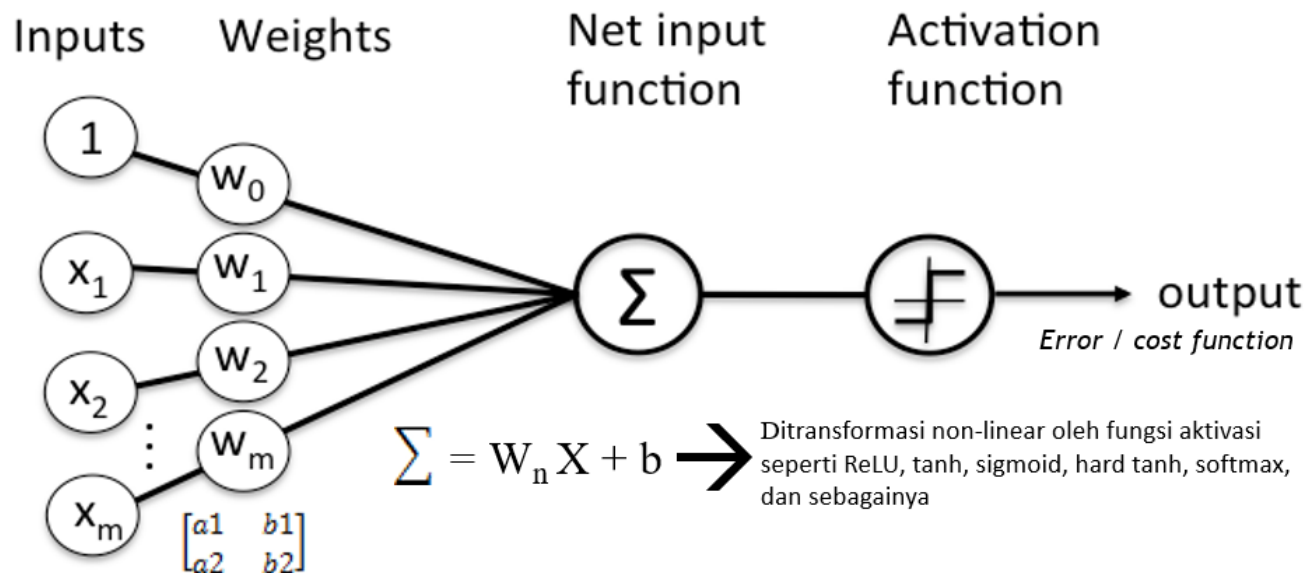
Neural networks terdiri dari 3 jenis layer yaitu:

- Visible: Mengandung variabel yang dapat kita amati
- Hidden: Mengekstrak fitur abstrak dari data yang ingin diamati.
- Output layer: Identitas terkait objek yang di observasi.



Classification: Neural Networks (3)

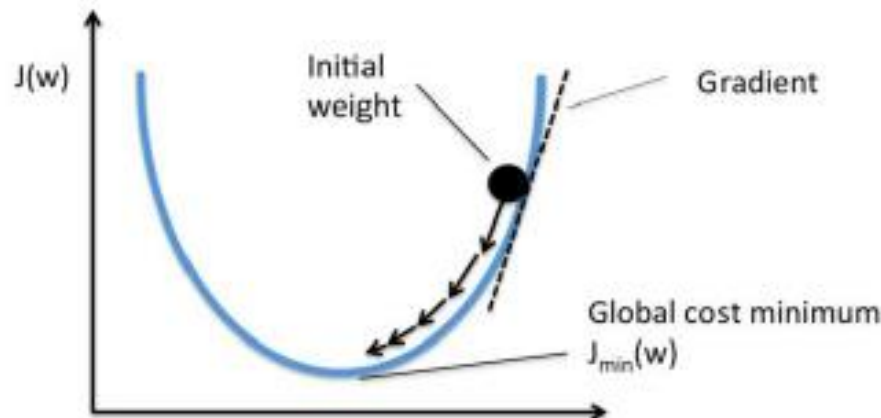
Neural-network terdiri dari sekumpulan neuron yang masing-masing melakukan operasi seperti berikut:



Keseluruhan proses di atas disebut forward propagation.

Classification: Neural Networks (4)

- Error keseluruhan dari forward propagation akan menjadi dasar dilakukannya back propagation.
- Cost dari forward propagation akan digunakan untuk mengukur gradient menggunakan gradient-descent.
- Gradient-descent bertujuan untuk melakukan pembaharuan bobot agar menghasilkan kesalahan yang paling minimum.

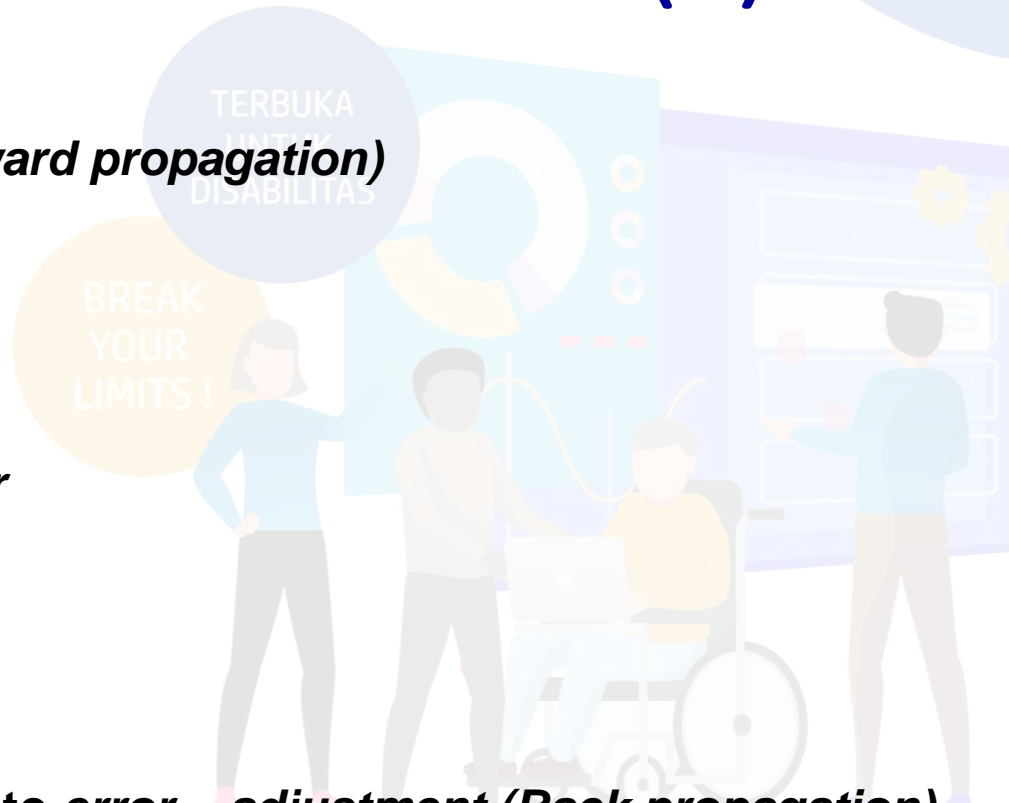


Classification: Neural Networks (5)

input * weight = guess (Forward propagation)

Ground-truth – guess = error

error * weight's contribution to error = adjustment (Back propagation)



Algoritma *Classification*

1. Nearest centroid

- a) Menghitung centroid untuk setiap kelas
- b) Menghitung jarak antara test sample dan setiap kelas centroid
- c) Memprediksi kelas dengan metode centroid terdekat

2. K-nearest neighbor

- a) Dalam hal ini setiap data baru akan dibandingkan dengan data training.
- b) Lalu 3 data training terdekat (misalkan kita ambil $k = 3$) dengan data baru akan diambil.
- c) Misalkan ketiga data tersebut masuk ke dalam kelompok 1, 2 dan 1, maka data baru tersebut dimasukan ke dalam kelompok 1 (seperti voting, karena suara yang terbanyak adalah 1, maka keputusannya adalah 1).
- d) Menggunakan prediksi dengan majority vote dengan jumlah yang ganjil.

Use case

1. Spektrometri massa protein

Untuk mengklasifikasi penyakit dan deteksi dini kanker.

2. Bike-Sharing

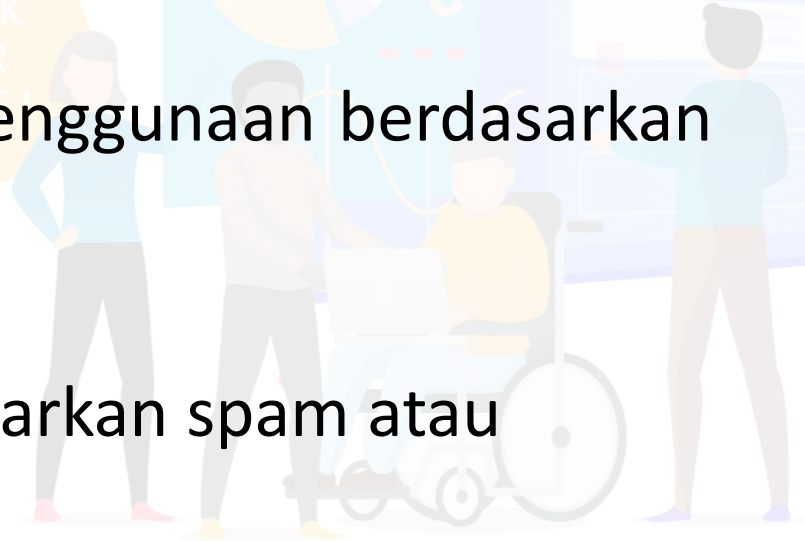
Untuk mengklasifikasi tingkat penggunaan berdasarkan hari, suhu, cuaca, dan waktu.

3. Email spam filter

Mengklasifikasikan email berdasarkan spam atau bukan.

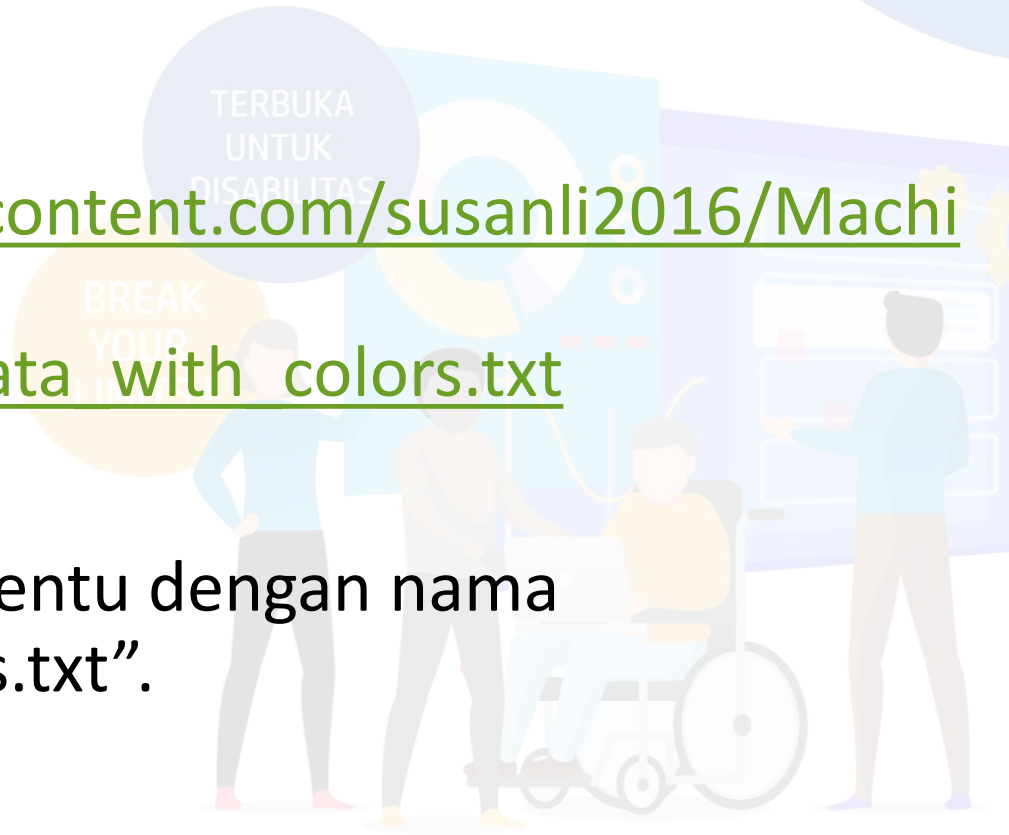
TERBUKA
UNTUK
SEMUA

BREAK
YOUR
LIMITS



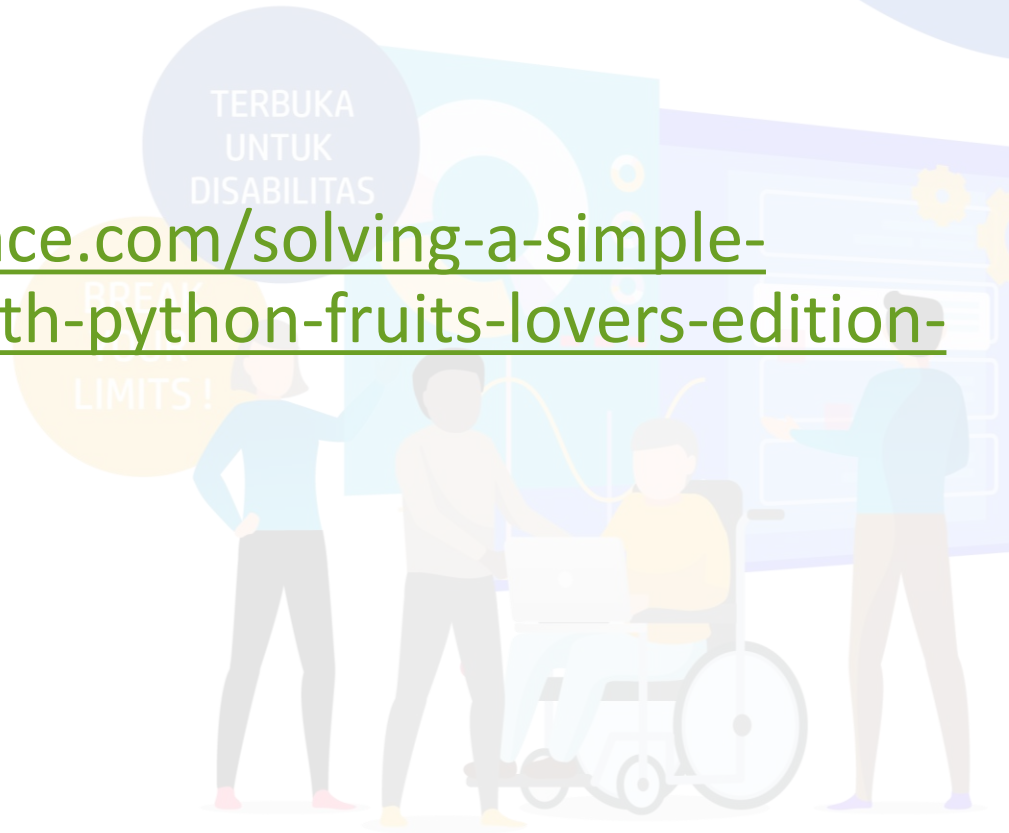
Tugas 2: Klasifikasi

- Akses https://raw.githubusercontent.com/susanli2016/Machine-Learning-with-Python/master/fruit_data_with_colors.txt
- Copy-paste ke notepad.
- Simpan di direktori tertentu dengan nama “fruit_data_with_colors.txt”.



- Ikuti tutorial di:

<https://towardsdatascience.com/solving-a-simple-classification-problem-with-python-fruits-lovers-edition-d20ab6b071d2>



Referensi

- Jain, AK and Dubes, RC, 1948, Algorithm for Clustering Data, Prentice Hall
- Matloff, N, Statistical Regression and Classification: from Linear Model to Machine Learning, Chapman & Hall
- <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>
- <http://heather.cs.ucdavis.edu/draftregclass.pdf>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1274262/>
- <https://towardsdatascience.com/machine-learnings-algorithms-how-they-work-and-use-cases-for-each-type-part-i-of-iii-67997c6dda84>
- <https://towardsdatascience.com/solving-a-simple-classification-problem-with-python-fruits-lovers-edition-d20ab6b071d2>
- <https://blog.floydhub.com/introduction-to-k-means-clustering-in-python-with-scikit-learn/>







DIGITAL
TALENT
SCHOLARSHIP

IKUTI KAMI



DIGITAL
TALENT
SCHOLARSHIP

-  [digitalent.kominfo](https://www.facebook.com/digitalent.kominfo)
-  [digitalent.kominfo](https://www.instagram.com/digitalent.kominfo)
-  [DTS_kominfo](https://twitter.com/DTS_kominfo)
-  Digital Talent Scholarship 2019

Pusat Pengembangan Profesi dan Sertifikasi
Badan Penelitian dan Pengembangan SDM
Kementerian Komunikasi dan Informatika
Jl. Medan Merdeka Barat No. 9
(Gd. Belakang Lt. 4 - 5)
Jakarta Pusat, 10110

