Nama : Sheva Ivanda Pratama

NIM : 24060120140089

Kelas : C

Session 1

What is data mining?

Didefinisikan sebagai bidang studi interdisipliner yang menggunakan data untuk berbagai tujuan penelitian dan pelaporan untuk memperoleh wawasan dan makna dari data tersebut.

Data science terms

Artificial Intelligence(AI)

Kecerdasa manusia oleh mesin

Contoh: NLP, Amazon purchase

Machine Learning(ML)

Sebuah pendekatan untuk mencapai AI

Contoh: Google maps

Deep Learning(DL)

Sebuah teknik untuk menerapkan pembelajaran mesin

Contoh: self driving car, robotic

Data science

Metode ilmiah, algoritme, dan sistem untuk mengekstrak pengetahuan atau wawasan dari data besar

Data analysis

Proses pemeriksaan, pembersihan, transformasi, dan pemodelan data

Data analytics

Penemuan, interpretasi dan komunikasi pola yang bermakna adalah data

Data Mining

Proses menemukan pola dalam kumpulan data besar yang melibatkan metode di persimpangan pembelajaran mesin, statistik, dan sistem basis data

What is Data?

Data mengacu pada kumpulan fakta yang biasanya diperoleh sebagai hasil dari pengalaman, pengamatan, atau eksperimen

Big Data with 8 V’s

1. Volume
2. Value
3. Veracity
4. Visualisation
5. Variety
6. Velocity
7. Viscosity
8. Virality

Data vs Information vs Knowledge

Data: sesuatu yang tidak memiliki arti

Information: data yang diproses dan memiliki arti

Knowledge: informasi yang diolah berdasarkan pengalaman dan eksperimen menjadi sebuah pengetahuan

Data vs Information vs Knowledge in Soup

Data: kumpulan bahan-bahan yang ada di konter

Information: kemudian Anda menyiapkan semuanya dengan mencuci, mengupas, dan memotong sayuran, memotong ayam dan membuka kaleng kaldu

Knowledge: Sekarang sup sudah siap diletakkan di mangkuk dan disajikan.

Skills For Data Science

Entry Level Skills

Python, R, SQL.

Soft Skills

Communication skills, curiosity, collaboration.

On-the Job Skills

Machine Learning & AI, deep learning, data mining.

What is data science life cycle?

Data science life cycle adalah serangkaian langkah-langkah ilmu data berulang yang Anda ambil untuk menyampaikan proyek atau analisis

Data science lifecycle

1. Business understanding
2. Data mining
3. Data cleaning
4. Data exploration
5. Feature Engineering

6. Predictive modelling
7. Data visualization

Machine Learning

Supervised Learning

- Analisis yang dilakukan untuk memprediksi kejadian di masa depan atau data atau tren lainnya.
- Itu menggunakan fungsi pembelajaran terawasi yang digunakan untuk memprediksi nilai target
- Data masukan adalah label
- Gunakan set data pelatihan
- Gunakan untuk prediksi

Unsupervised Learning

- Pada dasarnya digunakan untuk menghasilkan korelasi, tabulasi silang, frekuensi dll.
- Teknologi ini digunakan untuk menentukan kesamaan dalam data dan menemukan pola yang ada
- Untuk mengembangkan subkelompok yang menawan di sebagian besar data yang tersedia
- Data masukan tidak diberi label
- Gunakan hanya input dataset
- Digunakan untuk analisis untuk peringkasan dan transformasi data menjadi informasi yang berarti

Data Science Tools & Platform

Comercial and open platforms

- Orange
- Rapidminer
- IBM

Languages and programming platforms

- Java
- Python
- R
-

Big Data Programming tools

- Hadoop
- Spark
- Cloudera

Data Science is Everywhere

1. Manufacturing
2. E-Commerce
3. Healthcare
4. Sport
5. Finance
6. Banking
7. Transporation
8. Agriculture

Top Data Science Companies

1. Microsoft
2. Amazon
3. Visa
4. Google
5. Netflix
6. Facebook

Session 2

Natural Language Processing (NLP)

Bidang di persimpangan ilmu komputer, linguistik, kecerdasan buatan, dan banyak lagi

Spoken Language Understanding

People → speech recognition → Natural language understanding → dialogue management → natural language generation → text-to-speech → people

Why NLP is hard?

Ambigu: kata atau kalimat memiliki beberapa arti

Variavilitas: arti yang sama dapat diungkapkan dalam berbagai cara

Language Modelling (LM)

A central task in NLP:

- Machine translation
- Summarization
- Spell checker
- Dialogue systems

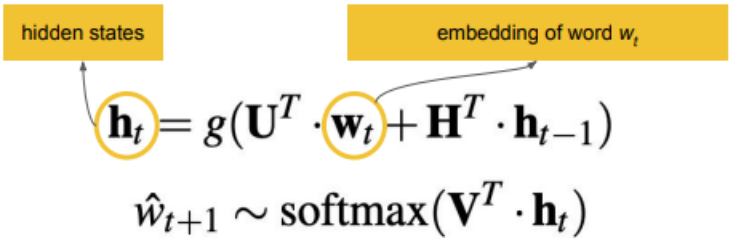
Word2vec

Approcximate softmax:

- Negative sampling
  Hanya pilih sejumlah kecil "negatif" untuk memperbarui parameter.
- Hierarchical softmax layers

Recurrent Neural Network LM

Gunakan jumlah konteks yang tak terbatas. Pada setiap langkah waktu t, RNN menghitung sebagai berikut:



Sequence-to-Sequence Model

Biasanya digunakan untuk tugas NLP yang menghasilkan teks, misalnya, terjemahan mesin atau ringkasan

- Encoder: Ubah input mentah menjadi representasi tersembunyi
- Decoder: Hasilkan output dari representasi tersembunyi

Attention mechanism:

Berikan bobot yang berbeda (“perhatian”) untuk input yang berbeda

Transformer Model

Non-recurrent encoder-decoder model

- Long-distance context has “equal opportunity”
- Allows parallelization

Components:

- Multi-headed self attention
- Feed-forward layers
- Layer norm and residuals
- Positional encoding

Transfer Learning with Large Language Models (LLMs)

Karena LLMs dilatih pada sejumlah besar data, kami dapat memanfaatkan pengetahuan mereka untuk tugas target tertentu.

Sangat berguna untuk tugas target di mana kami memiliki data berlabel terbatas atau nol

Intermediate-Task Transfer

Idea:

1. Pralatih model pada data yang tidak berlabel
2. Sempurnakan model pada kumpulan data perantara berlabel besar
3. Sempurnakan lagi pada kumpulan data berlabel target yang lebih kecil

RoBERTa → Finetune on intermediate task → Finetune on target task

(Traditional) Few-Shot Learning

N way K shot learning

Methods:

- Fine-tuning
- KNN
- Meta-learning

Challenges in Modern Few-Shot Learning

Many things to consider:

- Examples to be used
- Order of examples
- Prompt selection (design)

Zero-Shot Learning

Pada bahasa yang sama, format ulang tugas target di tugas sumber

Bahasa sumber berbeda dengan bahasa target

- Gunakan LM multibahasa sebagai model terlatih
- Penyesuaian pada data berlabel bahasa sumber daya tinggi
- Evaluasi data bahasa target
- Gunakan terjemahan mesin jika diperlukan

Challenges:

- Data prapelatihan LM multibahasa mungkin berisi sangat sedikit atau tidak ada data dalam bahasa sumber daya rendah
- Kinerja model MT yang buruk