

Machine Learning

Indah Agustien Siradjuddin

Pengenalan Machine Learning

Semester Gasal 2019-2020

Machine Learning

Definisi *Machine Learning*

- *The science (and art) of programming computers so they can learn from data*
- Field of study that gives computers the ability to learn without being explicitly programmed
- A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E

Aplikasi *Machine Learning*

Pengenalan Pola

Identifikasi E-Mail Spam

Identifikasi (Filter) E-mail Spam

- Spam vs. Ham Email
- Set Data Pelatihan dan *Training Instances*
- *Attribute* vs. *feature*
- Machine Learning :
 - Task T
 - Experience E
 - Performance P

Conventional vs Machine Learning

Conventional :

- ekstrak fitur penting yang merepresentasikan objek, misal : fitur untuk spam dan ham
- buat aturan untuk pengenalan
- buat code berdasarkan aturan, dan test code

Machine Learning :

- ekstrak fitur penting yang merepresentasikan objek
- implementasikan algoritma/pendekatan *machine learning*

Rambu penting dalam *Machine Learning*

1. *Supervision*
2. Iterasi Proses Pembelajaran
3. Generalisasi
4. Permasalahan di *Machine Learning*
5. Validasi

Supervision

1. *Supervised Learning*
2. *Unsupervised Learning*
3. *Semisupervised Learning*
4. *Reinforcement Learning*

Supervised Learning

Sumber :

Training Data : data + target/solution/class/label

Aplikasi :

- Filter *spam email*
- Prediksi *housing price*
- Pengenalan wajah, dll

Algoritma:

- *Regression*
- *Decision Tree*
- *Neural Networks*
- *LDA (optional)*

Unsupervised Learning

Sumber :

Training Data : data (*unlabeled data*)

Implementasi :

- *clustering*
- segmentasi

Algorithms :

- K-Means Clustering
- Dimension reduction - PCA

Semisupervised Learning

Kombinasi *supervised learning* dan *unsupervised learning*

Training Data : Unlabeled Data + labeled Data

Implementasi : Google Photos

Reinforcement Learning

Algoritma Pembelajaran :

- Pengamatan terhadap *Environment*
- Output berdasarkan aturan
- Penalty : *reward* atau *punishment*

Iterasi Proses Pembelajaran

1. *Batch Learning*
2. *Incremental Learning*

Batch Learning

offline learning : Iterasi pelatihan dilakukan dengan seluruh data pada dataset

Incremental Learning

online learning/stochastic learning: iterasi dilakukan pada setiap data pada dataset (tiap data atau group/mini batch)

Generalisasi

Pemberian keputusan ketika terdapat data baru

1. *Instance-based Learning*
2. *Model-based Learning*

Instance-based Learning

Sumber :

Generalisasi dilakukan dengan :

- Pencocokan data dengan setiap data (*instance*) yang terdapat pada dataset
- Data baru akan diklasifikasikan/dikelompokkan menjadi kelas yang sama dengan data yang terdekat
- similaritas

Model-based Learning

Sumber :

Generalisasi dilakukan dengan :

- data pelatihan digunakan untuk membangun fungsi diskriminan (model)
- data baru akan diklasifikasikan berdasarkan model

Permasalahan di dalam Machine Learning

1. *Bad Data*
2. *Bad Algorithm*

Bad Data

Insufficient Training Data

Sumber : Membutuhkan banyak data pada data pelatihan

Nonrepresentatif Data

Setiap data seharusnya dapat digunakan untuk generalisasi data, oleh karena itu data yang sama tidak untuk pelatihan

Poor Quality of Data

Dataset yang ada terkadang terdapat *Noises*, *outliers*, bahkan terdapat atribut-atribut yang tidak terdapat nilai

Irrelevant Features

Garbage In Garbage Out

Bad Algorithm

Overfitting Data

- Akurasi bagus untuk data training, akan tetapi
- memiliki akurasi rendah untuk generalisasi (testing)

Underfitting Data

- Akurasi rendah bahkan untuk data pelatihan

Validasi

- Data dibagi menjadi pelatihan dan pengujian
- Overfitting
- validasi : *cross validation*

Referensi