

Jan Wira Gotama Putra

Pengenalan Konsep Pembelajaran Mesin dan Deep Learning

Edisi 1.3

July 25, 2019

Untuk Tuhan, Bangsa, dan Almamater

Kata Pengantar

Buku ini ditujukan sebagai bahan pengantar (atau penunjang) mata kuliah *machine learning* untuk mahasiswa di Indonesia, khususnya tingkat sarjana (tidak menutup kemungkinan digunakan untuk tingkat pascasarjana). Buku ini hanya merupakan komplemen, bukan sumber informasi utama. Buku ini memuat materi dasar *machine learning*, yang ditulis sedemikian rupa sehingga pembaca mampu mendapatkan **intuisi**. Pembaca masih harus membaca buku-buku lainnya untuk mendapatkan pemahaman lebih dalam.

Walaupun tidak sempurna, mudah-mudahan buku ini mampu memberi inspirasi. Anggap saja membaca buku ini seperti sedang membaca “*light novel*”. Penulis ingin buku ini bisa menjadi *pointer*; i.e. dengan membaca buku ini, diharapkan kawan-kawan juga mengetahui harus belajar apa (lebih jauhnya) dalam bidang *machine learning*. Setelah membaca buku ini, pembaca diharapkan mampu membaca literatur *machine learning* yang dijelaskan secara lebih matematis (kami memberi rekomendasi bacaan lanjutan).

Di Indonesia, penulis banyak mendengar baik dari teman, junior, senior, dll; suatu pernyataan “kuliah mengajarkan teori saja, praktiknya kurang, dan tidak relevan dengan industri”. Menurut saya di satu sisi itu benar; tapi di sisi lain, karena pemikiran macam itu terkadang kita tidak benar-benar mengerti permasalahan. Ketika mengalami kendala, kita buntu saat mencari solusi karena fondasi yang tidak kokoh. Banyak orang terburu-buru “menggunakan *tools*” karena lebih praktikal. Penulis ingin mengajak saudara/i untuk memahami konsep *machine learning* secara utuh sebelum memanfaatkan. Ada perbedaan yang mendasar antara hanya mampu menggunakan *tools* dan mengerti konsep secara utuh.

Buku ini menjelaskan algoritma *machine learning* dari sudut pandang “agak” matematis. Pembaca disarankan sudah memahami/mengambil setidaknya mata kuliah statistika, kalkulus, aljabar linear, pengenalan kecerdasan buatan, dan logika fuzzy. Penulis merasa banyak esensi yang hilang ketika materi *machine learning* hanya dijelaskan secara deskriptif karena itu buku ini ditulis dengan bahasa “agak” matematis. Walaupun demikian, penulis berusaha menggunakan notasi matematis seminimal dan sesederhana

mungkin, secukupnya sehingga pembaca mampu mendapatkan intuisi. Saat membaca buku ini, disarankan membaca secara runtun. Gaya penulisan buku ini **santai/semiformal** agar lebih mudah dipahami, mudah-mudahan tanpa mengurangi esensi materi.

Buku ini ditulis menggunakan template monograph (L^AT_EX) dari Springer yang dimodifikasi. Dengan demikian, mungkin ada kesalahan pemenggalan kata. Tentunya, buku tidak lepas dari kekurangan, misalnya kesalahan tipografi. Kami sarankan pembaca untuk membaca secara seksama, termasuk menginterpretasikan variabel pada persamaan.

Petunjuk Penggunaan

Struktur penyajian buku ini dapat dijadikan acuan sebagai struktur kuliah *machine learning* untuk satu semester (bab 1 untuk sesi pertama, dst). Agar dapat memahami materi per bab, bacalah keseluruhan isi bab secara utuh sebelum mempertanyakan isi materi. Penulis sangat menyarankan untuk membahas soal latihan sebagai tambahan materi (bisa juga sebagai PR). Soal latihan ditujukan untuk mengarahkan apa yang harus dibaca/dipahami lebih lanjut.

Pembaca dipersilahkan menyebarkan (*share*) buku ini untuk alasan **NON KOMERSIAL** (pendidikan), tetapi **dimohon kesadarannya untuk tidak menyalin /meniru isi buku ini**. Bila ingin memuat konten diktat ini pada media yang pembaca kelola, dimohon untuk mengontak pengarang terlebih dahulu. Tidak semua istilah bahasa asing diterjemahkan ke Bahasa Indonesia supaya makna sebenarnya tidak hilang (atau penulis tidak tahu versi Bahasa Indonesia yang baku).

Bab lebih awal memuat materi yang relatif lebih “mudah” dipahami dibanding bab berikutnya. Buku ini memberikan contoh dimulai dari contoh sederhana (beserta contoh data). Semakin menuju akhir buku, notasi yang digunakan akan semakin simbolik, beserta contoh yang lebih abstrak. Penulis sangat menyarankan untuk **membaca buku ini secara sekuensial**.

Kutipan

Buku ini tergolong *self-published work*, tetapi sudah di-*review* oleh beberapa orang. Kami yakin para *reviewers* adalah orang yang berkompeten. Silahkan merujuk buku ini sesuai dengan paduan cara merujuk *self-published work* (apabila diperbolehkan untuk merujuk *self-published work* pada pekerjaan kamu).

Notasi Penting

Karakter *bold* kapital merepresentasikan matriks (**X**, **Y**, **Z**). Dimensi matriks ditulis dengan notasi $N \times M$ dimana N merepresentasikan banyaknya baris dan M merepresentasikan banyaknya kolom. Elemen matriks direpresentasikan oleh $\mathbf{X}_{i,j}$, $\mathbf{X}_{[i,j]}$, atau $x_{i,j}$ untuk baris ke- i kolom ke- j (penggunaan akan menyesuaikan konteks pembahasan agar tidak ambigu). Karakter *bold* merepresentasikan vektor (**x**). Elemen vektor ke- i direpresentasikan oleh

x_i atau $\mathbf{x}_{[i]}$ tergantung konteks. Ketika penulis menyebutkan vektor, yang dimaksud adalah **vektor baris** (*row vector*, memiliki dimensi $1 \times N$, mengadopsi notasi Goldberg [1]). Perhatikan, literatur *machine learning* lainnya mungkin tidak menggunakan notasi *row vector* tetapi *column vector*. Kami harap pembaca mampu beradaptasi. Simbol “ \cdot ” digunakan untuk melambangkan operator *dot-product*.

Kumpulan data (atau himpunan) direpresentasikan dengan karakter kapital (C, Z), dan anggotanya (*data point*, *data entry*) ke- i direpresentasikan dengan karakter c_i . Perhatikan, elemen vektor dan anggota himpunan bisa memiliki notasi yang sama (himpunan dapat direpresentasikan di komputer sebagai *array*, jadi penggunaan notasi vektor untuk himpunan pada konteks pembicaraan kita tidaklah salah). Penulis akan menggunakan simbol $\mathbf{x}_{[i]}$ sebagai elemen vektor apabila ambigu. Fungsi dapat direpresentasikan dengan huruf kapital maupun non-kapital $f(\dots), E(\dots), G(\dots)$. Ciri fungsi adalah memiliki parameter! Pada suatu koleksi vektor (himpunan vektor) \mathbf{D} , vektor ke- i direpresentasikan dengan \mathbf{d}_i , dan elemen ke- j dari vektor ke- i direpresentasikan dengan $\mathbf{d}_{i[j]}$, $\mathbf{D}_{i,j}$, atau $\mathbf{D}_{[i,j]}$ (karena sekumpulan vektor dapat disusun sebagai matriks).

Karakter non-kapital tanpa *bold* atau indeks (a, b, c, x, y, z) merepresentasikan *random variable* (statistik) atau variabel (matematik). Secara umum, saat *random variable* memiliki tertentu, dinotasikan dengan $x = X$ (nilai tertentu dinotasikan dengan huruf kapital), kecuali disebutkan secara khusus saat pembahasan. Probabilitas direpresentasikan dengan karakter kapital (P), dengan karakter non-kapital merepresentasikan *probability density* (p). Penulis yakin pembaca dapat menyesuaikan interpretasi simbol berdasarkan konteks pembahasan. Untuk menginterpretasikan notasi lain, selain yang diberikan pada paduan ini, mohon menyesuaikan dengan ceritera pembahasan.

Ucapan Terima Kasih

Penulis ingin mengucapkan terima kasih pada Bapak/Ibu/Saudara/i atas kontribusi pada pengembangan dan penulisan buku ini: Adhiguna Surya Kuncoro, Arief Yudha Satria, Candy Olivia Mawalim, Chairuni Aulia Nusapati, Genta Indra Winata, Hayyu Luthfi Hanifah, I Gede Mahendra Darmawiguna, dan Tifani Warnita.

Catatan lain

Buku ini adalah *ongoing project*. Versi terakhir dan terakurat dapat dilihat pada <https://wiragotama.github.io/>. Buku ini lebih baik dibaca versi pdf-nya (agar gambar bisa di-*zoom*).

Tokyo, Jepang

Jan Wira Gotama Putra
<https://wiragotama.github.io/>

Daftar Isi

Bagian I Pengetahuan Dasar	1
1 Pengenalan	3
1.1 Kecerdasan Buatan	3
1.2 Intelligent Agent	6
1.3 Konsep Belajar	8
1.4 Statistical Learning Theory	8
1.5 Training, Development, Testing Set	10
1.6 Supervised Learning	12
1.7 Regresi	15
1.8 Semi-supervised Learning	15
1.9 Unsupervised Learning	15
1.10 Proses Belajar	17
1.11 Tips	18
1.12 Contoh Aplikasi	19
Soal Latihan	19
2 Fondasi Matematis	21
2.1 Probabilitas	21
2.2 Probability Density Function	23
2.3 Expectation dan Variance	25
2.4 Bayesian Probability	25
2.5 Gaussian Distribution	27
2.6 Teori Keputusan	29
2.7 Teori Informasi	31
2.8 Matriks	33
2.9 Bacaan Lanjutan	34
Soal Latihan	35
3 Data Analytics	37
3.1 Pengenalan Data Analytics	37
3.2 Nilai Atribut dan Transformasi	39
3.3 Ruang Konsep	40
3.4 Linear Separability	41
3.5 Seleksi Fitur	42

3.6	Classification, Association, Clustering	43
3.7	Mengukur Kinerja	44
3.8	Evaluasi Model	44
3.9	Kategori Jenis Algoritma	46
3.10	Tahapan Analisis	46
	Soal Latihan	47
Bagian II Algoritma Pembelajaran Mesin		49
4	Algoritma Dasar	51
4.1	Naive Bayes	51
4.2	K-means	54
4.3	K-nearest-neighbor	56
	Soal Latihan	56
5	Model Linear	59
5.1	Curve Fitting dan Error Function	59
5.2	Binary Classification	62
5.3	Log-linear Binary Classification	62
5.4	Multi-class Classification	64
5.5	Multi-label Classification	67
5.6	Pembelajaran sebagai Permasalahan Optimisasi	69
5.7	Batasan Model Linear	73
5.8	Overfitting dan Underfitting	74
5.9	Regularization	76
5.10	Transformasi Data	77
5.11	Bacaan Lanjutan	78
	Soal Latihan	79
6	Pohon Keputusan	81
6.1	Inductive Learning	81
6.2	ID3	82
6.3	Isu pada ID3	86
6.4	Pembagian Ruang Konsep	86
	Soal Latihan	87
7	Support Vector Classifier	89
7.1	Maximal Margin Classifier	89
7.2	Support Vector Classifier	94
7.3	Support Vector Machine	95
7.4	Klasifikasi lebih dari dua kelas	96
7.5	Tips	97
	Soal Latihan	97
8	Hidden Markov Model	99

XII DAFTAR ISI

8.1	Probabilistic Reasoning	99
8.2	Generative Model	102
8.3	Part-of-speech Tagging	103
8.4	Hidden Markov Model Tagger	106
8.5	Algoritma Viterbi	109
8.6	Proses Training Hidden Markov Model	111
	Soal Latihan	114
9	Seleksi Fitur dan Metode Evaluasi	115
9.1	Feature Engineering	115
9.2	High Dimensional Data	116
9.3	Feature Selection	116
9.4	Cross Validation	120
9.5	Replicability, Overclaiming dan Domain Dependence	122
	Soal Latihan	123
10	Clustering	125
10.1	K-means, Pemilihan Centroid, Kemiripan Data	126
10.2	Hierarchical Clustering	127
10.3	Evaluasi	128
	Soal Latihan	130
	Bagian III Artificial Neural Network	131
11	Feedforward Neural Network	133
11.1	Definisi Artificial Neural Network	133
11.2	Single Perceptron	134
11.3	Permasalahan XOR	136
11.4	Multilayer Perceptron	138
11.5	Interpretability	142
11.6	Binary Classification	143
11.7	Multi-class Classification	143
11.8	Multi-label Classification	144
11.9	Deep Neural Network	144
11.10	Tips	147
11.11	Regularization and Dropout	149
11.12	Vanishing and Exploding Gradients	150
11.13	Rangkuman	150
	Soal Latihan	151
12	Autoencoder	153
12.1	Representation Learning	153
12.2	Singular Value Decomposition	155
12.3	Ide Dasar Autoencoder	156

12.4	Resisting Perturbation	159
12.5	Representing Context: Word Embedding	161
12.6	Tips	169
	Soal Latihan	169
13	Arsitektur Neural Network	171
13.1	Convolutional Neural Network	171
13.2	Recurrent Neural Network	176
13.3	Part-of-speech Tagging Revisited	181
13.4	Sequence to Sequence	184
13.5	Arsitektur Lainnya	193
13.6	Architecture Ablation	193
13.7	Transfer Learning	193
13.8	Multi-task Learning	196
	Soal Latihan	200
	Bagian IV Aplikasi dan Topik Tambahan	201
14	Penerapan Pembelajaran Mesin	203
14.1	Sistem Rekomendasi	204
14.2	Peringkasan Dokumen	207
14.3	Konklusi	210
14.4	Saran Buku Lanjutan	211
	Soal Latihan	213
	Referensi	215

