

Jan Wira Gotama Putra

Pengenalan Konsep Pembelajaran Mesin dan Deep Learning

Edisi 1.2

December 27, 2018

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 ($\mathbf{X}, \mathbf{Y}, \mathbf{Z}$). 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 (\mathbf{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.

Tokyo, Jepang

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

Daftar Isi

Bagian I Pengetahuan Dasar

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	14
1.8	Semi-supervised Learning	14
1.9	Unsupervised Learning	15
1.10	Proses Belajar	16
1.11	Tips	17
1.12	Contoh Aplikasi	18
	Soal Latihan	18
2	Fondasi Matematis	19
2.1	Probabilitas	19
2.2	Probability Density Function	21
2.3	Expectation dan Variance	23
2.4	Bayesian Probability	23
2.5	Gaussian Distribution	25
2.6	Teori Keputusan	27
2.7	Teori Informasi	29
2.8	Matriks	31
2.9	Bacaan Lanjutan	32
	Soal Latihan	33

3	Data Analytics	35
3.1	Pengenalan Data Analytics	35
3.2	Nilai Atribut dan Transformasi	37
3.3	Ruang Konsep	38
3.4	Linear Separability	39
3.5	Seleksi Fitur	40
3.6	Classification, Association, Clustering	41
3.7	Mengukur Kinerja	42
3.8	Evaluasi Model	42
3.9	Kategori Jenis Algoritma	44
3.10	Tahapan Analisis	44
	Soal Latihan	45

Bagian II Algoritma Pembelajaran Mesin

4	Algoritma Dasar	49
4.1	Naive Bayes	49
4.2	K-means	52
4.3	K-nearest-neighbor	54
	Soal Latihan	54
5	Model Linear	57
5.1	Curve Fitting dan Error Function	57
5.2	Binary Classification	60
5.3	Log-linear Binary Classification	60
5.4	Multi-label Classification	62
5.5	Pembelajaran sebagai Permasalahan Optimisasi	64
5.6	Batasan Model Linear	69
5.7	Overfitting dan Underfitting	70
5.8	Regularization	72
5.9	Transformasi Data	73
5.10	Bacaan Lanjutan	74
	Soal Latihan	75
6	Pohon Keputusan	77
6.1	Inductive Learning	77
6.2	ID3	78
6.3	Isu pada ID3	82
6.4	Pembagian Ruang Konsep	82
	Soal Latihan	83

7	Support Vector Classifier	85
7.1	Maximal Margin Classifier	85
7.2	Support Vector Classifier	90
7.3	Support Vector Machine	91
7.4	Klasifikasi lebih dari dua kelas	92
7.5	Tips	93
	Soal Latihan	93
8	Hidden Markov Model	95
8.1	Probabilistic Reasoning	95
8.2	Generative Model	98
8.3	Part-of-speech Tagging	99
8.4	Hidden Markov Model Tagger	102
8.5	Algoritma Viterbi	104
8.6	Proses Training Hidden Markov Model	106
	Soal Latihan	110
9	Seleksi Fitur dan Metode Evaluasi	111
9.1	Feature Engineering	111
9.2	High Dimensional Data	112
9.3	Feature Selection	112
9.4	Cross Validation	116
9.5	Replicability, Overclaiming dan Domain Dependence	118
	Soal Latihan	119
10	Clustering	121
10.1	K-means, Pemilihan Centroid, Kemiripan Data	122
10.2	Hierarchical Clustering	123
10.3	Evaluasi	124
	Soal Latihan	126

Bagian III Artificial Neural Network

11	Feedforward Neural Network	129
11.1	Definisi Artificial Neural Network	129
11.2	Single Perceptron	130
11.3	Permasalahan XOR	132
11.4	Multilayer Perceptron	134
11.5	Interpretability	138
11.6	Binary Classification	139
11.7	Multi-label Classification	139
11.8	Deep Neural Network	141
11.9	Tips	142
11.10	Regularization and Dropout	144

11.11	Vanishing and Exploding Gradients	145
11.12	Rangkuman	146
	Soal Latihan	147
12	Autoencoder	149
12.1	Representation Learning	149
12.2	Singular Value Decomposition	151
12.3	Ide Dasar Autoencoder	152
12.4	Resisting Perturbation	155
12.5	Representing Context: Word Embedding	156
12.6	Tips	165
	Soal Latihan	165
13	Arsitektur Neural Network	167
13.1	Convolutional Neural Network	167
13.2	Recurrent Neural Network	172
13.3	Part-of-speech Tagging Revisited	177
13.4	Sequence to Sequence	179
13.5	Arsitektur Lainnya	189
13.6	Architecture Ablation	189
13.7	Topik Khusus: Multi-task Learning	189
	Soal Latihan	193
<hr/>		
Bagian IV Aplikasi dan Topik Tambahan		
<hr/>		
14	Penerapan Pembelajaran Mesin	197
14.1	Sistem Rekomendasi	198
14.2	Peringkasan Dokumen	201
14.3	Konklusi	204
14.4	Saran Buku Lanjutan	205
	Soal Latihan	207
	Referensi	209