

Jan Wira Gotama Putra

Pengenalan Konsep Pembelajaran Mesin dan Deep Learning

Edisi 1.0

March 11, 2018

Untuk Tuhan, Bangsa, dan Almamater

Kata Pengantar

Buku ini ditujukan sebagai bahan penunjuang (atau pengantar) 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. Beberapa *reviewers* merasa materi buku ini relatif cukup berat, karena itu ada baiknya membaca buku pengantar yang lebih “ringan” sebelum membaca buku ini.

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 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.

Buku ini menjelaskan algoritma *machine learning* dari sudut pandang “agak” matematis. Pembaca disarankan sudah memahami/mengambil setidaknya mata kuliah statistika, kalkulus, aljabar linier/geometri, 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. Saat membaca buku ini, disarankan membaca secara runtun. Gaya penulisan buku ini **santai/semiformal** agar lebih mudah dipahami, dengan notasi matematis dibuat seminimal mungkin, 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 (karena dipenggal berdasarkan jumlah karakter).

Petunjuk Penggunaan

Struktur penyajian buku ini dapat dijadikan acuan sebagai struktur kuliah *machine learning* untuk satu semester (bab 1 untuk sesi pertama, dst), sementara materi mungkin masih perlu ditambahkan diluar buku ini. 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. Agar dapat memahami materi per bab, bacalah keseluruhan isi bab secara utuh sebelum mempertanyakan isi materi.

Pembaca dipersilahkan menyebar 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 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 adalah tidak 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	5
1.3	Konsep Belajar	7
1.4	Statistical Learning Theory	8
1.5	Training, Development, Testing Set	10
1.6	Supervised Learning	11
1.7	Regresi	14
1.8	Semi-supervised Learning	14
1.9	Unsupervised Learning	14
1.10	Proses Belajar	16
1.11	Tips	17
1.12	Contoh Aplikasi	17
	Soal Latihan	18
2	Fondasi Matematis	19
2.1	Probabilitas	19
2.2	Probability Density Function	21
2.3	Expectation dan Variance	22
2.4	Bayesian Probability	23
2.5	Gaussian Distribution	24
2.6	Teori Keputusan	26
2.7	Teori Informasi	28
2.7.1	Entropy	29
2.7.2	Relative Entropy dan Mutual Information	30
2.8	Matriks	31
2.9	Bacaan Lanjutan	32
	Soal Latihan	32

3	Data Analytics	33
3.1	Pengenalan Data Analytics	33
3.2	Nilai Atribut dan Transformasi	35
3.3	Ruang Konsep	36
3.4	Linear Separability	37
3.5	Seleksi Fitur	38
3.6	Classification, Association, Clustering	39
3.7	Mengukur Kinerja	40
3.8	Evaluasi Model	40
3.9	Kategori Jenis Algoritma	42
3.10	Tahapan Analisis	42
	Soal Latihan	42

Bagian II Algoritma Pembelajaran Mesin

4	Algoritma Dasar	45
4.1	Naive Bayes	45
4.2	K-means	47
4.3	K-nearest-neighbor	50
	Soal Latihan	50
5	Model Linear	53
5.1	Curve Fitting dan Error Function	53
5.2	Overfitting dan Underfitting	55
5.3	Binary Classification	57
5.4	Log-linear Binary Classification	58
5.5	Multi-class Classification	59
5.6	Transformasi	60
5.7	Pembelajaran sebagai Permasalahan Optimisasi	61
5.8	Regularization	64
5.9	Bacaan Lanjutan	65
	Soal Latihan	66
6	Pohon Keputusan	67
6.1	Inductive Learning	67
6.2	ID3	68
6.3	Isu pada ID3	72
6.4	Hubungan Decision Tree dan Model Linear	72
	Soal Latihan	73
7	Hidden Markov Model	75
7.1	Probabilistic Reasoning	75
7.2	Generative Model	78
7.3	Part-of-speech Tagging	79

7.4	Hidden Markov Model Tagger	82
7.5	Algoritma Viterbi	84
7.6	Proses Training Hidden Markov Model	86
	Soal Latihan	89
8	Clustering	91
8.1	K-means, Pemilihan Centroid, Kemiripan Data	92
8.2	Hierarchical Clustering	93
8.3	Evaluasi	94
	Soal Latihan	96

Bagian III Neural Networks

9	Artificial Neural Network	99
9.1	Definisi	99
9.2	Single Perceptron	100
9.3	Permasalahan XOR	102
9.4	Multilayer Perceptron	104
9.5	Interpretability	107
9.6	Binary Classification	108
9.7	Multi-label Classification	109
9.8	Deep Neural Network	110
9.9	Tips	112
9.10	Regularization and Dropout	113
9.11	Vanishing and Exploding Gradients	114
9.12	Rangkuman	114
	Soal Latihan	115
10	Dimensionality Reduction dan Representation Learning ...	117
10.1	Curse of Dimensionality	117
10.2	Singular Value Decomposition	119
10.3	Ide Dasar Autoencoder	120
10.4	Representing Context: Word Embedding	122
10.4.1	Vector Space Model	123
10.4.2	Sequential, Time Series, dan Compositionality	124
10.4.3	Distributed Word Representation	125
10.4.4	Distributed Sentence Representation	127
10.5	Tips	130
	Soal Latihan	130
11	Arsitektur Neural Network	131
11.1	Convolutional Neural Network	131
11.1.1	Convolution	133
11.1.2	Pooling	134

XIV Daftar Isi

11.1.3 Rangkuman	136
11.2 Recurrent Neural Network	136
11.3 Part-of-speech Tagging Revisited	141
11.4 Sequence to Sequence	144
11.4.1 Encoder	145
11.4.2 Decoder	146
11.4.3 Beam Search	146
11.4.4 Attention-based Mechanism	148
11.4.5 Variasi Arsitektur Sequence to Sequence	150
11.4.6 Rangkuman	151
11.5 Arsitektur Lainnya	153
Soal Latihan	153

Bagian IV Aplikasi dan Topik Tambahan

12 Penerapan Pembelajaran Mesin	157
12.1 Sistem Rekomendasi	158
12.1.1 Content-based Filtering	158
12.1.2 Collaborative Filtering	160
12.2 Peringkasan Dokumen	161
12.2.1 Pipelined Approach	163
12.2.2 Single-view Approach	163
12.3 Konklusi	164
12.4 Saran Buku Lanjutan	165
Soal Latihan	167
Referensi	169