Statistika dan Machine Learning: Satu Ilmu Dua Wajah

Rizal Zaini Ahmad Fathony (`rizalzaf@gmail.com`) [1]

Hampir semua orang yang bergelut di dunia ilmu pengetahuan sedikit banyak pasti pernah berinteraksi dengan statistika (ilmu statistik). Tidak dapat dipungkiri bahwa statistika mempunyai peranan penting sebagai katalis perkembangan ilmu-ilmu lain, baik ilmu alam (seperti astronomi dan biologi) ataupun ilmu sosial (seperti ekonomi, demografi, sosiologi, dsb.). Statistika dipakai oleh disiplin ilmu lain sebagai alat untuk mengambil kesimpulan, menguji hipotesis/teori, memahami fenomena, menganalisis eksperimen, menentukan keputusan, dsb.

Machine learning saat ini menjadi cabang ilmu pengetahuan yang populer dibicarakan di media. Didapuk sebagai salah satu cabang dari ilmu kecerdasan buatan (*artificial inteligence*), hampir semua orang pernah berinteraksi, memakai ataupun mendengar sistem komputer yang dibangun memakai teknik *machine learning*. Mulai dari melihat *tag*-otomatis foto di *Facebook*, menggunakan rekomendasi pencarian di *Google*, meng-klik rekomendasi produk sejenis di *online shopping*, menikmati servis email tanpa-spam, sampai dengan mendengar berita *AlphaGo* yang mengalahkan pemain profesional top di permainan *Go*.

Meskipun aplikasi dari disiplin ilmu statistika dan *machine learning* kelihatan sangat berbeda, dua ilmu tersebut sangat berkaitan. Baik statistika maupun *machine learning* merupakan ilmu tentang data. Teori-teori di disiplin ilmu statistika dan *machine learning* sebagian besar juga saling tumpang tindih. Hal ini bisa dilihat dari isi sebuah buku tentang statistika, "*All of statistics: a concise course in statistical inference*" (Wasserman, 2013) dan buku tentang *machine learning*, "Machine learning: a probabilistic perspective" (Murphy, 2012). Jika kita lihat sekilas kata-kata kunci dari daftar isi buku "*All of statistics*", akan sangat susah mencari kata kunci yang tidak tercantum juga di buku "*Machine learning*". Topik-topik mulai dari *probability distribution*, *linear & logistic regression*, *maximum likelihood estimation*, *markov model*, sampai *Bayesian MCMC* dibahas oleh kedua buku. Jadi, apa persamaan dan perbedaan kedua disiplin ilmu tersebut? Kita akan bahas lebih detail di tulisan ini.

Pondasi dasar: teori peluang

Pondasi dasar dari statistika dan *machine learning* adalah ilmu teori peluang. Semua teknik-teknik dalam statistika dan *machine learning* dibangun di atas teori peluang, yang merupakan bahasa matematika untuk mengukur derajat ketidakpastian. Ilmu-ilmu dasar yang juga penting bagi statistika dan *machine learning* diantaranya adalah aljabar linear, kalkulus dan teknik optimisasi. Dalam mendalami statistika dan *machine learning*, latar belakang kuat di ilmu-ilmu dasar tersebut akan sangat membantu.

Disiplin ilmu statistika sudah dirintis sejak abad ke 17 di mana dasar-dasar teori peluang dirumuskan oleh berbagai ahli. Berbagai distribusi peluang seperti distribusi normal/Gaussian, dan distribusi Poisson dirumuskan di sekitar abad ke 18 dan 19. Pondasi dari teori statistika modern dituangkan oleh statistisi dari Inggris, Ronald Fisher di awal abad ke 20. Kontribusi Fisher sangat penting dalam perkembangan statistika termasuk diantaranya adalah teknik estimasi terkenal *maximum likelihood*.

Dibandingkan statistika, disiplin ilmu *machine learning* masih tergolong relatif lebih muda, baru dirumuskan di sekitar akhir abad 20. *Machine learning* didefinisikan sebagai disiplin ilmu yang mempelajari algoritma komputer yang bisa belajar (*learn*) tanpa harus diprogram secara eksplisit, melainkan dengan sendirinya belajar dari data. Karena *machine learning* juga merupakan ilmu yang menggunakan data, pondasi dasar *machine learning* banyak 'meminjam' dari disiplin ilmu statistika yang pada abad 20 sudah relatif lebih matang.

Pengambilan kesimpulan dan interpretasi vs. prediksi

Kalau topik-topik yang dipelajari statistika dan *machine learning* kurang lebih sama, ilmu-ilmu dasar yang dibutuhkan-pun juga sama, lalu di mana letak perbedaan statistika dan *machine learning*? Perbedaan utamanya terletak pada fokus yang berbeda dari kedua disiplin ilmu tersebut. Statistika lebih fokus ke pengambilan kesimpulan dan interpretasi dari model, sedangkan *machine learning* lebih fokus ke penggunaan model untuk prediksi data baru. Untuk lebih jelasnya mari kita lihat contoh sederhana dari alur kerja praktisi statistika dan *machine learning* dalam memecahkan masalah.

Salah satu contoh problem sederhana yang sering dianalisis oleh praktisi statistika diantaranya adalah problem di

mana peneliti mempunyai variabel dependen kategorik y yang ingin dianalisis dengan beberapa variabel independen $x_1, \dots x_m$. Contoh aplikasinya misalnya analisis tentang kelulusan mahasiswa di mana $y = 1$ bermakna lulus dan $y = 0$ bermakna tidak lulus. Variabel-variabel x_j bisa berupa profil mahasiswa, total jam belajar, jam tidur, keaktivan di kelas, dan variabel-variable lain yang mungkin berhubungan dengan kelulusan. Teknik yang sering dipakai untuk memodelkan problem seperti ini diantaranya model regresi logistik.

Alur kerja praktisi statistika dalam membangun model regresi logistik biasanya dimulai dengan pengecekan asumsi, menjalankan estimasi parameter (*maximum likelihood*) dengan software, mengecek nilai parameter dan signifinasinya, dan memilih variabel yang signifikan. Untuk mengecek dan membandingkan model mana yang lebih bagus (variabel mana saja yang sebaiknya dimasukkan ke model), ukuran seperti Akaike Information Creiterion (AIC) akan digunakan. Setelah modelnya fix, praktisi statistika akan menganalisis interpretasi dari parameter yang dihasilkan, seperti variabel apa yang mempengaruhi kelulusan, apakah mempengaruhi secara positif/negatif, dan seberapa besar pengaruhnya (bisa dicek dari nilai parameter untuk variabel tersebut). Pada akhirnya akan praktisi statistika akan menarik kesimpulan tentang hubungan variabel-variabel tersebut dengan kelulusan mahasiswa.

Kita lihat contoh problem yang dihadapi praktisi *machine learning* dengan tipe yang mirip, satu variabel dependen kategorik y dan beberapa variabel independen $x_1, \dots x_m$. Contoh aplikasinya misalnya untuk mendeteksi email spam. Nilai variabel $y = 1$ berarti email spam, $y = 0$ berarti bukan email spam. Variabel-variabel x_j berisi karakteristik dari email tersebut, misalnya jumlah kata di email, apakah email mengandung attachment, apakah ada gambar di email itu, atau apakah email mengandung kata-kata tertentu, misalnya kata 'promo' atau 'obat'.

Model regresi logistik juga termasuk salah satu model yang sering dipakai praktisi machine learning. Mereka juga melakukan estimasi parameter dengan *maximum likelihood* via software. Praktek menambahkan regularisasi seperti L1 (lasso) atau L2 (ridge) ke model sangat lumrah dilakukan di *machine learning* untuk meningkatkan generalisasi dari model. Praktisi *machine learning* biasanya tidak begitu fokus ke pemilihan variabel terutama jika jumlah variabelnya tidak banyak, semua variabel masuk ke model. Fokus pemilihan dan pembandingan model lebih ke bagaimana memilih konstanta regularisasi yang tepat. Untuk melakukan itu praktisi *machine learning* membagi dataset untuk modeling (training data) ke beberapa *fold* dan melakukan teknik *cross validation*. Model yang memiliki tingkat kesalahan prediksi paling kecil di tahap *cross validation* akan dipilih. Selanjutnya, model tersebut akan disimpan, di-*deploy* ke server dan dibikin *programming interface* yang nantinya akan dipanggil sekiranya ada email baru yang perlu diprediksi apakah spam atau bukan.

Dari dua alur kerja diatas, walaupun sama-sama menggunakan model regresi logistik, terlihat jelas perbedaan fokus dati praktisi statistika dan *machine learning*. Selain dari fokus yang berbeda (interpretasi versus prediksi), ada beberapa perbedaan yang terlihat. Diantaranya adalah bagaimana memperlakukan parameter hasil estimasi. Tidak seperti praktisi statistika, praktisi *machine learning* kurang peduli tentang nilai dari parameter-parameter tersebut. Ada preferensi di mana praktisi *machine learning* ingin agar nilai parameter-parameter tersebut relatif kecil dengan menambahkan regularisasi, namun berapapun nilainya tidak menjadi permasalahan, asalkan model tersebut bisa memprediksi data baru dengan akurat. Perbedaan yang lain ada dalam cara pemilihan model. Praktisi statistika cenderung memilih model berdasarkan teori analitik seperti AIC, sedangkan praktisi *machine learning* cenderung memilih model berdasarkan performa empirikal di tahap *cross validation*.

Apa konsekuensi dari perbedaan ini? Arah kompleksitas pengembangan yang berbeda dari statistika dan *machine learning*. Langkah selanjutnya setelah pemodelan regresi logistik bagi praktisi statistika akan berusaha mendapatkan kesimpulan dan interpretasi yang lebih pas, seperti dengan menganalisis lebih lanjut sumber *variance* dari model dengan ANOVA, atau memodelkan berdasarkan distrbusi peluang yang lain dengan Generalized Linear Model (GLM). Karena pentingnya interpretasi dari model yang didapat, praktisi statistika cenderung memilih model-model linear untuk analisis. Faktanya, GLM merupakan model yang paling populer di disiplin ilmu statistika.

Karena fokus peneliti *machine learning* lebih ke akurasi prediksi, semakin banyak informasi (variabel independen) yang masuk ke model, cenderung memberikan prediksi yang lebih bagus. Untuk kasus permodelan text, seperti contoh email spam di atas, variabel-variable yang lumrah dipakai oleh praktisi *machine learning* adalah diperoleh dari teknik *bag-of-words*. Dengan teknik ini, setiap kata atau kombinasi kata yang ada di kamus menjadi satu variabel yang nilainya 1 jika variabel kata tersebut ada di email, atau 0 jika tidak ada. Hasilnya, jumlah variabel bisa ribuan bahkan jutaan. Model linear juga kurang begitu banyak dipakai oleh praktisi *machine learning* karena akurasi prediksi akan cenderung meningkat dengan model non-linear. Non-linearitas bisa didapatkan dengan banyak cara, di antaranya adalah dengan menumpuk beberapa model linear ke beberapa *layer* seperti yang dilakukan di model *Neural Network*, atau memproyeksikan variabel ke dimensi lebih tinggi seperti yang di lakukan oleh teknik kernel trick yang lumrah diterapkan di model

Support Vector Machine (SVM). Kompleksitas model tidak menjadi masalah bagi praktisi *machine learning* asalkan bisa meningkatkan akurasi prediksi dan ada *resource* komputer untuk menjalankan estimasi model.

Area-area sama-sama di dalami oleh peneliti statistika dan *machine learning*

Meskipun alur kerja praktisi statistika dan *machine learning* begitu berbeda seperti contoh di atas, mereka sama-sama menggunakan model regresi logistik dan menggunakan teori dan rumus yang sama. Faktanya, banyak sekali teori yang sama-sama dipelajari baik di bidang ilmu statistika ataupun *machine learning*. Di bagian ini kita akan melihat beberapa contohnya. Dari sisi paling dasar misalnya, statistika dan *machine learning* sama-sama membahas konsep *random variable*, distribusi-distribusi statistik, *expected value*, variansi, sampai pada konsep distribusi prior dan posterior.

Teknik inferensi model parametrik dengan *maximum likelihood estimation (MLE)* juga dipelajari di kedua ilmu, sekaligus teori-teori MLE seperti konsistensi dan *sufficient statistic*. Algoritma untuk mendapatkan estimasi MLE secara numerik seperti menggunakan metode *gradient descent* and *Quasi Newton* juga dipelajari di *machine learning* dan statistika komputasi. Teknik MLE untuk kasus di mana suatu model bergantung pada variabel yang tidak diobservasi melalui *Expectation - maximization (EM) algorithm* juga sangat populer di kalangan peneliti statistika dan *machine learning*, termasuk juga penerapan *EM algorithm* untuk estimasi mixture model.

Model-model linear seperti regresi linear, regresi logistik, dan GLM, beserta variasi regularisasi dari model-model tersebut seperti regressi lasso dan regresi ridge dipelajari di kedua disiplin ilmu. Teknik reduksi variabel secara linear seperti Principal Component Analysis (PCA) dan Independent Component Analysis (ICA) sering juga digunakan oleh praktisi kedua disiplin ilmu. Model linear Support Vector Machine (SVM) yang mempunyai karekteristik sample sparsity –dimana parameter model hanya bergantung pada sebagian kecil dari data– sangat populer di bidang *machine learning* dan juga mulai dipelajari di bidang ilmu statistika.

Konsep-konsep Bayesian statistics memegang peranan sangat penting baik di disiplin ilmu statistika dan *machine learning*. Banyak sekali model-model yang dipelajari di kedua ilmu yang menggunakan prinsip *Bayesian statistics* seperti Bayesian linear regression, Bayesian logistic regression, Bayesian GLM, Latent Dirichlet Allocation (LDA), dan beberapa Bayesian non-parametrik model seperti *Gaussian Process* dan *Dirichlet Process*. Konsep-konsep *distribution sampling* seperti *importance sampling*, Markov Chain Monte Carlo (MCMC), dan Gibbs sampling sama-sama dipelajari di statistika dan *machine learning*.

Area-area lain yang sama-sama dipelajari di bidang ilmu statistika dan *machine learning* diantaranya, *probability density estimation*, model non-parametik, analisis cluster, dan model Marckov chain.

Area penting bagi peneliti statistika yang kurang didalami peneliti *machine learning*

Fokus yang berbeda dari disiplin ilmu statistika dan *machine learning* mengakibatkan arah konsentrasi yang berbeda pula dari kedua ilmu tersebut. Berikut ini adalah contoh-contoh area yang menjadi fokus peneliti statistika namun kurang didalami oleh peneliti *machine learning*. Sebagian besar dari area-area di bawah ini sangat penting untuk menunjang keakuratan pengambilan kesimpulan dan interpretasi model namun tidak begitu penting untuk prediksi.

***Sampling* (dari populasi)**. Teori pengambilan sampel (*sampling*) sangat penting peranannya pada fase pengumpulan data sebelum nantinya diproses dan dianalisis, terutama untuk aplikasi statistika di bidang ilmu sosial. Teknik pengambilan sampel yang benar akan memberikan garansi pada ke-valid-an penarikan kesimpulan yang nantinya akan diambil saat melakukan analisis lebih lanjut. Di sisi lain, data-data yang di olah oleh praktisi *machine learning* kebanyakan berupa data transaksional dimana tidak diperlukan pengambilan sampel. Sebagai contoh, kasus kasus klasifikasi email spam. Data untuk membentuk model didapatkan dari email-email sebelumnya. Tantangan di *machine learning* biasanya adalah kebanyakan dari data yang ada tidak mengandung label (variabel dependen y). Untuk kasus ini diperlukan input dari manusia untuk memberikan label ke data, sebagai contoh, memberikan label apakah suatu email adalah spam atau bukan.

Uji hipotesis. Pengujian hipotesis adalah salah satu aspek paling penting di bidang ilmu statistika. Praktisi statistika menggunakan teknik-teknik pengujian hipotesis untuk menarik kesimpulan apakah hipotesis awal yang mereka bentuk dalam suatu permasalahan didukung oleh data ataukah tidak. Pengujian hipotesis tidak banyak didalami peneliti *machine learning* karena fokus mereka yang lebih ke prediksi daripada pengambilan keputusan.

Analisis varians (ANOVA). ANOVA dan generalisasinya (seperti MANOVA, dan MANCOVA) merupakan tekniki yang dipakai luas di bidang ilmu statistika. Teknik ini menganalisa dan membandingkan variasi dari dua grup berbeda. Sebagaimana dengan uji hipotesis, analisis varians juga kurang didalami oleh peneliti *machine learning*.

Model-model linear *advanced*. Beberapa model linear yang lebih *advanced* dikembangkan oleh peneliti statistika untuk menganalisis lebih lanjut model yang lebih kompleks. Sebagai contoh adalah *path analysis* dimana interaksi antar variabel dideskripsikan dalam bentuk graph. Contoh lainnya adalah survival analysis yang memodelkan rata-rata waktu sebelum suatu *event* akan terjadi. Peneliti statistika juga mengembangkan model linear yang lebih kompleks untuk menganalisis kasus-kasus dimana asumsi standar dari model yang sederhana tidak terpenuhi. Sebagai contoh adalah two stages least square regression, dan generalized estimating equation (GEE).

Area penting bagi peneliti *machine learning* yang kurang didalami peneliti statistika

Di sisi lain, banyak juga area-area yang menjadi fokus di disiplin ilmu *machine learning*, namun kurang didalami oleh peneliti statistika. Area-area tersebut sebagian besar berguna untuk meningkatkan akurasi dari prediksi namun mengakibatkan model yang dibentuk menjadi kompleks dan susah untuk di-interpretasikan. Berikut ini beberapa contoh diantaranya.

Kernel trick. Teknik ini terkait dengan proyeksi variabel. Berbalikan dengan teknik-teknik proyeksi variabel ke dimensi lebih rendah yang populer di statistika dan *machine learning* seperti melalui PCA, *kernel trick* berkaitan dengan proyeksi variabel ke dimensi lebih tinggi. *Kernel trick* memungkinkan sebuah model linear seperti regresi logistik atau SVM untuk mendapakkan non-linearitas dengan secara tidak langsung memproyeksikan variabel ke dimensi yang lebih tinggi (bahkan dimensi tak terhingga) tanpa harus melakukan transformasi variabel secara eksplisit.

Neural networks* dan *deep learning. Cara lain untuk mendapatkan non-linearitas adalah dengan menumpuk beberapa model linear ke beberapa layer. Teknik inilah yang dilakukan oleh *neural network*. Teknik *deep learning* yang sangat populer saat ini menggunakan banyak layer yang dimana di setiap layer berfungsi untuk membentuk representasi menengah yang lebih *compact* dari data. Beberapa teknik tambahan seperti *parameter sharing*, *convolution* dan *recurrence* berperan untuk menambah keakuratan dari representasi tersebut dan menunjang kesuksesan applikasi *deep learning* di berbagai area seperti *computer vision (CV)* dan *natural language processing (NLP)*.

Inferensi *semi-supervised*. Di beberapa kasus, terkadang data yang tersedia sangat banyak, namun hanya sebagian kecil dari data tersebut memiliki label. Untuk mendapatkan label membutuhkan ahli yang biayanya mahal. Sebagai contoh adalah kasus untuk memprediksi apakah suatu review di website e-commerce asli dari pembeli ataukah palsu (misalnya komentar bayaran). Data untuk variabel independen (x) sangat banyak tersedia yang bisa didapatkan dari komentar-komentar yang sudah ada. Namun untuk mendapatkan label (y), perlu seorang ahli yang membaca dan menganalisis review untuk menentukan review tersebut palsu atau tidak. Teknik inferensi *semi-supervised* mencoba untuk mengikutsertakan data-data tanpa label ke dalam pembentukan model.

Probabilistic graphical models (PGM). PGM memodelkan *conditional dependency* dari koleksi beberapa *random variable*. PGM banyak dipakai oleh peneliti *machine learning* untuk menangaini *structured prediction*, kasus dimana model tidak hanya melakukan prediksi satu variabel y, tetapi melakukan prediksi vektor $\mathbf{y}$ yang mempunyai struktur. Sebagai contoh adalah kasus aplikasi di area NLP dimana kita diberikan satu kalimat sebagai isi dari variabel x dan harus memprediksi tiap kata di kalimat itu apakah menjadi subjek, predikat, objek, ataukah keterangan. Beberapa model PGM yang populer diantaranya adalah hidden Markov model (HMM), conditional random field (CRF), dan latent Dirichlet allocation (LDA).

Cara berfikir matematis vs. algoritmis

Peneliti disiplin ilmu statistika kebanyakan berada di *Department of Statistics* atau *Department of Mathematics and Statistics*, sedangkan peneliti *machine learning* kebanyakan berada di *Department of Computer Science*. Peneliti statistika rata-rata memulai belajar dari latar belakang matematika, sedangkan peneliti *machine learning* rata-rata memulai dari latar belakang algoritma.

Dalam menghadapi suatu masalah, peneliti statistika lebih melihat dari sisi formulasi matematika untuk memodelkan masalah. Peneliti *machine learning*, selain memperhitungkan model matematika, rata-rata juga memikirkan bagaimana peforma dari algoritma yang akan digunakan untuk mengestimasi model. Performa dari algoritma biasanya di ukur dengan big O notation. Sebagian besar publikasi riset peneliti *machine learning* mendeskripsikan algoritma estimasi model beserta performa/kompleksitas dari algoritma tersebut. Sebagai contoh adalah algoritma inferensi model *latent dirichlet allocation (LDA)* (Blei et al., 2003) yang memiliki kompleksitas $O(N^2k)$, dimana N adalah jumlah sampel dan k adalah jumlah variabel. Artinya kurang lebih adalah, jika jumlah sampel bertambah menjadi dua kali lipat, algoritma tersebut akan berjalan 4 kali lebih lambat dari sebelumnya; dan jika jumlah variabel bertambah dua kali lipat, algoritma akan menjadi 2 kali lebih lambat juga.

Analisis komplksitas algoritma ini sangat penting untuk melihat apakah algoritma itu akan cocok diimplementasikan ke data yang lebih besar. Algoritma inferens yang memiliki kompleksitas $O(N)$ tentunya akan lebih

dipilih daripada algorima dengan kompleksitas $O(N^2)$. Hal ini dikarenakan jika jumlah sampel bertambah menjadi seribu kali lipat misalnya, algoritma dengan kompleksitas $O(N)$ hanya akan menjadi lebih lambat seribu kali lipat juga, semisal yang awalnya memakan waktu 1 detik menjadi 1000 detik atau sekitar 17 menit. Di sisi lain, algoritma dengan kompleksitas $O(N^2)$ menjadi satu juta kali lebih lambat, dari 1 detik menjadi 1 juta detik atau sekitar 12 hari. Sangat terasa perbedaannya.

Kultur jurnal vs. konferensi

Selain dari perbedaan cara berfikir, terdapat perbedaan kultur dalam proses publikasi hasil riset peneliti statistika dan *machine learning*. Tempat utama untuk publikasi pengambangan riset statistika adalah di jurnal, seperti *Annals of Statistics*, *Biometrika*, dan *Journal of the American Statistical Association (JASA)*. Disisi lain, sebagai mana area riset lain di bawah *Department of Computer Science*, tempat utama publikasi riset oleh peneliti *machine learning* adalah di konferensi. Publikasi di konferensi-konferensi utama seperti *Neural Information Processing Systems (NIPS)* dan International Conference on Machine Learning (ICML) dan konefensi-konferensi lain (UAI, AISTATS, COLT) menjadi fokus utama peneliti *machine learning*. Dampaknya adalah progress pengembangan *machine learning* berjalan lebih cepat dari statistika, karena alur pemrosesan publikasi mulai dari *submission* sampai penerbitan hanya memakan waktu beberapa bulan untuk konferensi, dibandingkan jurnal yang bisa memakan waktu beberapa tahun. Peneliti *machine learning* dituntut untuk mengejar deadline konferensi untuk menuangkan ide baru mereka, sebelum orang lain yang punya ide mirip mempublikasikannya. Jika ide yang sudah dipublikasikan di konferensi perlu penjelasan lebih detail, jurnal *machine learning* seperti *Journal of Machine Learning Research (JMLR)* adalah tempatnya.

Bahasa pemrograman

Bahasa pemrograman yang populer di kalangan peneliti statistika adalah R. Hampir semua peneliti statistika menggunakan R untuk menuangkan ide dan teori mereka. Untuk praktisi statistika, selain R, bahasa yang pupuler diantaranya adalah SAS, Stata, dan Python.

Di kalangan peneliti dan praktisi *machine learning*, Python menjadi bahasa yang paling populer, disusul oleh C++ untuk implementasi model yang membutuhkan performa tinggi, MATLAB untuk implementasi ide secara cepat, dan Lua yang populer di praksisi *deep learning*. Bahasa *R* juga populer di kalangan praktisi *machine learning*.

Sebagian kecil dari peneliti dan praktisi statistika dan *machine learning* menggunakan bahasa pemrograman Julia yang memungkinkan implementasi ide secara cepat tanpa harus mengorbankan performa.

Penutup

Seperti yang telah dibahas di bagian-bagian sebelumnya, disiplin ilmu statistika dan *machine learning* mempunyai banyak persamaan dan juga perbedaan. Kedua disiplin ilmu sama-sama berdasarkan teori peluang dan membahas dasar-dasar teori dan model yang sama. Perbedaan keduanya terletak pada fokus yang berbeda. Statistika lebih fokus ke arah pengambilan kesimpulan, sedangkan *machine learning* fokus ke prediksi data baru. Dari persamaan dan perbedaan tersebut, tidak salah kalau statistika dan *machine learning* disebut sebagai dua wajah berbeda dari satu kesatuan disiplin ilmu.

Referensi

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Breiman, Leo et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

Harrell, Frank. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Wasserman, Larry. Rise of the machines.

Wasserman, Larry. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Beberapa artikel di wikipedia.