BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topic 12

TEKNIK LAIN DATA SAINS

LEARNING OUTCOMES

Setelah mempelajari materi ini peserta kuliah diharapkan mampu melakukan penalaran/ penarikan kesimpulan berdasar / melalui :

- Association and co-occurrences;
- Behavior profiling;
- Link prediction;
- Data reduction;
- Latent information mining;
- Movie recommendation;
- Bias-variance decomposition of error;
- Ensembles of models; Causal reasoning from data.

OUTLINE MATERI :

1. Co-occurrences dan Asosiasi
2. Mengukur Surprise: Lift and Leverage
3. Contoh contoh kasus
4. Profiling: Menemukan perilaku tipikal
5. Link Prediction dan Social Recommendation
6. Bias, Variance, and Ensemble Methods
7. Kesimpulan

PENDAHULUAN

Masalah bisnis mendefinisikan tujuan sekaligus konstrain pada solusi. Data dan pemahaman atas domain menyediakan bahan baku dan ilmu data menyediakan kerangka kerja untuk menguraikan masalah menjadi subproblem, serta tool dan Teknik penyelesaian. Prinsip-prinsip dasar yang telah dibahas menjadi fundasi bagi sebagian besar ilmu data.

Seperti halnya masalah rekayasa lainnya, seringkali lebih efisien untuk merumuskan masalah baru ke dalam masalah yang sudah dikenal yang telah memiliki tool penyelesaian, daripada membangun solusi baru dari awal. Ilmu data menyediakan banyak tool untuk menyelesaikan persoalan tertentu. Telah dibahas sebelumnya metode untuk menemukan korelasi / menemukan variabel informatif, menemukan entitas serupa, klasifikasi, estimasi kelas probabilitas, regresi, dan pengelompokan.

Dlam data sains tool tool ini merupakan tool yang paling sering dijumpai, namun masih banyak terdapat tool lain yang belum pernah dibahas, dimana konsep dasar yang sama yang yang telah dipelajari juga menjadi dasar bagi tool tool ini.

Co-occurrences dan Asosiasi

Tinjau aplikasi yang membantu operasi pengiriman produk ke pelanggan online dari banyak pusat distribusi di seluruh dunia. Tidak setiap pusat distribusi menyimpan setiap produk. Pusat distribusi regional yang lebih kecil biasanya hanya menyimpan produk yang lebih sering dibeli oleh konsumer di sekitarnya. Pusat distribusi regional dibangun untuk mengurangi biaya pengiriman. Namun, sering sekali terjadi bahwa untuk banyak pesanan harus dilakukan pengiriman dari pusat distribusi utama. Ini dapat terjadi karena ketika orang memesan barang-barang populer, mereka sering menyertakan barang-barang yang kurang populer. Ini adalah masalah bisnis yang dapat dicoba untuk diatasi oleh Teknik asosiasi. Jika dapat dideteksi bahwa barang-barang yang kurang populer sering dipesan bersama barang-barang yang paling populer, maka barang barang ini harus pula disimpan di pusat distribusi regional, sehingga dapat dilakukan penghematan besar dalam biaya pengiriman.

Pengelompokan co-occurence adalah pencarian melalui data untuk kombinasi item yang statistiknya "menarik." Terdapat beberapa cara berbeda untuk menyelesaikan tugas ini. Co-occurrence dapat dipandang sebagai aturan: "Jika A terjadi maka B kemungkinan akan terjadi juga.

Persoalan ini perlu dicermati pertama tama menyangkut kontrol kompleksitas: ada kemungkinan terdapat sejumlah besar kookurren, banyak di antaranya mungkin hanya karena kebetulan, daripada ke pola yang dapat digeneralisasikan. Cara mudah untuk mengendalikan kompleksitas ini adalah dengan menentukan aturan / batasan bahwa peristiwa tersebut harus terjadi dengan persentase minimum— katakanlah 0,01% dari semua transaksi. Presentasi minimum ini disebut *support*.

Istilah "likely/kemungkinan" adalah istilah yang sering ditemukan dalam asosiasi. Jika seorang pelanggan membeli A maka dia kemungkinan akan membeli B. Gagasan ini dapat ditinjau dari sisi probabilitas bahwa B terjadi ketika A terjadi; p (B | A). Besaran ini disebut kepercayaan atau kekuatan (strength) aturan. Strength atau kekuatan perlu pula ditentukan ambang batas, seperti 5% (artinya 5% atau lebih dari kejadian pembeli barang A juga membeli B).

Mengukur Surprise: Lift and Leverage

Biasanya diinginkan bahwa dalam hubungan asosiasi terjadi *surprise*. Sebuah asosiasi disebut surprise jika hubungan itu bertentangan dengan sesuatu yang sudah kita ketahui atau yakini.

Gagasan intuitif untuk mengukur *surprise*: disebut *lift*. Gagasannya menjawab pertanyaan seberapa seringkah asosiasi ini terjadi dibanding kejadian secara kebetulan? Kita akan lebih surprise jika kita menemukan asosiasi yang terjadi lebih sering daripada yang disebabkan oleh suatu kebetulan. *Lift* dihitung hanya dengan menerapkan gagasan dasar probabilitas.

$$\text{Lift}(A,B) = \frac{p(A,B)}{p(A)\,p(B)}$$

Persamaan 12-1. Lift

Ini hanya satu cara yang mungkin untuk menghitung seberapa besar kemungkinan dari suatu hubungan yang ditemukan. Alternatifnya adalah dengan melihat perbedaan kuantitas ini daripada rasio. Ukuran ini disebut *leverage*.

$$\text{Leverage}(A, B) = p(B, A) - p(A)\,p(B)$$

Persamaan 12-2. Leverage

Contoh Mie dan Telur

Asosiasi sering digunakan dalam analisis keranjang belanja untuk menemukan dan menganalisis co-ocurrence barang yang dibeli. Perhatikan contoh konkret.

Misalkan kita mengoperasikan toko kecil di mana orang membeli beras, aqua, telur, sabun dll. Pertama, mari kita nyatakan aturan asosiasi yang mewakili kita yakini dari data histori: "Pelanggan yang membeli mie juga cenderung membeli telur"; atau lebih singkat, “mie ⇒ telur.” Selanjutnya, mari kita hitung *lift* asosiasi ini.

Kita sudah tahu p (mie) = 0,3. p (telur) = 0,4. Jika dua item ini sama sekali tidak terkait (independen), kemungkinan bahwa mereka akan dibeli bersama-samaadalah hasil kali dari keduanya: p (mie) × p (telur) = 0,12.

Sebagaimana disebutkan di atas, 20% dari transaksi termasuk keduanya, dan ini adalah probabilitas: p (mie, telur) = 0,2. Jadi *lift*= 0,2 / 0,12, yaitu sekitar 1,67. Ini berarti bahwa membeli telur dan mie bersama-sama sekitar 1,67 kali lebih mungkin daripada yang diharapkan secara kebetulan.

Leverage =p (mie, telur) - p (mie) × p (telur), yaitu 0,2 - 0,12, atau 0,08.

Ada dua statistik penting lainnya yang harus juga dihitung yaitu *support* dan *strength*. Support adalah prevalensi dalam data pembelian dua barang bersama, p (mie, telur), yaitu 20%. *Strength* adalah probabilitas bersyarat, p (telur | mie), yaitu 67%.

Facebook “Likes”

Meskipun menemukan asosiasi sering digunakan pada data keranjang belanja — dan kadang-kadang disebut analisis keranjang belanja — tekniknya jauh lebih umum. Kembali ke data tentang “Like" oleh sejumlah pengguna Facebook. Dengan menganalogikan pada data keranjang belanja, kita dapat memandang masing-masing pengguna memiliki "keranjang" berisi “like”, dengan menggabungkan semua “like” dari setiap pengguna. Pertanyaannya adalah apakah frekwensi “like” tertentu terjadi lebih sering daripada secara kebetulan? Ini adalah sebagai contoh yang menarik untuk mengilustrasikan temuan asosiasi, tetapi prosesnya sebenarnya bisa memiliki aplikasi bisnis yang penting. Jika Anda seorang pemasar yang ingin memahami konsumen di pasar tertentu, Anda mungkin tertarik untuk menemukan pola dari hal-hal yang disukai orang. Jika Anda memikirkan data secara analitis, Anda akan menerapkan secara persis jenis pemikiran yang telah diilustrasikan sejauh ini: Anda ingin mengetahui hal-hal apa yang terjadi bersama secara lebih sistematis daripada yang diharapkan secara kebetulan.

Asosiasi “like” ditemukan menggunakan sistem penambangan asosiasi populer Magnum Opus. Magnum Opus memungkinkan mencari asosiasi yang memberikan daya angkat tertinggi atau daya ungkit tertinggi, sementara mengenyampingkan asosiasi yang mencakup terlalu sedikit kasus.

Family Guy & The Daily Show -> The Colbert Report Support=0.010; Strength=0.793; Lift=31.32; Leverage=0.0099

Spirited Away -> Howl's Moving Castle
Support=0.011; Strength=0.556; Lift=30.57; Leverage=0.0108

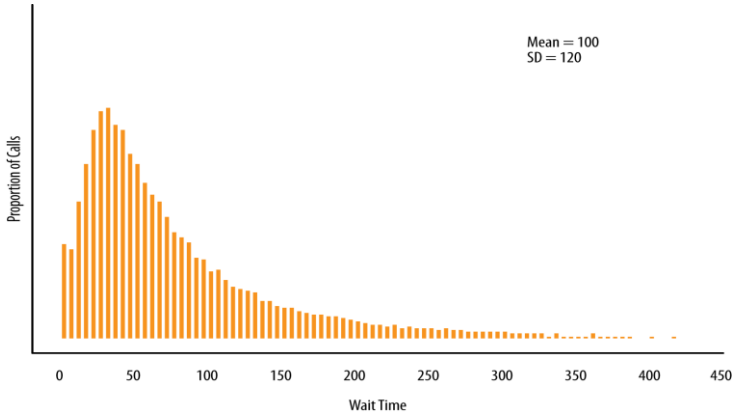
Selena Gomez -> Demi Lovato
Support=0.010; Strength=0.419; Lift=27.59; Leverage=0.0100

Profiling: Menemukan perilaku tipikal

Profiling digunakan untuk mengkarakterisasi perilaku khas dari individu, kelompok, atau populasi. Contoh pertanyaan profiling: Bagaimanakah gaya penggunaan kartu kredit dari segmen pelanggan tertentu? Jawabnnya mungkin rata-rata pengeluaran, tetapi deskripsi sederhana semacam itu mungkin tidak mewakili perilaku dengan baik untuk keperluan kita. Sebagai contoh, deteksi penipuan (fraud) sering menggunakan profiling untuk mengkarakterisasi perilaku normal dan kemudian mencari contoh yang menyimpang secara substansial yang sebelumnya telah menunjukkan penipuan. Membuat profil penggunaan kartu kredit untuk mendeteksi penipuan mungkin memerlukan deskripsi yang kompleks mungkin tentang rata-rata penggunaan di hari kerja dan akhir pekan, pembelian internasional, dan sebagainya. Perilaku dapat dijelaskan secara umum di seluruh populasi, pada tingkat kelompok kecil, atau bahkan untuk setiap individu. Misalnya, setiap pengguna kartu kredit mungkin diprofilkan sehubungan dengan penggunaan internasionalnya.

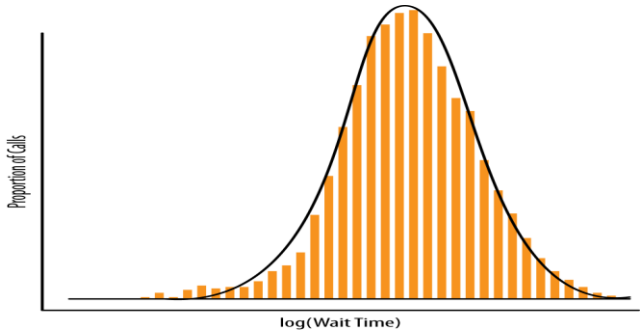
Perhatikan kasus berikut: asumsikan bahwa waktu tunggu pelanggan mengikuti distribusi Normal atau Gaussian. Ini adalah "profil" dari waktu tunggu yang (dalam hal ini) hanya memiliki dua parameter penting: mean dan standar deviasi. Ketika kita menghitung rata-rata dan standar deviasi, kita menemukan profil "terbaik" atau model waktu tunggu dengan asumsi bahwa itu terdistribusi secara normal. Dalam hal ini "terbaik" adalah gagasan yang sama yang kita diskusikan untuk regresi logistik, misalnya, mean yang kita hitung dari pembelanjaan memberi kita mean distribusi Gaussian yang paling mungkin menghasilkan data ("kemungkinan maksimum") model).

Pandangan ini mengilustrasikan mengapa perspektif ilmu data dapat membantu bahkan dalam skenario sederhana: jauh lebih jelas sekarang apa yang kita lakukan ketika kita menghitung rata-rata dan standar deviasi. Kita perlu mempertimbangkan dengan hati-hati apa yang kita inginkan dari hasil sains data kita. Dalam kasus ini kita ingin menampilkan waktu tunggu "normal" pelanggan. Jika kita memplot data dan mereka tidak terlihat seperti mereka berasal dari Gaussian (kurva lonceng simetris), kita mungkin ingin mempertimbangkan kembali hanya melaporkan rata-rata dan standar deviasi.



Gambar 12-1. Distribusi Waktu Tunggu nasabah suatu bank ke bank's call center.

Untuk memperlihatkan lebih dalam apa yang mungkin dilakukan oleh manajer yang paham tentang data, mari kita melangkah lebih jauh. Kita tidak akan membahas detailnya di sini, tetapi trik umum untuk menangani data yang miring dengan cara ini adalah dengan mengambil logaritma (log) dari waktu tunggu. Gambar 12-2 menunjukkan distribusi yang sama seperti Gambar 12-1, kecuali menggunakan logaritma dari waktu tunggu. Kita sekarang melihat bahwa setelah transformasi yang sederhana, waktu tunggu sangat mirip dengan kurva lonceng klasik.



Gambar 12-2. Distribusi waktu tunggu setelah melakuan pendefinsian kembali data.

Gambar 12-2 juga menunjukkan distribusi Gaussian aktual (kurva lonceng) yang sesuai dengan distribusi berbentuk lonceng, seperti yang dijelaskan di atas. Ini sangat cocok, dan dengan demikian kita memiliki justifikasi untuk melaporkan rata-rata dan standar deviasi sebagai ringkasan statistik dari profil (log) waktu tunggu.

Link Prediction dan Social Recommendation

Kadang-kadang, alih-alih memprediksi properti (nilai target) dari suatu item data, lebih berguna untuk memprediksi hubungan antara item data. Contoh umum dari ini adalah memprediksi bahwa sebuah tautan harus ada di antara dua individu. Prediksi tautan umum dalam sistem jejaring sosial: Karena Anda dan Karen berbagi 10 teman, mungkin Anda ingin menjadi teman Karen? Prediksi tautan juga dapat memperkirakan kekuatan tautan. Misalnya, untuk merekomendasikan film kepada pelanggan, seseorang dapat memikirkan grafik antara pelanggan dan film yang telah mereka tonton atau nilai. Dalam grafik, kami mencari tautan yang tidak ada antara pelanggan dan film, tetapi yang kami prediksi harus ada dan harus kuat. Tautan-tautan ini membentuk dasar untuk rekomendasi.

Data Reduction, Latent Information, dan Movie Recommendation

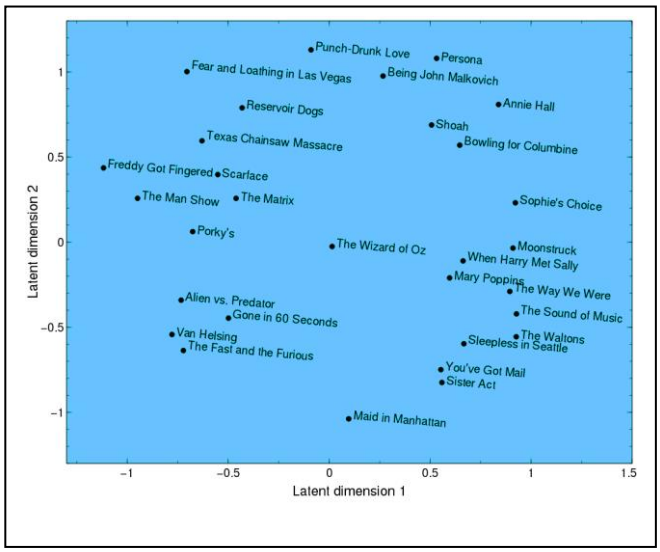
Untuk beberapa masalah bisnis, kami ingin mengambil sejumlah besar data dan menggantinya dengan set yang lebih kecil yang menyimpan banyak informasi penting dalam kumpulan yang lebih besar. Pengurangan data tersebut biasanya melibatkan pengorbanan beberapa informasi, tetapi yang penting adalah trade-off antara wawasan atau pengelolaan yang diperoleh terhadap informasi yang hilang. Ini sering kali merupakan nilai jual.

Perusahaan penyewaan film Netflix ™ menawarkan satu juta dolar kepada individu atau tim yang dapat memprediksi bagaimana konsumen menilai film. Tim pemenang menghasilkan teknik yang sangat rumit, tetapi sebagian besar keberhasilannya dikaitkan dengan dua aspek solusi: (i) penggunaan ensemble model, (ii) reduksi data.

Masalah yang harus diselesaikan pada dasarnya adalah masalah prediksi tautan, di mana secara khusus kami ingin memprediksi kekuatan tautan antara pengguna dan film — kekuatan yang mewakili seberapa besar pengguna akan menyukainya.

Salah satu pendekatan paling populer untuk memberikan rekomendasi adalah dengan mendasarkan model pada dimensi laten yang mendasari preferensi. Istilah "laten," dalam ilmu data, berarti "relevan tetapi tidak diamati secara eksplisit dalam data. Dimensi laten preferensi film termasuk kemungkinan penokohan seperti serius versus eskapis, komedi versus drama, orientasi terhadap anak-anak, atau orientasi jender. Bahkan jika ini tidak diwakili secara eksplisit

dalam data, mereka mungkin penting untuk menilai apakah pengguna tertentu akan menyukai film. Dimensi laten juga dapat mencakup hal-hal yang mungkin tidak jelas seperti kedalaman pengembangan karakter atau quirkiness, serta dimensi yang tidak pernah secara eksplisit diartikulasikan, karena dimensi laten akan muncul dari data.



Gambar 12-5. Kumpulan film yang ditempatkan dalam "ruang selera" yang ditentukan oleh dua dimensi laten terkuat yang ditambang dari data Tantangan Netflix. Lihat teks untuk diskusi terperinci. Pelanggan juga akan ditempatkan di suatu tempat di ruang angkasa, berdasarkan film yang sebelumnya telah dilihat atau dinilai. Pendekatan rekomendasi berbasis kesamaan akan menyarankan film terdekat kepada pelanggan sebagai rekomendasi kandidat.

Gambar 12-5 menunjukkan ruang laten dua dimensi sebenarnya ditambang dari data film Netflix, 8 serta koleksi film yang diwakili dalam ruang baru ini. Interpretasi dimensi laten seperti itu yang ditambang dari data harus disimpulkan oleh para ilmuwan data atau pengguna bisnis. Cara yang paling umum adalah mengamati bagaimana dimensi memisahkan film, lalu menerapkan pengetahuan domain.

Pada Gambar 12-5, dimensi laten yang diwakili oleh sumbu horizontal tampaknya memisahkan film-film menjadi film-film drama berorientasi pada film-film berorientasi kanan dan aksi di sebelah kiri.

Bias, Variance, and Ensemble Methods

Ensemble telah diamati untuk meningkatkan kinerja generalisasi dalam banyak situasi - tidak hanya untuk rekomendasi, tetapi secara luas di seluruh klasifikasi, regresi, estimasi probabilitas kelas, dan banyak lagi.

Salah satu cara untuk memahami mengapa ansambel bekerja adalah memahami bahwa kesalahan yang dibuat model dapat dicirikan oleh tiga faktor:

1. Inherent randomness,
2. Bias, and
3. Variance.

Data-Driven Causal Explanation and a Viral Marketing Example

Satu topik penting adalah penjelasan kausal dari data. Pemodelan prediktif sangat berguna untuk banyak masalah bisnis. Namun, jenis pemodelan prediktif yang telah kita bahas sejauh ini didasarkan pada korelasi daripada pengetahuan tentang sebab-akibat. Kita sering ingin melihat lebih dalam fenomena dan bertanya apa yang memengaruhi apa. Kami mungkin ingin melakukan ini hanya untuk memahami bisnis kami dengan lebih baik, atau kami mungkin ingin menggunakan data untuk meningkatkan keputusan tentang cara melakukan intervensi untuk menyebabkan hasil yang diinginkan.

Pertimbangkan contoh terperinci; "pemasaran viral. Salah satu penafsiran umum dari viral marketing adalah bahwa konsumen dapat dibantu untuk mempengaruhi satu sama lain untuk membeli suatu produk, dan dengan demikian seorang pemasar dapat memperoleh manfaat yang signifikan dengan "menyemai" konsumen tertentu (misalnya, dengan memberi mereka produk secara gratis), dan mereka kemudian akan menjadi "influencer" - mereka akan menyebabkan peningkatan kemungkinan bahwa orang yang mereka kenal akan membeli produk.

Analisis data yang naif bisa sangat menyesatkan. Ada berbagai metode untuk penjelasan kausal yang hati-hati dari data, dan mereka semua dapat dipahami dalam kerangka ilmu data umum. Analisis data kausal yang cermat membutuhkan pemahaman tentang investasi dalam memperoleh data, pengukuran kesamaan, penghitungan nilai yang diharapkan, korelasi dan penemuan variabel informatif, persamaan pemasangan data, dan banyak lagi.

KESIMPULAN

Ada banyak teknik khusus yang digunakan dalam ilmu data. Untuk mencapai pemahaman yang kuat tentang bidang ini, penting untuk mundur dari hal-hal spesifik dan berpikir tentang jenis tugas yang menerapkan teknik tersebut. Dalam buku ini, kami telah memfokuskan pada kumpulan tugas yang paling umum (menemukan korelasi dan atribut informatif, menemukan item data yang serupa, klasifikasi, perkiraan probabilitas, regresi, pengelompokan), menunjukkan bahwa konsep ilmu data memberikan landasan yang kuat untuk memahami baik tugas maupun metode untuk menyelesaikan tugas. Dalam bab ini, kami mempresentasikan beberapa tugas dan teknik sains data penting lainnya, dan mengilustrasikan bahwa mereka juga dapat dipahami berdasarkan fondasi yang disediakan oleh konsep fundamental kami.

Secara khusus, kami membahas: menemukan co-kejadian atau asosiasi yang menarik di antara barang-barang, seperti pembelian; profil perilaku khas, seperti penggunaan kartu kredit atau waktu tunggu pelanggan; memprediksi hubungan antara item data, seperti koneksi sosial potensial antara orang-orang; mengurangi data kami untuk membuatnya lebih mudah dikelola atau untuk mengungkapkan informasi tersembunyi, seperti preferensi film laten; menggabungkan model seolah-olah mereka ahli dengan keahlian yang berbeda, misalnya untuk meningkatkan rekomendasi film; dan menarik kesimpulan kausal dari data, seperti apakah dan sejauh mana fakta bahwa orang yang terhubung secara sosial membeli produk yang sama sebenarnya karena mereka saling mempengaruhi (penting untuk kampanye viral), atau hanya karena orang yang terhubung secara sosial memiliki selera yang sangat mirip (yang dikenal dalam sosiologi). Pemahaman yang kuat tentang prinsip-prinsip dasar membantu Anda untuk memahami teknik yang lebih kompleks sebagai contoh atau kombinasi dari mereka.

DAFTAR PUSTAKA

1. Foster Provost & Tom Fawcett (2013) Data Science for Business: What you need to know about data mining and data analytic thinking, O'Reilly, ISBN: 978-1-449-36132-7.
2. Sharda, R., Delen, D., Turban, E., (2018). Business intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.