BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topik 4

Model Prediktif

LEARNING OUTCOMES

Setelah mempelajari materi ini peserta kuliah diharapkan mampu:

- Identifying informative attributes;
- Segmenting data by progressive attribute selection

OUTLINE MATERI:

1. Model, Induksi dan dan Prediksi
2. Supervised Segmentation
3. Pohon dan Aturan Segmentasi
4. Kesimpulan

ISI MATERI

Model, Induksi dan dan Prediksi

Dalam data sains, model prediktif adalah rumus untuk memperkirakan nilai fitur yang tidak diketahui: yang disebut target. Rumusnya bisa matematis, atau bisa jadi pernyataan logika sebagai aturan (rule). Seringkali hibrida dari keduanya.

Hal ini berbeda dengan pemodelan deskriptif, di mana tujuan utama dari model ini bukan untuk memperkirakan nilai tetapi untuk mendapatkan wawasan tentang fenomena atau proses yang mendasarinya. Model deskriptif perilaku churn akan memberi tahu kita seperti apa pelanggan yang biasanya cenderung churn. Perbedaan antara jenis model ini tidak seketat yang diperkirakan; beberapa teknik yang sama dapat digunakan untuk keduanya, dan biasanya satu model dapat melayani kedua tujuan tersebut.

Pembelajaran yang diawasi (supervised learning) adalah pembuatan model di mana model menggambarkan hubungan antara sekumpulan variabel yang dipilih (atribut atau fitur) dengan variabel target. Model ini memperkirakan nilai dari variabel target sebagai fungsi (mungkin fungsi probabilistik) dari fitur.

Pemodelan dari data dikenal sebagai model induksi. Model adalah aturan umum dalam arti statistik (biasanya tidak menjamin 100% benar), dan prosedur yang membuat model dari data disebut algoritma induksi atau pembelajaran.

Induksi dapat dikontraskan dengan deduksi. Deduksi dimulai dengan aturan umum dan fakta spesifik, dan menurunkan fakta spesifik lainnya. Penggunaan model dapat dianggap sebagai proses (probabilistik) deduksi.

Data input untuk algoritma induksi, digunakan untuk menghasilkan model, data ini disebut data pelatihan atau “data training”. Sering disebut data berlabel karena nilai untuk variabel target (label) diketahui.

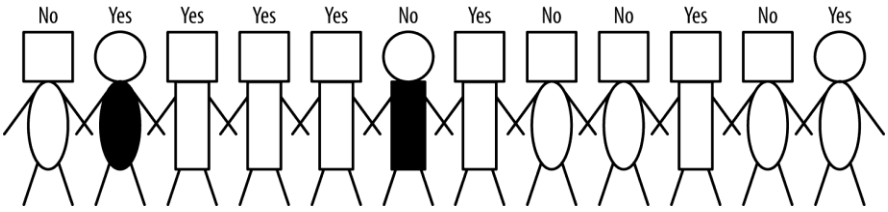
Bagaimana memilih satu atau lebih atribut / fitur / variabel yang akan membagi sampel dengan baik sesuai dengan variabel target yang kita minati?

Supervised Segmentation

Model prediktif berfokus pada memperkirakan nilai dari variabel target tertentu. Cara berpikir intuitif dalam mengekstraksi pola dari data dengan cara yang diawasi adalah mencoba mensegmentasikan populasi ke dalam subkelompok yang memiliki nilai variable target berbeda (sedangkan dalam subkelompok yang sama, sampel memiliki nilai variable target yang sama).

Ini membawa kita ke konsep fundamental: bagaimana kita bisa menilai apakah suatu variabel mengandung informasi penting tentang variabel target? Berapa banyak? Kita ingin secara otomatis mendapatkan pilihan variabel yang lebih informatif sehubungan dengan tugas tertentu (yaitu, memprediksi nilai variabel target). Bahkan lebih dari itu, kita mungkin ingin memberi peringkat variabel berdasarkan seberapa baik variable tersebut memprediksi nilai target.

Selecting Informative Attributes



Gambar 3.1 memperlihatkan masalah segmentasi sederhana: terdapat dua belas orang yang memiliki dua jenis kepala: persegi dan lingkaran; dan dua jenis tubuh: persegi panjang dan oval; dan dua orang memiliki tubuh abu-abu sementara yang lainnya berwarna putih.

Ini adalah atribut yang akan kita gunakan untuk menggambarkan setiap obyek dalam gambar. Di atas kepala setiap orang terdapat label target biner, Ya atau Tidak, yang menunjukkan (misalnya) apakah orang tersebut berpotensi menjadi peminjam yang gagal tau tidak. Atribut yang terkait dalam masalah ini dapat diringkas sebagai:

Atribut:

— Bentuk Kepala: persegi, lingkaran

— Bentuk tubuh: Persegi, oval

— Warna Tubuh: abu abu, putih

Target variable:

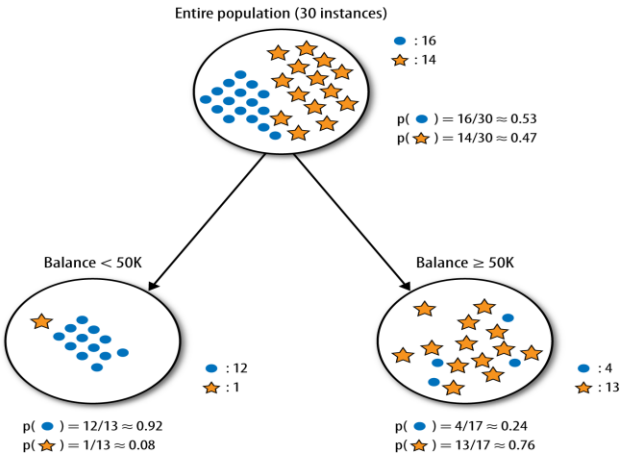
— Kredit Gagal: Ya, Tidak

Atribut manakah yang paling tepat untuk melakukan segmentasi obyek obyek ini ke dalam kelompok, yang membedakan kredit gagal (write-off) dari kredit lancar (non-write-off)? Kita ingin kelompok yang dihasilkan menjadi semurni mungkin. Dimana semakin murni semakin homogen dalam hubungan dengan variabel target. Jika setiap anggota grup memiliki nilai yang sama untuk target, maka grup tersebut murni. Jika terdapat setidaknya satu anggota saja dalam grup yang sama yang memiliki nilai variable target berbeda, maka grup tersebut tidaklah murni.

Kriteria pemisahan yang paling umum dipakai adalah konsep "information gain". Konsep ini didasarkan pada ukuran kemurnian yang disebut entropi. Kedua konsep diciptakan oleh salah satu pelopor teori informasi, Claude Shannon, dalam karya seminalnya di bidang ini (Shannon, 1948).

Entropi adalah ukuran ketidakteraturan yang dapat diterapkan ke satu set, seperti salah satu segmen kita. Perhatikan bahwa kita memiliki seperangkat properti dari anggota set, dan setiap anggota memiliki satu dan hanya satu properti. Dalam segmentasi yang diawasi, properti anggota akan berhubungan dengan nilai dari variabel target. Ketidak teraturan (disorder) berhubungan dengan bagaimana kemurnian segmen tersebut berkenaan dengan properti yang menjadi perhatian. Jadi, suatu segmen yang anggotanya terdiri dari banyak write-off dan non-write-off akan memiliki entropi yang tinggi.

Dalam konteks pemodelan prediktif, jika kita ingin mengetahui nilai atribut target, berapa banyak suatu atribut meningkatkan pengetahuan kita tentang nilai variabel target? Ini adalah konsep dasar mengenai "Information Gain" (IG).



Sebagai contoh, perhatikan pemisahan dalam Gambar 3-2 di atas. Ini adalah masalah dua kelas (• dan ★). Himpunan turunan terlihat "lebih murni" daripada induknya.

Jadi pemisahan ini mengurangi entropi secara substansial. Dalam istilah pemodelan prediktif, atribut "balance" menyediakan banyak informasi tentang nilai target.

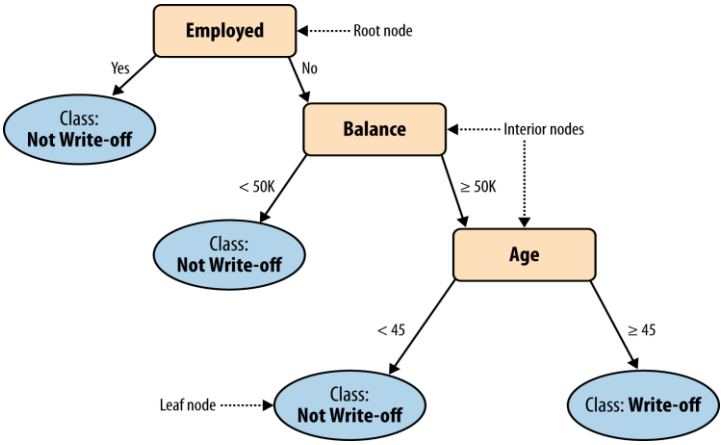
Variabel Numeric

Kita belum membahas bagaimana jika atributnya numerik. Variabel numerik dapat "didiskritkan" dengan memilih satu atau lebih titik pemisah dan kemudian memperlakukan hasil pemisahan sebagi grup yang masing masingnya diberi atribut kategori. Misalnya, penghasilan dapat dibagi menjadi dua rentang atau lebih. Information Gain (IG) dapat diterapkan untuk mengevaluasi segmentasi yang dibuat oleh diskretisasi atribut numerik. Bagaimana cara memilih titik pemisah untuk atribut numerik? Secara konseptual, kita akan mencoba semua titik pemisah yang wajar, dan memilih salah satu yang memberikan IG tertinggi.

Supervised Segmentation Menggunakan Model Tree-Structure

Pemilihan atribut saja tampaknya tidak cukup. Jika kita memilih variabel tunggal yang memberikan perolehan informasi paling banyak, kita membuat segmentasi yang sangat sederhana. Jika kita memilih beberapa atribut yang masing-masing memberikan beberapa perolehan informasi, tidak jelas cara menyatukannya.

Perhatikan segmentasi data yang mengambil bentuk "pohon," seperti yang ditunjukkan pada Gambar 3-3.



Pada gambar 3.3 digambarkan pohon terbalik. Pohon ini terdiri dari node, node interior dan node terminal, dan cabang yang berasal dari node interior. Setiap simpul interior dalam pohon berisi pengujian atribut, dengan setiap cabang dari simpul yang mewakili nilai yang berbeda dari atribut. Mengikuti cabang dari simpul akar ke bawah (ke arah panah), setiap jalur berakhir pada simpul terminal, atau daun (leaf). Pohon tersebut menciptakan segmentasi data: setiap titik data akan termasuk dalam satu dan hanya satu jalur di pohon, dan dengan demikian hanya satu dan hanya satu daun. Pohon adalah segmentasi yang diawasi, karena setiap daun mengandung nilai untuk variabel target. Dalam kasus ini setiap daun mengandung klasifikasi untuk segmennya. Pohon seperti ini disebut pohon klasifikasi atau pohon keputusan.

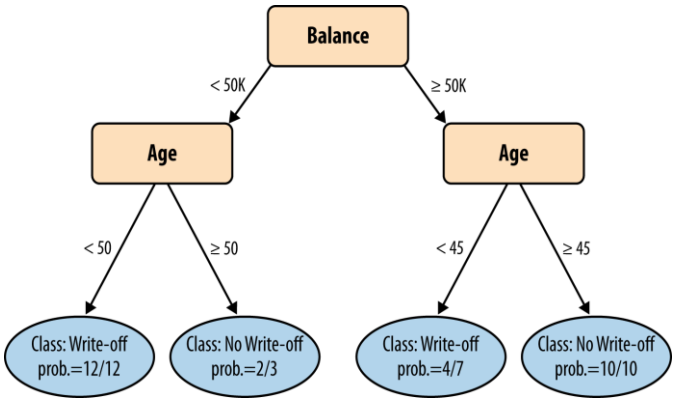
Pohon dan Aturan Segmentasi

Perhatikan sekali lagi pohon yang ditunjukkan pada Gambar 3-15. Anda mengklasifikasikan instans baru (tak terlihat sebelumnya) dengan memulai di simpul akar dan mengikuti uji atribut ke bawah sampai Anda mencapai simpul daun, yang menentukan kelas prediksi instans tersebut. Jika kita menelusuri satu jalur dari simpul akar ke daun, mencatat semua aturan pencabangan, kita menghasilkan aturan. Setiap aturan terdiri dari tes atribut di sepanjang jalur yang terhubung dengan kata sambung "AND".

Pohon klasifikasi ekivalen dengan himpunan aturan ini. Jika aturan ini terlihat berulang, itu karena memang seharusnya. Setiap pohon klasifikasi dapat dinyatakan sebagai seperangkat aturan dengan cara ini. Apakah pohon atau kumpulan aturan lebih dapat dimengerti adalah masalah pemahaman; dalam contoh sederhana ini, keduanya cukup mudah dimengerti. Ketika model menjadi lebih besar, orang akan lebih memilih pohon atau kumpulan aturan.

Estimasi Nilai Probabilitas

Dalam banyak masalah pengambilan keputusan, terkadang lebih diinginkan prediksi yang lebih informatif daripada hanya sekedar klasifikasi. Sebagai contoh, dalam masalah prediksi churn, daripada sekadar memprediksi apakah seseorang akan beralih ke perusahaan lain dalam waktu 90 hari setelah kontrak berakhir, mungkin akan lebih disukai untuk melakukan perkiraan seberapa besar kemungkinan bahwa ia akan meninggalkan perusahaan dalam waktu tersebut. Perkiraan semacam itu dapat digunakan untuk banyak tujuan. Secara singkat: Anda mungkin kemudian memberi peringkat prospek berdasarkan probabilitas churn, dan kemudian mengalokasikan anggaran insentif terbatas ke sampel dengan probabilitas tertinggi. Atau, Anda mungkin ingin mengalokasikan anggaran insentif pada instans dengan kerugian yang diperkirakan paling tinggi, dalam kasus mana Anda perlu (perkiraan) nilai probabilitas churn. Setelah Anda memiliki perkiraan probabilitas seperti itu, Anda dapat menggunakannya dalam proses pengambilan keputusan yang lebih canggih daripada contoh-contoh sederhana ini.



Gambar 3-4 menunjukkan secara umum model "pohon perkiraan probabilitas" untuk contoh prediksi kredit macet (write-off).

Jadi, dalam konteks segmentasi yang diawasi, kita ingin setiap segmen (daun model pohon) diberi perkiraan nilai kemungkinan menjadi anggota di kelas itu. Gambar 3-4 menunjukkan secara umum model "pohon perkiraan probabilitas" untuk contoh prediksi kredit macet (write-off). Dalam diagram tersebut, tidak hanya disajikan prediksi kelas tetapi juga perkiraan probabilitas keanggotaan di kelas tersebut.

Untungnya, induksi pohon yang telah kita diskusikan sejauh ini dengan mudah dapat menghasilkan pohon perkiraan probabilitas daripada hanya sekedar pohon klasifikasi sederhana. Ingat bahwa prosedur induksi pohon membagi ruang contoh ke dalam wilayah kemurnian kelas (entropi rendah). Jika kita puas untuk menetapkan probabilitas kelas yang sama untuk setiap anggota dari segmen yang terkait dengan daun pohon, kita dapat menggunakan jumlah instance di setiap daun untuk menghitung perkiraan probabilitas kelas. Sebagai contoh, jika sebuah daun mengandung n contoh positif dan m contoh negatif, kemungkinan setiap contoh baru yang positif dapat diperkirakan sebagai n / (n + m). Ini disebut perkiraan frekuensi berbasis kemungkinan keanggotaan kelas.

SIMPULAN

Sesi ini memperkenalkan konsep dasar pemodelan prediktif yang merupakan salah satu tugas utama dari data sains, di mana model dibangun untuk dapat memperkirakan nilai variabel target bagi instans baru yang tak terlihat. Dalam prosesnya, diperkenalkan salah satu gagasan mendasar ilmu data: menemukan dan memilih atribut informatif. Memilih atribut informatif dapat menjadi prosedur data mining yang sangat berguna. Mengingat banyaknya koleksi data, kita sekarang dapat menemukan variabel-variabel yang berkorelasi dengan atau memberi kita informasi tentang variabel lain yang menjadi fokus. Sebagai contoh, jika kita mengumpulkan data historis pelanggan yang telah atau tidak meninggalkan perusahaan (churn) tidak lama setelah masa kontraknya berakhir, pemilihan atribut dapat menemukan variabel-variabel demografis atau variabel berorientasi-akun yang memberikan informasi tentang kemungkinan pelanggan berpindah. Salah satu ukuran dasar dari informasi atribut disebut perolehan informasi, yang didasarkan pada ukuran kemurnian yang disebut entropi; ukuran lainnya adalah varians.

Memilih atribut informatif membentuk dasar dari teknik pemodelan umum yang disebut induksi pohon. Induksi pohon secara rekursif menemukan atribut informatif untuk subset dari data. Partisi ini "diawasi" karena berusaha mencari segmen yang memberikan informasi yang semakin tepat tentang kuantitas yang akan diprediksi (target). Model tree-structured yang dihasilkan akan membagi ruang dari semua kemungkinan instance menjadi sekumpulan segmen dengan nilai prediksi yang berbeda untuk target. Misalnya, ketika target adalah variabel "kelas" biner seperti churn versus tidak churn, atau write-off versus tidak write-off, setiap daun pohon sesuai dengan segmen populasi dengan probabilitas estimasi yang berbeda dari keanggotaan kelas.

Secara historis, induksi pohon telah menjadi prosedur data mining yang sangat populer karena mudah dimengerti, mudah diterapkan, dan murah secara komputasi. Penelitian tentang induksi pohon dimulai di tahun 1950-an dan 1960-an. Beberapa teknik induksi pohon populer yang paling awal termasuk CHAID (Chi-squared Automatic Interaction Detection) (Kass, 1980) dan CART (Classification and Regression Trees) (Breiman, Friedman, Olshen, & Stone, 1984), yang masih banyak digunakan. Selain itu, C4.5 dan C5.0 juga merupakan algoritma induksi pohon yang sangat populer (Quinlan, 1986, 1993). J48 adalah perbaikan dari C4.5 yang juga terdapat

dalam paket Weka (Witten & Frank, 2000; Hall et al., 2001) Dalam prakteknya, model struktur pohon bekerja dengan sangat baik, meskipun mereka mungkin bukan model yang paling akurat yang dapat dihasilkan dari satu set data tertentu. Dalam banyak kasus, terutama pada awal penerapan data mining, penting bahwa model dipahami dan dijelaskan dengan mudah. Ini dapat berguna tidak hanya untuk tim data sains tetapi untuk mengkomunikasikan hasil kepada pemangku kepentingan yang tidak memiliki pengetahuan tentang data mining.

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.