BINUS UNIVERSITY ONLINE LEARNING

People Innovation Excellence

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topik 5

Pencocokkan Model (Model Fitting)

LEARNING OUTCOMES

Setelah mempelajari materi ini peserta kuliah diharapkan mampu:

- Finding “optimal” model parameters based on data;
- Choosing the goal for data mining; Objective functions; Loss functions.
- Techniques: Linear regression; Logistic regression; Support-vector machines.

OUTLINE MATERI:

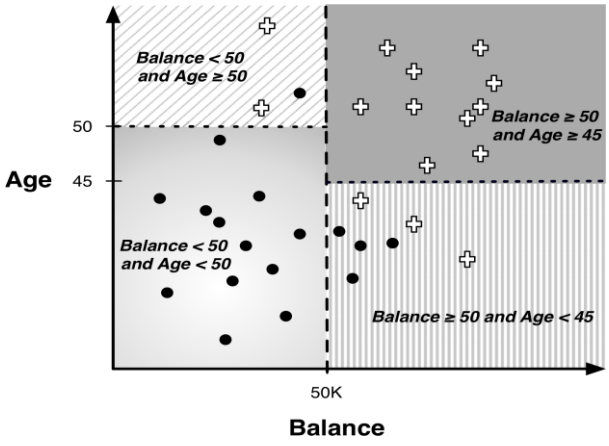
1. Klasifikasi Matematik
2. Mengoptimalkan fungsi obyektif
3. Support Vector Machine
4. Regresi Matematik
5. Kesimpulan

ISI MATERI

Metode alternatif untuk mempelajari model prediktif dari dataset adalah dimulai dengan menentukan struktur model dengan parameter numerik tertentu yang tidak ditentukan. Teknik data mining selanjutnya menghitung nilai parameter terbaik berdasarkan data pelatihan tertentu. Kasus yang sangat umum adalah di mana struktur model adalah parameter fungsi matematika atau persamaan dari satu set atribut numerik. Atribut yang digunakan dalam model dapat dipilih berdasarkan pengetahuan domain mengenai atribut mana yang dianggap informatif dalam memprediksi variabel target, atau dapat dipilih berdasarkan teknik data mining lainnya, seperti prosedur pemilihan atribut yang diperkenalkan pada Sesi 3. Pendekatan umum ini disebut pembelajaran parameter atau pemodelan parametrik.

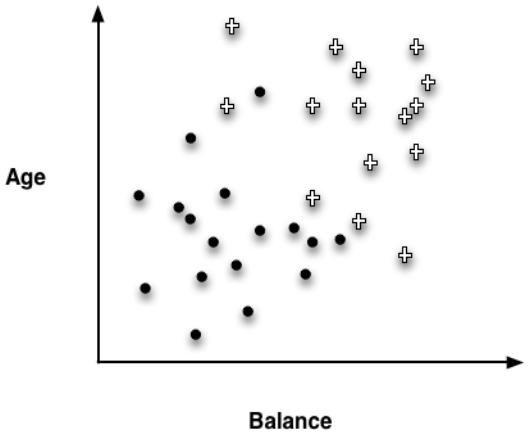
Sebagai contoh, akan disajikan beberapa teknik umum yang digunakan untuk memprediksi (memperkirakan) hal hal seperti nilai numerik yang belum diketahui, nilai-nilai biner yang tidak diketahui (seperti apakah dokumen atau halaman web relevan dengan permintaan), serta kemungkinan seperti default pada pemberian kredit, respons terhadap tawaran, pemalsuan akun, dan sebagainya. Apakah sebenarnya yang dimaksud ketika kita mengatakan sebuah model cocok dengan data yang tersedia? Ini adalah konsep inti dalam Sesi ini. Apakah model didukung oleh data?

Klasifikasi Matematik



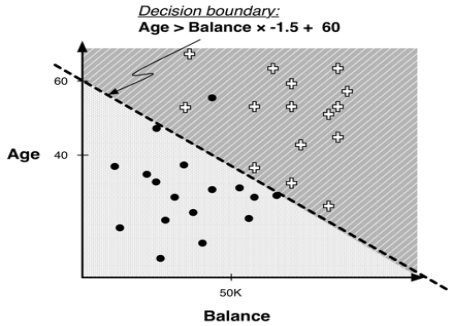
Gambar 4-1.Sebuah data set terbelah oleh pohon klasifikasi dengan 4 leaf nodes.

Kembali ke tampilan ruang-contoh dari model pohon di Sesi 3. Diagram ini direplikasi pada Gambar 4-1. Ruang instance dipecah menjadi daerah dengan batas-batas horisontal dan vertikal yang yang mempartisi ruang instance ke daerah yang sama. Tujuan utama membagi ke dalam wilayah homogen adalah agar dapat diprediksi nilai variabel target dari suatu sampel baru. Sebagai contoh, pada Gambar 4-1, jika pelanggan baru jatuh ke segmen kiri bawah, kita dapat menyimpulkan bahwa nilai target sangat mungkin menjadi "•". Demikian pula, jika jatuh ke segmen kanan atas, kita dapat memprediksi nilainya sebagai "+".



Gambar 4-2. Plot data mentah dari Gambar 4.1 tanpa garis keputusan

Tampilan ruang contoh sangat membantu karena jika kita menghapus garis paralel (Gambar 4-2) kita dapat melihat bahwa ada cara lain yang mungkin lebih baik untuk mempartisi ruang instance. Sebagai contoh, kita dapat memisahkan sampel hampir sempurna (berdasarkan kelas) jika kita diizinkan untuk memperkenalkan garis batas berupa garis lurus, tetapi tidak tegak lurus dengan sumbu sumbu(Gambar 4-3).



Gambar 4-3. Pemisahan dataset dari Gambar 4.2 oleh garis linier tunggal.

Ini disebut classifier linier yang pada dasarnya merupakan penjumlahan nilai-nilai untuk berbagai atribut.

Linear Discriminant Functions

Perhatikan bahwa persamaan umum garis lurus berbentuk y = mx + b, dimana m adalah koefisien kemiringan garis dan b adalah y- intercept. Garis pada Gambar 4-3 dapat dinyatakan dalam bentuk:

$$Age = (-1.5) \times Balance + 60$$

Sebuah instans x diklasifikasi sebagai + jika berada di atas garis, dan berlabel • jika terletak di bawah garis. Persamaan 4.1 memperlihatkan klaisifikasi ini.

Equation 4-1. Classification function

$$class(\,) = \begin{cases} + \text{ if } 1.0 \times Age - 1.5 \times Balance + 60 > 0 \\ \bullet \text{ if } 1.0 \times Age - 1.5 \times Balance + 60 \leq 0 \end{cases}$$

Fungsi ini disebut diskriminan linear karena pembedaan kelas didasari pada fungsi linier yang merupakan kombinasi linier dari atribut. Dalam kasus dua dimensi, kombinasi linear berhubungan dengan garis. Dalam tiga dimensi, batas keputusan adalah bidang, dan dalam dimensi yang lebih tinggi disebut hyperplane. Dengan demikian, model linear adalah jenis segmentasi supervisi multivariat yang berbeda. Tujuan segmentasi yang diawasi adalah memisahkan data ke wilayah dengan nilai variabel target yang berbeda. Fungsi diskriminan linear adalah model klasifikasi numerik. Sebuah model linier dapat dituliskan dalam bentuk persamaan.

Contoh dari persamaan 4.1 dapat ditulis dalam bentuk:

$$f(x) = 60 + 1.0 \times Age - 1.5 \times Balance$$

Untuk menggunakan model ini sebagai diskriminan linear, diberikan suatu sampel, diwakili oleh vektor fitur x, periksa apakah *f(x)* positif atau negatif.

Ini adalah contoh model berparameter; bobot (wi) adalah parameternya. Teknik data mining akan mencari parameter terbaik berdasarkan dataset yang diberikan.

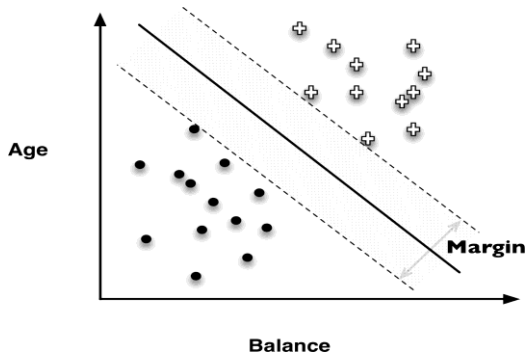
Bukanlah hal yang mudah untuk memilih parameter "terbaik" untuk memisahkan kelas. Ada banyak diskriminan linear berbeda yang dapat memisahkan kelas dengan sempurna. Mereka memiliki kemiringan dan intersep yang sangat berbeda, dan karenanya masing-masing mewakili model yang berbeda. Bahkan, ada banyak garis (model) yang mengklasifikasikan data pelatihan ini dengan sempurna. Mana yang harus dipilih?

Mengoptimalkan Fungsi Objective

Ide fundamental terpenting dalam data mining adalah – ukuran dalam memilih parameter. Prosedur umum adalah mendefinisikan fungsi obyektif yang mewakili ukuran tersebut. Menemukan nilai optimal untuk bobot dapat dilakukan dengan memaksimalkan atau meminimalkan fungsi obyektif. Sayangnya, menciptakan fungsi obyektif yang cocok dengan tujuan, biasanya tidak mungkin, sehingga para ilmuwan data sering memilih berdasarkan pada keyakinan dan pengalaman. Beberapa pilihan ini termasuk apa yang disebut Support Vector Machine (SVM), model linier untuk regresi, dan regresi logistik. Regresi logistik bukan regresi, yang merupakan estimasi dari nilai target numerik. Regresi logistik menggunakan model linear untuk estimasi probabilitas kelas, yang sangat berguna untuk banyak aplikasi. Regresi linier, regresi logistik, dan SVM semuanya merupakan contoh yang sangat mirip dari teknik fundamental dasar mencocokan model (linear) ke data. Perbedaan utamanya adalah masing-masing menggunakan fungsi obyektif yang berbeda.

Support Vector Machine (SVM)

SVM adalah salah satu jenis diskriminan linear. SVM mengklasifikasikan instance berdasarkan fungsi linear dari fitur-fitur. Ada dua ide utama terkait fungsi obyektif. Pertama cocokkan pita terlebar di antara kelas. Ini ditunjukkan oleh garis putus-putus paralel pada Gambar 4-8. Fungsi obyektif SVM menggabungkan ide bahwa pita yang lebih lebar akan lebih baik. Kemudian stelah pita terlebar ditemukan, diskriminan linear otomatis menjadi garis tengah melalui pita (garis utuh pada Gambar 4-8). Jarak antara dua garis paralel putus-putus disebut margin. Tujuan nya adalah untuk memaksimalkan margin.

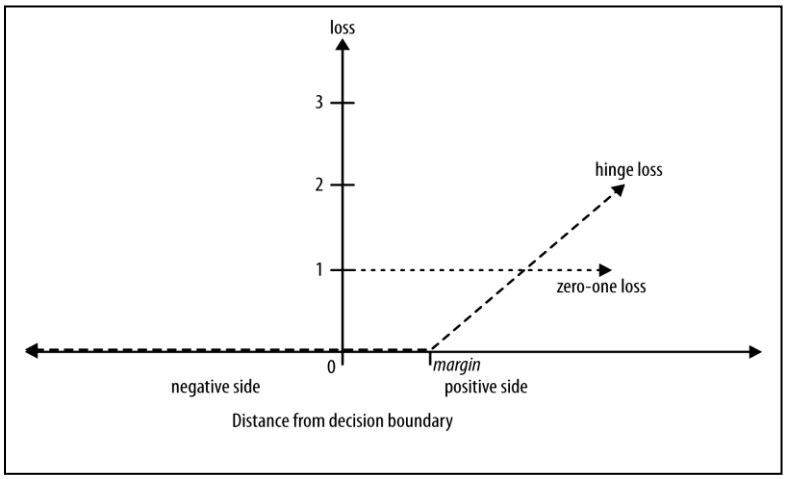


Gambar 4-8. Titik titik dari Gambar 4.2 dan classifier yang memaksimalkan margin.

Ide memaksimalkan margin secara intuitif beralasan. Data pelatihan hanyalah sampel dari populasi yang tidak diketahui distribusinya. Harapannya adalah distribusinya sama dengan data pelatihan. Beberapa contoh positif kemungkinan akan lebih dekat dengan batas diskriminan daripada contoh positif yang belum kita lihat. Hal yang sama berlaku untuk contoh negatif. Dengan kata lain, mereka mungkin berada di dalam margin. Memaksimumkan margin adalah cara maksimal untuk mengklasifikasikan titik-titik tersebut.

Gagasan kedua dari SVM terletak pada bagaimana menangani sampel yang berada pada sisi yang salah. Gambar 4-2 menunjukkan situasi di mana satu garis tidak dapat memisahkan data ke dalam kelas secara sempurna. Dari aplikasi dunia nyata yang kompleks beberapa titik data pasti akan salah diklasifikasi oleh model. Ini tidak menimbulkan masalah, karena klasifikasi tidak harus tepat untuk semua titik. Mungkin tidak ada garis pemisah yang sempurna.

Solusi SVM secara intuitif memuaskan. Alasannya adalah sebagai berikut. Dalam fungsi obyektif yang mengukur seberapa baik model, suatu titik sampel yang berada di sisi yang salah akan diberi penalti. Dalam kasus di mana data memang dapat dipisahkan secara linier, penalty akan sama dengan nol dan FO hanya akan memaksimalkan margin. Jika data tidak dapat dipisahkan secara linier, yang paling tepat adalah mencari keseimbangan antara maksimalkan lebar marjin dan penalty yang rendah. Penalti untuk sampel yang salah sebanding dengan jarak dari batas keputusan, jadi jika mungkin SVM hanya akan membuat kesalahan "kecil". Secara teknis, fungsi kesalahan ini dikenal sebagai “Hinge Loss”.



Gambar 4-9. Ilustrasi Dua Fungsi Loss. Sumbu x memperlihatkan jarak dari garis keputusan. Sumbu y menggambarkan loss yang diperolah oleh contoh negative sebagai fungsi jarak ke garis keputusan.(simetris untuk kasus positif instans). Jika negative instans terletak pada sisi positif, maka tidak mengakibatkan loss.Jika terletak terletak pada sisi positif (wrong),fungsi loss yang berbeda memberi penalty yang berbeda pula.

Regresi Matematik

Fungsi Loss

Istilah "Loss" digunakan dalam data sains sebagai istilah umum untuk memberikan penalty terhadap kesalahan (error). Fungsi Loss menentukan besarnya penalti yang diberikan kepada instance berdasarkan error dalam model prediksi. Beberapa fungsi Loss yang umum digunakan ditunjukkan pada Gambar 4-9.

SVM menggunakan hinge loss. Istilah ini digunakan karena grafiknya berbentuk seperti engsel. Hinge loss tidak memberikan penalti untuk sampel yang tidak berada di sisi yang salah dari margin. Hinge loss hanya menjadi positif ketika sebuah sampel berada di sisi yang salah dari batas dan di luar batas. Loss kemudian meningkat secara linier dengan jarak sampel dari margin, sehingga memberi penalty titik yang lebih jauh dari batas pemisah.

Zero-one memberikan nilai nol untuk keputusan yang benar dan satu untuk keputusan yang salah.

Squared error (SE) menentukan loss sebanding dengan kuadrat jarak dari batas. Loss SE biasanya digunakan untuk prediksi nilai numerik (regresi) bukan klasifikasi. Mengkuadratkan error memiliki efek menghukum prediksi yang sangat meleset.

Struktur model regresi linier persis sama dengan fungsi diskriminan linear Persamaan 4-2:

$f(\mathbf{x}) = w0 + w1x1 + w2x2 + \cdots$

Estimasi Probabilitas Suatu Kelas dan "Logistic Regression"

Pada beberapa masalah, ingin diperkirakan nilai probabilitas bahwa sebuah sampel baru masuk dalam kelas tertentu. Dalam banyak kasus, digunakan perkiraan probabilitas dalam konteks pengambilan keputusan yang mencakup faktor-faktor lain seperti biaya dan manfaat. Misalnya, pemodelan prediktif dari data konsumen besar digunakan secara luas dalam deteksi penipuan di banyak industri, terutama perbankan, telekomunikasi, dan perdagangan online. Diskriminan linier dapat digunakan untuk mengidentifikasi akun atau transaksi palsu.

Pengertian estimasi yang akurat untuk keanggotaan kelas adalah subyek perdebatan yang cukup hangat. Secara garis besar suatu estimasi harus: (i) dikalibrasi dengan baik, artinya jika Anda mengambil 100 kasus yang kemungkinan keanggotaan kelasnya diperkirakan 0,2, maka sekitar 20 di antaranya akan benar-benar menjadi anggota kelas yang diduga. (ii) Memiliki sifat diskriminatif, yaitu memberikan perkiraan probabilitas yang sangat berbeda untuk sampel yang berbeda.

Fungsi Nonlinear, Support Vector Machine, and Neural Network

Sejauh ini Sesi ini berfokus pada fungsi-fungsi numerik yang paling umum digunakan dalam data sains: model linear. Model ini mencakup berbagai teknik yang berbeda.

Konsep dasar jauh lebih umum daripada hanya penerapan fungsi linear. Konsep dasar tidak berubah apabila dikembangkan pada fungsi numerik yang lebih kompleks dan parameternya dicari yang paling sesuai berdasarkan data. Dua teknik yang paling umum yang didasarkan pada parameter dari fungsi-fungsi nonlinier yang rumit adalah SVM nonlinier dan jaringan syaraf tiruan (JST).

Nonlinier SVM pada dasarnya merupakan cara sistematis untuk menerapkan "trik" yang lebih kompleks. SVM memiliki apa yang disebut "fungsi kernel" yang memetakan fitur asli ke ruang fitur dengan dimensi yang lebih tinggi. Kemudian sebuah model linier yang sesuai ditentukan dalam ruang fitur baru ini.

JST juga mengimplementasikan fungsi numerik nonlinear kompleks. JST menawarkan sentuhan yang menarik. JST dapat dianggap sebagai "tumpukan" model. Di tumpukan paling bawah adalah fitur asli. Dari fitur-fitur ini dipelajari berbagai model yang relatif sederhana. Kemudian, setiap lapisan berikutnya dalam tumpukan menerapkan model sederhana ke keluaran lapisan berikutnya.

SIMPULAN

Sesi ini memperkenalkan jenis kedua teknik pemodelan prediktif yang dikenal dengan pencocokan fungsi atau “function fitting” yang atau pemodelan parametrik. Dalam hal ini model adalah persamaan yang ditentukan sebagian: fungsi numerik dari atribut data, dengan beberapa parameter numerik yang tidak ditentukan. Tugas dari proses data mining adalah “mencocokkan” model dengan data dengan mencari parameter terbaik.

Ada banyak jenis teknik pencocokkan fungsi, tetapi kebanyakan menggunakan struktur model linier yang sama: jumlah bobot sederhana dari nilai atribut. Parameter yang akan dicari melalui data mining adalah bobot dari atribut. Teknik pemodelan linier terdiri dari diskriminan linear seperti svm, regresi logistik, dan regresi linier tradisional. Secara konseptual perbedaan utama antara teknik-teknik ini adalah konsep yang digunakan dalam memahami "fungsi obyektif," dan setiap teknik menggunakan fungsi obyektif yang berbeda.

Dua jenis pemodelan data telah diperkenalkan, induksi pohon dan pencocokkan fungsi, dan telah membandingkannya (dalam “Contoh: Regresi Logistik versus Induksi Pohon” di halaman 102).

Sesi ini juga fokus pada konsep dasar untuk mengoptimalkan kecocokan model pada data. Namun, melakukan hal ini terlampau ketat mengarah ke masalah mendasar yang paling penting dalam data mining yaitu overfitting. Mengenali dan menghindari overfitting adalah topik umum yang penting dalam data sains.

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.