BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topik 6

OVERFITTING DAN CARA MENGATASINYA

LEARNING OUTCOMES

Setelah mempelajari materi ini peserta kuliah diharapkan mampu:

Memahami Fundamental concepts: Generalization; Fitting and overfitting; Complexity control, Cross-validation; Attribute selection; Tree pruning; Regularization.

OUTLINE MATERI:

1. Overfitting

2. Overfitting dalam tree induction

3. Overfitting pada fungsi matematik

4. Kesimpulan

ISI MATERI

Setiap dataset adalah sampel terbatas dari sebuah populasi. Model yang dihasilkan diharapkan dapat diterapkan tidak hanya pada data pelatihan tetapi juga untuk keseluruhan populasi dari mana data diambil. Sampel mungkin representatif, tetapi teknik data mining yang digunakan mungkin tidak berhasil membuat model yang umum di luar data pelatihan.

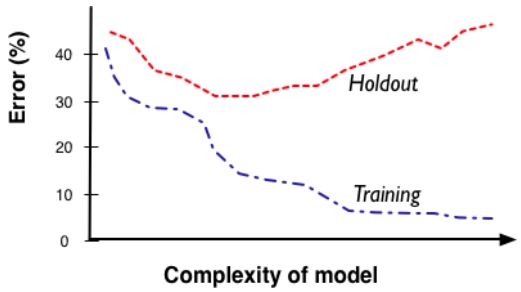
Overfitting

Overfitting adalah prosedur data mining yang cenderung mencocokkan model dengan data secara ketat,sehingga mengorbankan generalisasi pada data yang sebelumnya tak terlihat. Sebelum membahas apa yang harus dilakukan untuk mengatasi, kita perlu tahu bagaimana cara mengenalinya.

Holdout Data and Fitting Graphs

Sebuah alat analitik sederhana yang dikenal dengan fitting graph sering digunakan untuk menunjukkan keakuratan model sebagai fungsi dari kompleksitas model. Untuk menguji overfitting, perlu diketahui konsep evaluasi dasar dalam data mining yaitu holdout data.

Masalah pada bagian sebelumnya adalah model dievaluasi pada data pelatihan. Evaluasi pada data pelatihan tidak memberikan penilaian tentang seberapa baik model tersebut digeneralisasi untuk kasus-kasus yang tidak terlihat. Yang perlu kita lakukan adalah "menyimpan" (holdout) sebagian data yang diketahui nilai dari variabel target, tetapi yang tidak digunakan untuk membangun model. Data ini akan disembunyikan dari model. Model akan memprediksi nilai targetnya. Kemampuan generalisasi akan diprediksi dengan membandingkan nilai yang diprediksi dengan nilai sebenarnya yang disembunyikan. Kemungkinan ada perbedaan antara keakuratan model pada data pelatihan dan akurasi model,yang dihitung dari holdout data. Hold out data sering disebut testing data.



Gambar 5-1. Contoh fitting graph. Setiap titik pada kurva menunjukkan penduga akurasi dari model dengan kompleksitas tertentu. Pendugaan akurasi pada training data dan testing data bervariasi sesuai kompleksitas model. Ketika kompleksitas model rendah, akurasinya rendah, namun ketika model semakin kompleks tingkat akurasi untuk data training meningkat, keakuratan data test (generalisasi) semakin rendah.

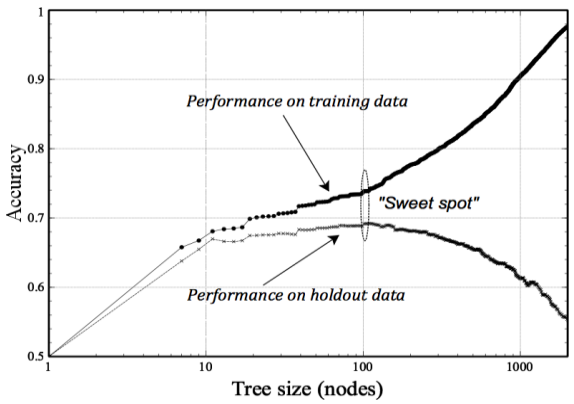
Keakuratan model tergantung pada kompleksitasnya. Suatu model bisa rumit dengan cara yang berbeda. Pertama mari kita gunakan perbedaan ini antara data pelatihan dan data test untuk menentukan grafik fitting lebih tepat. Fitting Graph (lihat Gambar 5-1) menunjukkan perbedaan antara akurasi prosedur pemodelan pada data pelatihan dan akurasi pada data test karena perubahan kompleksitas model. Umumnya, akan ada lebih banyak overfitting karena model menjadi lebih kompleks.

Overfitting dalam Kasus Tree Induction

Kembali ke kasus membangun model struktur pohon untuk klasifikasi. Kasus itu menerapkan kemampuan mendasar untuk menemukan atribut yang penting. Mari kita asumsikan untuk ilustrasi bahwa dataset tidak memiliki dua instance dengan vektor fitur yang sama persis tetapi nilai target yang berbeda. Jika kita terus membagi data, akhirnya subhimpunan akan murni; semua instance dalam subset yang dipilih akan memiliki nilai yang sama untuk variabel target. Ini akan menjadi daun pohon kita. Mungkin ada beberapa contoh di daun, semua dengan nilai yang sama untuk variabel target. Jika perlu, kita dapat tetap membagi atribut, dan membagi lagi data hingga menghasilkan satu kejadian di setiap simpul daun.

Apa yang baru saja dilakukan pada dasarnya membuat versi tabel pencarian yang dibahas di bagian sebelumnya sebagai contoh ekstrim, overfitting! Setiap contoh pelatihan yang diberikan kepada pohon untuk klasifikasi akan menurun, akhirnya mendarat di daun yang sesuai; daun yang berhubungan dengan subset dari data yang mencakup contoh pelatihan khusus ini. Apa yang akan menjadi ketepatan dari pohon ini di set pelatihan? Ini akan sangat akurat, memprediksi dengan benar kelas untuk setiap contoh pelatihan.

Apakah ini akan menyamaratakan? Mungkin. Pohon ini harus sedikit lebih baik daripada tabel pencarian karena setiap contoh yang tidak terlihat sebelumnya akan sampai pada beberapa klasifikasi, bukan hanya gagal untuk mencocokkan; pohon akan memberikan klasifikasi trivial bahkan untuk contoh yang belum pernah dilihat sebelumnya. Oleh karena itu, penting untuk memeriksa secara empiris seberapa baik akurasi pada data pelatihan cenderung sesuai dengan akurasi pada data uji.



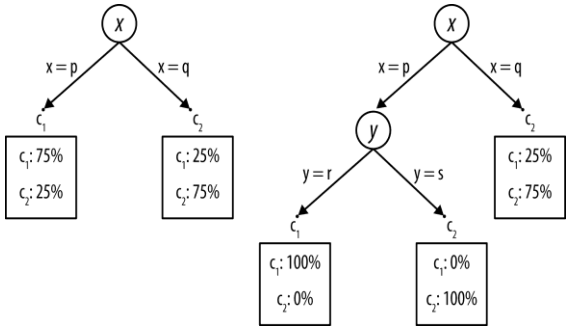
Gambar 5.2 memperlihatkan fitting graph untuk tree induction.

Overfitting pada Fungsi Matematik

Ada berbagai cara untuk menjadikan sebuah fungsi matematika semakin rumit atau sebaliknya. Ada banyak buku membahas tentang topik ini. Cara yang jauh lebih jelas di mana fungsi dapat menjadi terlalu rumit.

Holdout Evaluation Dan Cross-Validation

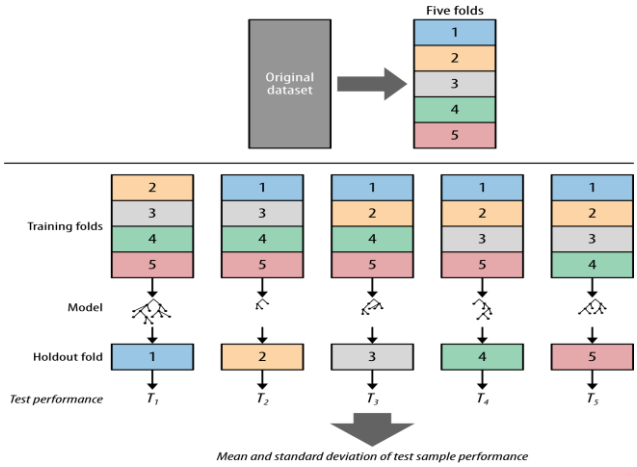
Sementara holdout set memberi kita perkiraan kinerja generalisasi, namun ini hanya perkiraan tunggal. Dapatkah kita mempercayai satu perkiraan akurasi model?



Gambar 5.3 Contoh Overfitting dalam Model Klasifikasi Pohon

Validasi silang (cross validation) adalah prosedur pelatihan dan pengujian yang lebih canggih. Tidak hanya perkiraan sederhana dari kinerja generalisasi, tetapi juga beberapa statistik perkiraan kinerja, seperti mean dan varians akan digunakan sehingga dapat dipahami bagaimana kinerja pada seluruh dataset. Varians ini sangat penting untuk menilai keyakinan dalam perkiraan kinerja.

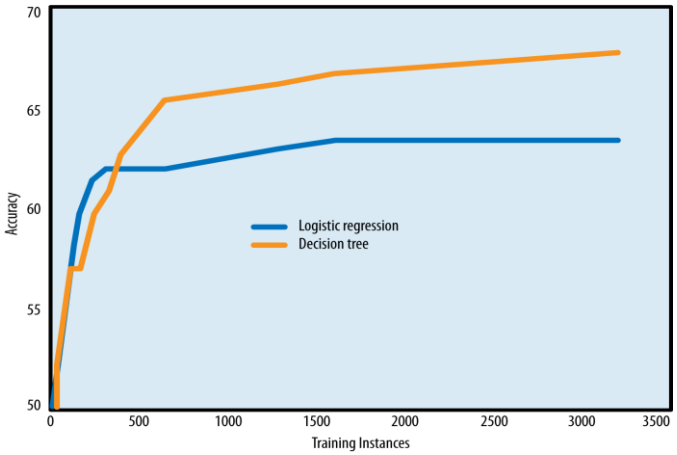
Validasi silang juga membuat penggunaan dataset menjadi lebih baik. Tidak seperti membagi data menjadi satu pelatihan dan satu set holdout, cross-validasi menghitung perkiraan atas semua data dengan melakukan beberapa pemisahan dan secara sistematis menukar sampel untuk pengujian.



Gambar 5.4 Ilustrasi untuk validasi silang…..

Learning Curves

Jika ukuran set pelatihan berubah, mungkin kinerja generalisasi juga turut berubah. Plot dari kinerja generalisasi terhadap jumlah data pelatihan disebut kurva belajar. Kurva belajar adalah alat analitis penting lainnya.



Gambar 5.5 Kurva belajar untuk induksi pohon dan regresi logistik

Kurva belajar untuk induksi pohon dan regresi logistik ditunjukkan pada Gambar 5-5 untuk masalah churn telekomunikasi. Kurva belajar biasanya memiliki bentuk yang khas. Awalnya curam karena prosedur pemodelan menemukan keteraturan paling jelas dalam dataset. Kemudian karena prosedur pemodelan diizinkan untuk melatih pada dataset yang lebih besar dan lebih besar, ia menemukan model yang lebih akurat. Namun, keuntungan marjinal memiliki lebih banyak data berkurang, sehingga kurva belajar menjadi kurang curam. Dalam beberapa kasus, kurva mendatar sepenuhnya karena prosedur tidak dapat lagi meningkatkan akurasi bahkan dengan lebih banyak data pelatihan.

Penting untuk memahami perbedaan antara kurva belajar dan fitting graph. Kurva belajar menunjukkan kinerja generalisasi; kinerja hanya pada data pengujian, diplot terhadap jumlah data pelatihan yang digunakan. Fitting graph menunjukkan kinerja generalisasi serta kinerja pada data pelatihan, tetapi diplot terhadap kompleksitas model. Fitting graph umumnya ditunjukkan untuk jumlah data pelatihan yang tetap.

Metode Umum Menghindari Overfitting

Secara umum, jika kita memiliki kumpulan model dengan kompleksitas yang berbeda, kita dapat memilih yang terbaik hanya dengan memperkirakan kinerja generalisasi masing-masing. Tapi bagaimana kita bisa memperkirakan kinerja generalisasi mereka? Pada data uji (berlabel)? Ada satu masalah besar dengan itu: data uji harus benar-benar independen dari pembuatan model sehingga kita bisa mendapatkan perkiraan independen dari akurasi model. Sebagai contoh, kita mungkin ingin memperkirakan kinerja bisnis utama atau untuk membandingkan model terbaik yang dapat kita bangun dari satu keluarga (katakanlah, pohon klasifikasi) terhadap model terbaik dari keluarga lain (katakanlah, regresi logistik).

Namun, bahkan jika kita menginginkan hal-hal ini, kita masih bisa melanjutkan. Kuncinya adalah menyadari bahwa tidak ada yang istimewa tentang pelatihan / tes split pertama yang kita buat. Katakanlah kita menyimpan set tes untuk penilaian akhir. Kita dapat mengambil set pelatihan dan membaginya lagi menjadi bagian pelatihan dan bagian pengujian. Kemudian kita dapat membangun model pada subset pelatihan ini dan memilih model terbaik berdasarkan pada bagian pengujian ini. Mari kita sebut yang pertama dengan sub-pelatihan dan yang terakhir validasi ditetapkan untuk kejelasan. Set validasi terpisah dari set tes akhir, di mana kita tidak akan pernah membuat keputusan pemodelan. Prosedur ini sering disebut pengujian holdout bertingkat.

Kembali ke contoh pohon klasifikasi kita, kita dapat menginduksi pohon dengan banyak kerumitan dari set subtraining, kemudian kita dapat memperkirakan kinerja generalisasi untuk masing-masing dari set validasi. Ini akan sesuai dengan memilih bagian atas terbalik; Kurva holdout berbentuk U pada Gambar 5-1. Katakanlah model terbaik dengan penilaian ini memiliki kerumitan 122 node (“sweet spot”). Kemudian kita dapat menggunakan model ini sebagai pilihan terbaik kita, mungkin memperkirakan kinerja generalisasi yang sebenarnya pada set tes ketidaksepakatan akhir. Kita juga bisa menambahkan satu sentuhan lagi. Model ini dibangun di bagian data pelatihan kita, karena kita harus menahan set validasi untuk memilih kerumitan. Tetapi setelah kita memilih kerumitannya, mengapa tidak menginduksi pohon baru dengan 122

node dari keseluruhan, set pelatihan asli? Kemudian kita mungkin mendapatkan yang terbaik dari kedua dunia: menggunakan pemisahan subtraining / validasi untuk memilih kompleksitas terbaik tanpa mencemari set tes, dan membangun model kompleksitas terbaik ini di seluruh set pelatihan (subtraining plus validation).

SIMPULAN

Data mining melibatkan trade-off mendasar antara kompleksitas model dan kemungkinan overfitting. Sebuah model yang kompleks mungkin diperlukan jika fenomena yang menghasilkan data itu sendiri kompleks, tetapi model kompleks memiliki risiko data pelatihan yang berlebihan.

Semua tipe model bisa menjadi terlalu berlebihan. Tidak ada pilihan atau teknik tunggal untuk menghilangkan overfitting. Strategi terbaik adalah mengenali overfitting dengan menguji dengan holdout set. Beberapa jenis kurva dapat membantu mendeteksi dan mengukur overfitting. Grafik yang sesuai memiliki dua kurva yang menunjukkan kinerja model pada pelatihan dan pengujian data sebagai fungsi dari kompleksitas model. Fitting graph pada data pengujian biasanya memiliki perkiraan U atau bentuk-U terbalik (tergantung pada apakah kesalahan atau akurasi diplot). Akurasi rendah ketika model sederhana, meningkat ketika kompleksitas meningkat, mendatar, kemudian mulai menurun lagi karena overfitting. Sebuah kurva pembelajaran menunjukkan kinerja model pada pengujian data yang diplot terhadap jumlah data pelatihan yang digunakan. Biasanya kinerja model meningkat dengan jumlah data, tetapi tingkat peningkatan dan kinerja asimtotik akhir bisa sangat berbeda antar model.

Metodologi eksperimental umum yang disebut validasi silang menetapkan cara sistematis membagi satu set data sehingga menghasilkan beberapa ukuran kinerja. Nilai-nilai ini memberi tahu ilmuwan data tentang perilaku rata-rata yang dihasilkan model serta variasi yang diharapkan.

Metode umum untuk mengendalikan kompleksitas model untuk menghindari overfitting disebut model regularisasi. Teknik termasuk pemangkasan pohon (memotong pohon klasifikasi kembali ketika sudah menjadi terlalu besar), seleksi fitur, dan menggunakan penalti kompleksitas eksplisit ke dalam fungsi obyektif yang digunakan dalam pemodelan.

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.