

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topic 8

MENGEVALUASI KUALITAS MODEL

LEARNING OUTCOMES

Setelah mempelajari materi ini peserta kuliah diharapkan mampu mengidentifikasi dan memahami:

- What is desired from data science results;
- Expected value as a key evaluation framework;
- Appropriate comparative baselines.
- Various evaluation metrics; Estimating costs and benefits;
- Calculating expected profit; Creating baseline methods for comparison.

OUTLINE MATERI :

1. Mengevaluasi klasifier
2. Ukuran akurasi sederhana
3. Confusion Matrix
4. Balanced/ Unbalanced Class
5. Unbalanced Cost and Benefit dan Nilai Ekspektasi
6. Kesimpulan

ISI MATERI

Untuk memberikan nilai tambah pada sebuah aplikasi, adalah penting bagi para ilmuwan data dan pemangku kepentingan lainnya untuk mempertimbangkan dengan hati-hati apa yang ingin dicapai lewat data mining proses. Menghubungkan hasil data mining dengan ke tujuan datamining dilakukan sering diabaikan. Ini dapat terlihat dari laporan laporan statistik tanpa pemahaman yang jelas, atau alasan mengapa sesuatu ukuran digunakan.

Setiap aplikasi berbeda. Kita tidak dapat menawarkan metrik tunggal yang "tepat" untuk semua masalah klasifikasi, regresi, atau masalah apa pun yang mungkin dihadapi.

Sesi ini akan memperkenalkan serangkaian kerangka kerja dan metrik untuk tugas-tugas klasifikasi (dalam Sesi ini) dan instance scoring (misalnya, berdasarkan kemungkinan seseorang menanggapi tawaran). Teknik spesifik harus dilihat sebagai contoh yang menggambarkan konsep pemikiran umum secara mendalam tentang penggunaan aplikasi. Untungnya, teknik khusus ini berlaku cukup luas. Akan dijelaskan pula kerangka kerja umum menyangkut evaluasi, penggunaan nilai ekspektasi, yang dapat diterapkan pada berbagai aplikasi yang sangat luas.

Mengevaluasi Klasifier

Perhatikan bahwa suatu model untuk masalah klasifikasi mengambil sampel yang belum diketahui kelasnya dan memprediksi kelas yang tepat untuk sampel tersebut. Perhatikan suatu klasifikasi biner, dimana terdapat dua kelas yang diberi atribut "positif" dan "negatif." Bagaimana mengukur kinerja model tersebut?

Contoh positif perlu dipahami sebagai salah satu yang perlu diperhatikan, dan contoh negatif sebagai sesuatu yang tidak menarik atau biasa. Secara konvensi, kelas positif adalah kelas yang keanggotaannya langka, atau setidaknya lebih jarang daripada kelas negatif. Karena itu, jumlah kesalahan yang dibuat pada contoh negatif (kesalahan positif palsu) dapat mendominasi, meskipun biaya setiap kesalahan yang dilakukan pada contoh positif (kesalahan negatif palsu) akan lebih tinggi.

Ukuran Akurasi Sederhana dan Masalahnya

Akurasi dari suatu model klasifikasi adalah metrik yang populer karena sangat mudah dihitung. Sayangnya, ukuran ini dianggap terlalu sederhana dalam aplikasi data mining. Bagian ini membahas konsep ini dan beberapa alternatif yang ada.

Istilah "akurasi suatu classifier" kadang-kadang digunakan secara tidak resmi untuk mengartikan ukuran umum kinerja suatu model classifier. Akurasi untuk secara teknis adalah proporsi keputusan yang benar:

$$\text{accuracy} = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}}$$

Akurasi sama dengan $1 - \text{error rate}$. Akurasi adalah metrik umum yang digunakan dalam data mining yang direpresentasi oleh angka tunggal dan sangat mudah diukur. Sayangnya, karena kesederhanaannya ukuran ini memiliki beberapa masalah yang dikenal umum. Untuk memahami masalah ini, akan dilakukan analisis terhadap berbagai jenis keputusan yang benar dan salah yang dibuat oleh classifier.

Confusion Matrix

Setiap contoh dalam set tes memiliki label kelas aktual serta label kelas yang diprediksi oleh classifier (kelas yang diprediksi). Untuk menyederhanakan, kita akan berurusan dengan masalah dua kelas yang memiliki 2×2 confusion matrix.

Matriks ini memisahkan keputusan yang dibuat oleh pengklasifikasi. Dengan cara ini berbagai jenis kesalahan dapat ditangani secara terpisah. Dalam confusion matrix, diagonal utama berisi jumlah keputusan yang benar. Kesalahan classifier dikategorikan sebagai positif palsu (FP/ False Positive contoh negatif diklasifikasikan sebagai positif) dan negatif palsu (FN/False Negative; positif diklasifikasikan sebagai negatif).

Unbalanced Classes (Kelas Tak Imbang)

Sebagai contoh mengapa evaluasi model perlu diperhatikan secara intimewa, perhatikan masalah klasifikasi di mana kemunculan salah satu kelas jarang terjadi. Ini adalah situasi umum dalam terapan, karena pengklasifikasi sering digunakan untuk menyaring populasi besar entitas normal atau tidak menarik untuk menemukan jumlah yang relatif kecil dari yang tidak biasa; misalnya, mencari pelanggan yang ditipu, memeriksa jalur perakitan untuk komponen yang rusak, atau menargetkan konsumen yang benar-benar akan menanggapi penawaran. Karena kelas yang tidak biasa atau menarik jarang di antara populasi umum, distribusi kelas tidak seimbang atau miring.

Sayangnya, karena distribusi kelas menjadi lebih miring, evaluasi berdasarkan nilai akurasi tidak memberikan arti.

Sesi 5 menyebutkan “base rate” dari suatu kelas, yang berhubungan dengan seberapa baik performance sebuah classifier dengan hanya memilih kelas itu untuk setiap contoh. Dengan domain berbentuk miring, “base rate” untuk sebagian besar kelas bisa sangat tinggi.

Bahkan ketika kemiringan tidak terlalu mencolok, dalam domain di mana satu kelas lebih dominan daripada yang lain, akurasi bisa sangat menyesatkan. Intinya adalah bahwa akurasi hanyalah salah satu ukuran.

Ketidakseimbangan Cost and Benefit

Masalah lain dengan penggunaan akurasi sederhana sebagai metrik adalah bahwa ukuran ini tidak membedakan antara positif palsu dan negatif palsu. Dengan domain dunia nyata ini jarang terjadi. Kedua jenis kesalahan ini memiliki konsekuensi biaya yang sangat berbeda.

Kembali ke contoh masalah seluler-churn, perhatikan kembali biaya pemberian insentif kepada seseorang agar tidak berpindah (retensi) namun kenyataannya ia berpindah (kesalahan positif palsu). Bandingkan ini dengan biaya kehilangan pelanggan karena tidak ada insentif yang ditawarkan (negatif palsu). Berapa pun biaya yang mungkin Anda putuskan untuk masing-masing, tidak mungkin mereka akan sama; dan karenanya masing masing error harus

diperlakukan atau dihitung secara terpisah. Memang, sulit membayangkan setiap domain di mana pembuat keputusan dapat dengan aman bersikap tidak peduli apakah ia membuat kesalahan positif atau kesalahan negatif palsu. Idealnya, kita harus memperkirakan biaya atau manfaat dari setiap keputusan yang dapat dibuat oleh pembuat klasifikasi. Setelah dikumpulkan, ini akan menghasilkan perkiraan keuntungan (atau perkiraan manfaat atau perkiraan biaya) untuk penggolong.

Generalisasi ke Masalah Umum

Kita telah menggunakan pemodelan klasifikasi untuk mengilustrasikan banyak masalah sains data secara konkret. Sebagian besar masalah ini berlaku pula melampaui masalah klasifikasi. Prinsip umum yang dikembangkan di sini adalah ketika diterapkan ke aplikasi yang sebenarnya, penting untuk memperhatikan pertanyaan pertanyaan: apa yang penting dalam aplikasi ini? Apa tujuannya? Apakah hasil data mining dapat dievaluasi dengan tepat sesuai tujuan sebenarnya?

Kerangka Kerja Analitik Utama: Nilai Ekspektasi

Salah satu tool konseptual yang sangat berguna untuk membantu pemikiran analitik data adalah Nilai Ekspektasi. Perhitungan nilai ekspektasi menyediakan kerangka kerja yang sangat berguna dalam data analitik. Secara khusus, tool ini menghadirkan pemikiran analitik data dalam (i) struktur masalah, (ii) elemen analisis yang dapat diekstraksi dari data, dan (iii) elemen analisis yang perlu diperoleh dari luar (misalnya, pengetahuan bisnis dari ahli dalam bidangnya).

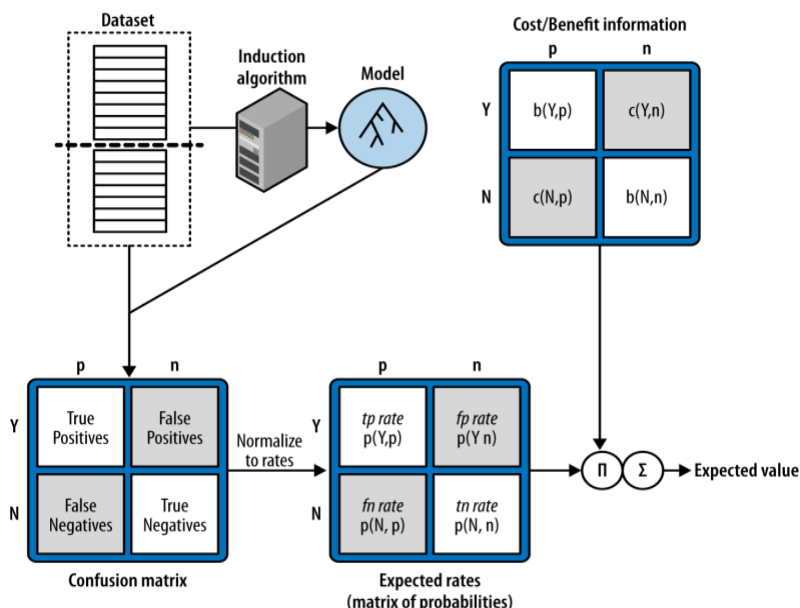
Selanjutnya akan dilistrasikan penggunaan nilai ekspektasi sebagai kerangka analitis dalam dua skenario data sains yang berbeda. Kedua skenario ini sering membingungkan tetapi sangat penting untuk dapat membedakannya. Untuk itu perlu diingat perbedaan antara mining (penambangan atau induksi) model, dan penggunaan model. Ini menunjukkan bagaimana penghitungan nilai yang diharapkan dapat mengungkapkan bagaimana kita akan menggunakan model. Pembuatan eksplisit ini membantu untuk mengatur perumusan masalah dan analisis. Kini

dilanjutkan ke aplikasi penting lainnya dari kerangka nilai ekspektasi, untuk menganalisis apakah model yang dihasilkan dari data sudah sesuai.

Penggunaan Nilai Ekspektasi Dalam Mengevaluasi Suatu Classifier

Pada titik ini fokus dialihkan dari keputusan yang bersifat individual ke koleksi keputusan. Secara khusus, perlu dievaluasi serangkaian keputusan yang dibuat oleh model ketika diterapkan pada sekumpulan contoh. Evaluasi semacam itu diperlukan untuk membandingkan satu model dengan model lainnya. Sangat mungkin bahwa masing-masing model akan membuat beberapa keputusan lebih baik daripada model lainnya. Yang perlu diperhatikan adalah, secara agregat, seberapa baik performance masing-masing model: berapakah nilai ekspektasinya?

Kita dapat menggunakan kerangka nilai ekspektasi yang baru dijelaskan untuk menentukan keputusan terbaik untuk masing-masing model tertentu, dan kemudian menggunakan nilai yang sama dengan cara yang berbeda untuk membandingkan model.



Seluruh penghitungan keuntungan yang diharapkan pada tingkat (model) agregat, dalam proses menghitung probabilitas. Gambar 7-2 menunjukkan diagram skematik perhitungan nilai yang diharapkan dalam konteks induksi dan evaluasi model.

Tingkat Error (Error Level)

Ketika menghitung nilai ekspektasi untuk masalah bisnis, analis sering dihadapkan dengan pertanyaan: dari mana sebenarnya nilai probabilitas berasal? Ketika mengevaluasi model pada data testing, jawabannya adalah: nilai probabilitas ((error) kesalahan dan keputusan yang benar) dapat diperkirakan dari penghitungan dalam matriks confusion dengan menghitung tingkat error dan keputusan yang benar. Setiap sel dari matriks confusion berisi hitungan jumlah keputusan yang sesuai dengan kombinasi yang sesuai (diprediksi, aktual), yang akan kita nyatakan sebagai $count(h, a)$ (kita menggunakan h untuk "berhipotesis" karena p sudah ada yang menggunakan). Untuk perhitungan nilai yang diharapkan, nilai $count$ ini digunakan untuk menentukan nilai probabilitas yang diperkirakan, $p(h, a)$. Ini dilakukan dengan membagi setiap hitungan dengan jumlah total instance:

$$p(h, a) = count(h, a) / T$$

Costs and benefits

Untuk menghitung laba yang diharapkan, dibutuhkan nilai $cost$ (biaya) dan $manfaat$ (benefit) yang sesuai dengan setiap pasangan keputusan. Ini akan membentuk entri dari matriks biaya-manfaat dengan dimensi yang sama (baris dan kolom) dengan confusion matrix. Namun untuk setiap pasangan (yang diprediksi, dan aktual), matriks biaya-manfaat, berisi biaya atau manfaat akibat keputusan dalam sel itu (lihat Gambar 7-3). Klasifikasi yang benar (benar positif dan benar negatif) bersesuaian dengan $manfaat$ $b(Y, p)$ dan $b(N, n)$. Klasifikasi yang salah (positif palsu dan negatif palsu) bersesuaian dengan "manfaat" $b(Y, n)$ dan $b(N, p)$. Yang sebenarnya menjadi biaya (manfaat negatif), dan sering secara eksplisit disebut sebagai biaya $c(Y, n)$ dan $c(N, p)$.

		Actual	
		p	n
Predicted	Y	$b(Y,p)$	$c(Y,n)$
	N	$c(N,p)$	$b(N,n)$

Gambar 7-3. Contoh Matrix cost-benefit.

Jika diberikan matriks biaya dan manfaat, ini akan dikalikan sel dengan sel terhadap matriks probabilitas, kemudian dijumlahkan menjadi nilai akhir yang mewakili total laba yang diharapkan.

Dengan menggunakan formula ini, kita dapat menghitung dan membandingkan keuntungan yang diharapkan untuk berbagai model dan strategi penargetan lainnya. Yang kita butuhkan adalah dapat menghitung confusion matriks atas serangkaian data tes, untuk menghasilkan matriks biaya-manfaat. Formula ini cukup untuk membandingkan antar classifier.

Kini telah tersedia satu cara untuk menghadapi contoh. Daripada menghitung akurasi untuk model yang bersaing, sebaiknya dihitung nilai ekspektasi. Lebih jauh, menggunakan formulasi alternatif ini, kita dapat membandingkan dua model meskipun satu analisis diuji menggunakan distribusi representatif dan yang lain diuji menggunakan distribusi kelas-seimbang. Dalam setiap perhitungan, kita dapat mengganti prior. Menggunakan distribusi seimbang sesuai dengan prior dari $p(p) = 0,5$ dan $p(n) = 0,5$.

KESIMPULAN

Bagian penting dari data sains adalah tersedianya alat evaluasi model yang tepat. Hal ini sangat sulit untuk dilakukan dengan benar dan seringkali memerlukan beberapa iterasi. Sering para ilmuwan tergoda untuk menggunakan langkah-langkah sederhana, seperti akurasi sederhana, karena ini mudah untuk dihitung. Namun, dalam domain dunia nyata, alat sederhana jarang menangkap apa yang sebenarnya penting untuk masalah yang dihadapi, dan sering menyesatkan. Sebaliknya, ilmuwan data sains harus memberikan pemikiran yang cermat tentang bagaimana model akan digunakan dalam praktik dan menyusun metrik yang tepat.

Penghitungan nilai ekspektasi adalah kerangka kerja yang baik untuk keperluan ini. Ini akan membantu membuat framework evaluasi, dan jika model akhir yang digunakan menghasilkan hasil yang tidak dapat diterima, itu akan membantu mengidentifikasi apa yang salah.

Karakteristik data harus diperhitungkan dengan hati-hati ketika mengevaluasi hasil data sains. Sebagai contoh, masalah klasifikasi nyata sering menyajikan data dengan distribusi kelas yang sangat tidak seimbang (yaitu kelas tidak berdistribusi dengan prevalensi yang sama). Menyesuaikan proporsi kelas mungkin berguna (atau bahkan perlu) untuk mempelajari model dari data; namun, evaluasi harus menggunakan populasi yang asli dan realistis sehingga hasilnya mencerminkan apa yang sebenarnya akan dicapai.

Untuk menghitung nilai keseluruhan yang diharapkan dari suatu model, biaya dan manfaat keputusan harus ditentukan. Jika ini memungkinkan, ilmuwan data dapat menghitung biaya per model yang diharapkan untuk setiap model dan memilih model mana yang menghasilkan biaya terendah yang diharapkan atau keuntungan terbesar.

Penting juga untuk mempertimbangkan apa yang harus dibandingkan dengan model berbasis data terhadap, untuk menilai apakah kinerjanya baik atau lebih baik. Jawaban atas pertanyaan ini sangat terkait dengan pemahaman bisnis, tetapi ada berbagai praktik terbaik umum yang harus diikuti oleh tim sains data. Ide-ide dari Sesi ini diilustrasikan dengan aplikasi konsep yang disajikan dalam Sesi-Sesi sebelumnya. Konsepnya lebih umum, dan berhubungan dengan konsep fundamental pertama: data harus dianggap sebagai aset dan perlu dipertimbangkan cara berinvestasi yang benar. Hal ini diilustrasikan dengan membahas secara singkat bahwa seseorang tidak hanya dapat membandingkan model yang berbeda dan baseline yang berbeda, tetapi juga membandingkan hasil dengan sumber data yang berbeda. Sumber data yang berbeda mungkin memiliki biaya terkait yang berbeda, dan evaluasi yang cermat dapat menunjukkan yang dapat dipilih untuk memaksimalkan laba atas investasi.

Sebagai ringkasan terakhir, Sesi ini telah membahas angka-angka kuantitatif tunggal sebagai perkiraan ringkasan kinerja model. Mereka dapat menjawab pertanyaan seperti “Berapa banyak laba yang dapat diharapkan?” Dan “Haruskah saya menggunakan model A atau model B?” Jawaban semacam itu berguna tetapi hanya menyediakan “nilai satu titik” yang berada di bawah asumsi tertentu. Sering diperlukan untuk memvisualisasikan perilaku model di bawah berbagai kondisi.

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.