BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topic 10

BUKTI DAN PROBABILITAS

LEARNING OUTCOMES

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

- Explicit evidence combination with Bayes' Rule;
- Probabilistic reasoning via assumptions of conditional independence.
- Naive Bayes classification;
- Evidence lift.

OUTLINE MATERI :

1. Contoh kasus target pemasaran online
2. Menggabungkan Bukti secara Probabilistik
3. Penerapan hukum Bayes
4. Kelebihan dan kekurangan Naive Bayes
5. Kesimpulan

PENDAHULUAN

Sejauh ini kita telah mendiskusikan beberapa metode yang berbeda dalam pengggunaan data untuk membantu menarik kesimpulan tentang beberapa kuantitas yang tidak diketahui, seperti klasifikasi. Sesi ini mengajak kita memeriksa cara lain untuk menarik kesimpulan seperti itu. Kita dapat memberi istilah pada hal-hal yang kita ketahui tentang data sebagai bukti yang menentang atau mendukung nilai yang berbeda untuk target. Hal-hal yang diketahui tentang sampel/ instans direpresentasikan sebagai fitur dari instans. Jika diketahui kekuatan bukti yang diberikan oleh setiap fitur, kita bisa menerapkan prinsip prinsip dasar dalam menggabungkan bukti probabilistik untuk mencapai kesimpulan mengenai nilai target. Kita akan menentukan kekuatan setiap bukti berdasarkan data pelatihan.

Contoh Kasus: Menentukan Target Untuk Iklan Online

Perhatikan satu aplikasi klasifikasi bisnis lain: menargetkan iklan tampilan online kepada konsumen, berdasarkan pada halaman web yang telah mereka kunjungi di masa lalu. Sebagai konsumen, kita telah terbiasa mendapatkan sejumlah besar informasi dan layanan di Web yang tampaknya gratis.

Iklan bergambar berbeda dari iklan berbasir pencarian (misalnya, iklan yang muncul dengan hasil pencarian Google). Dalam sebagian besar halaman web, pengguna tidak mengetikkan frase yang terkait dengan apa yang sebenarnya dia cari. Oleh karena itu, penargetan iklan ke pengguna perlu didasarkan pada jenis inferensi lainnya. Kita telah berbicara tentang jenis inferensi tertentu: menyimpulkan nilai variabel target instance dari nilai fitur instance. Oleh karena itu, kita dapat menerapkan teknik yang sudah kita bahas untuk menyimpulkan apakah pengguna tertentu tertarik dengan iklan yang ditayangkan. Dalam sesi ini kita akan memperkenalkan cara berbeda dalam melihat masalah, yang memiliki penerapan yang luas dan cukup mudah diterapkan.

Definisikan masalah penargetan iklan lebih tepat. Apa yang akan menjadi sampel? Apa yang akan menjadi variabel target? Apa yang akan menjadi fitur? Bagaimana kita akan mendapatkan data pelatihan?

Asumsikan bahwa kita bekerja untuk penyedia konten yang sangat besar ("penerbit") yang memiliki berbagai macam konten, melihat banyak konsumen online, dan memiliki banyak peluang untuk menampilkan iklan kepada konsumen ini. Misalnya, Yahoo! memiliki sejumlah besar "properti" web yang didukung iklan yang berbeda, yang dapat kita anggap sebagai "potongan konten" berbeda. Selain itu, baru-baru ini (sejak tulisan ini ditulis) Yahoo! setuju untuk membeli situs blogging Tumblr, yang memiliki 50 miliar posting blog di lebih dari 100 juta blog. Masing-masing ini juga dapat dilihat sebagai "potongan konten" yang memberikan beberapa pandangan ke dalam kepentingan konsumen yang membacanya. Demikian pula, Facebook mungkin mempertimbangkan setiap "Like" yang dibuat oleh konsumen sebagai bukti terkait selera konsumen, yang mungkin juga membantu menargetkan iklan.

Untuk mempermudah, asumsikan kita melakukan kampanye iklan yang kita ingin targetkan beberapa bagian dari konsumen online yang mengunjungi situs kita. Kampanye ini untuk jaringan hotel kelas atas, Luxhote. Tujuan dari Luxhote adalah agar orang-orang memesan kamar. Kita telah menjalankan kampanye ini di masa lalu, memilih konsumen online secara acak. Kita sekarang ingin menjalankan kampanye yang ditargetkan, dengan harapan mendapatkan lebih banyak pemesanan per dolar yang dihabiskan untuk tayangan iklan.

Anggaplah seorang konsumen adalah sebuah instance. Variabel target kita adalah: apakah konsumen akan memesan kamar Luxhote dalam waktu satu minggu setelah melihat iklan Luxhote? Melalui “cookie browser”, bekerja sama dengan Luxhote kita dapat mengamati konsumen mana saja yang memesan kamar. Sebagai data pelatihan, kita memiliki nilai biner untuk variabel target untuk setiap konsumen. Dalam penerapan, kita akan memperkirakan probabilitas bahwa konsumen akan memesan kamar setelah melihat iklan, dan kemudian, karena anggaran kita memungkinkan, menargetkan beberapa bagian dari konsumen dengan probabilitas tertinggi.

Kita disuguhkan dengan pertanyaan kunci: apa yang akan menjadi fitur yang akan kita gunakan untuk menggambarkan konsumen, sehingga kita mungkin dapat membedakan mereka yang lebih atau kurang mungkin menjadi pelanggan yang baik untuk Luxhote? Untuk contoh ini, kita akan mempertimbangkan konsumen untuk dideskripsikan oleh kumpulan potongan konten yang telah kita amati bahwa dia telah melihat (atau Disukai) sebelumnya, seperti yang tercatat melalui cookie browser atau beberapa mekanisme lainnya. Kita memiliki banyak jenis konten: keuangan,

olahraga, hiburan, blog memasak, dll. Kita mungkin memilih beberapa ribu potongan konten yang sangat populer, atau kita dapat mempertimbangkan ratusan juta. Kita yakin bahwa beberapa di antaranya (mis., Blog keuangan) lebih cenderung dikunjungi oleh prospek yang baik untuk Luxhote, sementara potongan konten lainnya terlihat kurang menjanjikan (mis., Laman penggemar traktor-tarik).

Namun, untuk latihan ini kita tidak ingin bergantung pada anggapan kita tentang konten tersebut, kita juga tidak memiliki sumber daya untuk memperkirakan potensi bukti untuk setiap bagian konten secara manual. Lebih jauh lagi, sementara manusia cukup pandai menggunakan pengetahuan dan akal sehat untuk mengenali apakah suatu bukti termasuk "mendukung (pros)" atau "melawan (cons)," manusia terkenal buruk dalam memperkirakan kekuatan bukti yang tepat. Kita ingin data historis kita untuk memperkirakan arah dan kekuatan bukti. Kita selanjutnya akan menjelaskan kerangka kerja yang sangat luas baik untuk mengevaluasi bukti, dan untuk menggabungkannya untuk memperkirakan kemungkinan hasil keanggotaan kelas (di sini, kemungkinan bahwa konsumen akan memesan kamar setelah melihat iklan).

Ternyata terdapat banyak masalah lain yang mirip dengan contoh yang diangkat di atas, misalnya: masalah estimasi probabilitas klasifikasi / kelas di mana masing-masing contoh digambarkan oleh sekumpulan bukti, kemungkinan diambil dari total koleksi bukti yang sangat besar. Sebagai contoh, klasifikasi dokumen teks. Setiap dokumen adalah kumpulan kata-kata, dari total kosakata yang sangat besar. Setiap kata mungkin dapat memberikan beberapa bukti mendukung atau melawan klasifikasi, dan kita ingin menggabungkan bukti. Teknik yang kita perkenalkan selanjutnya adalah yang digunakan dalam banyak sistem deteksi spam: dimana sebuah instance adalah pesan email, kelas target adalah spam atau bukan-spam, dan fitur-fiturnya adalah kata-kata dan simbol dalam pesan email.

Menggabungkan Bukti secara Probabilistik

Untuk membahas ide menggabungkan bukti probabilistik, kita perlu memperkenalkan beberapa notasi probabilitas. Anda tidak harus belajar (atau ingat) teori probabilitas — pengertiannya cukup intuitif, dan kita tidak akan melampaui dasar-dasarnya. Notasi memungkinkan kita untuk menjadi tepat. Mungkin terlihat seperti ada banyak matematika dalam hal-hal berikut, tetapi Anda akan melihat bahwa itu cukup mudah.

Seperti disebutkan di atas, kita ingin menggunakan beberapa data berlabel, seperti data dari kampanye iklan yang ditargetkan secara acak, untuk mengaitkan koleksi berbeda dari bukti E dengan probabilitas yang berbeda. Sayangnya, ini memperkenalkan masalah utama. Untuk koleksi tertentu dari bukti E, kita mungkin belum melihat cukup banyak kasus dengan koleksi bukti yang sama untuk dapat menyimpulkan kemungkinan keanggotaan kelas dengan keyakinan apa pun. Bahkan, kita mungkin tidak melihat koleksi bukti khusus ini sama sekali! Dalam contoh kita, jika kita mempertimbangkan ribuan situs web berbeda, apa peluang dalam data pelatihan kita, kita telah melihat konsumen dengan pola kunjungan yang sama persis sebagai konsumen yang akan kita lihat di masa mendatang? Itu sangat kecil. Oleh karena itu, apa yang akan kita lakukan adalah mempertimbangkan berbagai bukti yang berbeda secara terpisah, dan kemudian menggabungkan bukti. Untuk membahas ini lebih lanjut, kita perlu beberapa fakta tentang menggabungkan probabilitas.

Joint Probability dan Kebebasan

Rumus umum untuk menggabungkan probabilitas dengan memperhitungkan keterikatan antara peristiwa adalah :

Persamaan 9-1. Joint probability menggunakan probabilitas bersyarat

$$p(AB) = p(A) \cdot p(B \mid A)$$

Aturan Bayes

Perhatikan bahwa di dalam $p(AB) = p(A)p(B|A)$ urutan A and B dapat dipertukarkan:

$$p(AB) = p(B) \cdot p(A \mid B)$$

Ini membeikan:

$$p(A) \cdot p(B \mid A) = p(AB) = p(B) \cdot p(A \mid B)$$

Dan:

$$p(A) \cdot p(B \mid A) = p(B) \cdot p(A \mid B)$$

Jika kedua ruas dibagi dengan $p(A)$ diperoleh:

$$p(B \mid A) = \frac{p(A \mid B) \cdot p(B)}{p(A)}$$

Sekarang, perhatikan B sebagai hipotesis yang akan dicari nilai kemungkinannya, dan A merupakan bukti yang diamati. Dengan mengganti nama dengan H untuk hipotesis dan E untuk bukti, kita mendapatkan:

$$p(H \mid E) = \frac{p(E \mid H) \cdot p(H)}{p(E)}$$

Ini adalah Aturan Bayes yang terkenal itu,. Aturan Bayes mengatakan bahwa kita dapat menghitung probabilitas dari hipotesis H diberikan bukti E dengan melihat probabilitas dari bukti dengan syarat probabilitas terjadinya hipotesis diketahui, serta probabilitas tanpa syarat dari hipotesis dan bukti.

Metode Bayes

Aturan Bayes, dikombinasikan dengan prinsip dasar pemikiran penting tentang independensi bersyarat, adalah fondasi untuk sejumlah besar teknik ilmu data tingkat lanjut yang tidak akan dibahas dalam kuliah ini.

Yang penting, tiga kuantitas terakhir mungkin lebih mudah untuk ditentukan yaitu, p (H|E). Perhatikan contoh (yang disederhanakan) dari diagnosis medis. Anggaplah Anda seorang dokter dan seorang pasien tiba dengan bintik-bintik merah. Anda menebak (berhipotesis) bahwa pasien menderita campak. Kita ingin menentukan kemungkinan diagnosis hipotesis kita (H = campak), diketahui bukti (E = bintik merah). Untuk memperkirakan secara langsung p (campak|bintik

merah) kita perlu memikirkan semua penyebab yang berbeda seseorang mendapat gejala bintik-bintik merah dan berapa proporsi mereka terkena campak.

Namun, sebagai gantinya, pekerjaan melakukan perkiraan ini dapat ditempuh dengan menggunakan sisi kanan Aturan Bayes.

- p (E|H) adalah probabilitas bahwa seseorang memiliki bintik-bintik merah dan diketahui ia terserang campak. Seorang ahli penyakit menular dapat mengetahui ini atau memperkirakannya secara relatif akurat.
- p(H) merupakan probabilitas seseorang menderita campak, tanpa melihat bukti apa pun; Nilai ini adalah prevalensi campak dalam populasi.
- p(E) adalah probabilitas dari bukti: Berapakah kemungkinan seseorang memiliki bintik merah (hanya prevalensi bintik merah dalam populasi), yang tidak memerlukan penalaran yang rumit tentang penyebabnya.

Aturan Bayes telah membuat perkiraan p(H|E) jauh lebih mudah. Kita membutuhkan tiga jenis informasi, yang jauh lebih mudah untuk diperkirakan daripada nilai aslinya.

p(E) mungkin masih sulit untuk dihitung. Namun, dalam banyak kasus, hal itu tidak harus dihitung, karena kita tertarik untuk membandingkan probabilitas berbagai hipotesis yang berbeda yang diberikan bukti yang sama.

Penerapan Aturan Bayes

Sebagian besar ilmu data didasarkan pada metode "Bayesian", yang memiliki alasan inti berdasarkan Aturan Bayes. Menggambarkan metode Bayesian secara luas jauh melampaui ruang lingkup kuliah ini. Kita akan memperkenalkan ide-ide yang paling mendasar, dan kemudian menunjukkan bagaimana mereka menerapkan teknik dasar Bayesian yang paling dasar. Mari menulis ulang Aturan Bayes lagi, tetapi sekarang kembali ke masalah klasifikasi.

Persamaan 9-2. Aturan Bayes untuk klasifikasi

$$p(C = c \mid E) = \frac{p(E \mid C = c) \cdot p(C = c)}{p(E)}$$

Dalam Persamaan 9-2, kita memiliki empat kuantitas. Di sisi kiri adalah kuantitas yang ingin kita perkirakan. Dalam konteks masalah klasifikasi, ini adalah probabilitas bahwa variabel target C mengambil kelas minat c diberikan bukti E (vektor nilai fitur). Ini disebut probabilitas posterior.

Aturan Bayes menguraikan probabilitas posterior ke dalam tiga kuantitas yang kita lihat di sisi sebelah kanan. Kita ingin dapat menghitung jumlah ini dari data:

1. p(C=c) adalah probabilitas “sebelumnya” dari kelas, yaitu, probabilitas yang akan kita tetapkan ke kelas sebelum melihat bukti apa pun. Dalam pemikiran Bayesian umumnya, ini bisa berasal dari beberapa tempat. Itu bisa berupa (i) suatu "subyektif" sebelumnya, yang berarti bahwa itu adalah keyakinan dari pembuat keputusan tertentu berdasarkan semua pengetahuan, pengalaman, dan pendapatnya; (ii) keyakinan "sebelumnya" berdasarkan pada beberapa aplikasi sebelumnya dari Aturan Bayes dengan bukti lain, atau (iii) probabilitas tak bersyarat yang disimpulkan dari data. Metode spesifik yang kita perkenalkan di bawah ini mengambil pendekatan (iii), menggunakan sebagai kelas sebelum “tingkat dasar” c — prevalensi c dalam populasi secara keseluruhan. Ini dihitung dengan mudah dari data sebagai persentase dari semua contoh yang ada di kelas c.
2. p(E |C = c) adalah kemungkinan melihat bukti E — fitur-fitur khusus dari contoh yang diklasifikasikan — ketika kelas C = c. Orang mungkin melihat ini sebagai "generatif" pertanyaan: jika dunia ("proses menghasilkan data") menghasilkan sebuah instance dari kelas c, seberapa sering akan terlihat seperti E? Kemungkinan ini dapat dihitung dari data sebagai persentase contoh kelas c yang memiliki vektor fitur E.
3. Akhirnya, p(E) adalah kemungkinan bukti: seberapa umum representasi fitur E di antara semua contoh? Ini mungkin dihitung dari data sebagai persentase kemunculan E di antara semua contoh.

Memperkirakan ketiga nilai ini dari data pelatihan, kita bisa menghitung perkiraan untuk posterior p(C=c|E) untuk contoh tertentu yang digunakan. Ini dapat digunakan secara langsung sebagai perkiraan kemungkinan kelas, mungkin dalam kombinasi dengan biaya dan manfaat. Alternatifnya, p (C=c|E) dapat digunakan sebagai skor untuk memeringkat contoh. Atau, kita bisa memilih sebagai klasifikasi p maksimum (C = c | E) di seluruh nilai yang berbeda c.

Sayangnya, kita kembali ke kesulitan utama yang kita sebutkan di atas, yang membuat Persamaan 9-2 tidak dapat digunakan langsung dalam penggalian data. Pertimbangkan E sebagai vektor umum nilai atribut <e1, e2, ⋯, ek>, kemungkinan besar, kumpulan kondisi tertentu. Menerapkan Persamaan 9-2 secara langsung akan membutuhkan pengetahuan p(E|c) sebagai p(e1 ∧ e2 ∧ ⋯ ∧ ek|c). Ini sangat spesifik dan sangat sulit untuk diukur. Kita mungkin tidak pernah melihat contoh spesifik dalam data pelatihan yang sama persis dengan E yang diberikan dalam data pengujian kita, dan bahkan jika kita melakukannya mungkin kita tidak akan melihat cukup banyak dari mereka untuk memperkirakan probabilitas dengan keyakinan apa pun.

Metode Bayesian untuk ilmu data berurusan dengan masalah ini dengan membuat asumsi independensi probabilistik. Metode yang paling banyak digunakan untuk menangani komplikasi ini adalah dengan membuat asumsi yang sangat kuat tentang independensi.

Kelebihan dan kekurangan Naive Bayes

Naive Bayes adalah penggolong yang sangat sederhana, namun masih mempertimbangkan semua bukti fitur. Ini sangat efisien dalam hal ruang penyimpanan dan waktu komputasi. Pelatihan hanya terdiri dari menyimpan jumlah kelas dan kejadian fitur karena setiap contoh terlihat. Seperti disebutkan, p (c) dapat diperkirakan dengan menghitung proporsi contoh kelas c di antara semua contoh. p (ei | c) dapat diestimasi dengan proporsi contoh di kelas c untuk mana fitur ei muncul. Terlepas dari kesederhanaan dan asumsi independensinya yang ketat, penggolongan Naive Bayes bekerja dengan sangat baik untuk klasifikasi pada banyak tugas di dunia nyata. Ini karena pelanggaran asumsi independensi cenderung tidak melukai kinerja klasifikasi, karena alasan yang intuitif memuaskan. Secara khusus, pertimbangkan bahwa dua bukti itu sebenarnya sangat bergantung — apa artinya itu? Secara kasar, itu berarti bahwa ketika kita melihatnya, kita juga cenderung melihat yang lain. Sekarang, jika kita memperlakukan mereka sebagai independen, kita akan melihat satu dan mengatakan "ada bukti untuk kelas" dan

melihat yang lain dan mengatakan "ada lebih banyak bukti untuk kelas." Jadi, sampai batas tertentu kita akan penghitungan ganda bukti. Namun, selama bukti-bukti pada umumnya menunjukkan kita pada arah yang benar, untuk klasifikasi, penghitungan ganda tidak akan cenderung menyakiti kita. Bahkan, itu akan cenderung membuat perkiraan probabilitas lebih ekstrim dalam arah yang benar: probabilitas akan berlebihan untuk kelas yang benar dan meremehkan untuk kelas yang salah (es). Namun untuk klasifikasi, kita memilih kelas dengan perkiraan probabilitas terbesar, sehingga membuatnya lebih ekstrim dalam arah yang benar adalah OK.

Ini memang menjadi masalah, meskipun, jika kita akan menggunakan perkiraan probabilitas itu sendiri — jadi Naive Bayes harus digunakan dengan hati-hati untuk pengambilan keputusan yang sebenarnya dengan biaya dan manfaat, seperti yang dibahas di Sesi 7. Praktisi menggunakan Naive Bayes secara teratur untuk menentukan peringkat, di mana nilai sebenarnya dari probabilitas tidak relevan — hanya nilai relatif untuk contoh di kelas yang berbeda.

Keuntungan lain dari Naive Bayes adalah bahwa ia secara alami merupakan "pembelajar tambahan." Pembelajar tambahan adalah teknik induksi yang dapat memperbarui model satu contoh pelatihan pada suatu waktu. Tidak perlu memproses ulang semua contoh pelatihan sebelumnya ketika data pelatihan baru tersedia.

Pembelajaran incremental sangat menguntungkan dalam aplikasi di mana label pelatihan terungkap dalam perjalanan aplikasi, dan kita ingin model untuk mencerminkan informasi baru ini secepat mungkin. Misalnya, pertimbangkan untuk membuat penggolongan email sampah pribadi. Ketika saya menerima email sampah, saya bisa mengklik tombol "sampah" di browser saya. Selain menghapus email ini dari Kotak Masuk saya, ini juga membuat titik data pelatihan: contoh positif dari spam. Akan sangat berguna jika model yang mengklasifikasikan email saya dapat diperbarui dengan cepat, dan dengan demikian segera mulai mengklasifikasikan email serupa lainnya sebagai spam. Naive Bayes adalah basis dari banyak sistem pendeteksian spam yang dipersonalisasi, seperti yang ada di Mozilla Thunderbird.

Naive Bayes tercakup dalam hampir semua teknik penambangan data dan berfungsi sebagai penggolongan garis dasar umum yang dapat dibandingkan dengan metode yang lebih canggih.

Kita telah membahas Naive Bayes menggunakan atribut biner. Ide dasar yang disajikan di atas dapat diperpanjang dengan mudah ke atribut kategoris multi nilai, serta atribut numerik.

Kesimpulan

Pada sesi sesi sebelumnya dibahas teknik pemodelan yang pada dasarnya digunakan untuk menjawab pertanyaan: "Cara terbaik apa yang dapat digunakan untuk membedakan (segmen) target?" Baik pohon klasifikasi maupun persamaan linear membuat model dengan cara ini, dengan meminimalkan kerugian atau entropi, yang merupakan fungsi diskriminabilitas. Ini disebut metode diskriminatif, di mana metode ini mencoba langsung untuk membedakan target yang berbeda.

Sesi ini memperkenalkan keluarga metode baru yang pada dasarnya bertanya tentang: "Bagaimana target yang berbeda menghasilkan nilai fitur?" Metode ini mencoba untuk memodelkan bagaimana data dihasilkan. Pada fase penggunaan, ketika dihadapkan dengan contoh baru untuk diklasifikasikan, diterapkan aturan Bayes untuk menjawab pertanyaan: "Kelas mana yang paling mungkin menghasilkan instans ini?" Dalam ilmu data pendekatan pemodelan ini disebut generatif, dan meletakan dasar bagi metode populer yang dikenal sebagai metode Bayesian yang dilandasi oleh Aturan Bayes.

Fokus utama sesi ini diletakkan pada metode Bayesian yang umum dan sederhana yang disebut penggolongan Naive Bayes. Karena kesederhanaannya itu maka metode ini sangat cepat dan efisien, namun efektif.

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.