

Sebelum kita melanjutkan ke bagian lain dari buku ini, saya akan menjelaskan kepada Anda dengan jelas. Jadi, ambil secangkir kopi dan mari kita mulai!

Jika Anda sudah familiar dengan dasar-dasar machine learning, Anda mungkin ingin langsung melompat ke Bab 2. Jika Anda belum yakin, coba jawab semua pertanyaan yang tercantum di akhir bab sebelum melanjutkan.

Apa Itu Machine Learning?

Machine learning adalah ilmu (dan seni) dalam memprogram komputer sehingga mereka dapat belajar dari data.

Berikut adalah definisi yang sedikit lebih umum:

[Machine learning adalah] bidang studi yang memberikan kemampuan pada komputer untuk belajar tanpa harus diprogram secara eksplisit.

—Arthur Samuel, 1959

Dan definisi yang lebih berorientasi pada rekayasa:

Program komputer dikatakan belajar dari pengalaman E dalam melakukan tugas T dan mengukur kinerja P, jika kinerjanya pada T, sebagaimana diukur oleh P, meningkat dengan pengalaman E.

—Tom Mitchell, 1997

Filter spam Anda adalah program machine learning yang, diberikan contoh email spam (diberi tanda oleh pengguna) dan contoh email reguler (non-spam, juga disebut "ham"), dapat belajar untuk memberi tanda spam. Contoh yang digunakan sistem untuk belajar disebut set pelatihan. Setiap contoh pelatihan disebut instansi pelatihan (atau sampel pelatihan). Bagian dari sistem machine learning yang belajar dan membuat prediksi disebut model. Jaringan saraf dan hutan acak adalah contoh model.

Dalam hal ini, tugas T adalah memberi tanda spam untuk email baru, pengalaman E adalah data pelatihan, dan pengukuran kinerja P perlu ditentukan; misalnya, Anda dapat menggunakan rasio email yang diklasifikasikan dengan benar. Ukuran kinerja khusus ini disebut akurasi, dan sering digunakan dalam tugas klasifikasi.

Jika Anda hanya mengunduh salinan semua artikel Wikipedia, komputer Anda memiliki banyak data, namun tiba-tiba tidak lebih baik dalam melakukan tugas apapun. Ini bukanlah machine learning.

Mengapa Menggunakan Pembelajaran Mesin?

Bayangkan bagaimana jika Anda menulis filter spam menggunakan teknik pemrograman tradisional. Pertama, Anda akan meneliti seperti apa biasanya tampilan spam. Anda mungkin akan memperhatikan bahwa beberapa kata atau frasa (seperti "4U", "kartu kredit", "gratis", dan "luar biasa") cenderung muncul secara sering di baris subjek. Mungkin juga Anda akan melihat beberapa pola lain dalam nama pengirim, isi email, dan bagian lain dari email tersebut. Kemudian, Anda akan menulis algoritma deteksi untuk setiap pola yang Anda perhatikan, dan program Anda akan menandai email sebagai spam jika sejumlah pola ini terdeteksi. Anda akan menguji program Anda dan mengulangi langkah-langkah 1 dan 2 sampai program itu cukup baik untuk diluncurkan.

Namun, karena masalahnya sulit, program Anda kemungkinan besar akan menjadi daftar panjang aturan yang kompleks—sulit untuk dipelihara. Di sisi lain, filter spam yang berbasis pada teknik pembelajaran mesin secara otomatis belajar kata-kata dan frasa mana yang merupakan prediktor yang baik untuk spam dengan mendeteksi pola kata yang tidak lazim dalam contoh-contoh spam dibandingkan dengan contoh-contoh ham. Program tersebut jauh lebih singkat, lebih mudah dipelihara, dan kemungkinan besar lebih akurat.

Gambar 1-2. Pendekatan pembelajaran mesin

Bagaimana jika para pengirim spam menyadari bahwa semua email mereka yang mengandung "4U" diblokir? Mereka mungkin mulai menulis "For U" sebagai gantinya. Sebuah filter spam yang menggunakan teknik pemrograman tradisional akan perlu diperbarui untuk menandai email "For U". Jika para pengirim spam terus mencari cara menghindari filter spam Anda, Anda akan perlu terus menulis aturan baru selamanya. Sebaliknya, filter spam yang berbasis pada teknik pembelajaran mesin secara otomatis memperhatikan bahwa "For U" telah menjadi tidak lazim dalam spam yang ditandai oleh pengguna, dan mulai menandainya tanpa campur tangan Anda (Gambar 1-3).

Gambar 1-3. Beradaptasi secara otomatis terhadap perubahan

Bidang lain dimana pembelajaran mesin unggul adalah untuk masalah yang terlalu kompleks bagi pendekatan tradisional atau tidak memiliki algoritma yang dikenal. Sebagai contoh, pertimbangkan pengenalan ucapan. Katakanlah Anda ingin memulai dengan hal yang sederhana dan menulis program yang mampu membedakan kata-kata "satu" dan "dua". Anda mungkin menyadari bahwa kata "dua" dimulai dengan suara tinggi ("T"), sehingga Anda bisa membuat algoritma tertanam yang mengukur intensitas suara tinggi dan menggunakannya untuk membedakan satuan dan puluhan - tetapi jelas

Teknik ini tidak akan dapat diterapkan pada ribuan kata yang diucapkan oleh jutaan orang yang sangat berbeda dalam lingkungan bising dan dalam puluhan bahasa. Solusi terbaik (setidaknya saat ini) adalah menulis algoritma yang belajar sendiri, dengan memberikan banyak contoh rekaman untuk setiap kata.

Terakhir, pembelajaran mesin dapat membantu manusia belajar. Model pembelajaran mesin dapat diperiksa untuk melihat apa yang telah dipelajarinya (meskipun untuk beberapa model hal ini bisa sulit). Sebagai contoh, begitu filter spam telah dilatih dengan cukup banyak spam, bisa dengan mudah diperiksa untuk mengungkap daftar kata-kata dan kombinasi kata-kata yang diyakininya sebagai prediktor terbaik dari spam. Terkadang ini akan mengungkap korelasi atau tren baru yang tidak terduga, dan dengan demikian membawa pada pemahaman yang lebih baik tentang masalah tersebut. Menyelami dalam jumlah data besar untuk menemukan pola tersembunyi disebut penambangan data, dan pembelajaran mesin sangat unggul dalam hal ini.

Secara ringkas, pembelajaran mesin cocok untuk:

- Masalah-masalah di mana solusi yang ada memerlukan banyak penyesuaian atau daftar aturan panjang (model pembelajaran mesin seringkali dapat menyederhanakan kode dan bekerja lebih baik dibanding pendekatan tradisional)
- Masalah kompleks di mana menggunakan pendekatan tradisional tidak menghasilkan solusi yang baik (teknik pembelajaran mesin terbaik mungkin bisa menemukan solusi)
- Lingkungan yang fluktuatif (sistem pembelajaran mesin dapat dengan mudah dilatih ulang dengan data baru, selalu menjaganya tetap terkini)
- Mendapatkan wawasan tentang masalah kompleks dan jumlah data besar

Mengapa Menggunakan Pembelajaran Mesin? | 7