Predgovor

Od kada sam držala prvi kurs o mašinskom učenju na Stanfordu 2017. godine, mnogo ljudi mi je tražilo savet kako da primene modele mašinskog učenja u svojim organizacijama. Ta pitanja mogu biti opšta, kao što su „Koji model treba da koristim?“ „Koliko često treba da obučavam svoj model?“ „Kako mogu da otkrijem promene u distribuciji podataka?“ „Kako da osiguram da su karakteristike korišćene tokom obučavanja konzistentne sa karakteristikama korišćenim tokom inferencije (predviđanja)?“

Ta pitanja mogu biti i specifična, kao na primer „Uveren sam da će prelazak sa paketnog predviđanja (batch prediction) na predviđanje u realnom vremenu (online prediction) poboljšati performanse našeg modela, ali kako da ubedim svog menadžera da mi to dozvoli?“ ili „Ja sam najstariji naučnik podataka u mojoj kompaniji i nedavno mi je povereno da postavim našu prvu platformu za mašinsko učenje; odakle da počnem?“

Moj kratki odgovor na sva ova pitanja uvek je: „To zavisi“. Moji detaljni odgovori često traži sate diskusije kako bih razumela odakle dolazi osoba koja pita, šta zaista pokušava da postigne i prednosti i mane različitih pristupa za njihov specifičan slučaj.

Sistemi mašinskog učenja su istovremeno složeni i jedinstveni. Složeni su jer se sastoje od mnogih različitih komponenti (ML algoritmi, podaci, poslovna logika, merni pokazatelji evaluacije, osnovna infrastruktura itd.) i uključuju mnogo različitih subjekata (naučnici podataka inženjeri mašinskog učenja, poslovni lideri, korisnici, pa čak i društvo u širem smislu). Sistemi mašinskog učenja su jedinstveni jer zavise od podataka, a podaci se drastično razlikuju od slučaja do slučaja.

Na primer, dve kompanije mogu biti iz iste oblasti (e-trgovina) i imati isti problem koji žele da reše pomoću mašinskog učenja (sistem preporuka), ali njihovi rezultujući ML sistemi mogu imati različitu arhitekturu modela, koristiti različite skupove karakteristika, biti evaluirani na različitim metrikama (mernim pokazateljima) i doneti različiti povraćaj investicije.

Mnogo blogova i tutorijala o mašinskom učenju fokusira se na odgovaranje na jedno specifično pitanje. Iako fokus pomaže u prenošenju poente, to može stvoriti utisak da je moguće razmatrati svako od ovih pitanja izolovano. Međutim,

promene u jednoj komponenti verovatno će uticati na druge komponente. Zato je neophodno razmatrati sistem u celini prilikom pokušaja donošenja bilo kakve projektne odluke.

Ova knjiga zauzima holistički pristup sistemima mašinskog učenja. Uzima u obzir različite komponente sistema i ciljeve različitih učesnika koji su uključeni. Sadržaj ove knjige ilustrovan je koristeći stvarne studije slučaja, mnoge od onih na kojima sam lično radila, podržan je brojnim referencama i pregledan od strane praktikanata mašinskog učenja u akademiji i industriji. Sekcije koje zahtevaju dubinsko znanje o određenoj temi – npr., obrada u paketima nasuprot obradi u stvarnom vremenu infrastruktura za skladištenje i računske operacije, odgovorna veštačka inteligencija – dalje su pregledane od strane stručnjaka čiji rad se fokusira na tu temu. Drugim rečima, ova knjiga je pokušaj da se daju nijansirani odgovori na pomenuta pitanja i više od toga.

Kada sam prvi put napisala beleške za predavanje koje su postavile temelj za ovu knjigu, mislila sam da ih pišem za svoje studente kako bih ih pripremila za zahteve njihovih budućih poslova kao naučnika podataka i inženjera mašinskog učenja. Međutim, ubrzo sam shvatila da sam i ja mnogo naučila tokom tog procesa. Početni nacrti koje sam podelila sa ranih čitaocima pokrenuli su mnogo razgovora koji su testirali moje pretpostavke, primorali me da razmotrim različite perspektive i upoznala sam se sa novim problemima i novim pristupima.

Nadam se da će se ovaj proces učenja nastaviti sada kada je knjiga u vašim rukama, jer imate iskustva i perspektive koje su jedinstvene za vas. Slobodno sa mnom podelite bilo koji komentar koji imate o ovoj knjizi, preko MLOps Discord servera koji vodim (gde možete pronaći i druge čitaoce ove knjige), Twittera, LinkedIna ili drugih kanala koje možete pronaći na mojoj veb stranici.

Kome je namenjena knjiga

Ova knjiga je za svakoga ko želi da iskoristi mašinsko učenje za rešavanje stvarnih problema. ML u ovoj knjizi odnosi se i na duboko učenje (deep learning) i klasične algoritme, sa naglaskom na sisteme mašinskog učenja na velikoj skali, kao što su oni viđeni u srednjim do velikim preduzećima i brzo rastućim startapovima. Sistemi manje skale imaju tendenciju da budu manje složeni i mogu manje koristiti od sveobuhvatnog pristupa izloženog u ovoj knjizi.

Budući da mi je struka inženjerstvo, jezik ove knjige je usmeren ka inženjerima, uključujući inženjere mašinskog učenja, naučnike podataka inženjere podataka inženjere platformi mašinskog učenja i menadžere inženjeringa. Možda se možete poistovetiti sa jednim od sledećih scenarija:

- Dobili ste poslovni problem i puno sirovih podataka. Želite da konstruišete ove podatke i izaberete prave metrike za rešavanje ovog problema.
- Vaši početni modeli dobro rade u oflajn eksperimentima i želite da ih primenite.
- Imate malo povratnih informacija o performansama vaših modela nakon što su primenjeni i želite da pronađete način da brzo otkrijete, otklonite i adresirate bilo koji problem sa kojim se vaši modeli mogu suočiti u produkciji.
- Proces razvoja, evaluacije, primene i ažuriranja modela za vaš tim je uglavnom bio ručni, spor i sklon greškama. Želite da automatizujete i poboljšate ovaj proces.
- Svaki ML slučaj korišćenja u vašoj organizaciji je primenjen koristeći sopstveni tok rada i želite da postavite temelj (npr. skladište modela, skladište karakteristika, alate za monitoring) koji se može deliti i ponovo koristiti preko slučajeva korišćenja.
- Brinete da vaši sistemi mašinskog učenja mogu imati pristrasnosti i želite da učinite vaše sisteme odgovornim.

Možete imati koristi od knjige ako pripadate jednoj od sledećih grupa:

- Razvojni programeri alata koji žele da identifikuju oblasti koje su nedovoljno iskorišćene u proizvodnji ML-a i shvate kako da pozicionirate vaše alate u ekosistemu.
- Pojedinci koji traže poslove u vezi sa ML-om u industriji.
- Tehnički i poslovni lideri koji razmišljaju o usvajanju ML rešenja za poboljšanje vaših proizvoda i/ili poslovnih procesa. Čitaoci bez jake tehničke pozadine mogli bi imati najviše koristi od Poglavlja 1, 2 i 11.

Šta ova knjiga nije

Ova knjiga nije uvod u ML. Postoje mnoge knjige, kursevi i resursi dostupni za teoriju ML-a i stoga se ova knjiga odmiče od ovih koncepta kako bi se fokusirala na praktične aspekte ML-a. Da budemo specifični, knjiga pretpostavlja da čitaoci imaju osnovno razumevanje sledećih tema:

- *ML modeli* kao što su klasterisanje, logistička regresija, stabla odluka, kolaborativno filtriranje i različite arhitekture neuronskih mreža uključujući napredne, rekurentne, konvolutivne i transformatorske
- *ML tehnike* kao što su nadgledano nasuprot nenadgledanom učenju, gradijentni spust, cilj/funkcija gubitka, regularizacija, generalizacija i podešavanje hiperparametara

- *Metrike* kao što su tačnost, F1 skor, preciznost, odziv, ROC, srednja kvadratna greška i log-verovatnoća
- *Statistički koncepti* kao što su varijansa, verovatnoća i normalna/sa dugim repom distribucija
- *Uobičajeni ML zadaci* kao što su modelovanje jezika, detekcija anomalija, klasifikacija objekata i mašinski prevod

Nije potrebno da ove teme znate u detalje – za koncepte čije se tačne definicije teže pamte, npr. F1 skor, uključujemo kratke beleške kao referencu – ali trebalo bi imati grubu ideju šta oni znače.

Iako ova knjiga pominje tekuće alate da bi ilustrovala određene koncepte i rešenja, knjiga nije udžbenik. Tehnologije se razvijaju tokom vremena. Alati brzo ulaze i izlaze iz mode, ali osnovni pristupi rešavanju problema trebalo bi da traju malo duže. Ova knjiga pruža okvir kako biste procenili koji alat najbolje funkcioniše za vaše slučajeve upotrebe. Kada postoji alat koji želite da koristite, obično je lako pronaći udžbenika za njega na internetu. Kao rezultat toga, ova knjiga sadrži malo delova koda, a umesto toga se fokusira na pružanje velikog izlaganja o kompromisima, prednostima i nedostacima i konkretnim primerima.

Kako se kretati kroz ovu knjigu

Poglavlja u ovoj knjizi su organizovana tako da odražavaju probleme sa kojima se naučnici podataka mogu susresti tokom životnog ciklusa ML projekta. Prva dva poglavlja postavljaju osnovu za uspostavljanje uspeha ML projekta, počevši od osnovnog pitanja: da li vašem projektu treba ML? Pokriva i izbor ciljeva za vaš projekat i kako postaviti vaš problem na način koji vodi ka jednostavnijim rešenjima. Ako ste već upoznati sa ovim razmatranjima i nestrpljivi da stignete do tehničkih rešenja, slobodno preskočite prva dva poglavlja.

Poglavlja 4 do 6 pokrivaju fazu ML projekta pre implementacije i primene: od kreiranja podataka za trening i inženjeringa osobina do razvoja i evaluacije vaših modela u razvojnom okruženju. Ovo je faza u kojoj su posebno potrebni ekspertiza i u ML-u i u domenu problema.

Poglavlja 7 do 9 pokrivaju implemenatciju i primenu i potonju fazu ML projekta. Saznaćemo kroz priču sa kojom se mnogi čitaoci mogu povezati da postavljanje modela nije kraj procesa primene. Primenjen model treba da bude nadgledan i kontinualno ažuriran prema promenama okruženja i poslovnim zahtevima.

Poglavlja 3 i 10 fokusiraju se na infrastrukturu potrebnu da omogući saradnicima iz različitim predznanjima da rade zajedno na isporuci uspešnih ML sistema. Poglavlje 3 se fokusira na sisteme podataka, dok se Poglavlje 10 fokusira na

kompjutersku infrastrukturu i ML platforme. Dugo sam razmišljala o tome koliko duboko da idem u sisteme podataka i gde da ih uvedem u knjigu. Sistemi podataka, uključujući baze podataka, formate podataka, premeštanje podataka i procesore obrade podataka, tendenciono su slabo pokriveni u ML kursu i stoga mnogi naučnici podataka mogu misliti o njima kao na nešto niskog nivoa ili nebitno. Posle konsultovanja sa mnogim kolegama, odlučila sam da zato što ML sistemi zavise od podataka, prethodno pokrivanje osnova sistema podataka pomoći će nam da budemo na istoj stranici kada raspravljamo o pitanjima podataka u ostatku knjige.

Iako pokrivamo mnoge tehničke aspekte ML sistema u ovoj knjizi, ML sisteme grade ljudi, za ljude i mogu imati veliki uticaj na živote mnogih. Bio bi propust napisati knjigu o ML u proizvodnji bez poglavlja o ljudskoj strani toga, što je fokus Poglavlja 11, poslednjeg poglavlja.

Zapamtite da je „naučnik podataka“ uloga koja se dosta razvijala u proteklih nekoliko godina i bilo je mnogo diskusija da se odredi šta ova uloga treba da obuhvati – ući ćemo u neka od ovih razmatranja u Poglavlju 10. U ovoj knjizi, koristimo „naučnik podataka“ (data scientist) kao opšti pojam koji obuhvata svakoga ko radi na razvoju i primenu ML modela, uključujući ljude čiji su nazivi poslova možda ML inženjeri, inženjeri podataka, analitičari podataka itd.

GitHub skladište i zajednica

Ova knjiga je praćena GitHub skladištem koji sadrži:

- Pregled osnovnih ML koncepta
- Spisak referenci korišćenih u ovoj knjizi i drugih naprednih, ažuriranih resursa
- Delove koda korišćenog u ovoj knjizi
- Listu alata koje možete koristiti za određene probleme sa kojima se možete susresti u vašim radnim procesima

Vodim i Discord server o MLOps-u gde ste pozvani da diskutujete i postavljate pitanja o knjizi.

Konvencije korišćene u ovoj knjizi

Sledeće tipografske konvencije su korišćene u ovoj knjizi:

Kurziv

Označava nove termine, URL-ove, email adrese imena fajlova i ekstenzije fajlova.

`Konstantna širina`

Koristi se za programsku listu, kao i unutar paragrafa za pozivanje na programske elemente kao što su imena varijabli ili funkcija, baze podataka, tipovi podataka, environment varijable, naredbe i ključne reči.

Ovaj element ukazuje na upozorenje ili oprez.

Upotreba primera koda

Kao što je pomenuto, dodatni materijal (primeri koda, vežbe itd.) dostupan je za preuzimanje na *https://oreil.ly/designing-machine-learning-systems-code*.

Ako imate tehničko pitanje ili problem u korišćenju primera koda, molimo da pošaljete email na *bookquestions@oreilly.com*.

Ova knjiga je ovde da vam pomogne da obavite svoj posao. Generalno, ako je primer koda ponuđen sa ovom knjigom, možete ga koristiti u svojim programima i dokumentaciji. Ne treba da nas kontaktirate za dozvolu osim ako reprodukujete značajan deo koda. Na primer, pisanje programa koji koristi nekoliko delova koda iz ove knjige ne zahteva dozvolu. Prodaja ili distribucija primera iz O'Reilly knjiga zahteva dozvolu. Odgovaranje na pitanje citiranjem ove knjige i navođenjem primera koda ne zahteva dozvolu. Uključivanje značajne količine primera koda iz ove knjige u dokumentaciju vašeg proizvoda zahteva dozvolu.

Cenimo, ali generalno ne zahtevamo, navođenje. Navođenje obično uključuje naslov, autora izdavača i ISBN. Na primer: „Mašinsko učenje: projektovanje sistema“ od Chip Huyen (Mikro knjiga). Copyright 2024 Huyen Thi Khanh Nguyen, 978-86-7555-474-5.“

Ako smatrate da vaša upotreba primera koda izlazi izvan fer upotrebe ili dozvole navedene iznad, slobodno nas kontaktirajte na *permissions@oreilly.com*.