مبانی یادگیری ماشین

نیم‌سال دوم ۰۲-۰۱

دانشکده‌ی مهندسی کامپیوتر — مدرس: دکتر حامد ملک

تمرین سری سوم — موعد تحویل: ۱۴۰۲/۰۱/۳۰

مسائل تئوری

مسئله‌ی ۱.

دو مشکل رایج در آموزش مدل های یادگیری ماشین (به طور خاص یادگیری عمیق)، گرادیان محوشونده[1] و انفجار گرادیان[2] هستند. دو مشکل و راهکاری که برای رفع هر کدام وجود دارد را توضیح دهید.

مسئله‌ی ۲.

فرض کنید در یک مجموعه داده، یکی از ویژگی ها مقدار عددی پیوسته دارد. در مساله ای که به دنبال حل آن به کمک این مجموعه داده هستیم، مقدار دقیق آن ویژگی اهمیت ندارد و ویژگی‌ها بسته به بازه‌ای که در آن قرار دارند تاثیرگذارند. چه راهکاری برای تغییر مقادیر ویژگی وجود دارد؟ توضیح دهید.

مسئله‌ی ۳.

یک مجموعه داده آموزشی در اختیار داریم که وجود یک بیماری خاص را در افراد بررسی کرده. فرض کنید در مجموعه داده 95% افراد به بیماری مبتلا نیستند و تنها 5% بیمار هستند. اگر بخواهیم از این مجموعه داده برای آموزش مدلمان استفاده کنیم با چه مشکلی ممکن است مواجه شویم و برای مقابله با آن چه راهکارهایی پیشنهاد می دهید؟

مسئله‌ی ۴.

یک مسئله تشخیص قیمت خانه داریم که در آن اطلاعات خانه‌ها از استان های مختلف آمده است. در این مسئله اگر برای رمزگذاری[3] کردن ویژگی های غیر عددی مانند استان چه کاری باید انجام بدهیم؟ اگر به هر استان یک عدد نسبت بدهیم(مثلا 10 برای تهران) چه مشکلی ممکن است رخ بدهد؟

مسئله‌ی ۵.

برای مقابله برای هر کدام از مشکلات بیش‌برازش[4] و کم‌برازش[5] سه روش نام ببرید. همچنین ضرورت استفاده از مجموعه آزمایشی در مقایسه با مجموعه اعتبارسنجی[6] را در هنگام تقسیم داده آموزشی بیان کنید.

---

[1] Vanishing Gradients
[2] Exploding Gradients
[3] Encode
[4] Overfitting
[5] Underfitting
[6] Validation

مسائل کدی

مهندسی ویژگی

در این بخش از تمرین یک مجموعه داده در اختیار شما قرار گرفته است که مربوط به تقاضای وام مشتریان از بانک است. شما پیش‌بینی کنید که آیا متقاضی وام می‌تواند آن را پس از مدت قرارداد پرداخت کند یا خیر.

معنای ستون‌های دیتاست در زیر آمده است:

- id: به هر مشتری یک آیدی اختصاص داده شده
- Initial List Status: وضعیت فهرست وام که W به معنی انتظار و F به معنی ارسال شده است
- Funded Amount Investor: مبلغ وام مورد تائید سرمایه گذار
- Total Revolving Credit Limit: سقف مجموع اعتبار گردان[۷]
- Debit to Income: نسبت بازپرداخت ماهانه بدهی تقسیم بر درآمد متقاضی
- Total Current Balance: مجموع موجودی جاری از همه حساب‌ها
- Recoveries: ریکاوری ناخالص پس از شارژ کردن
- Revolving Utilities: میزان اعتباری که یک نماینده نسبت به موجودی گردان استفاده می‌کند
- Status: وضعیت تخصیص وام، مقدار ۱ به معنی تخصیص نیافتن و مقدار ۰ به معنی تخصیص یافتن است
- Total Collection Amount: کل تعداد مجموعه هایی که به آن بدهکار است
- Revolving Balance: موجودی گردان حساب متقاضی
- Funded Amount: مبلغ تامین شده
- Total Received Interest: مجموع سود دریافتی
- Interest Rate: نرخ بهره
- Last week Pay: شخص وام گیرنده پس از ثبت نام، چه مدت زمانی اقساط ماهانه[۸] را پرداخت کرده است
- Public Record: تعداد سوسابقه عمومی
- Batch Enrolled: شماره دسته ثبت‌شده
- Open Account: تعداد حساب های باز متقاضی
- Loan Amount: مقدار وام اقدام شده
- Accounts Delinquent: تعداد حساب هایی که شخص وام‌گیرنده به آن‌ها بدهی دارد

[7] Revolving
[8] EMI

- Collection Recovery Fee: هزینه جمع آوری پس از شارژ
- Total Accounts: تعداد کل حساب‌های نمایندگان
- Application Type: حساب شخص وام گیرنده به صورت انفرادی یا مشترک است
- Delinquency - two years: تعداد روزهای بزهکاری که بیشتر از ۳۰ روز بوده است در دو سال گذشته
- Collection 12 months Medical: کل مجموعه ها در ۱۲ ماه اخیر - به جز مجموعه های پزشکی
- Months: مدت وام به ماه
- Sub Grade: زیردرجه توسط بانک
- Inquires - six months: تعداد استعلام در ۶ ماه گذشته
- Total Received Late Fee: مجموع هزینه تاخیر دریافتی

برای این بخش استفاده از مدل‌های آماده مجاز است و میتوانید از کتابخانه learn-scikit استفاده کنید. ضمناً، با توجه به اینکه انتخاب مدل محدودیتی ندارد، لازم است که نحوه ی عملکرد مدل را به صورت کامل در گزارش تمرین ذکر کنید.

ارزیابی مدل

هدف از این بخش، ارزیابی مدل پیاده‌سازی شده در قسمت قبل برای ارزیابی عملکرد مدل است.

الف) در بخش ابتدایی نیاز است تا با انتخاب معیار ارزیابی[9] مناسب، مدل خود را ارزیابی کنید. دلیل انتخاب معیار ارزیابی خود را شرح دهید و توضیح دهید چرا برای ارزیابی مدل شما معیار مناسبی است؟

ب) منحنی یادگیری[10] را رسم کنید و خروجی آن را تفسیر کنید. آیا نیاز است که مدل مجددا با پارامترهای دیگری تمرین داده شود؟ آیا مقادیر استفاده شده برای بخش‌بندی آموزش، اعتبارسنجی و آزمایش به درستی تنظیم شده‌بودند؟

نکات تمرین

- در صورت هرگونه تقلب نمره صفر برای شما لحاظ می‌گردد.
- استفاده از زبان غیر از پایتون مجاز نیست.
- تمرین تحویل حضوری خواهد داشت.

[9] Evaluation Metric
[10] Learning Curve