دانشگاه تهران، دانشکده ریاضی، آمار و علوم کامپیوتر
درس یادگیری ماشین
استاد درس: دکتر باباعلی و ساجدی

تمرین سری دوم

مهلت تحویل: پایان روز ۱۱ آبان ۱۴۰۳

طراح تمرین: ارشیا یوسفی

نکات ضروری

۱. لازم است کد تمرین به همراه گزارش تحویل داده شود که شامل پاسخ سوالات نظری، توضیحات و تحلیل کد و تصاویر و نمودارهای لازم می باشد.

۲. کد تمرین باید در زبان پایتون، به صورت notebook و شامل نتایج اجرا باشد.

۳. فایل گزارش باید به فرمت pdf باشد. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، لذا لطفا در حد نیاز توضیح دهید.

۴. استفاده از کتابخانه های از پیش آماده مجاز است، مگر اینکه نقیض آن گفته شود. در گزارش کتابخانه ها ذکر شوند.

۵. نمره هر تمرین از ۱۰۰ است (مگر آنکه نمره امتیازی در تمرین مشخص شده باشد). تا سه روز بعد از مهلت اعلام شده برای تمرین امکان بارگذاری وجود دارد اما به ازای هر روز ۱۰ درصد از نمره تمرین کسر خواهد شد.

۶. مشاهده شباهت نامتعارف در گزارش و کد به منزله تقلب می باشد و طرفین مشمول کسر نمره خواهند شد.

۷. برای این تمرین، می‌توانید گزارش و پاسخ‌هایتان را به زبان‌های فارسی و یا انگلیسی تحویل دهید.

مسئله ۱. (مسئله رگرسیون خطی با منظم‌سازی)[1]

همان‌طور که می‌دانید یکی از مشکلاتی که مدل‌های یادگیری ماشین امکان مواجهه با آن را دارند مسئله‌ی بیش‌برازش (overfitting) می‌باشد. یکی از راه‌حل‌های این مسئله اضافه کردن یکی قسمت کنترل‌کننده به تابع هزینه می‌باشد.
(الف) یکی از انواع این روش‌ها رد برازن خطی یا regression Ridge یا $l_۲$regularization- می‌باشد که تابع هزینه‌ی آن به شکل زیر است که در آن w یک بردار $n \times ۱$ و X یک ماتریس $m \times n$ می‌باشد.

$$L(w) = \|Xw - y\|_۲^۲ + \lambda \|w\|_۲^۲$$

این معادله را حل کنید (جواب فرم بسته) و توضیح دهید افزودن این ترم چگونه پیچیدگی مدل را کنترل می‌کند. مشتق آن را با مشتق رگرسیون معمول مقایسه کنید.

[۳۰ نمره]

مسئله ۲. (محاسبه رگرسیون خطی)

مجموعه داده زیر را در نظر بگیرید.

i	**میزان مطالعه** X_i	**نمره** y_i
۱	۱۶	۴۶
۲	۲۷	۸۰
۳	۱۱	۳۶
۴	۲۰	۵۲
۵	۳۰	۹۸
۶	۲۵	۷۵
۷	۵	۱۰
۸	۲۴	۷۰
۹	۲۱	۶۴
۱۰	۱۰	۳۰

۲-۱. در رابطه رگرسیون خطی شیب را در نظر نگیرید و در مدل زیر $\beta_۰$ را بیابید.

$$y = \beta_۰ + \epsilon_i$$

۲-۲. در رابطه رگرسیون عبارت عرض از مبدأ را در نظر نگیرید و در مدل زیر $\beta_۱$ را بیابید.

$$y = \beta_۱ x_i + \epsilon_i$$

۲-۳. حال فرض کنید یک مدل رگرسیون به صورت $\hat{y} = ۲۵ - ۰/۵x$ باشد. اگر یک نمره اضافی برای مشاهده جدید در $x = ۶$ به دست آمده باشد، آیا نمره آزمون برای مشاهده جدید لزوما ۲۲ خواهد بود؟ دلیل خود را توضیح دهید.

۲-۴. اگر مجموع مربعات خطا برای این مدل ۷ باشد و به اندازه ۱۶ مشاهده وجود داشته باشد، بهترین تخمین برای $\sigma^۲$ را ارائه دهید.

[۲۰ نمره]

---

[1] regression linear regularized

مسئله ۳. (پیاده‌سازی الگوریتم رگرسیون خطی)

یک مجموعه داده برای مسئله رگرسیون خطی در اختیار شما قرار گرفته است. داده‌ها به دو بخش آموزشی و تست در پوشه data تقسیم شده‌اند . در این مجموعه داده، ۱۳ ستون اول متغیرهای مستقل و ستون آخر متغیر وابسته است. معیار ارزیابی را MSE در نظر بگیرید.

۱. ابتدا داده‌ها را بر اساس مجموعه داده آموزشی استاندارد سازی کنید (از مجموعه داده تست استفاده نکنید).

$$X_{\text{standard}} = \frac{X - E[X]}{\sqrt{\text{Var}(X)}}$$

۲. ۳ رگرسیون چند جمله‌ای درجه ۱، درجه ۳ و درجه ۵ را با ۳ مقدار منظم سازی (λ) ۰.۰، ۱.۰ و ۱۰ بررسی کنید (جمعاً ۹ مدل می‌شود). سپس خطای آموزشی و خطای تست را با استفاده از تکنیک Cross-Validation Fold ۵ Repeated با تکرار ۱۰ گزارش دهید. سپس با نمایش نتایج از Boxplot (از مجموعه داده تست استفاده نکنید).

۳. نتایج را تحلیل کنید و تعیین کنید که کدام یک از مدل‌ها از نظر Underfit یا Overfit شده‌اند. همچنین با توجه به نتایج، آیا می‌توان تنظیمات بهتری ارائه داد؟ توضیح دهید.

۴. تنظیم بهینه را تعیین کنید و مدل با تنظیم بهینه را بر روی داده تست ارزیابی کنید. نتیجه را در فایلی با نام `prediction.csv` ذخیره کنید. توجه کنید که هر سطر بایستی پیش‌بینی سطر متناظر با داده تست باشد.

[۶۰ نمره]

[بارم کل: ۱۰ + ۱۰۰ نمره]

با آرزوی موفقیت