دانشگاه تهران

دانشکده علوم مهندسی

الگوریتم‌ها و محاسبات

یادگیری ماشین – دکتر سایه میرزایی

تمرین اول

اسفند ۹۹

سوال اول:

رگرسیون خطی ساده[1]، رگرسیون خطی چند متغیره[2] و رگرسیون چند جمله‌ای[3]

۱- برای **داده‌های آموزشی**[4] موجود در فایل $test.csv$ یک مدل رگرسیون خطی آموزش دهید. برای این کار **توابع خطا**[5] را با جستجو در منابع بررسی کنید (حداقل دو مورد) و سپس با استفاده از روش بهینه‌سازی گرادیان کاهشی بهترین ضرایب برای مدل را گزارش کنید.

۲- پس از یافتن ضرایب تابع، با استفاده از تابع خطایی که استفاده کرده‌اید مقدار کلی خطا را گزارش کنید.

۳- نقاط و تابع رگرسیون را رسم کنید.

۴- یک متغیر جدید به‌نام x_1 را با ضرب کردن متغیر x در عدد ۵ به‌دست آورید و سپس قسمت ۱ تا ۳ را برای داده‌های آموزشی جدید تکرار کنید.

۵- با استفاده از متغیر x^2 به‌جای x، رگرسیون چندجمله‌ای انجام دهید و قسمت ۱ تا ۳ را مجدداً تکرار کنید و نتایج آن‌را تحلیل کنید.

۶- صحت قسمت ۱ تا ۳ را با استفاده از روش تحلیلی $Normal\ Equation$ بررسی کنید. آیا ممکن است این روش مستقیم برای محاسبه ضرایب، پاسخی نداشته باشد؟ دلایل خود را توضیح دهید.

سوال دوم:

با استفاده از یک الگوریتم **جستجوی خط**[6] و **روشن نیوتون**[7] مقدار کمینه تابع $f(x)$ را به‌دست آورید. شرط توقف این الگوریتم را مشخص کرده و علت انتخاب آن را توضیح دهید سپس مقدار تابع و طول گام به‌دست آمده در هر مرحله را رسم کرده و در مورد تغییرات طول گام توضیح دهید. این روش را با گرادیان کاهشی مقایسه کنید.

$$f(x) = x^2 - 7x^3 + 8x^4 - 12$$

تفاوت روش‌های $Newton - Raphson$ و $Fisher\ Scoring$ را شرح دهید. چه زمانی این دو روش مشابه یکدیگر خواهند شد؟

---

[1] Simple Linear Regression
[2] Multivariate Linear Regression
[3] Polynomial Regression
[4] Training Data
[5] Loss Function
[6] Line Search
[7] Newton's Method

سوال سوم:

در این سوال قصد داریم با استفاده از رگرسیون چند متغیره کمیتی از یک دیتاست را پیش‌بینی کنیم. برای این منظور از این لینک در ابتدا با جزئیات دادگان Energy efficiency و ویژگی‌های هر داده آشنا شوید و آن‌ها را نرمال‌سازی کنید. قصد ما این است که دو کمیت ستون ۹ و ۱۰ که Heating Load و Cooling Load هستند، پیش‌بینی شوند. شما باید با استفاده از ویژگی‌های ۸ ستون اول بتوانید ستون ۹ و ۱۰ را پیش‌بینی کنید.

۱- هشت ستون اول هر سطر از داده‌ها را به عنوان داده آموزشی و ستون ۹ و ۱۰ را به عنوان داده هدف در نظر بگیرید، سعی کنید مدلی طراحی کنید که بتواند ستون ۹ و ۱۰ را پیش‌بینی کند. از داده ۶۰۰ به بعد به عنوان دادگان تست برای ارزیابی دقت مدلتان استفاده کنید. (تابع خطا و روش بهینه‌سازی مورد نظرتان انتخاب کرده و به صورت کامل در گزارش کار شرح دهید.)

۲- در الگوریتم‌های یادگیری ماشین یک روش مرسوم برای ارزیابی مدل ساخته شده، تقسیم کردن کل دادگان موجود به ۳ دسته می‌باشد. با دو دسته آموزش و تست آن آشنا هستیم، دسته سوم، دسته دادگان اعتبارسنجی هستند. در خصوص دادگان اعتبارسنجی تحقیق کنید و در گزارش کارتان شرح دهید.

۳- در این قسمت، بخش اول همین سوال را تکرار کنید با این تفاوت که ۱۰ درصد از کل دادگان آموزشتان را به عنوان دادگان اعتبارسنجی در نظر بگیرید. برای مثال به ازای آموزش از طریق ۴۵ داده آموزش، اعتبار مدلتان را بر روی ۵ داده بعدی به عنوان دادگان اعتبارسنجی بسنجید و به همین روال تا پایان دادگان آموزش ادامه دهید. (مقدار تابع خطا برای دادگان آموزش، اعتبارسنجی و تست را در هر مرحله گزارش کنید.)

۴- در برخی مواقع برای انجام رگرسیون با دادگان بسیار بزرگی مواجهیم که انجام محاسبات بر کل ویژگی‌های داده‌ها هم از نظر زمانی و هزینه مالی برای ما نامناسب می‌باشد. در این مواقع سعی میکنیم که از ویژگی های «موثر» دادگان برای رگرسیون استفاده کنیم. با تحقیق در منابع روش‌هایی برای شناسایی ویژگی‌های موثر داده ها پیشنهاد کنید.

۵- با استفاده از روش پیشنهادی خود در قسمت قبل سعی کنید ویژگی‌های موثر دیتاست این سوال برای پیش‌بینی ستون ۹ و ۱۰ را بیابید و بار دیگر قسمت اول را تکرار کنید و نتایج را با وقتی که از همه ویژگی‌ها استفاده کرده بودید به‌صورت کامل مقایسه کنید.

سوال چهارم:

هدف این سوال طراحی یک **طبقه‌بند لاجستیک**[8] است. داده‌های این سوال مربوط به یک بیماری خونی است و با نام «data_logistic.mat» در پیوست آمده، در این داده دو ویژگی خون «میزان گلوکز و اکسیژن خون» و در ستون سوم، لیبل مربوط به وجود بیماری در هر فرد آمده است که ۱ به معنای وجود بیماری و ۰ به معنای سالم بودن فرد مورد بررسی است. هدف این سوال طراحی طبقه‌بند برای پیش‌بینی وجود بیماری براساس این دو ویژگی خون است.

۱- نموداری رسم کنید که محورهای آن دو ویژگی داده‌ها باشد. بر روی این نمودار بیمار بودن یا نبودن را با رنگ‌های مختلف مشخص کنید.

[8] Logistic Regression

۲- می‌دانیم که با استفاده از تابع سیگموئید تابع فرضیه رگرسیون لاجستیک مطابق فرمول زیر تعریف می‌شود:

$$h_\beta(x) = \frac{1}{1 + e^{-(\beta^T x)}}$$

یکی از مهم‌ترین قسمت‌های هر طبقه‌بند **تابع هزینه**[9] است. در این سوال تابع هزینه را به‌صورت زیر تعریف می‌کنیم:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^i \log\left(h_\theta(x^i)\right) + (1 - y^i)\log\left(1 - h_\theta(x^i)\right)\right]$$

ابتدا دلیل استفاده از این تابع برای رگرسیون لاجستیک را بیان کنید و به‌طور کامل آن را بررسی کنید (ترجیحاً با رسم شکل) سپس با استفاده از گرادیان نزولی بردار θ را به گونه‌ای پیدا کنید که تابع هزینه کمینه شود. همچنین فرمول‌های محاسبه شده بر اساس گرادیان نزولی را در گزارش خود ذکر کنید. آیا می‌توان تابع دیگری به جای تابع سیگموئید استفاده کرد؟ (حداقل یک مورد را بررسی کنید.)

۳- با استفاده از L2 Norm تابع هزینه را تغییر دهید و Regularization را نیز به طبقه‌بند اضافه کنید و مجدداً θ بهینه را به‌دست آورید.

۴- بهترین دقت به‌دست آمده را گزارش کنید.

۵- چه روش‌هایی برای طبقه‌بندی بیش از دو کلاس وجود دارد؟ حداقل دو مورد را بررسی کنید.

سوال پنجم:

تعبیر احتمالاتی رگرسیون خطی و رگرسیون لاجستیک را به‌طور کامل همراه با فرمول‌ها تشریح کنید. همان‌طور که مشاهده می‌کنید انتخاب توزیع احتمالاتی رگرسیون‌های خطی متفاوت را به ما می‌دهد. با استفاده از مدل خطی تعمیم‌یافته[10] رگرسیون پواسون را به‌دست آورید (تمامی مراحل به همراه فرمول‌ها شرح داده شود).

[9] Loss Function
[10] Generalized Linear Model

نکات

- **تمرین‌ها را در سامانه ایلرن تحویل بدهید.**
- لطفا گزارش خود را به زبان فارسی تهیه کنید و تمامی نکات، فرض‌ها و فرمول‌ها در آن ذکر شوند. گزارش در روند تصحیح تمرین‌ها از اهمیت ویژه‌ای برخوردار است.
- کپی‌کردن کدهای آماده موجود در اینترنت و یا استفاده از کدهای همکلاسی‌ها تقلب محسوب می‌شود.
- استفاده از کتابخانه‌های آماده پایتون به جز *Numpy*، *Pandas* و *Matplotlib* غیرمجاز است، تنها برای بارگذاری داده‌ها mat. می‌توانید از کتابخانه‌های دیگر استفاده کنید.
- در صورت مشاهده تقلب نمرات تمامی افراد شرکت‌کننده در آن صفر لحاظ می‌شود.
- پس از به اتمام رسیدن مهلت تحویل تمرین، تاخیر تا یک هفته با کسر ۳۰ درصد نمره لحاظ خواهد شد.
- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق گروه کلاسی یا ایمیل mo.bakhtyari@ut.ac.ir با دستیار آموزشی در تماس باشید.