به نام خدا

دانشگاه تهران

دانشکدگان علوم

دانشکده ریاضی، آمار و علوم کامپیوتر

یادگیری ماشین

تمرین شماره 2

پاییز 1402

راهنمای تحویل

- **لطفا پاسخ تمرین‌هایی که سوال تئوری دارند را در یک فایل pdf و کد های خود را در قالب notebook که توضیحات لازم در آن نوشته شده است را در یک فایل زیپ به فرمت مقابل در سامانه‌ی کوئرا آپلود کنید:**

zip.[HW[Number]_[Lastname]_[StudentNumber

- **توجه کنید که اگر از ابزاری غیر از notebook استفاده کنید باید فایلی جداگانه به عنوان گزارش تمرین تهیه کنید.**
- **توجه کنید که هدف از این تمرین فرایند حل مسئله است و دقت نهایی موضوع اصلی نیست.**
- **در صورت وجود سوال می‌توانید با من از طریق (aidinkiani@ut.ac.ir)ایمیل یا گروه تلگرام در ارتباط باشید.**
- هدف از این تمرین یادگیری شماست. لطفا با صداقت و خودتان بنویسید:)

سوال 1: regularized linear regression

همان طور که می‌دانید یکی از مشکلاتی که مدل های یادگیری ماشین امکان مواجهه با آن را دارند مسئله‌ی بیش‌برازش (overfit)می‌باشد. یکی از راه‌حل‌های این مسئله اضافه کردن یکی قسمت کنترل کننده به تابع هزینه می‌باشد.

الف) یکی از انواع این روش‌ها رد برازش خطی l2-regularization یا Ridge regression می‌باشد که تابع هزینه‌ی آن به شکل زیر است که در آن w ماتریس $m \times 1$ و X یک ماتریس $n \times m$ می‌باشد.

$$L(w) = ||Xw - y||_2^2 + \lambda||w||_2^2$$

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 رگروسیون معمول مقایسه کنید.

ب) روش دیگر انجام این کار l1-regularization یا Lasso regression می‌باشد. که تابع هزینه‌ی آن به شکل زیر است:

$$L(w) = ||Xw - y||_2^2 + \lambda||w||_1^1$$

آیا جواب فرم بسته‌ای برای این معادله وجود دارد؟ در غیر این صورت الگوریتمی برای آن پیشنهاد دهید.(فرض کنید می‌توانیم به روش های عددی از تابع بالا مشتق بگیریم.(

توجه کنید که:

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

ج) توضیح دهید زیاد کردن مقدار λ چگونه در مدل تاثیر می‌گذارد. اگر λ بسیار بزرگ شود چه اتفاقی می‌افتد؟ اگر صفر شود چطور؟

سوال 2: پیاده سازی gradient descent برای مسئله‌ی تک ویژگی

دو فایل gradient.py و gradient-descent.ipynb که در فایل زیپ سوال با آپلود شده‌اند را تکمیل کنید. همچنین روابط پیدا کردن theta_1 و theta_0 را بنویسید.

امتیازی: داخل توابع فایل gd.py از حلقه استفاده نکنید.

سوال 3: انجامLinear regression

مجموعه داده‌ی داده شده مربوط به تعدادی خودرو می‌باشد که ویژگی‌هایی به شکل زیر می‌باشد:

Make: Company Of the car
Model: Name of the cae
Price: Selling Price of the car in INR
Year: Manufacturing Year of the car
Kilometes: Total kilometers Driven
Fuel TypeFuel type of the car
Transmission: Gear transmission of the car
Location: City in which car is being sold
Color: Color of the car
Owner: Number of previous owners
Seller Type: tells if car is sold by individual or dealer
Engine: engine capacity of the car in cc
Max Power: Max Power in bhp@rpm
emailMax Torque: MAx Torque in Nm@rpm
Drivetrain: AWD/RWD/FWD
Length: length of the car in mm
Width: width of the car in mm
Height: height of the car in mm
Seating Capacity: Maximum people that can fir in a car
Fuel Tank Capacity: Maximum fuel capacity of the car in litres

1- ابتدا قبل از هر چیزی داده‌ی تست خود را جدا کنید.(امتیازی: ابتدا داده‌ها را به رده‌های درآمدی تقسیم کنید و با استفاده از stratifyدر تابع train_test_split تستی متناسب با جامعه بسازید.)

2- کمی در دیتا کاوش کنید و سعی کنید آن را بفهمید. به این مرحله که از مهم ترین مراحل است EDA گفته می‌شود. به عنوان مثال می‌توانید histogram ویژگی های مختلف را رسم کنید. می‌توانید میانگینی قیمت مدل های متفاوت را به دست آورید. هر کار جالبی به نظرتان رسید را می‌توانید انجام دهید.

3- درصد خانه‌های خالی هر ستون را مشخص کنید و برای هر کدام راهی برای پر کردن آن پیشنهاد دهید و دلیل انتخاب خود را بیان کنید. توضیح مختصری در مورد انواع روش‌های پر کردن خانه‌های خالی در داده‌های عددی و غیر عددی ارائه کنید.(روش‌های متنوعی برای این کار وجود دارد ولی برای مثال می‌توانید اینجا را بخوانید.)

4- ماتریس همبستگی داده را رسم کنید. نتایج نشان داده شده را توضیح دهید.(چرا بعضی ویژگی ها همبسته هستند و اصلا همبستگی به چه معناست). مدل های یادگیری ماشین و خصوصا مدل های خطی با ویژگی های همبسته میانه‌ی خوبی ندارند. سعی کنید ویژگی‌های جدیدی از ویژگی های همبسته تولید کنید.

5- داده‌های غیر عددی را به روش مناسب encode کنید.(سعی کنید از تابع get_dummies استفاده نکنید زیرا ممکن است در test به مشکل بخورید. برای مطالعه‌ی بیشتر به اینجا مراجعه کنید.)

6- با استفاده از داده‌ی train و gridsearchCV بهترین پارامترها را برای مدل خطی عادی Lasso و Ridge به دست آورید. توجه کنید که تا قبل از مرحله‌ی ارزیابی نباید با داده‌ی test کار کنید و بهترین مدل را انتخاب کنید.

7- در مورد RMSE و R2 score تحقیق کنید و مقدار آن ها را برای داده‌ی test حساب کنید.(به عنوان مثال می‌توانید به این لینک مراجعه کنید)

8- مختصری درباره‌ی چرایی استفاده از k-fold cross validation توضیح دهید و برای مدل خود و با k = 5 مقادیر آن را محاسبه کرده و box plot آن را رسم کنید.