به نام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

تمرین سری اول یادگیری ماشین

دانشگاه تهران

پردیس دانشکده های فنی

سلام بر دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.

۲. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخ‌ها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.

۳. گزارش را حتما طبق **دستورالعمل ارسال شده**، ارسال نمایید.

٤. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.

٥. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW1_StudentNumber داشته باشد.

٦. از بین سوالات **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.

۷. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ (**۱۰ نمره امتیازی**) می توانید کسب کنید.

۸. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.

۹. در صورت داشتن سوال، از طریق ایمیل سوال خود را مطرح کنید.

سوالات ۱و ۲و ۳و۴ mahdavijoosaba@gmail.com

سوالات ۵ و۶ zeinab.yazdani@ut.ac.ir

سوال۱: (۱۵ نمره)

الف) درباره علت استفاده از اعتبار سنجی متقابل[1] و حداقل دو مورد از روش های آن توضیح دهید.

ب) متریک فاصله اقلیدسی در d بعد را در نظر بگیرید:

$$D(x,y) = \sqrt{\sum_{k=1}^{d}(x_k - y_k)^2}$$

فرض کنید عناصر هر بعد را در یک مقدار حقیقی غیرصفر ضرب می‌کنیم:

$$x'_k = a_k x_k \quad for\ k = 1,2,\dots,d$$

نشان دهید پس از ضرب نیز این متریک همچنان یک فاصله‌ی استاندارد است، یعنی ویژگی‌های یک فاصله‌ی استاندارد را دارا می باشد.

[1] Cross-Validation

سوال ۲: (۱۵ نمره)

الف) L1 Regulariaztion و L2 Regulariaztion را تعریف کرده و تفاوت های آن هارا توضیح دهید.

ب) یک مسئله رگرسیون خطی با مجموعه داده‌ی آموزشی $\{(x_1, y_1), \dots, (x_n, y_n)\}$ را در نظر بگیرید$(y_i \in \mathbb{R}, x_i \in \mathbb{R}^d)$. اگر از تابع هزینه زیر استفاده کنیم:

$$L(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2$$

که در آن λ یک ضریب ثابت مثبت است، فرم بسته مقدار بهینه w را به دست آورید.

سوال ۳: (۲۰ نمره)

در یک مسئله رگرسیون خطی، مجموعه داده‌ی $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ را در اختیار داریم. رابطه‌ی احتمالاتی میان x و y را به صورت زیر در نظر می‌گیریم:

$$y_i = wx_i + \epsilon_i$$
$$\epsilon_i = \mathcal{N}(0,1)$$

که در آن w پارامتر مدل و ϵ_i یک نویز گوسی با میانگین صفر و واریانس ۱ است.

با فرض i.i.d بودن داده‌ها، تابع log-likelihood را تشکیل دهید و نشان دهید که بیشینه کردن تابع log-likelihood روی پارامتر معادل است با کمینه کردن مجموع مجذور خطا، به عبارت دیگر نشان دهید:

$$\arg\max_w logP(D|w) = \arg\min_w \sum_{i=1}^{n} (y_i - wx_i)^2$$

سوال ۴: (۲۰ نمره)

در یک مسئله رگرسیون، می خواهیم رابطه بین ورودی و مقدار خروجی را به صورت زیر مدل کنیم:

$$y = \exp wx$$

در رابطه بالا، $y \in \mathbb{R}$ و $x \in \mathbb{R}$ است و $w \in \mathbb{R}$ پارامتر مدل است. فرض کنید مجوعه داده آموزشی $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ را در اختیار داریم.

الف) تابع هزینه مجموع مجذور خطا را برای مجموعه داده ی D تشکیل دهید.

ب) اگر بخواهیم با استفاده از روش کاهش گرادیان مقدار بهینه w را به دست آوریم، رابطه بروزرسانی w چه خواهد بود؟ به عبارت دیگر w_{t+1} چگونه از w_t به دست می آید.

ج) با انجام محاسبات نشان دهید که برای کمینه کردن تابع هزینه، مقدار بهینه پارامتر w باید در کدامیک از روابط زیر صدق کند؟

الف) $\sum_{i=1}^{n} x_i \exp wx_i = \sum_{i=1}^{n} x_i y_i \exp wx_i$

ب) $\sum_{i=1}^{n} \exp wx_i = \sum_{i=1}^{n} x_i y_i \exp wx_i$

ج) $\sum_{i=1}^{n} x_i \exp 2wx_i = \sum_{i=1}^{n} x_i y_i \exp wx_i$

سوال ۵: (شبیه سازی، ۲۰ نمره)

در این سوال الگوریتم رگرسیون خطی[1] به دو روش مختلف (گرادیان کاهشی[2] و معادلات نرمال[3]) به صورت دستی و بدون استفاده از کتابخانه، پیاده‌سازی خواهد شد. مجموعه‌داده مورد استفاده در این سوال یک مجموعه‌داده بسیار ساده شامل دو متغیر مستقل «سن» و «تجربه» و متغیر وابسته (خروجی) «درآمد» است. این مجموعه‌داده در مسیر "Datasets" در اختیار شما قرار داده شده است.

الف) EDA

داده‌ها را خوانده و ارتباط بین دو ویژگی موجود را با ویژگی خروجی بررسی کنید.

ب) پیش‌پردازش

درمورد نرمال‌سازی و استانداردسازی تحقیق کنید و باتوجه‌به به اطلاعاتی که به‌دست آورده‌اید یکی از آنها را به صورت دستی روی داده‌ها اعمال کنید.

ج) پیاده‌سازی

ج.۱) رگرسیون خطی را با گرادیان کاهشی پیاده‌سازی کنید و نتایج (خطای MSE[4] و خط به دست آمده توسط الگوریتم شما) را گزارش کنید.

ج.۲) رگرسیون خطی را با معادلات نرمال پیاده‌سازی کنید. نتایج را گزارش کرده و با قسمت قبل مقایسه کنید.

[1] Linear Regression
[2] Gradient Descent
[3] Normal Equation
[4] Mean Squared Error

سوال ۶: (شبیه سازی، ۲۰ نمره)

در این سوال الگوریتم رگرسیون لجستیک[1] بررسی خواهد شد. در این مسئله تحلیل شما از مجموعه‌داده و بررسی ویژگی‌های مختلف آن از اهمیت بالایی برخوردار است؛ بنابراین در این سوال می‌توانید از توابع آماده استفاده کنید. مجموعه‌داده مورد استفاده در این سوال مربوط به بیماری دیابت است که در ادامه اطلاعات مختصری درمورد ویژگی‌های آن ارائه شده است. این مجموعه داده در مسیر "Datasets" در اختیار شما قرار داده شده است.

جدول ۱- مشخصات مجموعه‌داده تشخیص دیابت

Pregnancies	Number of times the patient has been pregnant
Glucose	Two-hour plasma glucose concentration on an oral glucose tolerance test.
Blood Pressure	Diastolic blood pressure (mm Hg).
SkinThickness	Triceps skinfold thickness (mm).
Insulin	Two-hour serum insulin (mu U/ml).
BMI	Body mass index (weight in kg/(height in m)^2).
DiabetesPedigreeFunction/DPF	A function that assesses the likelihood of diabetes based on family history.
Age	in years.
Outcome	Class variable (0 if non-diabetic, 1 if diabetic). This is the target variable.

الف) EDA

[1] Logistic Regression

در برخورد با مجموعه‌داده‌های مختلف، بسیار مهم است که شما بتوانید به صورت دیداری اطلاعاتی را از ظاهر مجموعه داده کسب کنید تا بتوانید برای مراحل بعدی برنامه‌ریزی مناسب‌تری داشته باشید. در این قسمت برای درک بهتر دادگان، سعی کنید آنها را با ابزارهای مختلف نمایش دهید و به صورت ظاهری و نیز از نظر آماری ویژگی‌های مختلف و ارتباط آنها با خروجی را بررسی کنید. توجه کنید که تحلیل نمودارها در این سوال اهمیت بالایی دارد. بنابراین در این قسمت پس از خواندن داده‌ها سعی کنید مشخصات آن را بررسی کنید (وجود داده‌های گم‌شده[1]، بررسی پارامترهای آماری و تصویر سازی و رسم رابطه هر ویژگی با خروجی و...). بررسی کنید کدام ویژگی ها برای تصمیم گیری مفید تر هستند.

ب) پیش‌پردازش

یکی از مراحل مهم در برخورد با داده‌های دنیای واقعی، مرحله پیش‌پردازش است. در مورد پیش-پردازش‌های معمول قبل از استفاده از داده‌های خام تحقیق کنید. با ذکر دلیل بیان کنید انجام چه پیش‌پردازش‌هایی روی داده‌های این سوال به مسئله کمک می‌کند و این پیش‌پردازش‌ها را اعمال کنید. انجام درست قسمت قبل، در این قسمت به شما کمک زیادی می‌کند.

ج) طبقه‌بندی

[1] Missing data

ابتدا داده‌ها را به دو قسمت آموزش (۸۰٪) و آزمون (۲۰٪) تقسیم کنید؛ سپس مدل رگرسیون لجستیک را روی داده‌ها اعمال کنید و ماتریس درهم‌ریختگی[1]، صحت[2] و دقت[3] را برای هر دو مجموعه‌داده محاسبه کرده و نتایج را تحلیل کنید.

د) نرمال‌سازی

روش‌های نرمال‌سازی و استانداردسازی را روی داده‌ها اعمال کرده و نتایج این دو را با هم و نیز با نتایج قسمت قبل مقایسه کنید.

[1] Confusion Matrix
[2] accuracy
[3] precision