دانشگاه صنعتی شریف	**مقدمه ای بر یادگیری ماشین ۲۵۷۳۷**
دانشکده مهندسی برق	**گروه 2**
نیمسال پاییز 1401-1402	**مدرس: سید جمال الدین گلستانی**

تکلیف کامپیوتری شماره 1

موعد تحویل: یکشنبه 29 آبان 1401

توضیحات کلی

- در مورد هر تکلیف، تمام فایل‌های مربوط به سوالات کامپیوتری را در یک فایل به نام CHWiN.zip قرار دهید که i شماره تکلیف و N شماره دانشجویی شماست.
- سوالات خود را در مورد این تکلیف با دستیار آموزشی آقای امیرحسین عاملی در آدرس ایمیل amirahosseinameli@gmail.com مطرح کنید.

توجه: در دو مساله عملی این تکلیف، یادگیری بر اساس پاسخ ریاضی بدست آمده برای نقطه بهینه تابع خطای درجه دوم انجام میگیرد و برای بهینه سازی خطا از الگوریتمهای تکراری (iterative) استفاده نمیکنیم. در این دو مساله شما مجاز به استفاده از توابع و کتابخانه‌های آماده رگریسیون خطی نیستید و باید روابط ریاضی فوق الذکر را خودتان پیاده‌سازی کنید. البته میتوانید برای معکوس کردن ماتریس از توابع آماده استفاده نمایید.

در پایان فایل های نوت بوک به فرمت ipynb. را که هم شامل کدها و نتایج و هم شامل گزارش هست بفرستید. سعی کنید تمام چیزهایی که خواسته شده را داخل نوت بوک ها بنویسید اما اگر راحت تر بودید که بعضی سوالات تشریحی را به دلیل نیاز به فرمول نویسی یا موارد دیگر در Word یا ... بنویسید، می توانید این کار را انجام دهید اما در همان فایل نوت بوک بگویید که در کجا پاسخ این قسمت داده شده است.

مساله C1

این مساله ناظر به تخمین احتمال موفقیت یک داوطلب ورود به دوره کارشناسی ارشد بر اساس اطلاعاتی است که در فرم درخواست Application Form او وجود دارد. یک دیتا ست Data Set در فایل Q1_data.csv در اختیار شما قرار میگیرد که حاوی هشت ستون اطلاعات میباشد (علاوه بر ستون نخست که صرفا شماره داوطلب است). برای هر داوطلب، در ستون آخر احتمال موفقیت او که عددی بین 0 و 1 است آمده و در ستونهای یکم تا هفتم به ترتیب اطلاعات زیر قرار گرفته است:

- نمره GRE (از 340)
- نمره تافل (از120)
- کیفیت دانشگاه محل تحصیل دوره کارشناسی (از 5)
- امتیازStatement of Purpose (از 5)
- امتیاز معرفی نامه‌ها (ا ز5)
- معدل دوره کارشناسی (از 10)
- تجربه کار پژوهشی (0 یا 1)

الف- نخست بیست درصد آخر دیتاست (100 داده‌ی آخر از 502 داده) را به عنوان داده اعتبار سنجی Validation Set کنار بگذارید و تنها از هشتاد درصد نحست به عنوان داده آموزشی Training Set استفاده کنید.

ب-فرض کنید بخواهیم احتمال موفقیت را بر اساس هفت مشخصه feature فوق الذکر تعیین نماییم. بهترین بردار ضرایب W را برای مینیمم کردن خطای تجربی Empirical Risk (که به فرم Mean Square Error تعریف شده) بدست آورید.

ج- برای این بردار ضرایب، مقدار خطای تجربی را تعیین کنید. همچنین با استفاده از داده اعتبار سنجی، خطای واقعی True Risk را تخمین بزنید و با خطای تجربی بدست آمده مقایسه کنید.

اکنون فرض کنید که مساله یادگیری مورد بحث ما این باشد که احتمال موفقیت متقاضیان را بر اساس تنها یکی از هفت پارامتر فوق الذکر پیش بینی کنیم. به عبارت دیگر مایل هستیم تنها از یک مشخصه feature استفاده نماییم. برای این منظور نخست یکی از مشخصات را به عنوان بهترین مشخصه که میتواند مبنای پیش بینی قرار گیرد انتخاب میکنیم:

د- بر اساس داده آموزشی، هربار نمودار احتمال موفقیت را بر اساس یکی از مشخصه‌ها ترسیم نمایید. بدین ترتیب هفت نمودار بدست میاید که با مقایسه آنها میتوانید قضاوت خوبی نسبت به اینکه کدام مشخصه (به طور آماری) ارتباط قویتری با احتمال موفقیت متقاضیان دارد پیدا کنید. شما کدام مشخصه را انتخاب میکنید؟

ه- برای پیش‌بینی احتمال موفقیت بر حسب مشخصه‌ای که انتخاب کرده‌اید، بازهم از رگرسیون خطی استفاده میکنیم. ضرایب بهینه مربوط به رگرسیون خطی را برای این حالت بدست آورید.

و- برای این بردار ضرایب نیز مقدار خطای تجربی را تعیین کنید. همچنین با استفاده از داده اعتبار سنجی، خطای واقعی True Riskرا تخمین بزنید و با خطای تجربی بدست آمده مقایسه کنید.

ز- در نهایت خطای تجربی و تخمین خطای واقعی را که در بند قبل برای رگرسیون با استفاده از یک مشخصه بدست آمد، با آنچه در بند ج با استفاده از هر هفت مشخصه بدست آوردید مقایسه کرده مورد بحث قرار دهید.

مساله C2

در این مساله، دیتا ست مورد بحث تنها شامل یک مشخصه است که عددی حقیقی است. میخواهیم با استفاده از روش یادگیری خطی، رگرسیون چند جمله‌ای از درجه n=1 تا درجه n=15 را یادگیری نماییم و بامقایسه نتایج حاصله بهترین درجه n را برای چند جمله‌ای تعیین نماییم.

در این سوال سه دیتاست S ، V و T در اختیار شما قرار گرفته است. که به ترتیب در فایل های train_data.npy، validation_data.npy وtest_data.npy قرار دارند. به نحوی که می‌بینید، از S ، V و T به ترتیب برای آموزش، انتخاب بهترین درجه چندجمله ای و تخمین خطا برای بهترین چندجمله ای استفاده می‌کنیم.

الف ـ بر اساس داده آموزشی S ، به ازای هر یک از درجات چندجمله ای n=1 تا n=15 ، یک چندجمله ای با درجه n را با روش رگرسیون خطی یادگیری نمایید. چندجمله ای یادگیری شده به ازای n را h_n بنامید. خطای تجربی چندجمله‌ای‌های h_n برای داده آموزشی S را که با $L_S(h_n)$ نشان می‌دهیم برحسب n ترسیم نمایید.

ب ـ می‌دانید که برای انتخاب بهترین درجه چندجمله ای (n) نباید $L_S(h_n)$ را به ازای n های مختلف مقایسه کنیم (چرا؟)، تخمینی از $L(h_n)$ را بر اساس دیتاست V بدست می‌آوریم و آن را $L_V(h_n)$ می‌نامیم. در واقع $L_V(h_n)$ متوسط خطا برای نقاط دیتاست V می باشد. $L_V(h_n)$ را برای n=1,…,15 محاسبه و بر حسب n ، در کنار منحنی $L_S(h_n)$ در بند الف ترسیم کنید.

ج ـ نحوه تغییرات $L_S(h_n)$ و $L_V(h_n)$ را برحسب n با هم مقایسه کنید و علت تفاوت دو منحنی را توضیح دهید.

د ـ با استفاده از نتایج فوق نتیجه بگیرید که بهترین رگرسیون چندجمله‌ای در این مساله ازچه درجه‌ای است؟ برای این نتیجه‌گیری کدامیک از دو منحنی $L_S(h_n)$ و $L_V(h_n)$ را باید مورد استفاده قرار داد؟ چرا؟

ه ـ بهترین چندجمله ای بدست آمده در بند د را h^* بنامید. مقدار خطای حقیقی این چندجمله ای یا $L(h^*)$ را نمی‌دانیم و مجددا باید به محاسبه تخمینی از آن بسنده کنیم. به نحوی که بعدا در درس خواهید دید، $L_S(h^*)$ و $L_V(h^*)$ هیچکدام تخمین خوبی از $L(h^*)$ بدست نمی‌دهند. لذا متوسط خطای حاصل از h^* بر روی دیتاست T یا $L_T(h^*)$ را به عنوان تخمین $L(h^*)$ در نظر می‌گیریم. $L_T(h^*)$ را محاسبه نمایید.

و ـ آیا می‌توانید به طور شهودی توضیح دهید که چرا برای تخمین $L(h^*)$ خوب نیست از دیتاست V استفاده کنیم و لازم است از یک دیتاست ثالث یعنی T استفاده کرد؟

توضیح ـ دیتاست های S ، V و T را به ترتیب Training Set ، Validation Set و Test Set می‌نامیم.