دانشگاه صنعتی شریف	**مقدمه ای بر یادگیری ماشین ۲۵۷۳۷**
دانشکده مهندسی برق	**گروه ۱**
نیمسال بهار ۹۸-۹۷	**مدرس: سید جمال الدین گلستانی**

تکلیف شماره ۱

موعد تحویل: پنج شنبه ۹۷/۱۲/۹

توضیحات کلی

- در صورتی که برای عضو شدن در سایت‌های درس بر روی piazza.com و quera.ir یا برای آپلود کردن تکالیف خود دچار مشکل شدید، با آدرس ایمیل attarisadegh@yahoo.com تماس بگیرید.
- هر دو بخش کامپیوتری و تئوری هر تکلیف را بر روی سایت آپلود نمایید. تحویل به صورت کاغذی لازم نیست.
- در مورد هر تکلیف، تمام فایل‌های مربوط به سوالات کامپیوتری را در یک فایل به نام HWCiN.zip و تمام فایلهای مربوط به سوالات تئوری را در فایلی به نام HWTiN.zip قرار دهید که i شماره تکلیف و N شماره دانشجویی شماست.
- به دلیل قابلیت‌های سایت piazza.com، از این سایت برای مدیریت سوال‌های مطرح شده استفاده می‌گردد. سوالات خود را تنها از طریق این سایت بفرستید و از سایت quera.ir صرفا برای آپلود تکالیف خود استفاده کنید. در صورت ایمیل کردن تکالیف به دستیاران آموزشی، نمره‌ای به آن تعلق نمی‌گیرد.
- برای سوال‌های کامپیوتری از زبان برنامه نویسی پایتون یا متلب استفاده کنید.

سوالات تئوری

سوال T1:

فوق صفحه یا hyperplane (اختصارا HP) مشخص شده با رابطه‌ی $\omega^T x + b = 0\ ,\ \omega, x \epsilon \mathbb{R}^n\ ,\ b \epsilon \mathbb{R}$ را در فضای $X = \mathbb{R}^n$ در نظر بگیرید.

الف - نشان دهید که بردار ω بر این HP عمود است. به عبارت دیگر نشان دهید که به ازای هر دو بردار u و v در این HP، خط واصل بین u و v (یعنی بردار v-u) بر ω عمود است.

ب - نشان دهید که جهت بردار ω به سمت نیم فضای $\omega^T x + b > 0$ است. برای اینکار کافی است نشان دهید که اگر از هر نقطه x بر روی HP در جهت ω حرکت کنیم، یعنی به نقطه $u = x + \alpha\omega\ ,\ \alpha > 0$ برویم، u در نیم فضای مذکور قرار دارد.

ج – ملاحظه کنید که اگر ω را به $\omega' = \alpha\omega$ و b را به $b' = \alpha b$ تغییر دهیم که α یک عدد حقیقی است، HP تغییر نمی‌کند، اما اگر α منفی باشد، جای دو نیم فضا با هم عوض می‌شود.

د – فاصله یک نقطه دلخواه u را از فوق صفحه $\omega^T x + b = 0$ بدست آورید. با توجه به اینکه ω بر فوق صفحه عمود است، فاصله u از فوق صفحه برابر است با مسافتی که باید از نقطه u در جهت $u + \alpha\omega$ حرکت کرد تا به نقطه ای بر روی فوق صفحه رسید (α می‌تواند مثبت یا منفی باشد)

سوال T2:

فرض کنید $X = \mathbb{R}$ و $Y = \mathbb{R}$ باشد و مجموعه داده آموزشی به صورت $S = \{(0,1),(1,0),(2,4)\}$ در اختیار است. می‌خواهیم یک چندجمله‌ای درجه دوم $h(x) = a_0 + a_1 x + a_2 x^2$ بدست آوریم که بر اساس خطای mean square یعنی $l\big(h,(x,y)\big) = (h(x) - y)^2$ بهترین انطباق را با داده آموزشی S داشته باشد.

الف – تابع ریسک تجربی $L_s(h)$ را برحسب ضرایب a_2, a_1, a_0 بیان کنید.

ب – از این تابع مستقیما نسبت به ضرایب a_2, a_1, a_0 مشتق بگیرید و با صفر نهادن مشتقات و حل دستگاه معادله بدست آمده، ضرایب را بدست آورید.

ج – حال مساله را با استفاده از رابطه ماتریسی بدست آمده در درس حل نمایید و ضرایب بدست آمده را با بند 'ب' مقایسه کنید.

سوال T3:

مساله ۵ از فصل ۹ کتاب درسی

سوالات عملی

سوال C1: Linear Regression

داده‌های این سوال در فایل data_Q1.csv در اختیار شما قرار گرفته است. این داده‌ها شامل اطلاعاتی از وضعیت مسکن در مناطق مختلف کالفرنیاست و هدف ساختن مدلی برای پیشبینی قیمت مسکن بر اساس سایر ویژگی‌هاست. ابتدا ۲۵درصد از داده‌ها را برای validation (ارزیابی مدل) جدا کنید.

- نمودار قیمت برحسب ویژگی‌های مختلف را رسم کنید(در کل ۸ نمودار) به نظر شما اگه قرار باشد فقط از یک ویژگی برای پیشبینی قیمت استفاده شود این ویژگی کدام است؟(صرفا توجیه منطقی بیاورید و نیازی به اثبات نیست.)
- با استفاده از الگوریتمی که در کلاس بحث شده است مدل رگریسیون خطی را آموزش دهید و خطای تجربی empirical risk و خطای واقعی true risk این مدل را محاسبه کنید. ضرایب مدل‌های بدست آمده را در گزارش خود

ذکر کنید. توجه کنید که در این بخش مجاز به استفاده از توابع و کتابخانه‌های آماده رگریسیون خطی نیستید و باید قسمت‌های مختلف الگوریتم را خودتان پیاده‌سازی کنید.

- کد مربوط به خواندن داده ها از دیتاست داخل فایل های Question1.py و Question1.m قرار داده شده است. به دلخواه خود تنها یکی از این فایل ها را تغییر داده و ارسال نمایید. (data_Q1.csv باید کنار فایل های مذکور باشد)
- از تغییر نام فایل ها خود داری کنید.

سوال C2: Linear Regression for Polynomial Regression Tasks

برای این سوال دو دیتاست در اختیار شما قرار گرفته است. از داده‌های فایل train.csv برای اموزش مدل و از داده‌های فایل validation.csv برای ارزیابی(تخمین خطای واقعی) مدل استفاده کنید.

- به ازای هر یک از درجات چند جمله ای $n = 1$ تا $n = 15$ ، مطابق روشی که در کلاس بحث شده است، بهترین چند جمله ای عبوری از داده های آموزشی را بدست آورده و رسم کنید. برای تعیین بهترین جند جمله ای از تابع loss درجه دو استفاده کنید.
- نمودار خطای تجربی empirical risk و خطای واقعی true risk را برحسب درجه چندجمله‌ای رسم کنید.
- به ازای چندجمله‌ای با چه درجه ای خطای واقعی کمینه می‌شود؟
- علت تغیرات خطای واقعی و تجربی را نسبت به تغییر درجه چندجمله‌ای توضیح دهید.
- کد مربوط به خواندن داده ها از دیتاست داخل فایل های Question2.py و Question2.m قرار داده شده است. به دلخواه خود تنها یکی از این فایل ها را تغییر داده و ارسال نمایید. (train.csv و validation.csv باید کنار فایل های مذکور باشند)
- از تغییر نام فایل ها خود داری کنید.