دانشگاه صنعتی شریف	**مقدمه ای بر یادگیری ماشین ۲۵۷۳۷**
دانشکده مهندسی برق	**گروه ۲**
نیمسال پاییز ۰۰-۹۹	**مدرس: سید جمال الدین گلستانی**

تکلیف شماره ۱

موعد تحویل: جمعه ۲۵ مهر ۹۹

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 amirahosseinameli@gmail.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.

سوالات تئوری

سوال T۱:

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 $l(h,(x,y)) = (h(x) - y)^2$ 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 T۳:

مساله ۵ از فصل ۹ کتاب درسی

سوالات عملی

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.

برای حل تمرین های عملی به فایل HW1.ipynb مراجعه نمایید.

سوال۱C: Linear Regression

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 (۱۰۰ داده‌ی آخر از ۵۰۲ 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.

سوال C2: Linear Regression for Polynomial Regression Tasks

در این سوال 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 در اختیار شما قرار گرفته است. از داده‌های فایل train_data.npy برای اموزش مدل، و از داده‌های فایل‌های validation_data.npy وtest_data.npy برای تخمین خطای واقعی به نحوی که توضیح داده می‌شود، استفاده کنید.

الف- بر اساس داده‌های آموزشی، برای هریک از درجات رگرسیون چند جمله‌ای را برای هریک از درجات n=1 تا n=15 رگرسیون چندجمله‌ای را یادگیری نمایید و ضرایب چندجمله‌ای را در هر حالت تعیین نمایید.

ب- برای هریک از مقادیر n خطای تجربی را برای داده آموزشی و نیز تخمین خطای حقیقی را بر اساس فایل داده validation_data.npy پیدا نمایید ومنحنی تغییرات هر دو کمیت را در دیاگرامی بر حسب درجه چندجمله‌ای ترسیم کنید.

ج- نوع تغییراتی که هر یک از دو منحنی فوق بر حسب n دارد و نیز تفاوت آنها را بررسی کرده و علت را توضیح دهید.

د- با استفاده از دیاگرام فوق نتیجه بگیرید که بهترین رگراسیون چندجمله‌ای در این مساله ازچه درجه‌ای است؟ برای این نتیجه‌گیری کدامیک از دو منحنی بند ب را باید مورد استفاده قرار داد؟ چرا؟

ه- در پایان خطای حقیقی مربوط به چندجمله‌ای با بهترین درجه را بر اساس فایل داده test_data.npy تخمین بزنید. به نظر شما چرا برای این تخمین، این فایل داده بر فایل validation_data.npy رجحان دارد؟