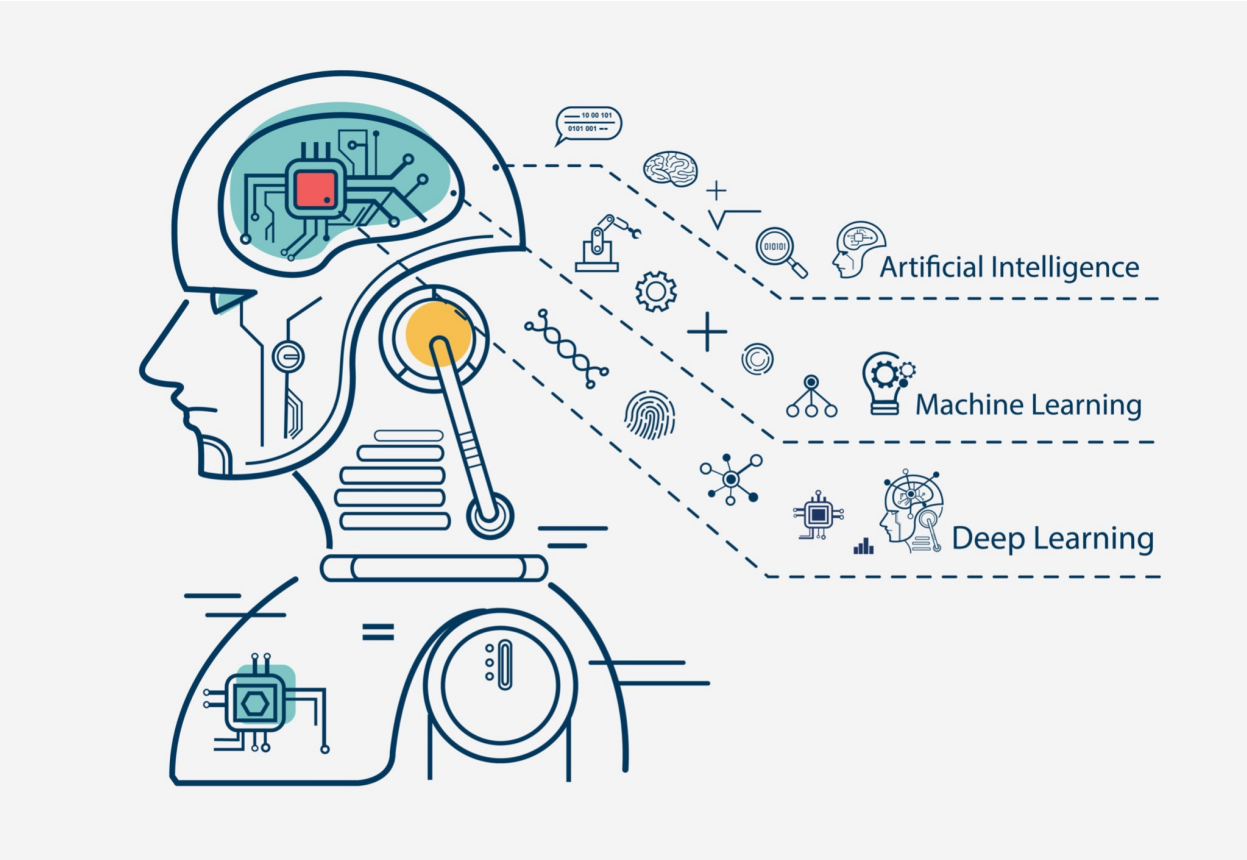

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

تمرین سری اول یادگیری ماشین

دانشکده مهندسی کامپیوتر

استاد درس: دکتر ناظرفرد

اسفند ۹۹

- تمامی مستندات شامل گزارش به همراه کدهای خود را در یک فایل فشره با فرمت zip ذخیره کرده و با عنوان stdNum_HW1.zip که stdNum شماره دانشجویی شما است در سامانه بارگزاری کنید.
- سوالات ستاره‌دار(*) نمره اضافی داشته و انجام آن‌ها اجباری نمی‌باشد.
- مهلت انجام تمرین تا ساعت **۲۳:۵۵ روز ۱۵ فروردین** می‌باشد.

سوال‌های تشریحی

سوال ۱) مفاهیم زیر را تعریف و مختصر توضیح دهید.

الف. یادگیری با نظارت[۱]

ب. یادگیری نیم‌نظارتی[۲]

ج. یادگیری بدون نظارت[۳]

د. یادگیری تقویتی[۴]

ه. یادگیری انتقالی[۵]

و. دسته‌بندی[۶]

ز. رگرسیون[۷]

ح. یادگیری برخط[۸]

ط. بیش برازش[۹]

ی. یادگیری فعال[۱۰]

ک. همبستگی[۱۱] و استقلال ویژگی‌ها[۱۲]

---

۱ Supervised Learning
۲ Semi-Supervised Learning
۳ Unsupervised Learning
۴ Reinforcement Learning
۵ Transfer Learning
۶ Classification
۷ Regression
۸ Online Learning
۹ Overfiting
۱۰ Active Learning
۱۱ Correlation
۱۲ Independence

سوال ۲) در این سوال هدف بررسی تغییرات بایاس و واریانس می‌باشد.

الف. با افزایش تعداد داده‌های آموزش، واریانس و بایاس مدل یاد گرفته شده چگونه تغییر می‌کند؟

ب. چهار راهکار برای مقابله با بیش‌برازش را بیان کنید و مختصری توضیح دهید.

سوال ۳) دو مجموعه داده از یک توزیع یکسانی نمونه‌برداری شده‌اند و در اختیار داریم. تعداد داده در یکی **۲ هزار** و در دیگری **۱۰۰ هزار** است. توسط یک الگوریتم، دو مدل جداگانه تولید میکنیم که هر کدام ۷۰ درصد داده‌ها به عنوان آموزش و ۳۰ درصد به عنوان تست در نظر گرفته شده است. نمودار خطای آموزش و تست برای دو مدل را با هم مقایسه کنید (۴ منحنی در یک نمودار رسم شود).

سوال ۴) خطای RMSE، MSE و MAE را تعریف کنید و بگویید تحت چه شرایطی از کدام خطا استفاده کنیم و برای آن دلیل بیاورید.

سوال ۵) اثر تکانه[۱۳] در روش گرادیان نزولی چیست؟ مختصر توضیح دهید. مزیت استفاده از این اثر را بیان کنید. تکانه زیاد و تکانه کم چه مشکلاتی پیش می‌أورد؟

سوال ۶) مجموعه داده آموزش شامل n داده به فرم (x_i, y_i) در اختیار داریم.(x_i، d بعدی است) تابع هزینه SSE بصورت زیر است:

$$J(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

الف. نشان دهید که رگرسیون خطی، با تابع هزینه SSE , w بهینه به صورت زیر است:

$$\hat{w} = (X^T X)^{-1} X^T y$$

ب. استفاده مستقیم از این رابطه مشکلاتی دارد. دو مشکل بالقوه‌ی استفاده‌ی مستقیم از این رابطه را ذکر کنید و راه حلی برای هر یک ارائه دهید.

ج. اگر یکی از ابعاد داده‌ها ترکیب خطی از سایر ابعاد داده‌ها باشد، با ذکر دلیل توضیح دهید که چرا نمی‌توان از رابطه بالا استفاده کرد. راه حل شما چیست؟

د. اگر یک جمله منظم ساز نرم ۲ به صورت $||w||^2$ به رابطه loss اضافه کنیم، فرم بسته w بهینه را بدست آورید. توضیح دهید اضافه کردن جمله منظم ساز چه مزیت‌هایی نسبت به فرم عادی به ما می‌دهد؟

ه. رگرسیون خطی وزن‌دار[۱۴] یک تعمیم روی رگرسیون خطی است که در آن، هر کدام از نقاط دادها یک ضریب وزن می‌گیرد:

[۱۳] momentum

[۱۴] Weighted Linear Regression

$$J(w) = \sum_{i=1}^{n} F_i (y_i - w^T x_i)^2$$

فرم بسته‌ی w بهینه را برای این تابع هزینه بدست آورید.

سوال ۷) در این سوال هدف یافتن رابطه‌ای برای روش گرادیان کاهشی[15] برای رگرسیون خطی و غیرخطی است. رابطه زیر را برای یک رگرسیون غیرخطی دو متغیره در نظر بگیرید:

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + \varepsilon \quad \text{Where } \varepsilon \sim N(0, \sigma^2)$$

الف. رابطه‌ای برای $P(y|x_1, x_2)$ بدست آورید.

ب. فرض کنید که مجموعه $\left(x_1^{(i)}, x_2^{(i)}, y^{(i)}\right) for\ i = 1, \ldots, n$ ، داده‌های آموزشی می‌باشند. رابطه $log\ likelihood$ را برای مجموعه داده‌های آموزشی بنویسید.

ج. با توجه به جواب بدست آمده در قسمت قبل، یک تابع به‌صورت $f(w_0, w_1, w_2, w_3)$ بنویسید که بتوان با مینیمم کردن آن، پارامترهای موجود را بدست آورد.

د. گرادیان $f(w)$ نسبت به بردار $w = [w_0, w_1, w_2, w_3]$ محاسبه نمایید.

[15] Gradient descent

- کدهای خود را به زبان پایتون و ترجیحا در محیط jupyter پیاده‌سازی کنید. می‌توانید تحلیل خودتان را به عنوان سلول‌های متنی در همان محیط ارائه کنید.
- نظم در نوشتن گزارش و کدها می‌تواند به کسب نمره‌ی بهتر به شما کمک کند. برنامه نوشته شده خوانا و کامنت گذاری مناسب داشته باشد.
- در پیاده‌سازی بخش‌های مختلف، امکان استفاده از کتابخانه‌های آماده مرتبط با الگوریتم‌های یادگیری ماشین را به طور کلی ندارید. **موارد مجاز در صورت سوال بخش‌ها ذکر شده است.**
- برای نمایش نمودارها و عملیات ماتریسی می‌توانید از کتابخانه‌های numpy و matplotlib استفاده کنید. همچنین برای خواندن داده‌ها به عنوان ورودی می‌توانید از pandas استفاده کنید.
- برای محاسبه معیارهای ارزیابی مانند دقت و ماتریس پریشانی می‌توانید از کتابخانه آماده استفاده کنید.
- مطابق قوانین دانشگاه هرگونه کپی‌برداری ممنوع می‌باشد و در صورت مشاهده نمره هر **دو طرف صفر** در نظر گرفته می‌شود.
- در صورت داشتن سوال می‌توانید با ایمیل تدریس یاران درس در تماس باشید:

hse.khalilian08@gmail.com , hamid.dargahi0072@gmail.com

سوال‌های پیاده‌سازی

مسئله ۱)

مجموعه داده اول: تخمین مقدار سیگنال

الف. داده‌ها را رسم کنید.

ب. با استفاده از روش گرادیان نزولی به ازای درجه‌های ۳، ۵ و ۷ نموداری بر روی داده‌ها برازش دهید و مقدار خطا را گزارش دهید. این الگوریتم را برای ۱۰۰۰ و ۱۰۰۰۰ بار تکرار کنید و در صورت مشاهده‌ی بیش‌برازش آن را گزارش دهید. این عمل را برای سه معیار خطای MSE، RMSE و MAE تکرار کنید و نتایج حاصل را مقایسه کنید.

خروجی مورد نظر: برای هر یک از موارد گفته شده یک نمودار شامل خطای آموزش و آزمون و یک نمودار برای اندازه قدم (α) ارائه کنید. در هر دو نمودار محور افقی را تکرارها و محور عمودی را مقدار خطا و اندازه گام درنظر بگیرید.

ج. قسمت قبل را با روش معادله‌ی نرمال و بدون درنظر گرفتن ضریب λ تکرار کنید و نتایج بدست آمده را مقایسه کنید.

د. به ازای درجه ۵ و با استفاده از معادله‌ی نرمال به ازای مقادیر λ برابر ۰، ۵۰ و ۵۰۰ نمودار را بر روی نقاط برازش کرده و مقدار خطای (MSE) را برای داده‌های آموزش و آزمون رسم کنید. تاثیر ضریب λ را برای بردار ضرایب θ بررسی کنید.

مسئله ۲)

مجموعه داده دوم: تخمین قیمت خانه

الف. داده‌ها را با استفاده از مختصات هرکدام رسم کنید.

ب. نمودار همبستگی بین ویژگی‌ها را رسم کنید.

ج. با توجه به نمودار به دست آمده آیا می‌توان یک یا چند ویژگی را حذف کنید. دلیل خود را ذکر کنید.

د. با استفاده از گرادیان نزولی یک نمودار بر روی داده‌ها برازش کنید. پارامترهای خود را به گونه‌ای انتخاب کنید که بهترین خروجی را بدست آورید. این عمل را با کل ویژگی‌ها و ویژگی‌های منتخب تکرار کنید. نمودار خطای آموزش و آزمون و نمودار طول گام را برای خروجی بدست آمده رسم کنید.

ه. قسمت قبل را با استفاده از معادله‌ی نرمال حل کنید.

* مسئله ۳)

مجموعه داده سوم: بررسی تعداد داده‌های آموزشی برای تخمین قیمت خانه

در فایل ضمیمه اطلاعات مربوط به ویژگی‌های ۱۲۰۰ خانه و قیمت هر کدام آمده‌است. ۶ ستون اول ویژگی و ستون آخر قیمت است.

الف. داده‌های ۱ تا ۱۰۰۰ را به عنوان داده‌های آموزش و ۱۰۰۱ تا ۱۲۰۰ را به عنوان داده‌ی تست درنظر بگیرید. در ابتدا داده‌های آموزش را ۱۰ درنظر بگیرید و با گام‌های ۱۰تایی آن را تا ۱۰۰۰ افزایش دهید و تاثیر آن را بر روی تغییرات خطا گزارش کنید.

ب. با ۵۰ داده‌ی تمرینی و استفاده از ضریب نامنظم‌ساز λ خطای تخمین روی داده‌های تست را بدست آورید. نمودار ضرایب θ رسم کرده و تغییرات آن را توضیح دهید. همچنین مقدار بهینه‌ی λ را گزارش کنید.

ج. در قسمت الف اثر تعداد داده های آموزش بر روی دقت خروج را مشاهده کردیم. در قسمت تئوری یادگیری ماشین مشاهده خواهیم کرد که اگر بخواهیم خطای تست و آموزش با احتمال $1-\delta$ بیش از ϵ از هم فاصله نداشته باشند به حداقل $N(\epsilon, \delta)$ داده نیاز داریم. در این قسمت صرفا میخواهیم شهودی نسبت به اثر تغییر تعداد داده ها، بر اطمینان ما از خطای بدست آمده داشته باشیم. مشابه قسمت قبل ، مدل را با ۱۰۰ داده آموزش میدهیم ، ولی برخلاف قبل، این کار را چندین بار انجام میدهیم و میانگین و واریانس خطاهای بدست آمده را محاسبه می‌کنیم. در هر دور ، ۱۰۰داده تصادفی از مجموعه دیتاست انتخاب می‌کنیم. (این کار را برای تعداد داده آموزش ۲۰۰ ، ۳۰۰ و... تا۱۰۰۰ انجام دهید). انتظار دارید با افزایش تعداد داده های آموزش، میانگین و واریانس خطا چطور تغییرکند؟ تغییرات میانگین و واریانس را گزارش کنید و تحلیل کنید.

موفق و برقرار باشید☺