دانشکده مهندسی کامپیوتر و فناوری اطلاعات

تمرین سری اول درس یادگیری ماشین

پاییز ۹۷

۱- شما باید سورس کد خود به همراه مستندات(پاسخ سوال‌ها و نتایج پیاده‌سازی که خواسته شده است) را در قالب یک فایل *RAR* که نام فایل *X_hw1* که *X* شماره دانشجویی شما است, تحویل دهید.

۲- پیاده‌سازی با متلب یا پایتون باید انجام شود.

۳- مهلت انجام این تمرین تا ساعت ۲۳:۵۵ روز سه‌شنبه ۲۴ مهر است.

سوال‌های تشریحی

سوال ۱) مفاهیم زیر را تعریف کنید و هر کدام را به طور مختصر توضیح دهید.

- یادگیری نظارتی
- یادگیری نیمه‌نظارتی
- یادگیری غیرنظارتی
- یادگیری تقویتی
- دسته‌بندی
- رگرسیون
- یادگیری برخط[1]
- یادگیری فعال[2]

سوال ۲) در مورد locally linear regression تحقیق کنید و در چند سطر به طور خلاصه در مورد آن توضیح دهید. (با ذکر منبع)

سوال ۳) کم کردن فضای فرضیه چه تاثیری بر بیش‌برازش دارد؟ به طور مختصر توضیح دهید.

سوال ۴) در یک شبکه عصبی, افزایش یا کاهش تعداد نرون تاثیری بر بیش‌برازش و یا بایاس بالا[3] دارد؟ به طور مختصر توضیح دهید.

سوال ۵) خطای MSE و RMSE را تعریف کنید. در دیتاستی با داده پرت و ناهنجار, استفاده از کدام‌یک بهتر است؟ چرا؟

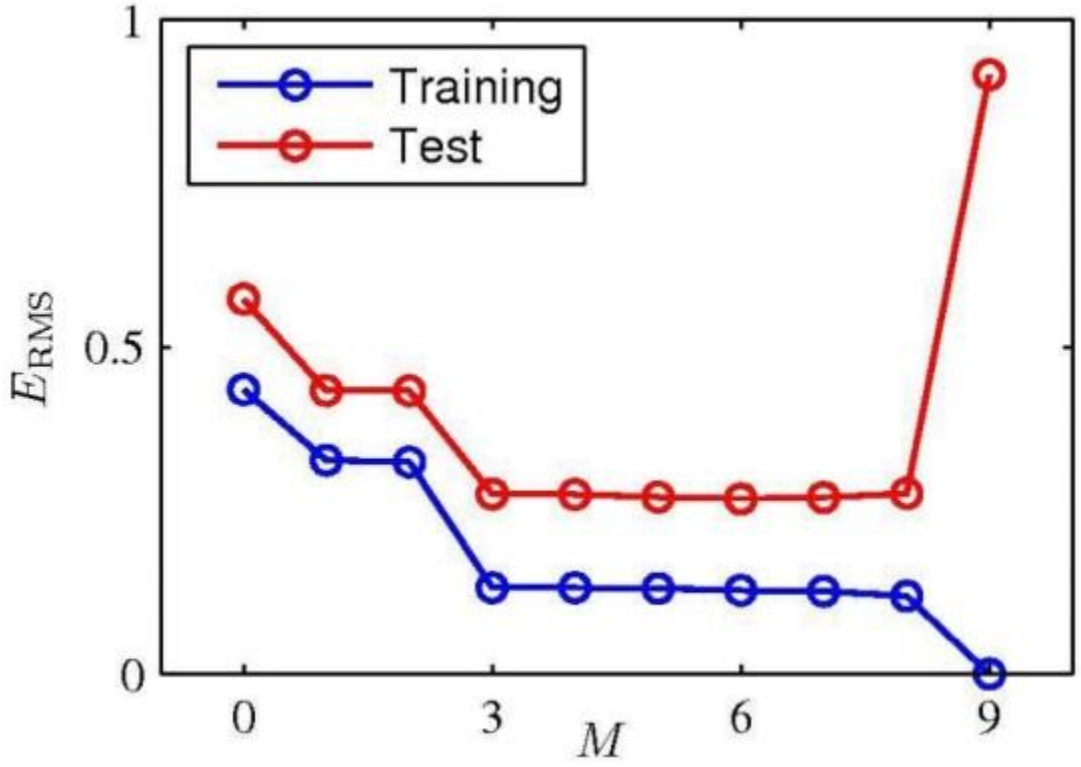
---

[1] Online learning
[2] Active learning
[3] High-bias

سوال ۶) زمانی که خطای پیش‌بینی از حد انتظار بالاتر باشد, چه کارهایی می‌توان انجام داد؟ چند مورد را بیان کنید.

سـوال ۷) اثر تکانه[۴] در روش گرادیان نزولی را توضیح دهید. مزیت و دلیل استفاده از این اثر را بیان کنید. تکانه زیاد و تکانه کم چه مشکلی پیش می‌آورد؟ برای هر دو حالت توضیح دهید.

سوال ۸) شکل زیر نشان‌دهنده خطای رگرسیون با درجه‌های مختلف چندجمله‌ای است. با توجه به این شکل در چه درجه‌هایی از M بایاس بالا و در چه درجه‌هایی بیش‌برازش اتفاق افتاده است؟



شکل ۱-۱

---

[4] momentum

سوال‌های برنامه نویسی

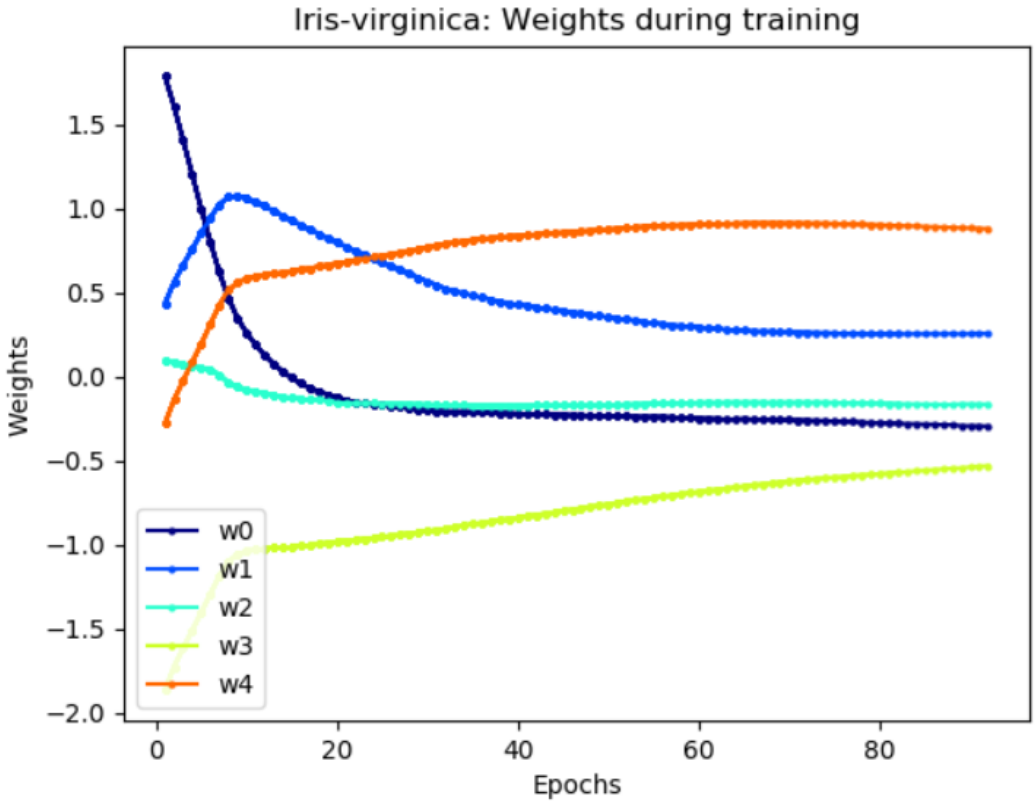
پیش از شروع به انجام سوال‌های برنامه‌نویسی یک بار متن سوال‌ها را به صورت کامل از اول تا آخر بخوانید.

ملاک اصلی انجام سوال‌های برنامه‌نویسی گزارش آن است و ارسال تمرین بدون گزارش فاقد ارزش است. برای پاسخ به هر یک از سوال‌ها، موارد زیر رعایت شود:

- کد فاقد کامنت، کاملا بی‌ارزش است و نمره‌ای به آن داده نمی‌شود.
- کدهای هر سوال باید در فایل‌هایی جدایی نوشته شوند.
- در گزارش، پاسخ به هر سوال از اول صفحه شروع شود.
- نمودارها عنوان داشته باشند و مشخص شود هر محور چه چیزی را نشان می‌دهد.
- تیترها و بخش‌هایی که برای پاسخ به هر سوال باید در گزارش نوشته شود:
  ۱) **عنوان:** در این بخش بیان کنید پاسخ به کدام یک از سوال‌ها را بیان می‌کنید.
  ۲) **شرایط آزمایش:** جدولی از پارامترها و تنظیمات خود را در اینجا بیان کنید و تا حد امکان روابط حاکم بر پارامترها و دلیل انتخاب تنظیمات انجام شده را شرح دهید.
  ۳) **نتیجه‌ی انجام آزمایش:** در این بخش نمودارهای مورد نیاز به همراه نتایج به‌دست‌آمده با توضیحات آن‌ها را ذکر کنید. (۱ تا ۲ پاراگراف)
  ۴) **نتیجه‌گیری:** نتایج حاصل از بررسی‌های خود را با دلایل آن به‌طور کامل توضیح دهید. در بیان دلایل اگر به مرجع خاصی اشاره شود بهتر است. (۲ تا ۳ پاراگراف)
- خروجی‌های خواسته شده برای هر سوال:
  ۱. نمودار خطای MSE برای مجموعه‌ی آموزش در هر بار تکرار الگوریتم(هر ایپاک).
  ۲. نمودار خطای MSE برای مجموعه‌ی ارزیابی در هر بار تکرار الگوریتم(هر ایپاک).
  ۳. نمودار وزن‌ها در هر بار تکرار الگوریتم(هر ایپاک). همه‌ی وزن‌ها(پارامترهای مدل) در یک نمودار رسم شوند، مانند نموداری دارای شباهت با شکل ۱-۲.
  ۴. نمودار منحنی برازش شده بر روی داده‌ها.
  ۵. خطای نهایی برای مجموعه‌ی آموزش، ارزیابی و تست.
  ۶. مقدار نهایی وزن‌های پیدا شده.
- مجاز به استفاده از هیچ کتابخانه آماده‌ای برای انجام کارهای خواسته شده نیستید و همه‌ی پیاده‌سازی‌ها باید توسط خودتان انجام شود. استفاده از مواردی مانند numpy و matplotlib برای رسم نمودار و

استفاده از ماتریس و آرایه‌های چند بُعدی و عملیات‌های محاسباتی مانند ضرب ماتریس‌ها، محاسبه‌ی معکوس ماتریس و عملیات‌های ریاضی مشابه مجاز است.

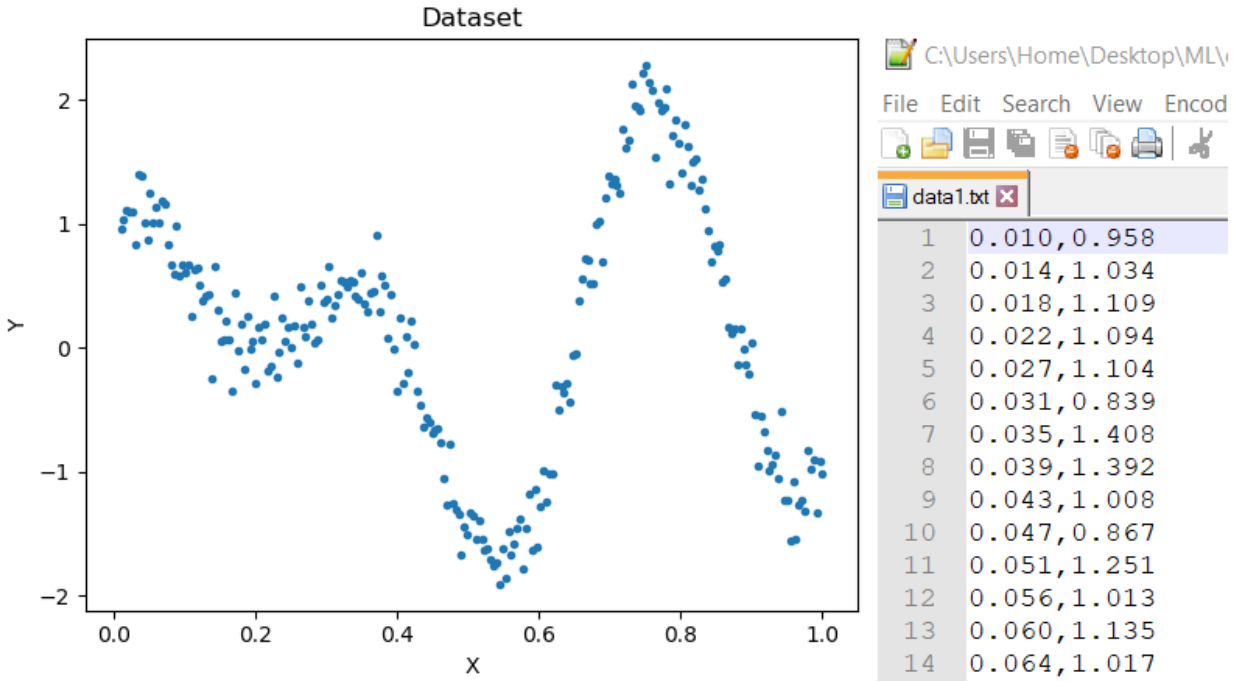
❖ مطابق قوانین دانشگاه هر نوع کپی‌برداری و اشتراک کار دانشجویان غیر مجاز بوده و شدیدا برخورد خواهد شد.



شکل ۱-۲

داده‌های موجود در فایل data1.txt را بخوانید. این فایل حاوی ۲۴۰ سطر است و از دو ستون تشکیل شده است. ستون اول از چپ مقادیر X و ستون دوم از سمت چپ مقادیر Y متناظر است. در شکل ۱-۳ بخشی از داده‌های درون این فایل و نقطه‌ها در دستگاه مختصات را مشاهده می‌کنید. برای انجام سوال‌های مختلف، دیتاست را به

سه مجموعه‌ی آموزش[5]، ارزیابی[6] و تست[7] تقسیم کنید. برای این کار می‌توانید ۶۰ درصد به آموزش، ۲۰ درصد به ارزیابی و ۲۰ درصد به تست اختصاص دهید.



شکل ۱-۳

الف) با استفاده از روش گرادیان نزولی با درجه ۵ **یا** ۷، به ازای **حداقل** سه مقدار مختلف تعداد تکرارهای الگوریتم (تعداد ایپاک)، منحنی بر نقطه‌های دیتاست برازش کنید و تاثیر تغییر تعداد تکرارهای الگوریتم را بررسی کنید. (سعی کنید حداقل سه مقدار متفاوتی که انتخاب می‌کنید به میزان قابل توجهی با هم متفاوت باشند تا تاثیر تغییر پارامتر تعداد ایپاک بر نتایج قابل مشاهده باشد.)

ب) با استفاده از روش گرادیان نزولی با درجه ۵ **یا** ۷، به ازای **حداقل** سه مقدار ضریب یادگیری مختلف، منحنی بر نقطه‌های دیتاست برازش کنید و تاثیر تغییر ضریب یادگیری را بررسی کنید. (سعی کنید حداقل سه مقدار

[5] Training Set
[6] Evaluation Set
[7] Test Set

متفاوتی که انتخاب می‌کنید به میزان قابل توجهی با هم متفاوت باشند تا تاثیر تغییر پارامتر ضریب یادگیری بر نتایج قابل مشاهده باشد.)

ج) با استفاده از روش گرادیان نزولی به ازای درجه‌های ۱،۳،۵،۷ یک منحنی بر روی نقطه‌های دیتاست برازش کنید و تاثیر افزایش درجه را بررسی کنید.

د) با استفاده از روش حداقل مربعات معمولی[8] به ازای دو درجه‌ی مختلف، منحنی بر نقطه‌های دیتاست برازش کنید. در این قسمت از سوال از بین خروجی‌های خواسته شده نیازی به رسم نمودارهای خطا و نمودار وزن‌ها نیست.

$$\theta = (x^T x)^{-1} x^T y$$

هـ) با استفاده از روش گرادیان نزولی همراه با Regularization به ازای یکی از درجه‌های ۷، ۸ یا ۹، با استفاده از حداقل سه مقدار مختلف λ، منحنی بر نقطه‌های دیتاست برازش کنید و تاثیر استفاده از مقدارهای مختلف λ را بررسی کنید. مقادیر مختلف λ چه تاثیری بر وزن‌های مدل دارد؟

و) با استفاده از یک کتابخانه آماده یک منحنی با درجه‌ی ۷ بر نقطه‌های دیتاست برازش کنید و آن را با خروجی پیاده‌سازی خودتان برای درجه‌ی ۷ که در قسمت‌های قبل آن را انجام دادید مقایسه کنید. در این قسمت از سوال از خروجی‌های خواسته شده نیازی به رسم نمودارهای خطای و نمودار وزن‌ها نیست.

[8] Ordinary Least Squares