دانشکده مهندسی برق و کامپیوتر

درس
یادگیری ماشین

تمرین کامپیوتری اول

مهلت تحویل: ۱۷ فروردین ۱۴۰۱

نیم سال دوم سال تحصیلی ۱۴۰۰ ـ ۱۴۰۱

۱. دو گوسی دو متغیره با پارامترهای زیر در نظر بگیرید:

$$\mu_1 = \begin{bmatrix} -3 \\ 5 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 3 \\ 5 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}$$

الف) برنامه‌ای بنویسید که از گوسی اول ۳۰ نمونه و از گوسی دوم ۴۰ نمونه گرفته و نمودار پراکندگی (scatter plot) آن‌ها را رسم کنید. (برای اینکار از توابع آماده کتابخانه‌های Numpy و Matplotlib استفاده نمایید.) (۱ نمره)

ب) از روی نمونه‌های قسمت الف، احتمال‌های P(X|C0)، P(X|C1)، P(X)، P(C1)، P(C0) را توسط برنامه به دست آورید. احتمال‌های P(X|C0)، P(X|C1)، P(X) را به صورت یک توزیع گوسی مدل نمایید. (۱ نمره)

ج) مجددا از هر یک از گوسی‌ها ۳ نمونه استخراج کنید و با احتمالات محاسبه شده در قسمت ب، دسته هر نمونه را مشخص نمایید. آیا با دسته‌ای که از آن استخراج شده‌اند همخوانی دارند. (۰/۵ نمره)

۲. در این تمرین قصد داریم بر روی داده‌های موجود در فایل insurance.csv یک برازش خطی به روش نزول گرادیانی انجام دهیم. این فایل شامل اطلاعات هزینه درمان اشخاص بر اساس سن، شاخص bmi و تعداد فرزندان است. آنچه مطلوب است، پیش‌بینی هزینه درمان هر شخص بر اساس سه پارامتر ذکر شده می‌باشد.

الف) هر سه روش GD، SGD و Mini_batch GD را بر روی تابع خطای میانگین مربعات پیاده‌سازی نموده و برای هر کدام از این روش‌ها، نمودار تغییرات خطا (Loss) بر روی کل داده‌ها را در هر گام به روزرسانی وزن‌ها رسم نمایید. علت اعوجاج های مشاهده شده در هر نمودار را توضیح دهید. (برای رسم نمودار توجه شود که محور عمودی نشان‌دهنده میزان خطا بر روی کل داده‌ها بوده و محور افقی نشان‌دهنده گام‌های به روزرسانی وزن‌ها است.) (۵ نمره)

ب) سه روش پیاده‌سازی شده در قسمت قبل را از نظر سرعت همگرایی و کمینه خطا با یکدیگر مقایسه کنید. (۰/۵ نمره)

د) یک‌بار دیگر عمل برازش را با در نظر گرفتن تابع خطای MAE و به روش SGD پیاده‌سازی نموده و با تابع خطای MSE مقایسه نمایید. در طول به روزرسانی وزن‌ها چند بار به نقاط مشتق ناپذیر برخورد کردید؟ اگر برخورد کردید راه حل شما چه بود؟ (۱ نمره)

ه) با آزمایش نشان‌دهید اگر نرخ یادگیری بزرگ باشد مدل همگرا نمی‌شود. (۰/۵ نمره)

و) با آزمایش نشان‌دهید در صورت کوچک بودن نرخ یادگیری، سرعت همگرایی کاهش می‌یابد. (۰/۵ نمره)

ز) با نرمال کردن داده‌های ورودی و به روش SGD یک‌بار دیگر عمل برازش را انجام داده و نمودار خطا را رسم نمایید. همچنین سرعت همگرایی را با حالت قبل (بدون نرمال سازی ورودی) مقایسه کنید. (۱ نمره)

۳. داده‌های موجود در پوشه data3 نشان دهنده درآمد یک شغل مدیریتی در طول سالیان متوالی است که به سه دسته آموزشی، ولیدیشن و تست تقسیم‌بندی شده و هر کدام در یک فایل csv جداگانه قرار دارد. ستون اول نشان دهنده سال و ستون دوم میزان حقوق است.

الف) با استفاده از توابع آماده موجود در کتابخانه scikit_learn یک خط بر روی داده‌های آموزشی برازش کنید. این خط را به همراه داده‌های آموزشی رسم کرده و میزان خطای MSE بر روی هر سه دسته داده و معادله خط را مشخص کنید. (۱ نمره)

راهنمایی ۱: به توجه به بزرگ بودن اعداد این مجموعه داده، بهتر است اعداد مربوط به سال را نرمال کرده(بین صفر و یک) و همچنین اعداد مربوط به دستمزد را بر ۱۰۰۰ تقسیم کنید.

ب) اکنون در هر مرحله با اضافه کردن یک ویژگی توان بالاتر از توان ۲ تا توان ۱۰، هر بار یک خط بر روی داده‌ها برازش کرده و علاوه بر رسم آن به همراه داده‌ها میزان خطای MSE و معادله آن خط را مشخص کنید. (۳ نمره)

راهنمایی ۲: شما باید در هر مرحله به ازای هر نقطه موجود در داده‌ها، ویژگی را به توان ۲ تا ۱۰ رسانده و به مجموعه ویژگی‌های مرحله قبل بیافزایید، به گونه ای که در مرحله آخر معادله خط به صورت زیر باشد.

$$Y = a_0 + a_1x^1 + a_2x^2 + a_3x^3 + \dots + a_{10}x^{10}$$

ج) معادلات بالا را از نظر بایاس و واریانس با یکدیگر مقایسه کنید. (۰/۵ نمره)

د) با اضافه کردن پارامتر رگولاریزیشن به معادله مرحله آخر و تنظیم مقدار آن با استفاده از داده‌ها ولیدیشن، مقدار خطا را بر روی داده‌ها تست گزارش کنید. (۱ نمره)

۴. داده‌های موجود در پوشه data4 نشان دهنده ساعات ابری بودن در طول اندازه گیری های یک بازه زمانی ۲۷ ماهه می‌باشد. این مجموعه داده به سه قسمت آموزش، ولیدیشن و تست تقسیم‌بندی شده و هر کدام در یک فایل csv قرار دارد.

الف) این داده‌ها را خوانده و با سه رنگ متفاوت نمایش دهید. (۰/۵ نمره)

ب) اکنون می خواهیم یک خط بر روی این داده‌ها برازش کنیم. با توجه به شکل خاص داده‌ها به نظر می‌رسد که نیاز باشد توسط یک تابع ابتدا ویژگی ورودی را تبدیل کرده و در فضای جدید این کار صورت پذیرد. با انتخاب تابع مناسب، خط برازش شده را به همراه داده‌ها نمایش داده و معادله آن خط را مشخص نمایید. همچنین میزان خطا میانگین مربعات را بر روی هر سه مجموعه داده مشخص نمایید. (۲ نمره)

راهنمایی: بهتر است از تابع cos(x/j) استفاده شود.

ج) با افزایش و کاهش مقدار j تغییرات خط برازش شده را به همراه داده‌ها نمایش داده و میزان خطای میانگین مربعات را اندازه گیری کنید. (بازه تغییرات j را از ۱ تا ۱۰ در نظر بگیرید) (۱ نمره)

نکات پیاده سازی و تحویل:

- انجام این تمرین به صورت یک نفره می باشد.
- در تمرین‌های برنامه‌نویسی علاوه بر کد نوشته شده، توضیح کد و همچنین تحلیل نتایج از اهمیت بالایی برخوردار است. برای این منظور می‌توانید در فایل ژوپیتر مربوط به کد خود و یا در یک فایل پی دی اف جداگانه، اطلاعات کامل از نحوه پیاده‌سازی، نتایج حاصله، علت نتایج، مطابقت نتایج با انتظارات و مقایسه ها را بنویسید.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است. لطفاً تمامی نکات و فرضهایی که برای پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید را در گزارش ذکر کنید.
- از نوشتار محاوره‌اي خودداري نمائيد.
- برای انجام تمرینها استفاده از زبان برنامه‌نویسی پایتون الزامی می‌باشد.
- در تمرین های برنامه نویسی حتما پیاده سازی خود را در محیط Jupyter Notebook و در یک فایل ipynb انجام دهید.
- در فایل Jupyter Notebook هر سوال از تمرین به همراه پاسخ آن مشخص شده و خروجی‌های مورد نیاز نیز ذخیره شده باشد . همچنین هرگونه نتیجه و یا تحلیلی که در شرح سوال از شما خواسته شده است را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر می‌شود.
- بعد از تکمیل پاسخ‌ها در فایل Jupyter Notebook مجددا kernel را راه‌اندازی کرده و فایل را اجرا نمایید به صورتیکه شماره‌ی هر سلول در فایل دقیقا مطابق با ترتیب سلول ها باشد.
- تکالیف کامپیوتری تا یک هفته بعد از موعد مقرر قابل تحویل مي‌باشند و به ازاء هر روز تأخیر 7 % از نمره کل کسر مي‌گردد.
- در صورت مشاهده تقلب امتیاز تمامی افراد شرکت‌کننده در آن، صفر لحاظ می‌شود.
- در صورت وجود سوال و یا ابهام میتوانید از طریق آی‌دی ها تلگرام زیر با دستیاران آموزشی در ارتباط باشید:
  @basir_ebr
  @givkashi
  @hamidravaee