به نام خدا

تکلیف اول یادگیری ماشین

نیمسال تحصیلی ۰۱ –۰۰

موعد تحویل: ۱۰ آبان ساعت ۲۳:۵۹

---

۱. داده‌های موجود در فایل data1.csv نقاط متعلق به دو توزیع نرمال با پارامترهای بیان شده می‌باشند. ستون اول مقادیر x1 و ستون دوم نشان‌دهنده مقادیر x2 هستند. ابتدا احتمال تعلق نقاط به هر یک از دو توزیع ذکر شده را محاسبه کرده و داده را به گوسی با بیشترین احتمال منتسب کنید.

الف) نقاط مربوط به توزیع اول را با قرمز و نقاط مربوط به توزیع دیگر را با آبی نمایش دهید. (برای نمایش نقاط می‌توانید از کتابخانه Matplotlib استفاده نمایید.) (۳ نمره)

ب) تابع شباهت (Likelihood Function) و لاگ تابع شباهت (Log Likelihood Function) را برای هر دو توزیع نرمال پیاده‌سازی کنید. (۲ نمره)

$\mu_1 = [\ 5\ ,\ 5\ ]$
$\mu_2 = [\ 2\ ,\ 2\ ]$
$\Sigma_1 = \Sigma_2 = [\ [\ 1\ ,\ 0\ ]\ ,\ [\ 0\ ,\ 1\ ]\ ]$

$$p(\boldsymbol{x}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right)$$

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.) (۷ نمره)

ب) سه روش پیاده‌سازی شده در قسمت قبل را از نظر سرعت همگرایی و کمینه خطا با یکدیگر مقایسه کنید. (۱ نمره)

ج) برای روش GD، برازش را یکبار با پارامترهای اولیه صفر و بار دیگر با مقدار دهی تصادفی انجام داده و این دو روش مقدار دهی را با یکدیگر مقایسه کنید. (۱,۵ 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؟(۲ نمره)

ه) با آزمایش نشان‌دهید اگر نرخ یادگیری بزرگ باشد مدل همگرا نمی‌شود. (۱ نمره)

و) با آزمایش نشان‌دهید در صورت کوچک بودن نرخ یادگیری، سرعت همگرای‌ی کاهش می‌یابد. (۱ 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.(۱,۵ نمره)

ملاحضات:

۱) حتما پیاده‌سازی‌های خود را در محیط Jupyter Notebook به زبان Python انجام دهید.

۲) در این تمرین لازم است توابع مورد نیاز را خودتان به زبان Python پیاده‌سازی نمایید، لذا برای این منظور از توابع آماده کتابخانه‌هایی همچون scikit-learn استفاده ننمایید.

۳) در صورت مشاهده تکالیف کپی بین دو دانشجو، به هر دو نفر نمره صفر داده می‌شود.

۴) نیازی به فایل پی دی اف جداگانه برای گزارش نمی‌باشد. توضیحات خود را در همان فایل ipynb و با ایجاد سلول جدید از نوع Markdown و به زبان فارسی بنویسید.

۵) در صورت داشتن هرگونه ابهام یا سوال می‌توانید با دستیاران آموزشی درس در ارتباط باشید و یا سوالات خود را در گروه تلگرامی درس مطرح کنید.