تمرین شماره 1

شما باید نسخه pdf گزارش مربوطه را قبل از مهلت مقرر ارسال کنید.

اگر شما راه حل های خود را دست نویس کنید ، باید صفحات را اسکن کنید ، آنها را در یک فایل pdf ادغام کنید و ارسال نمایید.

باید توضیحات خود را در گزارش بصورت دقیق بیان نمایید. کلمات غیر ضروری باعث از کم شدن نمره می شود.

شما می توانید از کتابخانه هایی با ابزارهای عمومی عمومی مانند matplotlib ، numpyو scipy برای پایتون استفاده نمایید با این حال ، شما باید الگوریتم ها را خودتان پیاده سازی کنید ، به این معنی که از پیاده سازی الگوریتم های موجود درSciKit learn و Tensorflow استفاده نکنید.

فقط می توانید از پایتون 3 استفاده کنید و باید راه حل خود را jupyter note-book ارسال کنید.

مطمئن شوید که تمام فایل های داده مورد نیاز برای اجرای کد شما در داخل پوشه و بارگذاری شده با مسیر قراردارد. کد باید بدون هیچ گونه تغییری اجرا قابل اجرا باشد.

دستورالعمل ارائه گزارش

1. گزارش شما باید مختصر و دقیق باشد.

2. کلیه بصری سازی ها (منحنی های یادگیری ، تناسب رگرسیون) را گزارش دهید.

3. کد خود را در گزارش وارد نکنید!

نمونه گیری

برنامه روزانه یک دانش آموز فارغ التحصیل به عنوان یک توزیع چند جمله ای ، p ، در مجموعه فعالیتهای زیر تعریف می شود:

• فیلم: 0.2

• INF8245E: 0.4

• بازی: 0.1

• مطالعه: 0.3

1. هر روز صبح ، دانش آموز از خواب بیدار می شود و به طور تصادفی از فعالیت های انجام شده برای بقیه روز نمونه می گیرد. به شرطی که فقط از توزیع یکنواخت روی (0،1) نمونه برداری کنید ، یک pseudocode برای نمونه از توزیع چند جمله ای داده شده بنویسید.

2. الگوریتم نمونه گیری خود را پیاده سازی کرده و از آن برای نمونه گیری از روتین دانشجویی به مدت 100 روز استفاده کنید. کسری (fraction) از روزهای سپری شده در هر فعالیت را گزارش کنید. اکنون از آن برای نمونه برداری به مدت 1000 روز استفاده کنید. کسری از روزهای سپری شده در هر فعالیت را گزارش کنید. این کسرها را با توزیع چند جمله ای اصلی مقایسه کنید.

انتخاب مدل

برای این آزمایش باید از Dataset-1 استفاده کنید. Dataset-1 شامل قطار ، اعتبار سنجی و فایل های آزمایشی است. ورودی و خروجی مقادیر اسکالر حقیقی هستند. مجموعه داده از یک چند جمله ای درجه n ایجاد می شود و یک نویز کوچک گوسی (small Gaussian noise) به هدف اضافه می شود.

1. یک چندجمله ای 20 درجه به داده ها fit نمایید.

(الف) آموزش و اعتبارسنجی RMSE (خطای میانگین مربع ریشه) را گزارش نمایید. از هیچگونه نظم دهی (regularization) استفاده نکنید.

(ب) مدل فیت شده را نمایش دهید (Visualize the fit).

(ج) آیا مدل overfitting یا underfitting ؟ چرا؟

2. اکنون منظم سازی L2 (regularization) را به مدل خود اضافه کنید. مقدار λ را از 0 تا 1 ، با گام 0.01 تغییر دهید.

(الف) برای مقادیر مختلف λ ، آموزش (training) RMSE و اعتبارسنجی RMSE را ترسیم کنید.

(ب) بهترین مقدار λ را بیابید و عملکرد تست (test performance) را برای مدل مربوطه گزارش دهید.

(ج) مدل فیت شده برای مدل انتخابی را نمایش دهید (Visualize the fit).

(د) آیا مدل overfitting یا underfitting ؟ چرا؟

3. به نظر شما درجه چند جمله ای اصلی (source polynomial) چند است؟ آیا می توانید از بصری سازی ارائه شده در سوال قبلی چنین نتیجه گیری کنید؟

Gradient Descent برای رگرسیون (Gradient Descent for Regression)

برای این آزمایش باید از Dataset-2 استفاده کنید. Dataset-2 شامل قطار ، اعتبار سنجی و فایل های آزمایشی است. ورودی و خروجی مقادیر اسکالر حقیقی هستند.

1. با استفاده از stochastic gradient descent (یک مثال در زمان) یک مدل رگرسیون خطی را به این مجموعه داده fit کنید.

(الف) با استفاده از گام 10^{-4}، آموزش و اعتبارسنجی RMSE (training and validation RMSE) را در برابر تعداد دوره ها (epochs)، تا زمان همگرایی ترسیم کنید.

2. اندازه گام های مختلف را امتحان کنید و با استفاده از داده های اعتبار سنجی ، بهترین گام را انتخاب کنید.

الف) در جدول عملکرد اعتبارسنجی (validation performance) با اندازه گام های مختلف گزارش کنید.

(ب) RMSE تست مدل انتخاب شده را گزارش دهید.

3. 5 تصویرسازی مختلف را که به طور تصادفی انتخاب شده اند برای نشان دادن چگونگی تکامل رگرسیون در طول روند آموزش گزارش دهید.

4. قسمت 1 را با استفاده از full-batch gradient descent تکرار کنید.

5. بر اساس آزمایش های خود تفاوت بین full-batch gradient and stochastic gradient descent را بیان کنید.

مجموعه داده های واقعی

برای این سوال، شما از مجموعه داده های سایت زیر استفاده خواهید کرد.

(http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime).

1. این یک مجموعه داده واقعی است و به همین دلیل ، ویژگی های خوبی را که انتظار داریم نخواهد داشت. اولین کار شما این است که این مجموعه داده را با پر کردن تمام مقادیر خالی قابل استفاده کنید.

(الف) از میانگین نمونه هر ستون برای پر کردن ویژگی های از دست رفته استفاده کنید. آیا این انتخاب خوبی است؟ توضیح دهید.

(ب) از چه چیز دیگری می توانید برای پر کردن ویژگی های از دست رفته استفاده کنید؟

(ج) اگر روش بهتری دارید ، آن را شرح دهید و از آن برای پر کردن داده های از دست رفته استفاده کنید.

توضیح دهید که چرا روش شما بهتر است.

(د) مجموعه داده های تکمیل شده را وارد کنید.

2. از 20٪ اول مجموعه داده برای آزمایش (تست) استفاده کنید و از 80٪ باقی مانده برای آموزش (training) به ترتیبی که در فایل مجموعه داده شده است استفاده کنید.

(الف) میانگین RMSE, 5-fold cross-validation را گزارش دهید.

(ب) آزمایش RMSE (test RMSE.) را گزارش دهید.

3. اکنون از Ridge-regression در داده های بالا استفاده می کنیم.

(الف) برای انتخاب بهترین λ ، میانگین RMSE را با استفاده از 5-fold cross-validation، برای مقادیر مختلف λ [محور x را λ و محور y را میانگین RMSE در نظر بگیرید] ترسیم کنید. نحوه انتخاب محدوده λ برای کاوش را توضیح دهید.

(ب) کدام مقدار λ بهترین fit را دارد؟

(ج) RMSE تست را با استفاده از مقدار λ که انتخاب کرده اید گزارش دهید.

(د) آیا می توان از اطلاعات به دست آمده در طول این آزمایش برای انتخاب ویژگی (feature) استفاده کرد؟ اگر چنین است ، چگونه توضیح دهید؟

(ه) RMSE تست بهترین fit را که با کاهش مجموعه ای از ویژگی ها بدست آورده اید ، گزارش دهید؟

(و) عملکرد مدل با ویژگی های کاهش یافته در مقایسه با مدل با استفاده از همه ویژگی ها چقدر متفاوت است؟ در مورد تفاوت نظر دهید.