|  |  |  |
| --- | --- | --- |
|  | به نام خدا |  |
| **دانشگاه تهران**  **دانشکده‌ مهندسی برق و کامپیوتر**  **شبکه های عصبی و یادگیری عمیق**  **مینی‌پروژه سری ۲** | | |

|  |  |
| --- | --- |
| مهسا قزوینی نژاد - نازنین صبری | نام و نام خانوادگی |
| ۸۱۰۱۹۸۳۲۰ - ۸۱۰۱۹۸۳۱۲ | شماره‌ دانشجویی |
| ۱۳ خرداد ۱۳۹۹ | تاریخ ارسال گزارش |

­

# سوال ۲ – نقصان دادگان

سوال ۱

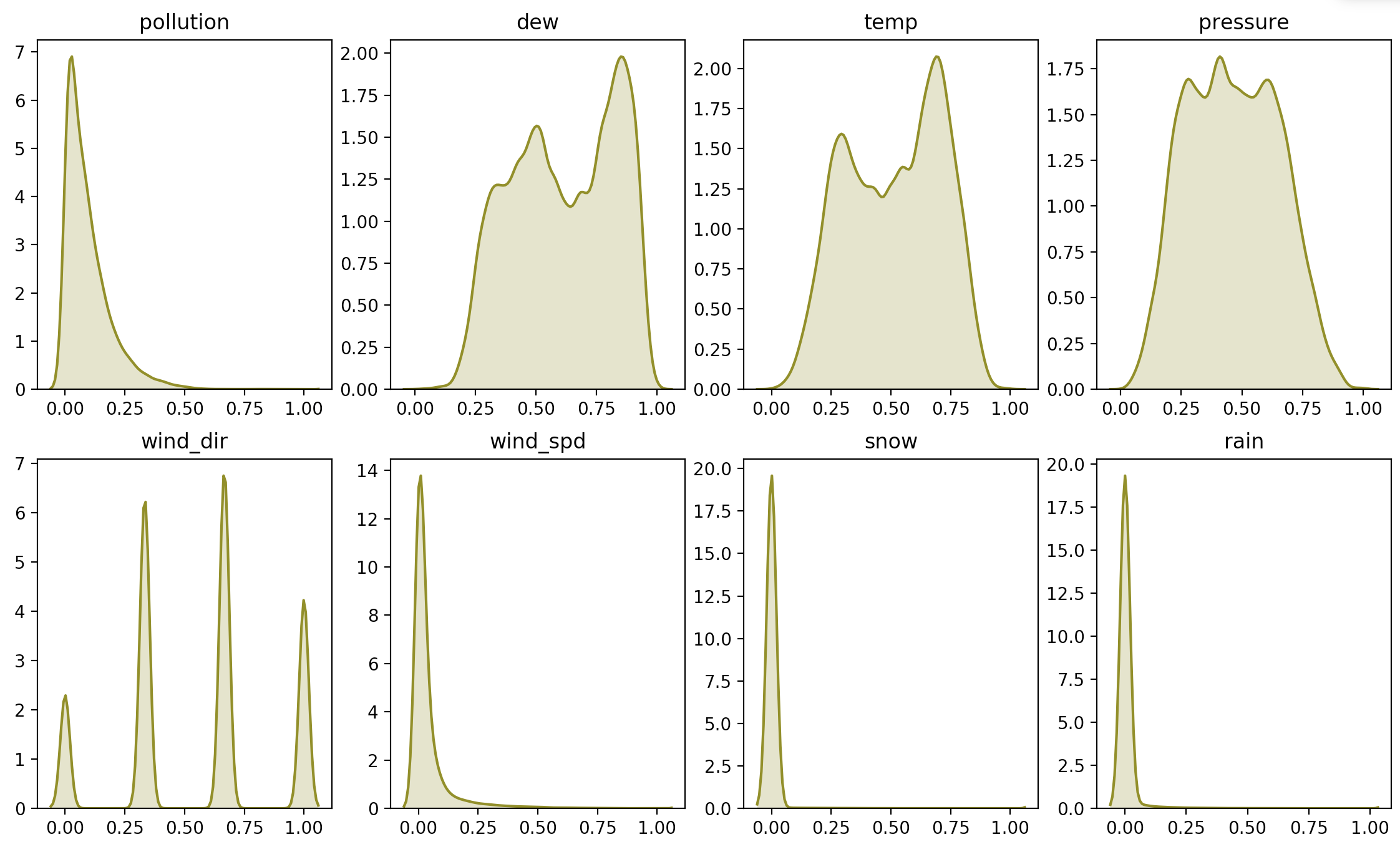
در ابتدای این سوال برای هر ستون به صورت مجزا، ۲۰ درصد از دادگان را حذف می‌کنیم. در مسئله‌های واقعی ممکن است به دلیل خرابی ابزار اندازه گیری و گاه خطای انسانی دادگان درست جمع آوری نشود یا از بین برود. برای بررسی شرایط مقابله با این مشکل در این سوال به این شکل بخشی از دادگان خود را از دسترس خارج می‌کنیم. در این قسمت به این صورت عمل می‌کنیم که به ازای هر ستون از دادگان تست آرایه‌ای تصادفی که هر خانه آن بین ۰ و ۱ است تولید می‌کنیم. سپس به ازای هر خانه از این جدول اگر مقداری کمتر از ۲۰ درصد یا همین tresh داشت آن را با مقدار np.nan جای‌گذاری می‌کنیم تا داده را از دست رفته نشان دهیم.

سوال ۲ و ۳

حال فرض می‌کنیم که داده‌‌های قسمت قبل را در دست داریم که در هر ستون شامل ۲۰ درصد np.nan می‌باشد. روش‌های مختلفی برای برطرف کردن نقصان داده یا از دست رفتن دادگان وجود دارد که شامل موارد زیر می‌باشد:

1. یک روش این است که کار خاصی برای پیش‌بینی دادگان از دست رفته انجام ندهیم و تنها آن داده‌ها را پاک کنیم (در برخی توابع نیز گزینه ignore کردن و در نظر نگرفتن این قبیل دادگان توسط الگوریتم وجود دارد.) این روش ساده است اما باعث از دست رفتن دادگان زیادی می‌شود و شاید برای یک ستون که تاثیری نیز در مقدار خروجی ما ندارد بخش زیادی از دادگان خود را از دست دهیم.
2. روش دیگر استفاده از مقدار صفر یا ثابت برای دادگان از دست رفته می‌باشد. مقدار صفر معمولا انتخاب بهتری است اما ممکن است در برخی از ستون‌‌ها کاربرد نداشته باشد مثلا انتخاب مقدار صفر برای ویژگی سن شاید بی معنی به نظر برسد. استفاده از این روش سریع و ساده است و برای ویژگی‌های دسته‌بندی نیز می‌تواند خوب عمل کند اما ممکن است باعث به وجود بایاس در نتیجه پایانی شود.
3. روش دیگر استفاده از مقادیر میانگین یا میانه می‌باشد. این مقادیر را باید به کمک دادگان از دست نرفته محاسبه کنیم و در نهایت مقادیر به دست آمده را جایگزین nan ها کنیم. این الگوریتم هم نسبتا ساده است و مثلا انتخاب ویژگی‌ میانگین باعث می‌شود که هر چه قدر هم که حجم از دست رفتن دادگان زیاد باشد، میانگین جا به جا نشود و توزیع آماری ویژگی تغییر زیادی نکند اما در مقابل این روش در ویژگی‌ها و ستون‌های مربوط به categorical قابل استفاده نمی‌باشد و دادگان آن ستون باید عددی باشد تا این معیارها قابل محاسبه باشد. در دادگان بزرگ می‌تواند خیلی دقیق نباشد و ارتباط بین ویژگی‌های مختلف را نمی‌تواند لحاظ کند.
4. روش دیگر knn است. در این روش از شباهت دیگر ویژگی‌ها برای پیش‌بینی ویژگی از دست رفته استفاده می‌شود. با این کار ارتباط بین ستون‌ها و ویژگی‌ها را نیز در نظر می‌گیریم و آن‌ها را مانند روش‌ های گذشته نادیده نمی‌گیریم. در این روش k نزدیک‌ترین همسایه از بین دادگان سالم برای هر کدام یک از دادگان از دست رفته در نظر گرفته می‌شود و سپس از بین مقدار آن‌ها میانگین گرفته می‌شود. بسته به نوع دیتا و شرایط ممکن است نتیجه خیلی بهتر از روش‌های گذشته به دست بیاید اما این روش سخت و هزینه‌بر می‌باشد چراکه مجبور است مجموعه تمام دادگان از دست نرفته برای آن ستون را همواره نگاه دارد تا بتواند فواصل را محاسبه کند. همچنین برخلاف SVM نسبت به outlier ها خیلی حساس می‌باشد و ممکن است خطای زیادی را وارد کند.
5. روش دیگر استفاده از EM می‌باشد. در این روش فرض می‌کنیم که توزیع دادگان شبیه به نرمال است و سعی می‌شود که با کمک تخمین پارامترهای توزیع‌ نرمال و چیدن آن‌ها کنار یکدیگر، این توزیع و دادگان از دست رفته را به خوبی تخمین بزند. این روش نیز زمان گیر است و پیاده سازی و اجرای هزینه‌بری دارد اما با توجه به توزیع دادگان‌مان فکر می‌کنیم که روی دادگان ما بتواند به خوبی جواب بدهد.

سوال ۴



شکل ۱-۲: نمودار توزیع در ستون‌های مختلف

شکل ۱-۲ نمودار‌های توزیع را در هر یک از ستون‌ها به تفکیک نشان می‌دهد و با توجه به این نمودارها و شکل توزیع هر یک از ستون‌ها به نظر می‌رسد که روش EM بتواند روش مناسبی برای ما باشد. در کنار این روش نیز روش‌های تخمین به کمک میانه و knn با k برابر با ۵ نیز بررسی شده که هر ۳ روش‌های قدرتمند و مناسبی به نظر می‌رسند.

از بین بردن دادگان و تخمین مجدد آن‌ها را همانطور که در سوال خواسته شده است روی ۱۲۰۰۰ داده اول که دادگان آموزش ما هستند انجام دادیم و شاید اگر از بقیه دادگان استفاده می‌کردیم به جهت در دست داشتن دادگان بیشتر، جواب بهتری نیز داشتیم.

سوال ۵

برای محاسبه MSE برای دادگان پیش‌بینی شده به ۲ صورت می‌توانیم عمل کنیم یکی این که مقدار MSE را برای کل ستون‌ها تنها دادگان حذف شده و از دست رفته را در نظر بگیریم که در این حالت برای ۸۰ درصد دادگان مقدار این خطا صفر است و نمی‌تواند معیار خوبی باشد. اما برای مقایسه این مقدار نیز آورده شده است. روش دیگر این است که مقدار MSE را در هر ستون تنها برای خانه‌هایی که در دیتای اصلی حذف شده‌اند و الگوریتم ما آن را پیش‌بینی کرده، محاسبه کنیم. مقدار MSE محاسبه شده در این حالت بیشتر است اما صرفا روش‌های مختلف در محاسبه MSE می‌باشد.

مقدار محاسبه شده را نیز برای ۳ روش EM (our predictor) که روش پیاده سازی شده در این قسمت می‌باشد و روش knn (knn predictor) با k=5 که به کمک کتابخانه و به جهت مقایسه آورده شده و روش میانگین (mean predictor) که ساده‌ترین روش پیاده سازی شده است و به جهت مقایسه می‌باشد، انجام دادیم. نتایج برای ۳ روش مختلف پیش‌بینی داده‌ها و با ۲ روش مختلف محاسبه MSE به صورت زیر می‌باشد.

|  |  |  |
| --- | --- | --- |
|  |  |  |

طبق محاسبات انجام شده و نشان داده شده به نظر می‌رسد که روش EM بهترین روش پیاده سازی شده بوده و توانسته خطا را به شکل مناسبی کاهش دهد. اما بعد از این روش به نظر می‌رسد که در برخی از ستون‌ها عملکرد تخمین به کمک میانگین بهتر از knn بوده و این بستگی به داده‌ها و توزیع آن‌ها دارد.