گزارش کار بخش پیاده سازی تمرین سوم - KNN

محمد لشکری ۴۰۰۱۱۲۰۸۷

۲۱ آذر ۱۴۰۰

مجموعه متغیرهای مستقل و وابسته به صورت زیر انتخاب شده‌اند:

```
X = {Age, Year_of_operation, Number_of_positive_cases}
```

```
Y = {Survival_status}
```

در مرحله پیش پردازش، دادگان آموزش و تست به حالت استاندارد درآمدند تا تأثیر تمام ویژگی‌ها در متر اقلیدسی برای پیش‌بینی یکسان باشد.

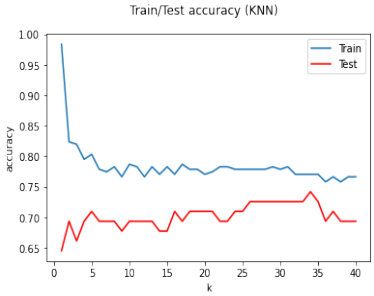
$$X_{train} := \frac{X_{train} - \mu}{\sigma},\ X_{test} := \frac{X_{test} - \mu}{\sigma}$$

که در آن μ و σ میانگین و واریانس X_{train} است.

۱ k- نزدیک ترین همسایه

۱.۱ اجرای الگوریتم به ازای مقادیر مختلف k

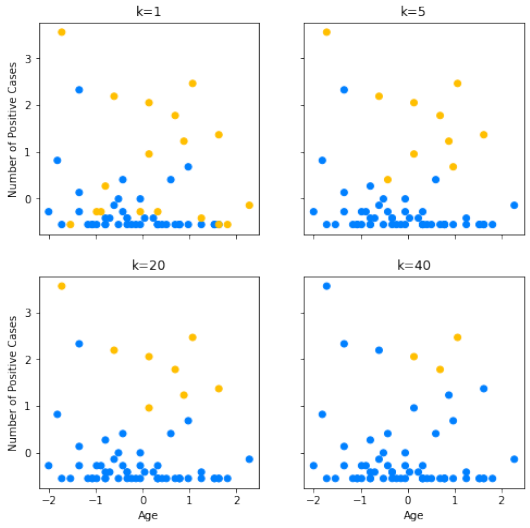
الگوریتم نزدیک ترین همسایگی به ازای اعداد بین ۱ تا ۴۰ به ازای پارامتر k اجرا شد که نتایج آن در شکل ۱.۱ قابل مشاهده است. اختلاف بین دقت دادگان آموزش و تست در عدد ۳۴ کمترین است.



شکل ۱.۱: نمودار دقت به ازای مقادیر مختلف k

۲.۱ دیداری سازی برای تحلیل و مقایسه

رنگ آبی نشان دهنده کلاس ۱ و رنگ نارنجی نشان دهنده کلاس ۲ است. همانطور که در شکل ۲.۱ مشاهده می‌شود با افزایش پارامتر k تمایل داده‌ها برای پیوستن به کلاس ۱ بیشتر می‌شود که علت این موضوع می‌تواند تعداد زیاد برچسب‌های ۱ در داده‌های آموزشی باشد.



شکل ۲.۱: نمودار دقت به ازای مقادیر مختلف k

۳.۱ بهبود سازی پارامتر k

در بخش ۱.۱ بهترین تعداد همسایه ۳۴ به دست آمد. اما با استفاده از Cross validation برای معیار دقت، این مقدار برابر ۱۱ شد. نتایج خطای تست روی هر مدل به دست آمده است و برای مدل ۳۴ نزدیک ترین همسایه، نتایج روی این دادگان خاص قابل قبول تر است که علت آن فراوانی بیشتر کلاس یک در دادگان تست است. همانطور که در بخش ۲.۱ گفته شد با افزایش k همگرایی مدل به کلاس یک بالاتر می‌رود.

پارامتر k	دقت روی داده‌های تست	میانگین دقت روی مجموعه اعتبارسنجی
۱۱	۰/۶۹	۰/۷۷
۳۴	۰/۷۵	۰/۷۴

جدول ۱.۱: دقت مدل برای مقادیر ۳۴و۱۱