دسته‌بندی به روش kـ نزدیک‌ترین همسایه

درس: مقدمه‌ای بر داده‌کاوی

استاد: دکتر مائده السادات طاهائی

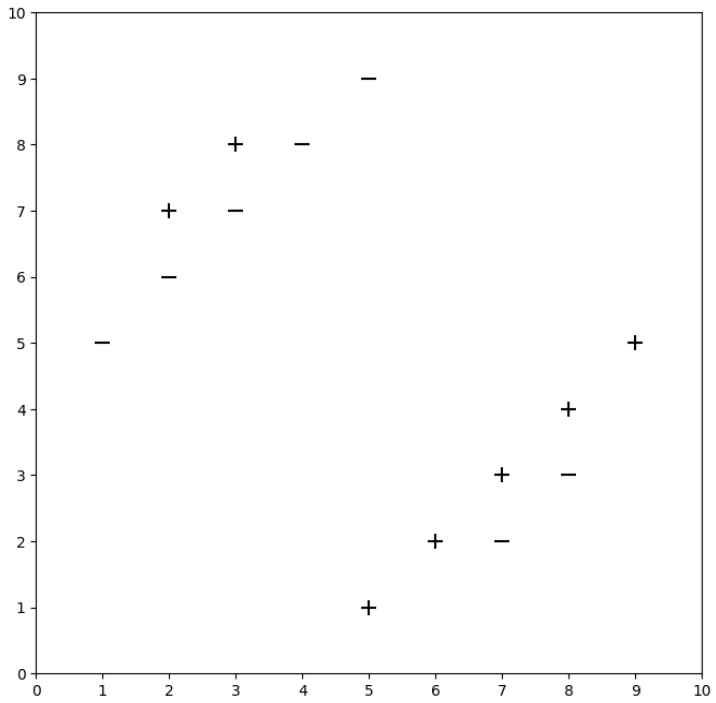
گردآورندگان: کاظم فرقانی، علیرضا کفاشها

مسائل دسته‌بندی به روش kـ نزدیک‌ترین همسایه و ارزیابی دسته‌بندی

دی ۱۴۰۳

پرسش ۱

در این سوال یک دسته‌بند KNN با متریک فاصله L_2 در نظر بگیرید. کلاس‌ها را تماما دوحالت (+/-) در نظر خواهیم گرفت. به سوالات زیر با توجه به مجموعه داده مشخص شده در تصویر پاسخ دهید.



شکل ۱: مجموعه داده مورد بررسی

الف) به ازای چه مقدار k خطای این دسته‌بندی کمینه می‌شود؟

ب) چرا استفاده از مقادیر بسیار زیاد یا بسیار کم برای k می‌تواند منجر به خطا شود؟

ج) مرز تصمیم برای دسته‌بند ۱-NN را برای این مجموعه داده در تصویر نشان دهید.

پاسخ ۱

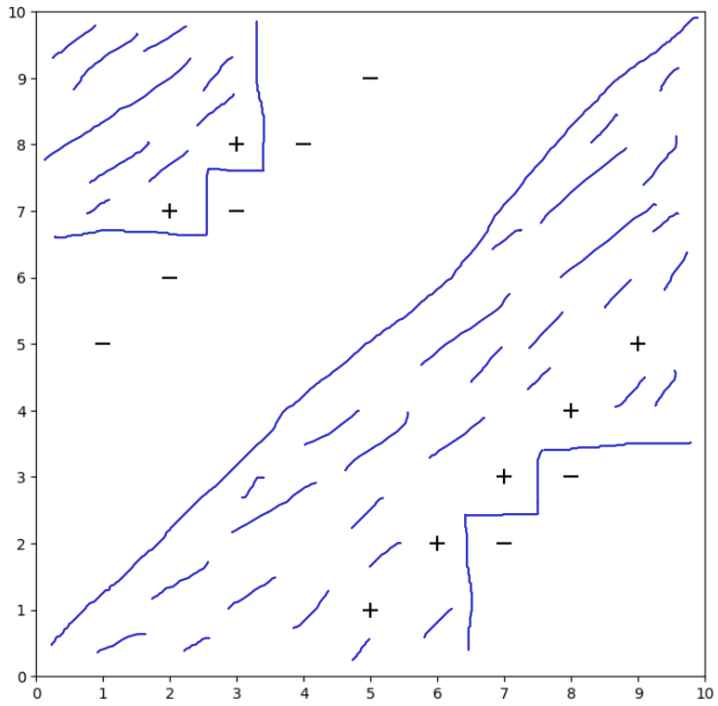
الف) مقدار k برای کمینه‌سازی خطا

چون هر نقطه‌ای همسایه خودش (نزدیک‌ترین همسایه خودش) به شمار می‌رود ۱-NN کمترین خطا را به دست می‌دهد که همان صفر است.

ب) تاثیر مقادیر بسیار زیاد یا بسیار کم k بر خطا

مقادیر زیاد k: خط پایین سمت راست تصویر که دو داده منفی روی آن قرار دارند را در نظر بگیرید؛ این خط به‌وسیله خطی متشکل از داده‌های مثبت، از سایر داده‌های منفی جدا می‌شود. در صورت افزایش k، دادگان مثبت نیز در دسته‌بندی در نظر گرفته می‌شوند که موجب افزایش خطا می‌شود.
مقادیر اندک k: در دو سمت مجموعه داده که دادگان dense هستند، می‌توان مشاهده کرد که اکثر داده‌هایی که فاصله اقلیدسی کمی از یکدیگر دارند، عضو کلاس‌های متفاوت هستند، که در صورت کوچک گرفتن اندازه همسایگی این امر موجب افزایش خطا خواهد شد.

ج) مرز تصمیم ۱-NN



شکل ۲: مرز تصمیم ۱-NN

این روش شامل بررسی چگالی (Density) با مقادیر مختلف K است، اما فقط برای یک متغیر قابل استفاده است (من متغیری را انتخاب خواهم کرد که بیشترین مقدارهای از دست رفته را دارد). متغیری که چگالی آن نزدیک‌ترین حالت به توزیع اصلی را دارد، بهترین گزینه برای انتخاب است.

پرسش ۲

فرض کنید مجموعه‌ای از داده‌ها در اختیار داریم که میزان وفاداری کاربران یک سرویس‌دهنده اینترنت را نشان می‌دهد. ویژگی‌های این مجموعه شامل رفتار کاربران در بازه‌های زمانی ماهانه است و ستون پایانی، ماندگاری کاربران در ماه بعدی را مشخص می‌کند.

یک روش پیشنهاد دهید که بتوان با استفاده از الگوریتم KNN مقادیر گمشده (NaN) را در ستون‌های مربوط به ویژگی‌های این جدول جایگزین کرد. این روش باید شامل نرمال‌سازی داده‌ها پیش از اجرای الگوریتم KNN باشد تا نتایج بهینه‌تری به دست آید. این فرآیند که با نام 'KNN Imputation' شناخته می‌شود، از مقادیر نزدیک‌ترین همسایگان برای پر کردن داده‌های گمشده استفاده می‌کند. لطفاً توضیح دهید که این روش چگونه عمل می‌کند.

پاسخ ۲

در واقع، یک روش برای بررسی بهترین مقدار K وجود دارد که نیازی به تقسیم داده‌ها به دو مجموعه آموزش و تست نیست.

کد نمونه

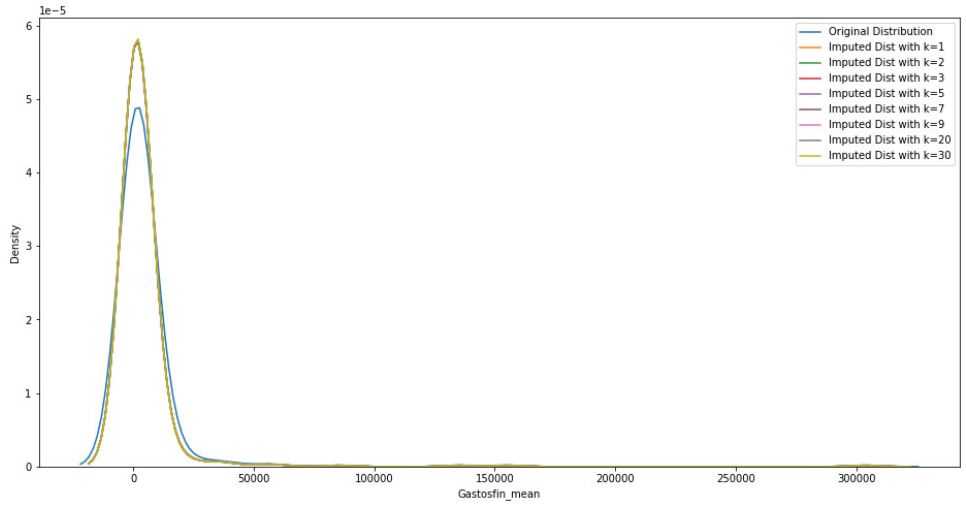
```
n_neighbors = [1, 2, 3, 5, 7, 9, 20, 30]
```

```
fig, ax = plt.subplots(figsize=(16, 8))
# Plot the original distribution
sb.kdeplot(df.variableselected, label="Original Distribution")
for k in n_neighbors:
    knn_imp = KNNImputer(n_neighbors=k)
    density.loc[:, :] = knn_imp.fit_transform(datos)
    sb.kdeplot(density.variableselected, label=f"Imputed Dist with k={k}")

plt.legend()
```

توضیحات

در مثالی که پایین‌تر نشان داده شده است، هر مقدار K دقت یکسانی دارد، اما این موضوع می‌تواند بسته به داده‌ها متغیر باشد.



شکل ۳: نمونه‌ای از نتیجه بررسی توزیع با مقادیر مختلف K

پرسش ۳

شما در حال طراحی یک مدل یادگیری ماشین برای تشخیص وجود یا عدم وجود گونه‌ای از سرطان هستید. در جلسه‌ای برای پیشنهاد دادن مدل خود جهت استفاده در کلینیک‌ها و بیمارستان‌ها، آیا معرفی معیار دقت (accu-racy) به عنوان معیار اصلی برای بررسی مدل‌تان کار منطقی می‌باشد؟ اگر بله، دلیل خود را ذکر کنید. اگر خیر، معیارهای دیگری که می‌توانند بیانگر بهتری برای عملکرد مدل شما باشند را ذکر کنید.

پاسخ ۳

با توجه به اینکه با مسئله تشخیص سرطان سر و کار داریم، اینجا بهتر است از Recall استفاده کنیم، چرا که هر نمونه Negative False توسط سیستم خیلی بیشتر آسیب‌زا است.

منابع و مراجع

- درس یادگیری ماشین، دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف، دکتر علی شریفی زارچی