مبانی یادگیری ماشین - تکلیف سری چهارم

مدرس: دکتر حامد ملک

پاییز ۱۴۰۱

ددلاین: ۲۵ آذر

مسائل

1. خوشه‌بندی به کمک الگوریتم K-means

الگوریتم K-means یک الگوریتم unsupervised است که یک مجموعه داده بدون label را به تعداد مشخصی خوشه تقسیم می‌کند.

1.1. پیاده‌سازی الگوریتم

در این بخش، باید الگوریتم K-means را پیاده‌سازی کنید و سپس برای بررسی صحت عملکرد آن، الگوریتم نوشته شده را با مقادیر مختلف k روی این دیتاست اجرا کنید. نمودار scatter داده‌ها را به همراه مرکز هر دسته رسم کنید.

1.2. فشرده‌سازی رنگ

در این بخش، باید به کمک الگوریتم k-means، تعداد رنگ‌های استفاده شده در این تصویر را کاهش دهید (مانند مثال زیر). عکس داده‌شده را با چند K متفاوت فشرده کنید.

مسائل تحلیلی

2. یک مدل مبتنی بر Random Forest با ۲۰ عدد درخت تصمیم گیری داریم که در هر نود از هر درخت، از یک ویژگی تصادفی برای گسترش دادن درخت استفاده می‌کنیم.

اگر تعداد درخت‌های تصمیم‌گیری را به ۱۰۰ عدد افزایش دهیم:

2.1. چه تغییراتی در عملکرد مدل در زمان Train دیده خواهد شد؟

2.2. چه تغییراتی در عملکرد مدل در زمان استفاده از آن دیده خواهد شد؟

2.3. اگر تعداد فیچرهایی که در هر گره از درخت‌ها، برای گسترش درخت انتخاب می‌کنیم، به ۳ تا افزایش یابد، چه تغییراتی در عملکرد مدل در زمان Train و زمان استفاده دیده خواهد شد؟

3. یکی از کاربردهای KNN در پیاده سازی Recommender System ها است. فرض کنید دیجی‌کالا یک دیتاست شامل افراد و ویژگی‌های فردی شان (به عنوان Feature) و کالاهای مورد علاقه آنها (به عنوان Label) دارد و قصد دارد از KNN برای پیشنهاد دادن محصولات به کاربرانی که به تازگی در آن ثبت نام کرده اند استفاده کند.

3.1. توضیح دهید وقتی کاربری در این سایت ثبت نام می‌کند، Recommender System این شرکت چه فرایندی را طی می کند تا سه محصول به کاربر پیشنهاد دهد؟

3.2. نقاط ضعف و قوت این روش چیست و آیا روشی برای رفع نقاط ضعف آن سراغ دارید؟

4. از ایده Random Forest می‌توان برای یادگیری Unsupervised در تشخیص Outlierها نیز استفاده کرد. توضیح دهید چگونه و با چه منطقی می توان داده‌های پرت را با کمک ایده RF پیدا کرد؟

نکات تمرین

- در صورت هرگونه تقلب نمره صفر برای شما لحاظ می‌گردد.
- استفاده از زبان غیر از پایتون مجاز نیست.
- این تمرین تحویل حضوری ندارد؛ بنابراین نوشتن مستندات بسیار مهم و بخش قابل توجهی از نمره است.

موفق باشید