به نام خدا

دانشکده‌ی علوم و مهندسی کامپیوتر — یادگیری ماشین

تکلیف دوم – پاییز ۱۳۹٤ — مهلت ارسال: ۱۵ دی ۱۳۹٤

**قبل از اینکه شروع به حل تمرین ها کنید، حتما یک بار فایل Policies.pdf (موجود در وبگاه) که حاوی نکات مهم در تحویل تکالیف هست را مطالعه فرمایید.**

**حل سوالاتی که امتیازی هستند، اجباری نبوده و نمره‌ی اضافی خواهد داشت.**

**طراح هریک از سوالات در زیر مشخص شده است .در صورتی که سوالی دارید می توانید سوالتان را در گروه مطرح کنید یا از طراح سوال بپرسید.**

**1: آقای کاهانی**

**2، 3 و 4: آقای زندی**

---

۱ -الف) درخت تصمیم را برای دسته‌بندی داده‌های جدول زیر به صورت دستی محاسبه کرده و نمایش دهید.

| DAY | OUTLOOK | TEMPERATURE | HUMIDITY | WIND | PLAY TENNIS |
|---|---|---|---|---|---|
| **D1** | Sunny | Hot | High | Weak | No |
| **D2** | Sunny | Hot | High | Strong | No |
| **D3** | Overcast | Hot | High | Weak | Yes |
| **D4** | Rain | Mild | High | Weak | Yes |
| **D5** | Rain | Cool | Normal | Weak | Yes |
| **D6** | Rain | Cool | Normal | Strong | No |
| **D7** | Overcast | Cool | Normal | Strong | Yes |
| **D8** | Sunny | Mild | High | Weak | No |
| **D9** | Sunny | Cool | Normal | Weak | Yes |
| **D10** | Rain | Mild | Normal | Weak | Yes |
| **D11** | Sunny | Mild | Normal | Strong | Yes |
| **D12** | Overcast | Mild | High | Strong | Yes |
| **D13** | Overcast | Hot | Normal | Weak | Yes |
| **D14** | Rain | Mild | High | Strong | No |

ب) الگوریتم‌های درخت تصمیم، K نزدیک‌ترین همسایه و طبقه‌بند بیزین را مقایسه کنید. ملاک‌های مقایسه:

- میزان حساسیت به نمونه های آموزشی

- مقاومت در برابر نمونه‌ی نویزی و نمونه‌ی پرت (Outlier)

ج) در برخی کاربردها با وجود اینکه بعضی از ویژگی‌ها ارزش زیادی در جداسازی کلاس‌ها دارند اما اندازه‌گیری آنها با هزینه ی زیادی همراه است. آیا می‌توان این هزینه را در تشکیل درخت تاثیر داد؟ چگونه؟

۲ -می‌خواهیم ابعاد داده‌های موجود در فایل data2.mat را کاهش دهیم.

الف) با استفاده از PCA ابعاد را به دو کاهش داده و نتیجه را طوری نمایش دهید که کلاس‌ها نیز مشخص باشند. (مثلا هر کلاس با یک رنگ)

ب) ابزار KPCA را به طور مختصر شرح دهید و تفاوت آن را با PCA بیان نمایید.

پ) از KPCA با کرنل گاوسی استفاده کنید و ابعاد را به دو کاهش دهید. نتایج را طوری نمایش دهید که کلاس‌ها نیز مشخص باشند.

ت) در کدام نگاشت مولفه‌های نهایی از نظر خطی مستقل‌تر هستند؟ علت را توضیح دهید.

ث) به نظر شما کدام یک از این ابزارها دقت بیشتری برای دسته‌بندی فراهم آورده است؟ آیا همیشه این‌طور است؟

ج) با توجه به اینکه PCA و KPCA به صورت بدون ناظر عمل می‌کنند، چطور می‌توانند دقت دسته‌بندی را بهبود بخشند؟

تذکر: برای PCA می‌توانید از خود تابع PCA در متلب بهره ببرید. اما با توجه به اینکه ابزار KPCA در متلب موجود نیست، پیاده سازی آن ضمیمه شده است که کارتان را بسیار ساده می‌کند.

۳ -مجموعه‌ای از داده‌های تصادفی در فایل data3.txt قرار دارد. می‌خواهیم با استفاده از خوشه‌بندی سلسله‌مراتبی این داده ها را خوشه بندی نماییم.

الف) یکی از معیارهای موجود برای جداسازی خوشه‌ها Inconsistency Coefficient است. این معیار را شرح دهید و مزیت آن را نسبت به معیار ثابت بودن فاصله‌ی خوشه‌ها بیان نمایید.

ب) داده‌ها را با استفاده از معیار Inconsistency Coefficient خوشه‌بندی نمایید. فاصله‌ی خوشه‌ها را یک بار به صورت Single و یک بار به صورت Centroid اندازه بگیرید. نتایج را نمایش داده و مقایسه نمایید.

۴ -قصد داریم تا با استفاده از متد KNN، ابزاری برای یک سایت املاک تهیه کنیم که قیمت خانه ها را تخمین می زند . داده‌ها در فایل housing.mat موجود هستند.

الف) مهم‌ترین مزیت این کار نسبت به یک روش پارامتری چیست؟

ب) با میزان K از ۱ الی ۵ این کار را انجام دهید و دقت‌ها را گزارش نمایید. ۷۵ درصد از داده‌ها را برای آموزش و بقیه را برای تست قرار دهید.

ج) میزان K در کدام حالت بهتر است؟ چرا و از کدام منظر؟

موفق باشید