به نام خدا

دانشکده‌ی علوم و مهندسی کامپیوتر — یادگیری ماشین

پاسخ‌های تکلیف دوم – زمستان ۱۳۹٤

این پاسخ‌ها نمونه‌هایی هستند که حاوی نکات اصلی می‌باشند و لزوما تنها پاسخ‌های درست نیستند.

۱- الف- الف) برای تشکیل درخت تصمیم معیار های مختلفی وجود دارد، دو معیار بر مبنای شاخص Gini و Entropy به صورت زیر تعریف می شوند:

معیار تقسیم سازی بر مبنای شاخص Gini:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

برای گره t: $GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

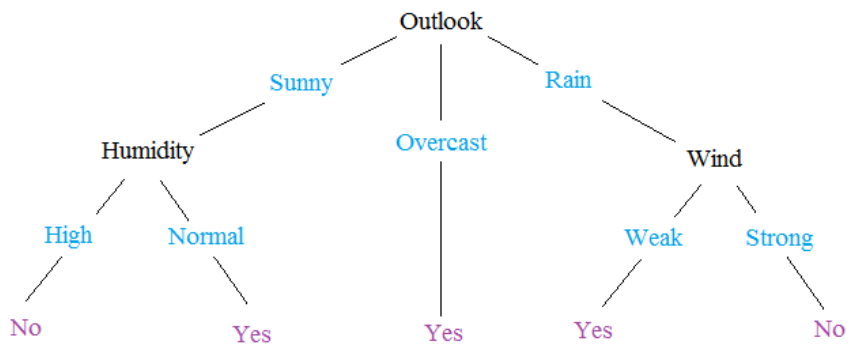
معیار تقسیم سازی بر مبنای شاخص Entropy:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

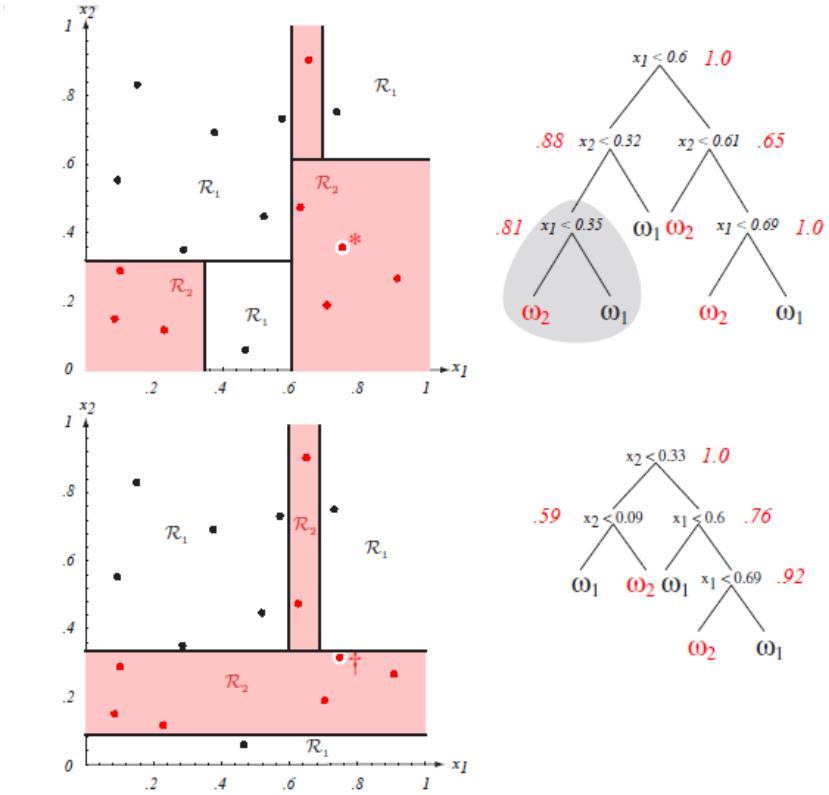
برای گره t: $Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$

* $p(j \mid t)$ احتمال رخداد کلاس j در گره t است.

درخت بدست آمده برای این داده ها:



ب) روش های درخت تصمیم و knn بر مبنای نمونه (Instance based) بوده و حساسیت زیادی به نمونه های آموزشی دارند در حالی که این حساسیت در روش های بر مبنای مدل (model based) مانند بیزین کمتر است. و البته هرچه مقدار k در روش knn بیشتر باشد این حساسیت کمتر خواهد شد. در مواجهه با نویز و داده ی پرت نیز وضعیت به همین شکل خواهد بود. شکل زیر تغییر زیاد در درخت حاصل را زمانی که تغییر کمی در مقدار یک نمونه (نمونه ی *) اعمال شده است را نشان می دهد.



شکل از کتاب Duda

ج) می توان وزنی را به معیار تقسیم (Split) اضافه کرد و برای هر ویژگی بر مبنای اهمیت آن وزن را کمتر و یا بیشتر در نظر گرفت.

۲- داده‌های اولیه سه‌بعدی هستند. تقریبا کلاس‌ها به شکل دو کره‌ی هم‌مرکز توزیع شده‌اند. برای مشاهده‌ی داده‌ها می‌توانید فایل showdata.m را اجرا نمایید. دقت نمایید که شکل سه بعدی است و می‌توانید آن را بچرخانید.

الف- این کار در فایل A.m صورت پذیرفته است. همان‌طور که مشخص است، داده‌های کلاس‌ها در هم ریخته‌اند و جدایی‌پذیر خطی نیستند. علت این پدیده این است که این داده‌ها در فضایی قرار دارند که هیچ نگاشت خطی نمی‌تواند آنها را خوب جدا نماید.

ب- ابزار Kernel PCA یا KPCA نوع تعمیم‌یافته‌ی PCA است که ابتدا داده‌ها را توسط یک نگاشت معمولا غیر خطی به فضایی با ابعاد بالاتر می‌برد و سپس از PCA در آن فضا استفاده می‌کند. با توجه به اینکه PCA تنها نیازمند ضرب داخلی داده‌ها است، نگاشت داده‌ها محاسبه نمی‌شود و تنها حاصل ضرب داخلی آنها در فضای نگاشت به وسیله‌ی یک تابع کرنل صورت می‌گیرد. وابسته به توزیع داده‌ها می‌توان از کرنلی برای محاسبه‌ی ضرب‌های داخلی بهره برد که منجر به کاهش بعد بیشتر ویا جداسازی بهتر آنها شود.

از نظر عمل‌کرد بسیار شبیه PCA است، فقط ابتدا تابع کرنل اعمال می‌شود. پس از این مرحله، ماتریسی داریم که ضرب داخلی هر جفت از داده‌ها در آن موجود است. مابقی مراحل مشابه PCA است. یکی دیگر از مزایای کرنل این است که در برخی موارد نگاشت مستقیم به فضای

بالاتر امکان‌پذیر نیست. به عنوان مثال، کرنل گاوسی ضرب داخلی را در فضایی با ابعاد بی‌نهایت محاسبه می‌نماید، اما نگاشت مستقیم بـه آن فضا قابل محاسبه نیست.

پ- فایل C.m بدین منظور ساخته شده است. مقدار پارامتر کرنل گاوسی به صورت دستی تغیر داده شده است تا حالتی به دسـت بیایـد کـه داده‌ها به راحتی جدا پذیر باشند.

ت- می‌دانیم که مولفه‌های به دست آمده از PCA کاملا مستقل خطی هستند. در مورد KPCA، ابتدا داده‌ها به فضایی با ابعاد بالاتر نگاشـت شده و سپس توسط PCA، مولفه‌های اصلی در آن فضا انتخاب می‌شوند و ابعاد کاهش می‌یابد. با توجه به اینکه در KPCA هم در نهایـت از PCA استفاده می‌شود، پس مولفه‌های استخراجی باز هم دارای استقلال خطی هستند. پس هر دو روش، به مولفه‌هایی مستقل خطی دسـت می‌یابند اما فضایی که مولفه‌ها را انتخاب می‌نمایند تفاوت دارد.

کافی است تا تابع cov را روی خروجی هر یک تبدیل‌ها اعمال کرده و مشاهده نمایید که ویژگی‌های استخراج‌شده مستقل خطی هستند.

ث- همانطور که نتایج نشان می‌دهند، روش KPCA داده‌ها را طوری کاهش بعد داده که می‌توان با دقـت ۱۰۰٪ داده‌هـا را جـدا نمـود. امـا همیشه این‌طور نیست. KPCA هنگامی مناسب است که روابط بین داده‌ها غیر خطی باشد. همچنین اگر این ابزار با کرنـل مناسـب بـه کـار نرود ممکن است که نتایج مطلوب را فراهم نیاورد.

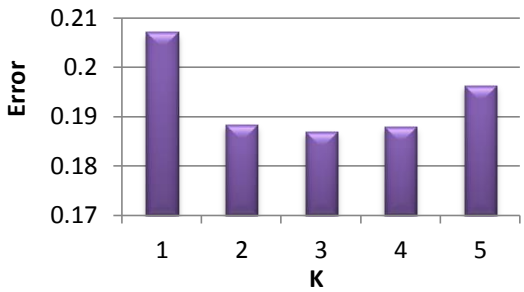
ج- معمولا داده‌های دو کلاس متفاوت، محل توزیع‌شان در فضا تفاوت دارد. هر دو ابزار سـعی دارنـد تـا راسـتاهـایی را بیابنـد کـه بیشـترین تغییرات در آنها وجود دارد. در نتیجه اگر فاصله‌ی کلاس‌ها از هم به نسبت فاصله‌ی داده‌های داخلشان بیشتر باشد، ابعاد نهایی شـامل بعـدی است که کلاس‌ها بر روی آن به خوبی از یکدیگر جدا خواهند شد.

۳- الف- Inconsistency coefficient یک معیار است که در خوشه‌بندی سلسله مراتبی کاربرد دارد. در این معیار به جای اینکه درخت داده‌هـا بر اساس تعداد اعضای یک خوشه، خوشه‌ها را جدا کند، به صورت وفقی عمل می‌کند. در واقع، این معیار می‌سنجد کـه آیـا دو شـاخه خـوب است که با یکدیگر ادغام شوند یا خیر. این معیار، برای هر لینک درخت به اختلاف ارتفاع آن لینک، از میانگین ارتفـاع لینـک‌هـای زیـرینش، بستگی دارد. لازم به ذکر است که، ارتفاع هر لینک برابر با فاصله‌ی دو فرزندش از یکدیگر است. ضریب ناسازگاری همچنین به کمک انحـراف معیار ارتفاع لینک‌های تحت بررسی، میزانش را نرمال می‌نماید. اگر اعضای داخلی خوشه‌ها خیلی نزدیک باشند، ولی خود خوشه‌ها دور از هم باشند، این لینک باید هرس شود، زیرا داده‌های بی‌ربطی را به هم مرتبط کرده است.

ب- فایل B.m خوشه‌بندی را به هر دو صورت انجام می‌دهد و نتایج را نمایش می‌دهد. در حالت single، تعداد بیشتری از خوشه‌ها با یکدیگر ترکیب شده‌اند. همان‌طور که مشاهده می‌شود با تغییرات کم در مقدار آستانه‌ی cutoff، خوشه‌بندی ممکن است خیلی تغییر کند. همچنـین در کل خیلی خوشه‌بندی مناسبی کسب نمی‌شود. برای داشتن خوشه‌های بهتر می‌توان پارامتر depth را تغییر داد که طبق آن میزان ضریب ناسازگاری تا عمق بیشتری برای هر لینک حساب می‌شود. در این حالت آمار بهتری از هر خوشه وجود دارد و در نتیجه خوشه‌ها بهتر ترکیب یا جدا می‌شود.

۴- الف- اصولا قاعده‌ی خاصی برای قیمت املاک وجود ندارد، اما معمولا این دانش را داریم که املاک مشابه باید قیمت مشابه داشته باشند. لـذا استفاده از روش‌های یادگیری غیرپارامتری مناسب به نظر می‌رسد.

ب- داده‌های آموزش و تست به صورت تصادفی انتخاب شده‌اند. برای اینکه میزان خطا نیز بهتر درک شود، میانگین خطا، تقسیم‌بـر میـانگین کل قیمت خانه‌ها شده است. همچنین از کرنل گاوسی برای تخمین قیمت‌ها استفاده شده است. (البته شما می‌توانستید از یک میانگین‌گیری ساده نیز استفاده نمایید). نمودار زیر مقادیر خطا به ازای Kهای مختلف را نشان می‌دهد. (برای هر K، ۵ بار الگوریتم اجرا شده و میانگین آن گزارش شده است.)



ج- افزایش میزان K، میزان محاسبات را برای یافتن همسایه‌ها افزایش می‌دهد. به علاوه طبق نمودار، در ایـن مثـال K=3 کمتـرین خطـای تست را دارد. در نتیجه بهترین میزان K در این مثال، (از بین ۱ الی ۵) مقدار ۳ است. لازم به ذکر است که اگر تعداد داده‌ها خیلی زیاد بـود، شاید بهتر بود که k=2 به دلیل محاسبات کمترش برگزیده شود.