دانیال شیرالی **شماره دانشجویی ۹۹۵۲۲۰۰۹**

۱. در خصوص کرنل های پرکاربرد روش SVM تحقیق کنید .به صورت کلی چرا ما از ایده کرنل در بحث SVM بهره میبریم . آیا میتوان در خصوص کرنل ها و استفاده ی آنها حکم کلی داد . به طور مثال بگوییم از کرنلRBF در این مواقع خاص استفاده میکنیم .

کرنل های پرکاربرد عبارتند از: Linear Kernel، Polynomial Kernel، Sigmoid Kernel، Gaussian Kernel، Bessel function kernel و ANOVA kernel.

Linear Kernel پایه‌ای ترین نوع کرنل است که ماهیت آن معمولاً یک بعدی است. ثابت شده است که در زمانی که تعداد ویژگی‌های زیادی داشته باشیم، Linear Kernel بهترین کرنل است.

Polynomial Kernel، نمای کلی تر از Linear Kernel است. اما مانند سایر کرنل ها مورد توجه قرار ندارد زیرا از کارآیی و دقت کمتری برخوردار است.

Sigmoid Kernel، بیشتر برای شبکه های عصبی ترجیح داده می شود. این کرنل شبیه به مدل دو لایه ای از پرسپترون شبکه عصبی است که به عنوان یک تابع فعال سازی برای سلول های عصبی کار می کند.

Gaussian Kernel، یک کرنل است که معمولاً مورد استفاده قرار می‌گیرد. وقتی که دانش قبلی از یک مجموعه داده وجود ندارد، از این کرنل استفاده می‌شود.

Bessel function kernel، عمدتا برای حذف cross term در توابع ریاضی استفاده می شود.

ANOVA kernel، همچنین به عنوان کرنل radial basis شناخته می شود. معمولاً در مشکلات رگرسیون چند بعدی عملکرد خوبی دارد.

از مزیت های استفاده از تابع‌های کرنل این است که اگر برای گسترش فضا مورد استفاده قرار بگیرد، در مقایسه با استفاده از توان های بالاتر ویژگی‌ها هزینه محاسباتی کمتری خواهد داشت. در حالتی که از کرنل استفاده شود تنها نیاز به محاسبه $\binom{n}{2}$ تابع کرنل داریم. اما زمانی که برای گسترش فضا از توان های چندم ویژگی ها استفاده می کنیم، هزینه محاسباتی ممکن است بسیار زیاد شود. بنابراین نمی‌توانیم برای استفاده یا عدم استفاده از کرنل ها حکم کلی بدهیم. حتی درصورت نیاز به استفاده از کرنل ها باید دیتاست و مسئله مورد نظر را بررسی کنیم تا بهترین کرنل را از بین کرنل های موجود انتخاب کنیم.

۵ (ج). بررسی کنید آیا استفاده از تبدیل هایی از قبیل log transformو یا تبدیل نمایی در اینجا کاربرد دارد . به صورت کلی چرا از این دست تبدیلات بهره میبریم (در این بخش شما مجاز هستید اگر تبدیل دیگری را مناسب میدانید اعمال کنید این بخش نمره امتیازی برای شما خواهد داشت . حتما دلیل استفاده از تبدیل استفاده شده را بیان کنید).

log transformation روشی است که به طور گسترده برای رسیدگی به داده های اریب مورد استفاده قرار می گیرد. این روش یکی از محبوب ترین transformations های است که در تحقیقات زیست پزشکی و روانشناختی مورد استفاده قرار می گیرد.

به دلیل سهولت استفاده و محبوبیت این روش، log transformation در اکثر بسته های نرم افزاری آماری از جمله SAS، Splus و SPSS گنجانده شده است. متأسفانه، محبوبیت آن باعث شده است که در معرض سوء استفاده(بکارگیری اشتباه آن) قرار گیرد (حتی توسط آماردان ها)، که منجر به تفسیر نادرست از نتایج تجربی خواهد شد. البته چنین سوءاستفاده و سوءتعبیرهایی، منحصر به این transformation خاص نیست. این یک مشکل مشترک در بسیاری از روشهای آماری رایج است.
یکی دیگر از کاربردهای رایج log transformation، کاهش تنوع داده ها، به ویژه در مجموعه داده‌هایی است که شامل مشاهدات دور از مرکز است. ازدیگر کاربردهای رایج log transformation، برای مطابقت داده ها با نرمال بودن آنهاست.

۷. در خصوص الگوریتم های مختلف ساخت درخت تصمیم (همانند CART، ID3و...) تحقیق کنید . به صورت کلی تفاوت الگوریتم های مختلف ساخت درخت تصمیم در چیست؟

الگوریتم‌های متعددی برای ساخت درخت تصمیم وجود دارند از جمله

ID3:Iterative Dichotomiser

C4.5: Classifier 4.5

CART: Classification And Regression Tree

ID4

ds CART: DempsterShafer Classification And Regression Tree

ID5R

EC4.5:Efficient Classifier 4.5

CHAID: Chi square Automatic Interaction Detection

RF: Random Forest

RT: Random Tree

الگوریتم ID3

این الگوریتم یکی از ساده ترین الگوریتم های درخت تصمیم است. در این الگوریتم درخت تصمیم از بالا به پایین ساخته می‌شود. این الگوریتم با این سوال شروع می‌شود: کدام ویژگی باید در ریشه درخت مورد آزمایش، قرار بگیرد؟ برای یافتن جواب از معیار بهره اطلاعات استفاده می‌شود. با انتخاب این ویژگی، برای هر یک از مقادیر ممکن آن یک شاخه ایجاد شده و نمونه های آموزشی بر اساس ویژگی هر شاخه مرتب می‌شوند. سپس عملیات فوق برای نمونه های قرار گرفته در هر شاخه تکرار می شوند تا بهترین ویژگی برای گره بعدی انتخاب شود.

الگوریتم C4.5

این الگوریتم یکی از تعمیم های الگوریتم ID3 است که از معیار نسبت بهره استفاده می کند. الگوریتم هنگامی متوقف می‌شود که تعداد نمونه ها کمتر از مقدار مشخص شده‌ای باشد. این الگوریتم از تکنیک پس هرس استفاده می‌کند و همانند الگوریتم قبلی داده‌های عددی را نیز می‌پذیرد.

الگوریتم CHAID

محققان آمار کاربردی، الگوریتم‌هایی را جهت تولید و ساخت درخت تصمیم توسعه دادند. الگوریتم CHAID در ابتدا برای متغیرهای اسمی طراحی شده بود. این الگوریتم با توجه به نوع برچسب کلاس از آزمون‌های مختلف آماری استفاده می‌کند. این الگوریتم هرگاه به حداکثر عمق تعریف شده‌ای برسد و یا تعداد نمونه‌ها در گره جاری از مقدار تعریف شده‌ای کمتر باشد، متوقف می‌شود. الگوریتم CHAID هیچگونه روش هرسی را اجرا نمی‌کند.

تفاوت الگوریتم های ساخت درخ تصمیم به صورت عمده در موارد زیر است:

- آیا این الگوریتم قابلیت آن را دارد که وزن‌های مختلف و غیر یکسانی را به برخی از ویژگی‌ها بدهد؟
- آیا می‌تواند مقادیر گسسته یا پیوسته را برای ویژگی‌ها درک کند؟
- آیا قادر است با وجود مقادیر گمشده نیز درخت تصمیم(decision tree) خود را بسازد؟
- ...

۱۰. در خصوص هرس کردن Pruningدرخت تصمیم تحقیق کنید . چرا ما به بحث هرس کردن درخت تصمیم نیاز دارد و چه کمکی به ما می‌کند.

اگر یک درخت تصمیم را صرفا براساس مینیمم کردن RSS تولید کنیم، درخت ساخته شده بر روی نمونه های آموزشی، جواب خوبی برمی گرداند اما بر داده‌های نمونه‌های تست، کارایی پایینی خواهد داشت. این مشکل به این دلیل رخ می‌دهد که اگر معیار صرفا مینیمم کردن RSS باشد، درخت تولید شده عمق زیاد و انشعاب های زیادی خواهد داشت و به اصطلاح پیچیده می‌گردد.

درعمل این امکان وجود دارد که یک درخت کوچکتر با تعداد انشعاب های کمتر، مقدار کمی بایاس را افزایش دهد، اما در عوض به شدت واریانس کمتری بر روی داده های تست داشته باشد. بدین منظور می‌توانیم از روش هرس کردن درخت استفاده کنیم.

یک روش مناسب برای هرس کردن یک درخت، این است که یک درخت T بزرگ را تولید کنیم و بعد آن را هرس کنیم. درواقع ما به دنبال زیردرختی از این درخت بزرگ T هستیم که دارای کمترین نرخ خطا بر روی داده های تست باشد. نرخ خطای تست برای هر زیردرخت را می توانیم با استفاده از cross-validation error تخمین بزنیم.

۱۳. تحقیق کنید چرا با وجود روش های جدید از قبیل یادگیری عمیق و شبکه عصبی ، هم چنان روشی مانند درخت تصمیم محبوب است ؟

مهم‌ترین دلیلی که باعث می‌شود در برخی موارد درخت تصمیم را به روش های جدیدی مثل یادگیری عمیق و یا شبکه عصبی ترجیح دهیم، آن است که درخت های تصمیم به صورت گرافیکی قابل تصویرسازی هستند و توسط افراد معمولی قابل تفسیر هستند. درواقع درخت های تصمیم قابل توضیح به اکثر مردم هستند و بیان آنها راحت تر از بیان نتایج اکثر متدهای موجود برای پردازش داده ها است.
یکی از مواردی که درخت های تصمیم را از سایر متدها مجزا می‌کند، آن است که آنها شمایی از عمل تصمیم گیری در ذهن انسان ها را نشان می‌دهد.

۱۴.(بخش امتیازی) در درخت های تصمیم ما قوانینی استخراج میکنیم و از این قوانین استفاده میکنیم. در خصوص روش های دیگری که به استخراج قوانین از روی دیتاست میپردازند (rule induction)همانند روش ... ، IREP، Ripper تحقیق کنید و آن ها را توضیح دهید . (حداقل دو روش)

RIPPER (Repeated Incremental Pruning to Produce Error Reduction)

RIPPER یکی روش های بسیار کارآمدی است که برای الگوریتم‌های rule learning به کار می‌رود. RIPPER یک استراتژی تقسیم و تخسیر را برای استخراج قوانین پیاده سازی می‌کند. درواقع در RIPPER برای هر کلاس IREP (Incremental Reduced Error Pruning) را برای تدوین مجموعه‌ای اولیه از قوانین اعمال می‌کند. سپس،یک مرحله اضافی بهینه سازی را در نظر می‌گیرد که در آن به ازای هر قانون بدست آمده در مرحله قبل، دو قانون جایگزین از آنها به وجود می‌آورد که به صورت replacement rule و revision rule نام گذاری شده‌اند. پس از آن، در مورد اینکه آیا مدل باید قاعده اصلی را حفظ کند، تصمیم گیری می‌شود.

CAMUR (Classifier with Alternative and MUltiple Rule-based models)

CAMUR بر مبنای الگوریتم RIPPER شکل گرفته است. در CAMUR با انجام مکرر محاسبه‌ی یک مدل کلاس بندی مبتنی بر قانون، چندین پایه قانون‌های معادل را استخراج می کند. CAMUR شامل یک

مخزن دانش ad-hoc (پایگاه داده) و یک ابزار پرس و جو است. CAMUR به عنوان یک برنامه جاوا مستقل و یک وب برنامه http://dmb.iasi.cnr.it/camur.php در دسترس است.

۱۵. بررسی کنید آیا از درخت های تصمیم میتوان برای حل مسائل سری زمانی استفاده کرد ؟

پیش بینی سری زمانی را می توان به عنوان یک مسئله یادگیری تحت نظارت در نظر گرفت. برای حل مسائل سری زمانی این امکان وجود دارد که از درخت های تصمیم استفاده کنیم. برای مثال پیش بینی سری زمانی می‌توان از Random Forest regressor استفاده نمود.

یک مثال از استفاده از classification tree برای سری های زمانی در مقاله [1] موجود است.

References

[1] C. A. Ahlame Douzal, "Classification trees for time series," *Pattern Recognition,* p. 10, 2011.