به نام خدا

دانشکده‌ی مهندسی کامپیوتر

سررسید تئوری: ۱۳ دی ماه چهارشنبه ۲۳:۵۹
سررسید عملی: ۱۵ دی ماه جمعه ۲۳:۵۹ پاییز ۱۴۰۲

یادگیری ماشین

مدرس: مهدی جعفری سیاوشانی تمرین ۵: Adaboost, Decision Tree & Learning Principles

- سررسید بخش تئوری این تمرین چهارشنبه ۱۳ دی ماه ساعت ۵۹ : ۲۳ است.
- سررسید بخش عملی این تمرین جمعه ۱۵ دی ماه ساعت ۵۹ : ۲۳ است.
- در صورت کشف تقلب، بار اول برای افراد درگیر تقلب، نمره‌ی همان سوال(های) خاص صفر در نظر گرفته می‌شوند. در صورت تکرار، نمره کل تمرین صفر در نظر گرفته می‌شود و در صورت تکرار، درس برای افراد حذف خواهد شد.
- تمامی پاسخ‌های خود را در یک فایل با فرمت (.pdf) HW5-[SID]-[Fullname].zip روی کوئرا قرار دهید.

پرسش‌ها

۱ قسمت تئوری

۱.۱ پرسش اول (۱۰ نمره)

فرض کنید یک بانک اعتباری روشی خوبی برای اعتبار دادن به کاربران جدید ندارد. با رسیدن مشتری‌های $x_۱, x_۲, \cdots, x_N$ این بانک با روش اولیه خود یک سری از این مشتری‌های را تایید می‌کند و پس از گرفتن کارت اعتباری عملکرد آن‌ها را بررسی می‌کند.
برای سادگی فرض کنید که به اولین N مشتری کارت اعتباری داده‌شده است.حال بانک برای بهبود الگوریتم خودش اطلاعاتی را که تا به حال جمع کرده به شما می‌دهد. این اطلاعات به صورت دوتایی‌های $(x_۱, y_۱), \cdots, (x_N, y_N)$ می‌باشد.
شما پیش از حتی گرفتن اطلاعات، با یک سری از فرمول‌های ریاضی یک تابع به آن‌ها ارائه می‌دهید که به صورت بی‌نقص کار می‌کند.

۱. (۲,۵ نمره) اندازه مجموعه فرضیه یا M چقدر است؟

۲. (۲,۵ نمره) با این اندازه M، باند هافدینگ چه چیزی را در مورد احتمال این که برای $N = ۱۰۰۰۰$ میزان خطای کارکرد واقعی کمتر از دو درصد است، می‌گوید؟

۳. (۲,۵ نمره) شما جوابتان را به بانک می‌دهید و به آن‌ها اطمینان می‌دهید که کارکرد بهتر از خطای دو درصد خواهد بود و اطمینان شما از جواب قسمت قبل بدست می‌اید. بانک از جواب شما برای اعتبار دادن به کاربران جدید استفاده می‌کند. متاسفانه ، بیش از نیمی از کارت‌های اعتبار آن‌ها به تاخیر می‌افتند. دلایل/ دلیل ممکن برای این اتفاق چیست؟

۴. (۲,۵ نمره) آیا راهی وجود دارد که بانک بتواند با اطمینان احتمالاتی شما از تابع‌تان استفاده کند؟ (راهنمایی: جواب مثبت است)

۲.۱ پرسش دوم (۲۰ نمره)

۱. (۲ نمره) فرض کنید با یک مسئله دسته‌بندی دو دسته‌ای مواجه هستیم و تعداد m ویژگی داریم که هر ویژگی دو مقدار ممکن را می‌تواند اختیار کند. در این صورت نشان دهید چند درخت تصمیم متفاوت می‌توانیم برای این مسئله داشته باشیم. (توجه داشته باشید که هر نود دو زیر نود دارد که در واقع مجموعه مقادیر به دو دسته تقسیم می‌شوند و عضویت در هر کدام از زیرشاخه‌های درخت را تشکیل می‌دهند)

۲. (۳ نمره) آیا درخت تصمیمی که به طور حریصانه و با کمک بهینه‌کردن معیارهایی نظیر Information Gain ساخته می‌شود، همیشه بهترین درخت ممکن است؟ علت این امر را نیز توضیح دهید.

۳. درخت‌های تصمیم در برخورد با ویژگی‌های چند مقداره از دو استراتژی استفاده می‌کنند. در استراتژی اول که Multiway Split نام دارد، هر نود تصمیم‌گیری می‌تواند k خروجی داشته باشد. در حالی که در استراتژی دوم که نامش Binary Split است، هر نود تصمیم‌گیری صرفا دو خروجی می‌تواند داشته باشد و در نتیجه اگر یک ویژگی داشته باشیم که k مقدار ممکن را بتواند اختیار کند، در هر نود تصمیم گیری در این استراتژی تنها می‌توانیم شرطی را چک کنیم که پاسخ به آن دو مقدار بله یا خیر باشد. جدول را در نظر بگیرید و با توجه به آن به سوالات زیر پاسخ دهید.

ردیف	جنسیت	نژاد	محل زندگی	حزب سیاسی
۱	مرد	سفیدپوست	کالیفرنیا	دموکرات
۲	مرد	سفیدپوست	کالیفرنیا	دموکرات
۳	مرد	سفیدپوست	تگزاس	جمهوری‌خواه
۴	مرد	سیاه‌پوست	تگزاس	جمهوری‌خواه
۵	مرد	سیاه‌پوست	اوهایو	دموکرات
۶	زن	سفیدپوست	کالیفرنیا	جمهوری‌خواه
۷	زن	سفیدپوست	تگزاس	جمهوری‌خواه
۸	زن	سیاه‌پوست	اوهایو	دموکرات
۹	زن	سیاه‌پوست	کالیفرنیا	دموکرات
۱۰	زن	سیاه‌پوست	اوهایو	جمهوری‌خواه

(آ) (۴ نمره) داده‌های جدول را در نظر بگیرید. حزب سیاسی برچسب هدف و بقیه ستون‌ها نیز ویژگی هستند. با استفاده از معیار Gini Impurity و استراتژی Multiway Split درخت تصمیم با عمق حداکثر دو را برای این داده‌ها به دست آورید.

(ب) (۴ نمره) حال با استفاده از معیار Gini Impurity و استراتژی Binary Split درخت تصمیم با عمق حداکثر دو را بر این داده‌ها به دست آورید.

(ج) (۳ نمره) این دو استراتژی را با یکدیگر مقایسه کنید و مزایا و معایب‌شان نسبت به هم را تحلیل کنید.

۴. (۴ نمره) یکی از مشکلات درخت تصمیم این است که در عین حالی که به دقت بالایی دست پیدا می‌کند اما واریانس خطای آن بالاست. توضیح دهید که جنگل تصادفی چگونه با وجود حفظ دقت بالای درخت تصمیم، واریانس خطای آن را کاهش می‌دهد.

۳.۱ پرسش سوم (۲۰ نمره)

در این سوال نشان می‌دهید که شیوه انتخاب پارامتر α_t در الگوریتم Adaboost معادل آن است که یک باند بالای نمایی برای تابع هزینه این الگوریتم را در هر دور به شکل حریصانه کمینه کنیم.

۱. (۴ نمره) فرض کنید $h_t : \mathbb{R}^m \to \{-۱, ۱\}$ دسته‌بند ضعیفی است که در گام t ام به دست آورده‌ایم و α_t وزن این دسته‌بند است. دیده‌ایم که دسته‌بند نهایی در الگوریتم Adaboost به شکل زیر است:

$$\hat{y} = sign(H_t(x)) = sign(\sum_{t=۱}^{T} \alpha_t h_t(x))$$

که H_t دسته‌بند نهایی در پایان گام t است. فرض کنید $\{(x_۱, y_۱), \cdots, (x_N, y_N)\} \subset \mathbb{R}^m \times \{-۱, ۱\}$ داده‌های آموزش ما باشند. نشان دهید خطای دسته‌بند نهایی می‌تواند توسط یک تابع هزینه نمایی از بالا محدود شود:

$$E = \frac{۱}{N} \sum_{i=۱}^{N} exp(-y_i H_t(x_i)) \geq \frac{۱}{N} \Sigma_{i=۱}^{N} \mathbb{I}(\hat{y}_i \neq y_i)$$

که $\mathbb{I}$ تابع indicator است.

۲. (۳ نمره) فرض کنید $D_t(i)$ توزیع وزن داده‌های آموزش در مرحله t است. $D_{t+۱}(i$ را بر حسب y_i, x_i, α_t, Z_t و دسته‌بند h_t بنویسید. T گام نهایی است و $t \in \{۱, \cdots, T\}$. به یاد داشته باشید که در Z_t عامل نرمال‌کننده توزیع $D_{t+۱}$ است:

$$Z_t = \sum_{i=۱}^{N} D_t(i) exp(-\alpha_t y_i h_t(x_i))$$

۳. (۳ نمره) نشان دهید که

$$E = \sum_{i=۱}^{N} \frac{۱}{N} exp(\sum_{t=۱}^{T} -\alpha_t y_i h_t(x_i))$$

۴. (۴ نمره) نشان دهید که

$$E = \prod_{t=۱}^{N} Z_t$$

۵. (۳ نمره) نشان دهید که Z_t می‌تواند به شکل زیر نوشته شود:

$$Z_t = (۱ - \epsilon_t) exp(-\alpha_t) + \epsilon_t exp(\alpha_t)$$

که ϵ_t خطای وزن‌دار دسته‌بند h_t است:

$$\epsilon_t = \sum_{i=۱}^{N} D_t(i) \mathbb{I}(h_t(x_i) \neq y_i)$$

۶. (۳ نمره) همه گام‌های بالا را به آن دلیل طی کردیم که کمینه کردن خطای ۰ یا ۱ دسته‌بند روی داده‌های آموزش دشوار است. اما کمینه کردن حریصانه باند بالای E خطا، ممکن است. نشان دهید که انتخاب حریصانه α_t در هر گام برای کمینه کردن Z_t ، به عبارت زیر منجر می‌شود:

$$\alpha_t = \frac{۱}{۲} \ln(\frac{۱ - \epsilon_t}{\epsilon_t})$$

۲ قسمت عملی

۱.۲ پرسش اول

در این سوال شما الگوریتم Adaboost را با دسته‌بندی‌های پایه decision stump پیاده‌سازی می‌کنید تا عمل دسته‌بندی روی مجموعه داده ساختگی adaboost-syndata انجام دهید. تولید داده مصنوعی (adaboost-syndata) به عهده خودتان است. مثلا می‌توانید تابعی تعریف کنید که به آن تعداد نمونه مورد نیاز را بدهید و داده برچسب‌دار در فضای دو بعدی خروجی بدهد (به طور رندوم یا هر طریقی که برای مساله مناسب‌تر می‌دانید). یک decision stump یک جداکننده خطی است که موازی یکی از محورهاست و نمونه‌های یک سویش را مثبت و نمونه‌های سوی دیگر را منفی دسته‌بندی می‌کند. مثلا فرض کنید ورودی‌ها دو بعدی هستند. در این صورت از decision stump های زیر استفاده می‌کنیم:

$$h(x) = \begin{cases} y & x_۰ \geq k \\ -y & o.w. \end{cases}$$

یا

$$h(x) = \begin{cases} y & x_۱ \leq k \\ -y & o.w. \end{cases}$$

که $y \in \{۱, -۱\}$ و k یک عدد حقیقی دلخواه است.

طراحی کد شما بر عهده خودتان است اما باید یک تابع Adaboost(X_train, y_train, X_test, num-iter) پیاده‌سازی کنید که یک بردار y_pred خروجی می‌دهد. num-iter تعداد دورهایی است که الگوریتم شما باید اجرا شود. دقت داشته باشید که داده‌ها دو بعدی هستند. بهتر است یک تابع داشته باشید که نمونه‌های یادگیری و وزن آن‌ها را ورودی می‌گیرد و خروجی بهترین decision stump را با توجه به وزن‌های فعلی و معیار decision stump و مکان آن، باز می‌گرداند. همچنین داشتن یک تابع برای به‌روزرسانی وزن‌ها و یک تابع برای پیش‌بینی برچسب‌ها با استفاده از decision stump های مراحل قبل توصیه می‌شود.

۱. (۷ نمره) در هر دور اجرای الگوریتم، خطای وزن‌دار ϵ_t دسته‌بند ضعیف خود h_t را روی داده‌های آموزش محاسبه کنید. برای $T = ۲۰$ نمودار ϵ_t را بر حسب t رسم کنید. بیشترین خطای یک دسته‌بند چقدر است؟ کجا رخ می‌دهد؟ نمودار را تفسیر کنید.

۲. (۶ نمره) در هر دور نرخ خطای دسته‌بند نهایی روی داده‌های آموزش H_t را محاسبه کنید. در $T = ۲۰$ نمودار این خطا را بر حسب t رسم کنید. نمودار را تفسیر کنید.

۳. (۶ نمره) در هر دور نرخ خطای دسته‌بند نهایی روی داده‌های تست H_t را محاسبه کنید. در $T = ۲۰$ نمودار این خطا را بر حسب t روی همان نمودار قسمت قبل رسم کنید. نمودار را تفسیر کنید.

۴. (۶ نمره) برای مقادیر $T = [۲۰, ۵۰, ۱۰۰, ۲۰۰, ۵۰۰, ۱۰۰۰, ۲۰۰۰, ۴۰۰۰]$ نرخ خطای آموزش و تست دسته‌بند نهایی را بیابید. نمودار این خطاها را رسم کنید. این نمودارها را تفسیر کنید.

۲.۲ پرسش دوم

داده‌های مورد استفاده در این تمرین مربوط به جمعیتی از قارچ‌ها است. از هر قارچ ۲۲ ویژگی استخراج شده است و هدف پیش‌بینی برچست داده‌هاست. هدف طراحی دست‌بند مناسب برای این دادگان است. پیش از شروع دادگان را به سه بخش ۷۰ درصدی آموزش و ۲۰ درصدی اعتبار سنجی و ۱۰ درصدی سنجش تقسیم کنید. دقت داشته باشید که در این تمرین می‌توانید از کتابخانه scikit-learn استفاده نمایید.

۱. (۶ نمره) یکی از چالش‌های این دادگان نمونه‌هایی هستند که برخی از ویژگی‌های آن‌ها miss شده است. با این پدیده با استراتژی‌های متفاوتی می‌توان برخورد کرد. استراتژی مورد نظر خود در برخورد با این نمونه در این مسئله را در گزارش خود ارائه دهید.

۲. (۷ نمره) با استفاده از درخت تصمیم و با مقادیر حداکثر عمق $\{۴, ۸, ۱۶, ۲۴, ۳۲\}$ دسته‌بند مناسب طراحی کنید. خطای آموزش و تست را گزارش کنید. (معیار انتخاب ویژگی و تقسیم در هر گره را معیار gini در نظر بگیرید.)

۳. (۱۲ نمره) حال با استفاده از جنگل تصادفی یک دسته‌بند مناسب طراحی کنید. دسته‌بندی را با استفاده از ترکیب نتیجه ۷ عدد درخت انجام دهید. بهترین هایپرپارامتر برای هر یک از درخت‌های جنگل را از میان $max\_depth \in \{۴, ۸, ۱۲, ۱۶\}$ و $feature\_numbers \in \{۳, ۵, ۷\}$ پیدا کنید. در نهایت دقت آموزش هر یک از تنظیمات جست‌وجو و دقت و صحت تست بهترین مدل انتخاب شده خود را گزارش کنید. (کلاس با برچسب e را به عنوان کلاس مثبت در نظر بگیرید)

موفق باشید