به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین تشریحی ۳

اردیبهشت ماه ۱۴۰۲

*فهرست

سوال ۱

مجموعه داده زیر را در نظر بگیرید:

Index	Age	Gender	Smoking	Exercise	Heart Disease
1	Young	Female	Yes	Low	Yes
2	Young	Female	No	High	No
3	Old	Female	No	High	No
4	Old	Male	Yes	High	Yes
5	Young	Female	No	Low	No
6	Young	Female	Yes	High	Yes
7	Old	Male	No	Low	No
8	Young	Male	Yes	High	Yes
9	Old	Female	Yes	High	Yes
10	Old	Female	Yes	Low	Yes

این مجموعه داده اطلاعات تعدادی از مراجعین مراکز درمانی را نشان می‌دهد که شامل سن، جنسیت، مصرف یا عدم مصرف دخانیات و میزان ورزش هر شخص را نشان می‌دهد و ابتلا یا عدم ابتلای هر شخص به بیماری قلبی مشخص شده است.

أ. به کمک الگوریتم ID3 و معیار Information gain، درخت تصمیمی که با استفاده از دادگان فوق آموزش می‌بیند را تا حداکثر عمق ۳ رسم کنید و محاسبات هر مرحله را ذکر کنید.

ب. توضیح دهید که آیا درخت تصمیم رسم شده توانایی تعمیم برای دادگانی که در مجموعه آموزش قرار ندارند را دارد یا خیر.

ت. با توجه به درخت تصمیم رسم شده مشخص کنید که هر کدام از نمونه‌های زیر در کدام کلاس قرار می‌گیرند.

Age	Gender	Smoking	Exercise
Young	Male	No	Low
Old	Male	No	High

د. اگر بدانیم برچسب داده‌های بخش (ج) به ترتیب Yes و Yes است، مقدار Precision را محاسبه کنید و ماتریس آشفتگی را برای این داده‌ها به دست آورید.

سوال ۲

مجموعه داده زیر تعدادی ایمیل به همراه برچسب هر ایمیل که نشان دهنده اسپم بودن یا نبودن آن ایمیل است را نشان می‌دهد:

Index	Text	Label
1	meet today ready	Not-spam
2	free phone today	Spam
3	free ticket	Spam
4	today ticket ready	Not spam
5	free ticket free	spam

أ. در مجموعه داده فوق، چه ویژگی‌هایی[1] برای دسته‌بندی داده‌ها مورد استفاده قرار می‌گیرند؟ پیش از محاسبه احتمالات مورد نیاز Naïve Bayes، چه فرضی را باید در مورد این ویژگی‌ها در نظر بگیریم؟

ب. آیا لغاتی که تنها در یکی از کلاس‌ها ظاهر شده‌اند، موجب بروز مشکل می‌شوند؟ در صورت بروز مشکل، راه حل چیست؟

ج. با توجه به پاسخ قسمت (ب) و استفاده از یک راه حل(در صورت وجود)، احتمال هر کلاس و احتمال تعلق هر لغت به کلاس‌ها را به دست آورید.

د. احتمال تعلق ایمیل "today phone ready" را به هر یک از دو کلاس محاسبه کنید. این ایمیل کدام برچسب را دریافت می‌کند؟

[1] Feature

سوال ۳

به سوالات زیر پاسخ دهید:

أ. فرض کنید برای ساخت درخت تصمیم از معیارهایی مانند Information gain استفاده نمی‌کنیم. در این روش در هر عمق صفت مورد استفاده برای تفکیک داده‌ها به صورت تصادفی از بین صفاتی که از ریشه تا شاخه مورد نظر استفاده نشده‌اند، انتخاب می شود. این روند تا زمانی ادامه می‌یابد که در یک مسیر از ریشه تا برگ همه‌ی صفات انتخاب شوند یا تمامی دادگان یک شاخه متعلق به یک کلاس باشند. دقت درخت تصمیم بر روی دادگان آموزش و تست در این روش چگونه خواهد بود؟

ب. آیا انتخاب روش هموارسازی[۱] مورد استفاده در Naïve Bayes (مانند Additive smothing، Laplace و ...) میتواند منجر به Overfit یا Underfit شدن مدل شوند؟ توضیح دهید.

ج. توضیح دهید که روش‌های Ensemble، مانند Bagging و Boosting، چه تفاوت‌هایی در رویکرد خود برای ترکیب خروجی مدل‌های مختلف دارند.

د. چگونه می‌توان استفاده از منحنی‌های ROC را فراتر از مسائل طبقه‌بندی دو کلاسه گسترش داد، و چه راهکارهایی برای تطبیق منحنی ROC با طبقه‌بندی چند کلاسه وجود دارد؟

ه. چگونه انتخاب تابع فعال‌سازی می‌تواند بر عملکرد یک شبکه عصبی تأثیر بگذارد؟ هر یک از توابع فعال‌ساز مانند sigmoid، ReLU، و tanh چه مزایا و معایبی ارائه می‌دهند؟

و. روش‌هایی مانند Data Augmentation نمونه‌های جدید را با اعمال تغییر شکل‌های[۲] متفاوت بر روی داده‌های موجود ایجاد می‌کنند. این روش‌ها چگونه می‌توانند به حل مشکل طبقه‌بندی مجموعه داده‌هایی با کلاس‌های نامتعادل کمک کنند و چه محدودیت‌ها و چالش‌هایی به همراه دارند؟

ز. برای انتخاب بهترین مقدار K در الگوریتم KNN چه روشی پیشنهاد می‌کنید و در صورت عدم انتخاب مقدار مناسب، عملکرد مدل چگونه دستخوش تغییر خواهد شد؟

---

[۱] Smoothing

[۲] Transformations

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_HW3_StudentID تحویل داده شود.

خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.

مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. مجموعا ۱۴ روز برای تمامی تمرین‌ها و پروژه‌ی درس به عنوان Grace day در نظر گرفته می‌شود و پس از پایان مجموعا ۱۴ روز، برای هر تمرینی که پس از زمان اختصاص یافته ارسال شود روزی ۱۵ درصد از نمره آن تمرین کسر خواهد شد.

- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:hosein7seifi@gmail.com

مهلت تحویل: ۱۴۰۲/۲/۲۰