دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی
پاییز ۱۳۹۹
استاد: محمدحسین رهبان

تمرین پنجم، بخش دوم **مقدمه‌ای بر یادگیری ماشین و درخت تصمیم‌گیری** مهلت ارسال: ۲۵ آذر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین (به استثنای هفته‌ی امتحان میانترم) تا سقف سه روز و در مجموع ۱۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهندبود.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری و یا استفاده از هر منبع خارج از کتاب و اسلایدهای درس، نام هم‌فکران و آدرس منابع مورد استفاده‌برای حل سوال مورد نظر را ذکرکنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.
- از مجموع ۱۳۰ نمره‌ی این تمرین، ۳۰ نمره امتیازی است.

سوالات نظری (۸۵ نمره)

۱. (۴۰ نمره) فرض کنید که داده‌های زیر به شما داده شده است. هدف، مشخص کردن این است که آیا شخص کامپیوتر خریداری می‌کند یا خیر. به سوالات زیر پاسخ دهید:

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no

(آ) با توجه به Information Gain یک درخت تصمیم‌گیری برای داده‌های بالا رسم کنید و دقت آن را نیز مشخص کنید.

(ب) یک گراف برای Naïve Bayes ارائه دهید و در آن فیچرها را مشخص کنید. توضیح دهید که مدلتان بر چه اساسی کار می‌کند و چگونه می‌توان با آن، لیبل یک داده‌ی جدید را مشخص کرد.

(ج) بر حسب Maximum Likelihood احتمالات این مدل را بدست آورید. فرض کنید که توزیع را باید به صورت توزیع دیریکله بدست آورید.

(د) برحسب Maximum A Posteriori احتمالات این مدل را بدست آورید. فرض کنید که Prior به شکل توزیع دیریکله است که تمام پارامترهای آن برابر ۲ است.

(ه) می‌دانیم به Model-Based Classificationها، Generative Model هم گفته می‌شود. به این دلیل که می‌توان از روی این مدل‌ها، داده‌ی جدید درست کرد. توضیح دهید که از روی گراف Naïve Bayes و با داشتن توزیع احتمالاتی فیچرها، چگونه می‌توان داده‌ی جدید تولید کرد و چرا نمی‌توان از درخت تصمیم‌گیری برای این کار استفاده کرد.

۲. (۲۰ نمره) فرض کنید در حال ساخت یک درخت تصمیم‌گیری هستیم و ورودی‌ها، ماشین‌های مختلفی هستند. هدف این است که بفهمیم آیا مصرف سوخت یک ماشین خوب است یا بد. در یک راس، می‌خواهیم برحسب فیچر E داده‌ها را دسته بندی کنیم. این فیچر، k مقدار مختلف اختیار می‌کند. فرض کنید که تعداد داده‌های خوبی که در آن‌ها $E = E_k$ ، برابر با p_k باشد. هم‌چنین تعداد داده‌های بد که $E = E_k$ نیز برابر با n_k باشد. ثابت کنید که $IG(E)$ مقداری مثبت خواهد بود مگر این که تمام $\frac{p_k}{p_k+n_k}$ ها به ازای هر k با هم برابر باشند. به صورت شهودی، این شرط چه معنایی می‌دهد ؟

۳. (۲۵ نمره) در مدل Naïve Bayes فرض می‌کنیم که با شرط داشتن لیبل، تمامی فیچرها از هم دیگر مستقل هستند. اما در واقعیت با مشخص بودن لیبل باز هم ممکن است که ارتباطی بین فیچرها وجود داشته باشد. یکی از راه‌هایی که می‌توان از اطلاعات این ارتباط استفاده کرد، این است که correlation بین فیچرها را محاسبه کنیم و از آن در بدست آوردن احتمال‌ها استفاده کنیم.

حال فرض کنید که یک مسأله‌ی classification داریم. لیبل یک داده را با Y و فیچرهای آن را با X نشان می‌دهیم. ضمنا می‌دانیم که : $X = (X_1, X_2, ..., X_n)$. در سوال ما، $n = 2$ است.

می‌دانیم که لیبل می‌تواند ۳ مقدار مختلف اختیار کند و داریم :

$$P(Y = 1) = P(Y = 2) = P(Y = 3) = \frac{1}{3}$$

در حالت Naïve bayes ، باید احتمال $P(X_i|Y)$ را هم می‌داشتیم تا مدل کامل باشد. اما اکنون موارد زیر به ما داده شده است :

$$\forall 1 \leq i \leq 3 : (X|Y = i) \sim N(\mu_i, \Sigma_i)$$

هم‌چنین می‌دانیم که :

$$\mu_1 = [0, 0]^T, \mu_2 = [1, 1]^T, \mu_3 = [-1, 1]^T$$

$$\Sigma_1 = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.7 \end{bmatrix} \quad (۱)$$

$$\Sigma_2 = \begin{bmatrix} 0.8 & 0.3 \\ 0.3 & 0.2 \end{bmatrix} \quad (۲)$$

$$\Sigma_3 = \begin{bmatrix} 0.7 & 0.2 \\ 0.2 & 0.8 \end{bmatrix} \quad (۳)$$

در واقع ما توزیع توأمان دو فیچرمان یعنی X_1 و X_2 را داریم. همچنین توزیع توأمان این دو متغیر تصادفی، توزیع نرمال ۲ متغیره است. حال اگر ورودی‌های زیر را داشته باشیم، لیبل داده‌ها را بدست آورید:

(آ) $x = [-0.5, 0, 5]$

(ب) $x = [0.5, 0.5]$

سوالات عملی (۳۵ + ۱۰ نمره)

۱. (۳۵ + ۱۰ نمره) در این قسمت شما باید با کمک Naïve Bayes یک مدل طراحی کرده و با داده‌های موجود، مدل خود را تمرین دهید. پس از اتمام تمرین دادن، باید برای داده‌های تست، لیبل آن‌ها را حدس بزنید.

توجه ۱ : در صورتی که شما در سوالی تنها یک دسته داده داشتید و دسته‌ی training از دسته‌ی test جدا نشده بود، شما باید ۱۰ درصد از داده‌ها را به شکل تصادفی انتخاب کنید. این ۱۰ درصد، داده‌های تست شما را تشکیل می‌دهند و مابقی ۹۰ درصد، داده‌های تمرین شما هستند.

توجه ۲ : در حل مسائل یادگیری ماشین، شما به هیچ وجه نباید مدل خود را بر اساس داده‌های تست تمرین دهید. در صورتی که چنین چیزی در کد شما دیده شود، نمره‌ی شما ۰ خواهد شد.

توجه ۳ : در حل مسائل یادگیری ماشین، آخرین کاری که انجام می‌شود باید اجرای مدل بر روی دادهای تست باشد. اگر شما مدلتان را بر روی داده‌های تست اجرا کنید و بعد مدلتان را تغییر دهید تا دقت بیشتری حاصل کنید، از داده‌های تست برای تمرین مدل خودتان استفاده کرده‌اید و راه شما فاقد اعتبار است.

در این قسمت، شما باید بر روی تعدادی متن، عملیات classification انحام دهید. داده‌های ما، به شکل یک متن هستند و هر متن یک category دارد که همان لیبل آن است. به مثال زیر توجه کنید :

text : tv future in the hands of viewers with home th... $\Rightarrow$ category : tech

حال شما تعدادی داده در فایل bbc-text.csv دارید. یک مدل طراحی کنید و آن را تمرین دهید. سپس متن‌های داده‌های تست را بخوانید و لیبل آن‌ها را پیش‌بینی کنید. در نهایت دقت مدلتان را چاپ کنید.

برای گرفتن نمره‌ی کامل، باید درصد دقت شما بیشتر از ۸۰ درصد باشد. اگر دقت کمتری داشته باشید، به نسبت دقت شما، از نمره‌ی شما کم می‌شود. ضمنا به ازای هر درصد دقت بیش از ۹۰ درصد، یک نمره‌ی مثبت می‌گیرید.