بخش تئوری

سوال اول

یکی از مباحثی که در درخت تصمیم مطرح میشود هرس درخت [1] برای جلوگیری از بیش برازش است. توضیح دهید چرا نمیتوان از مجموعه داده جدا برای هرس درخت استفاده کرد؟ منظور این است که داده‌هایی که برای هرس استفاده میشوند با مجموعه داده‌ای که برای ساخت درخت استفاده میشود یکسان نباشند.

اشکال استفاده از مجموعه جداگانه تاپل ها برای ارزیابی هرس این است که ممکن است نماینده تاپل های آموزشی مورد استفاده برای ایجاد درخت تصمیم اصلی نباشد. اگر مجموعه مجزای تاپل ها دارای انحراف باشند، استفاده از آنها برای ارزیابی درخت هرس شده نشانگر خوبی برای دقت طبقه بندی درخت هرس شده نخواهد بود. علاوه بر این، استفاده از مجموعه جداگانه ای از تاپل ها برای ارزیابی هرس به این معنی است که تعداد تاپل های کمتری برای ایجاد و آزمایش درخت وجود دارد. در حالی که این یک اشکال در یادگیری ماشینی به حساب می آید، ممکن است در داده کاوی به دلیل در دسترس بودن مجموعه داده های بزرگتر، چنین نباشد.

سوال دوم

با توجه به مطالب تدریس شده در کلاس، برای داده‌های زیر یک درخت تصمیم درست کنید. (ذکر تمام مراحل و توضیح آنها لازم است)

آیا به مهمانی دعوت میشود؟	وزن	قد	رنگ لباس
خیر	لاغر	۱۷۰	قرمز
بله	چاق	۱۶۲	آبی
خیر	چاق	۱۶۵	سبز
بله	لاغر	۱۷۲	سبز
بله	لاغر	۱۶۰	آبی

[1] Tree pruning

پاسخ:

در ابتدا آنتروپی کلی را حساب میکنیم:

$$E([3+,2-]) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.96$$

سپس باید Gain هر ویژگی را به دست بیاوریم تا ببینیم کدام ویژگی در ریشه قرار میگیرد.

رنگ لباس:

رنگ قرمز = $E([0+,1-]) = 0$

رنگ آبی= $E([2+,0-]) = 0$

رنگ سبز= $E([1+,1-]) = 1$

$$Gain\left(\text{رنگ لباس}\right) = E\left(\text{دعوت}\right) - \sum_{v\in(yes,no)} \frac{|D_v|}{|D|} E(D_v) = 0.96 - \frac{1}{5}\times 0 - \frac{2}{5}\times 0 - \frac{2}{5}\times 1$$
$$= 0.56$$

قد:

چون قد یک ویژگی پیوسته است باید آن را گسسته سازی کنیم. از راه best split point برای گسسته سازی استفاده میکنیم. قدها را به صورت صعودی مرتب کرده و نقطه‌ای انتخاب میشود که یک طرف آن به طور خالص از یک کلاس هستند که نقطه بین ۱۶۳ و ۱۶۵ انتخاب میشود :

160+ , 163+ | , 165- , 170- , 172+

$$\frac{163+165}{2} = 164$$

پس در واقع قدها به دو دسته کوچکتر از ۱۶۴ و بزرگتر از ۱۶۴ تقسیم میشوند.

قد کمتر از ۱۶۴ = $E([2+,0-]) = 0$

قد بیشتر از ۱۶۴= $E([1+,2-]) =$

$$Gain\left(\text{قد}\right) = E\left(\text{دعوت}\right) - \sum_{v\in(yes,no)} \frac{|D_v|}{|D|} E(D_v) = 0.96 - \frac{2}{5}\times 0 - \frac{3}{5}\times 0.9 = 0.42$$

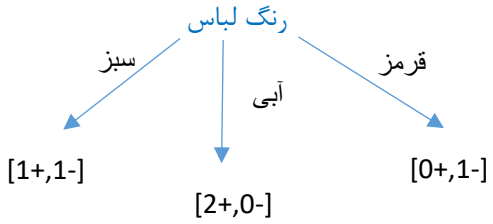
وزن:

چاق = E([1+,1-]) = 1

لاغر= E([2+,1-]) = 0.9

$$Gain\left(\text{وزن}\right) = E\left(\text{دعوت}\right) - \sum_{v\in(yes,no)} \frac{|D_v|}{|D|} E(D_v) = 0.96 - \frac{2}{5}\times 1 - \frac{3}{5}\times\ 0.9 = 0.02$$

چون رنگ لباس از همه Gain بالاتری دارد در ریشه قرار میگیرد.



چون شاخه آبی و قرمز دارای آنتروپی صفر هستند(خالص هستند) پس شاخه برگ محسوب شده و نیازی به ادامه دادن ندارند.

ساخت درخت را از شاخه سبز ادامه میدهیم.

قد:

قد کمتر از ۱۶۴ و رنگ سبز = E([0+,0-]) = 0

قد بیشتراز ۱۶۴ و رنگ سبز = E([1+,1-]) = 1

$$Gain\left(\text{قد}\right) = E\left(\text{سبز}\right) - \sum_{v\in(yes,no)} \frac{|D_v|}{|D|} E(D_v) = 1 - \frac{0}{2}\times 0 - \frac{2}{2}\times 1 = 0$$

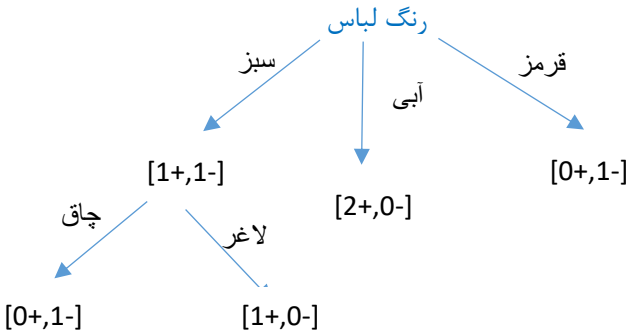
وزن:

چاق و رنگ سبز = E([0+,1-]) = 0

لاغر و رنگ سبز = E([1+,0-]) = 0

$$Gain\left(\text{وزن}\right) = E\left(\text{سبز}\right) - \sum_{v\in(yes,no)} \frac{|D_v|}{|D|} E(D_v) = 1 - 0 = 0$$

چون وزن Gain بیشتری دارد پس وزن برای شاخه بعدی انتخاب میشود.



و ساخت درخت به علت اینکه همه گره ها برگ هستند تمام میشود.

سوال سوم

در جدول داده شده زیر با استفاده از قانون بیز برچسب داده زیر را به دست آورید. در صورت صفر شدن احتمال از هموارسازی لاپلاس[۲] استفاده کنید.

(معدل = عالی ، مطالعه = بله ، حضور = خیر)

پاس شدن	حضور در کلاس‌ها	مطالعه برای امتحان	معدل
خیر	خیر	خیر	ضعیف
بله	بله	بله	ضعیف
خیر	خیر	خیر	متوسط
بله	بله	بله	متوسط
بله	خیر	خیر	عالی
بله	بله	بله	عالی

[2] laplace smoothing

$$P\left(\text{پاس} = \text{بله}\right) = \frac{4}{6}$$

$$P\left(\text{پاس} = \text{خیر}\right) = \frac{2}{6}$$

$$P\left(\text{معدل} = \text{عالی} \mid \text{پاس} = \text{بله}\right) = \frac{2}{4} = \frac{1}{2}$$

$$P\left(\text{مطالعه} = \text{بله} \mid \text{پاس} = \text{بله}\right) = \frac{3}{4}$$

$$P\left(\text{کلاس} = \text{خیر} \mid \text{پاس} = \text{بله}\right) = \frac{1}{4}$$

$$P\left(\text{معدل} = \text{عالی} \mid \text{پاس} = \text{خیر}\right) = \frac{0+1}{2+3} = \frac{1}{5}$$

$$P\left(\text{مطالعه} = \text{بله} \mid \text{پاس} = \text{خیر}\right) = \frac{0+1}{2+2} = \frac{1}{4}$$

$$P\left(\text{کلاس} = \text{خیر} \mid \text{پاس} = \text{خیر}\right) = \frac{2}{2} = 1$$

$$P\left(\text{x} \mid \text{پاس} = \text{بله}\right) = \frac{1}{2} \times \frac{3}{4} \times \frac{1}{4} = 0.09$$

$$P\left(\text{x} \mid \text{پاس} = \text{خیر}\right) = \frac{1}{5} \times \frac{1}{4} \times 1 = 0.05$$

$$P\left(\text{x} \mid \text{پاس} = \text{بله}\right) \times P\left(\text{پاس} = \text{بله}\right) = 0.09 \times \frac{4}{6} = 0.06$$

$$P\left(\text{x} \mid \text{پاس} = \text{بله}\right) \times P\left(\text{پاس} = \text{بله}\right) = 0.05 \times \frac{2}{6} = 0.01$$

پس برچسب داده مورد نظر **پاس شدن** است.

سوال چهارم

همانطور که میدانیم یکی از معیارها برای ارزیابی مدل‌های یادگیری نظارت شده صحت [3] است. اما این معیار در برخی موارد ممکن است معیار مناسبی برای ارزیابی نباشد. موقعیت‌هایی که این معیار برای ارزیابی به خوبی عمل نمیکند را توضیح دهید.

همانطور که میدانیم صحت به صورت زیر تعریف میشود:

$$accuracy = \frac{correct\ classifications}{all\ calssifications}$$

این معیار هنگامی مناسب است که ما از هر کلاس به میزان تقریبا مساوی داده داشته باشیم یا به عبارت دیگر مجموعه داده ما بالانس باشد. و هنگامی که مجموعه داده بالانس نباشد و به طور مثال یک کلاس مقدار بسیار بیشتری داده از کلاس‌های دیگر داشته باشد صحت معیار مناسبی برای ارزیابی کلی مدل نیست. برای مثال اگر ۹۰ درصد داده‌ها مربوط به کلاس A و تنها درصد مربوط به کلاس B باشند حتی اگر مدل در تشخیص کلاس B ضعیف عمل کند باز هم مدل صحت بالایی را نتیجه میدهد.

سوال پنجم

فرض کنید که برای انتخاب پارامتر α در مدل از روش 10 fold cross validation استفاده کرده‌ایم. بهترین روش برای انتخاب مدل نهایی و تخمین ارور کدام است؟

با استفاده از کل دیتا ست مدل را با اون پارامتر پیدا شده آموزش میدهیم و از ارور میانگین در کراس ولیدیشن به عنوان تخمین ارور استفاده میکنیم.

[3] Accuracy

سوال ششم

در الگوریتم boosting اگر هر کدام از موارد زیر رخ دهد ما یادگیری را متوقف میکنیم؟ برای پاسخ‌های خود دلیل بیاورید.

- میزان خطای طبقه‌بندی‌کننده ترکیبی در داده‌های آموزشی اصلی ۰ شود. خیر. بوستنیگ به بیش برازش[۴] مقاوم است. حتی بعد از اینکه خطای آموزش ۰ باشد خطای تست ممکن است کاهش یابد.

- میزان خطای طبقه‌بندی‌کننده ضعیف[۵] فعلی روی داده‌های تمرین وزن‌دار[۶] ۰ است. بله. طبقه‌بندی کننده ضعیف فعلی روی داده‌های تمرین وزن دار بسیار عالی عمل میکند پس میتوان گفت روی داده اصلی نیز عالی عمل میکند و نیازی به ترکیب این طبقه‌بندی کننده با طبقه‌بندی کننده‌های دیگر نیست

سوال هفتم

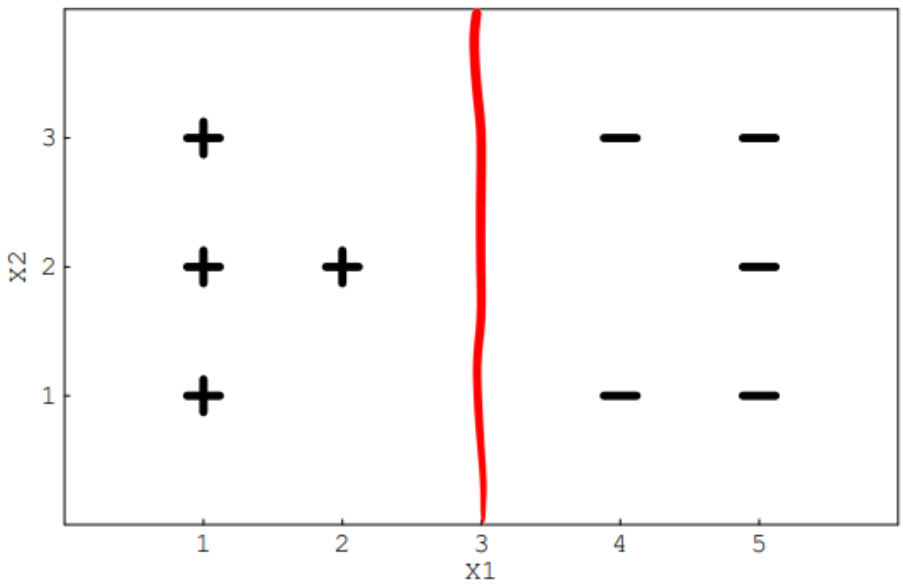
فرض کنید برای داده‌های زیر از طبقه‌بندی کننده svm خطی بدون کرنل استفاده میکنیم و پارامتر C در این طبقه بندی کننده بسیار بزرگ در نظر گرفته شده است. (اگر در مورد این پارامتر اطلاعی ندارید این لینک را مطالعه کنید).

الف) خطی که svm گفته شده با استفاده از آن داده ها را دسته میکند را رسم کنید و علت انتخاب این خط را توضیح دهید.

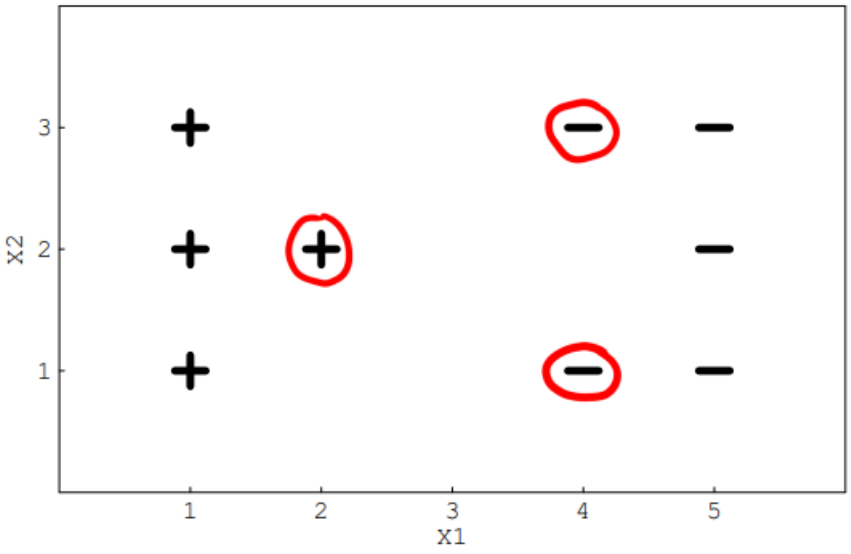
پاسخ:

به دلیل مقدار C بزرگ (کمترین مقدار ارور)، مرز تصمیم گیری همه را طبقه بندی می کند علاوه بر این، در میان جداکننده هایی که نمونه ها را به درستی طبقه بندی می کنند،بیشترین حاشیه (فاصله تا نزدیکترین نقطه) را خواهد داشت.

[4] overfit
[5] Weak
[6] Weighted training data



ب) در شکل بالا نقاطی را انتخاب کنید که حذف آنها باعث میشود خطی که svm داده ها را جدا میکند متفاوت از حالت الف شود. دلیل انتخاب این نقاط را توضیح دهید.



نقاط انتخاب شده همان بردارهای پشتیبان هستند بنابر حذف آنها باعث تغییر در مرز تصمیم گیری میشوند.

سوال هشتم

صحیح یا غلط بودن موارد زیر را با دلیل مشخص کنید :

الف) الگوریتم بیز ساده نمیتواند وابستگی بین متغیرها را مشخص کند. صحیح. زیرا در الگوریتم بیز ساده این فرض در نظر گرفته شده است که متغیرها از هم مستقل هستند.

ب) هنگامی که یک درخت تصمیم به سمت یک درخت پر پیش میرود احتمال اینکه نویز را هم پوشش دهد بیشتر میشود. صحیح. زیرا هر چقدر عمق درخت بیشتر میشود احتمال اینکه داده‌هایی که نویز باشند را هم پوشش دهد بیشتر میشود.

ج) در روش k نزدیک ترین همسایه[7] اگر k=1 الگوریتم نسبت به داده‌های نویز مقاوم تر از حالتی است که k=5 در نظر گرفته شود. غلط. زیرا وقتی k=1 در نظر گرفته شود داده های نویز تاثیر بیشتری بر طبقه بندی میگذارنند

سوال نهم

فرض کنید در حال طراحی یک سیستم برای تشخیص خستگی راننده در اتومبیل هستید. بسیار مهم است که مدل شما خستگی را تشخیص دهد تا از هر گونه حادثه ای جلوگیری شود. کدام یک از معیارهای زیر بهترین معیار برای ارزیابی هست : Accuracy, Precision, Recall, Loss Value دلیل انتخاب خود را شرح دهید.

Recall معیار مناسبتری است زیرا نباید هیچ موردی که راننده خسته است اشتباه تشخیص داده شود.

یادآوری:

$$recall = \frac{TP}{TP + FP}$$

[7] K nearest neighbor

سوال دهم

علاوه بر شاخص آنتروپی برای ساخت درخت تصمیم، شاخص دیگری نیز وجود دارد که میتوان به جای آنتروپی از آن برای ساخت درخت استفاده کرد. این شاخص را معرفی کنید و بگویید تفاوت آن با آنتروپی چیست؟ بالا یا پایین بودن این شاخص چه معنایی دارد و چگونه محاسبه میشود.

شاخص جینی که به عنوان ناخالصی جینی نیز شناخته می شود، میزان احتمال یک ویژگی خاص را محاسبه می کند که در صورت انتخاب تصادفی به اشتباه طبقه بندی می شود. اگر همه عناصر با یک کالس مرتبط باشند، می توان آن را خالص نامید. این شاخص یکی از چند پارامتر موجود برای split کردن و درست کردن درخت تصمیم است

عدد این شاخص بین صفر و یک قرار دارد که عدد صفر خالص بودن را نشان میدهد که وقتی اتفاق میفتد که تنها یک کلاس داریم یا تمام داده‌ها به یک کلاس خاص متعلق هستند. عدد یک توزیع تصادفی داده‌ها بین کلاسها را نشان میدهد و عدد ۵,۰ توزیع برابر دادهها بین کلاسها را نشان میدهد. فرمول این شاخص به صورت زیر است که pi نشان دهنده احتمال طبقه بندی یک عنصر برای یک کلاس مجزا است. پس هر چقدر عدد آن کوچکتر باشد میزان خالصی باالتر است و هر چه به یک نزدیک شود به توزیع رندم نزدیک میشود.

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

سوال یازدهم

درمورد مسائل رگرسیون به سوالات زیر پاسخ دهید :

الف) simple linear regression و multiple linear regression با یکدیگر مقایسه کرده و تفاوت و شباهت های آنها را بیان کنید.

رگرسیون خطی که می توان از آن به عنوان رگرسیون خطی ساده نیز یاد کرد، رایج ترین شکل تحلیل رگرسیون است و به دنبال خطی است که بر اساس مجموعه ای از معیارهای ریاضی به بهترین وجه با داده ها مطابقت دارد. به زبان ساده، از یک خط مستقیم برای تعریف رابطه بین دو متغیر استفاده می کند. بنابراین، اگر بخواهیم مقدار یک متغیر را بر اساس مقدار متغیر دیگری تخمین بزنیم، از این فرمول استفاده می کنیم.

اما، در نمونه ای که دو (۲) متغیر مستقل وجود دارد، به یک رگرسیون خطی چندگانه تبدیل می شود. این بدان معنی است که رگرسیون خطی چندگانه یا رگرسیون چندگانه زمانی است که دو یا چند متغیر توضیحی مستقل با متغیر وابسته رابطه خطی داشته باشند .

ب) یکی از راه‌های جلوگیری از بیش برازش استفاده از منظم‌سازی است که به دو نوع L1 و L2 تقسیم میشود. به نوع اول Lasso Regression و به نوع دوم Ridge regression گفته میشود. تفاوت این دو روش را از نوع بهینه سازی بیان کرده و نحوه کار آنها را توضیح دهید.

منظم سازی برای جلوگیری از برازش بیش از حد داده ها، به ویژه زمانی که اختلاف زیادی بین عملکرد مجموعه آموزش و مجموعه تست وجود دارد. با منظم‌سازی، تعداد ویژگی‌های مورد استفاده در تمرین ثابت نگه داشته می‌شود، اما مقدار ضرایب (w) کاهش می‌یابد.

Lasso Regression

این یک تکنیک منظم‌سازی است که در انتخاب ویژگی با استفاده از روش انقباض استفاده می شود که به آن روش رگرسیون جریمه‌شده[8] نیز گفته می شود. Lasso مخفف Least Absolute Shrinkage and Selection Operator است که هم برای منظم سازی و هم برای انتخاب مدل استفاده می شود. این روش عبارت جریمه را به تابع هزینه اضافه می کند. که این عبارت جریمه مجموع مطلق ضرایب است که باعث کاهش مقدار ضرایب به منظور کاهش ضرر میشود. این روش تمایل دارد که ضرایب را به صفر مطلق میل دهد.

$$L_{lasso} = argmin_{\hat{\beta}} \left(\|Y - \beta * X\|^2 + \lambda * \|\beta\|_1 \right)$$

Ridge Regression

تفاوت این روش با روش قبلی در این است که مقدار جریمه‌ای که به تابع هزینه اضافه میشود برابر با مجذور ضرایب است. بر خلاف Lasso این روش هیچگاه ضرایب را به سمت صفر مطلق سوق نمیدهد.

$$L_{ridge} = argmin_{\hat{\beta}} \left(\|Y - \beta * X\|^2 + \lambda * \|\beta\|_2^2 \right)$$

[8] Penalized

ج) در جدول زیر سن و فشار خون چند بیمار قلبی داده شده است. معادله رگرسیون به فرم $y = \beta_0 + \beta_1 x$ به دست آورید. همچنین با استفاده از معادله به دست آمده فشار خون یک بیمار ۴۰ ساله را پیش بینی کنید.(متغیر x نشان دهنده سن و متغیر y نشان دهنده فشار خون است)

Patient	A	B	C	D	E	F	G
x	42	74	48	35	56	26	60
y	98	130	120	88	182	80	135

$$x = \begin{bmatrix} 1 & 42 \\ 1 & 74 \\ 1 & 48 \\ 1 & 35 \\ 1 & 56 \\ 1 & 26 \\ 1 & 60 \end{bmatrix}$$

$$y = \begin{bmatrix} 98 \\ 130 \\ 120 \\ 88 \\ 182 \\ 80 \\ 135 \end{bmatrix}$$

$$x^T x\beta = x^T y$$

$$x^T x = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 42 & 74 & 48 & 35 & 56 & 26 & 60 \end{bmatrix} \times \begin{bmatrix} 1 & 42 \\ 1 & 74 \\ 1 & 48 \\ 1 & 35 \\ 1 & 56 \\ 1 & 26 \\ 1 & 60 \end{bmatrix} = \begin{bmatrix} 7 & 341 \\ 341 & 18181 \end{bmatrix}$$

$$x^T y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 42 & 74 & 48 & 35 & 56 & 26 & 60 \end{bmatrix} \times \begin{bmatrix} 98 \\ 130 \\ 120 \\ 88 \\ 182 \\ 80 \\ 135 \end{bmatrix} = \begin{bmatrix} 833 \\ 42948 \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 7 & 341 \\ 341 & 18181 \end{bmatrix}^{-1} \times \begin{bmatrix} 833 \\ 42948 \end{bmatrix} = \begin{bmatrix} 44.34 \\ 1.532 \end{bmatrix}$$

$$y = 1.532x + 44.34$$

تخمین فشار خون برای سن ۴۰ :

$$y = 1.532 \times 40 + 44.34 = 105.62$$