



دانشگاه صنعتی امیرکبیر

(پلیتکنیک تهران)

دانشکده مهندسی کامپیوتر

درس داده کاوی

تمرین دوم ۲

امیرمهدی زرین نژاد

۹۷۳۱۰۸۷

نیمسال دوم ۱۴۰۰

سوالات تشریحی

سوال ۱) مثال از مسائل دسته‌بندی و رگرسیون

دسته‌بندی:

- در دسته‌بندی می‌توان مثلا از روی تعداد زیادی تصویر حیوانات (مثلا سگ یا گربه) یاد گرفت و با دریافت یک تصویر جدید از آن حیوانات تشخیص داد که در کدام دسته حیوان قرار می‌گیرد. (مثل تشخیص حیوانات و طبیعت در گوشی‌های هوشمند و ...)
- یا در تصویر برداری با گوشی هوشمند مناظر طبیعی را از دیگر دسته مناظر تشخیص داد و براساس منظره، نور و رنگ تصویر برداری را تنظیم کرد.
- به همین صورت می‌توان چهره افراد مختلف را دسته‌بندی کرد و تصاویر مربوطه‌شان را جدا کرد و گفت که در تصویر چه فردی وجود دارد. (مانند گالری گوشی‌های هوشمند که افراد مختلف را از یکدیگر تفکیک می‌کند و یا سیستم تشخیص چهره که در سیستم‌های امنیتی استفاده می‌شود)
- تشخیص ایمیل‌های اسپم از نمونه کاربردهای دسته‌بندی است. به این صورت که ارائه دهنده سرویس ایمیل سیستمی دارد که دسته‌های مختلف ایمیل را یاد گرفته و با دریافت ایمیل جدید می‌تواند تشخیص دهد که ایمیل اسپم است یا در دسته‌های پرایمری، آپدیت، سوشیال و ... قرار می‌گیرد.
- در هواشناسی مثلا اینکه تشخیص دهیم هوا گرم است یا سرد است و یا اینکه در دسته‌ی ابری قرار می‌گیرد یا نیمه‌ابری یا آفتایی و
- در پزشکی می‌توان از دسته‌بندی برای تشخیص بیماران کرونایی (و یا دیگر بیماری‌های ریوی) استفاده کرد به این صورت که تصویر ریه افراد بیمار را همراه با دیگر مشخصات پزشکی‌شان مدل کنیم و با دریافت یک تصویر جدید و اطلاعات فرد بیمار تشخیص دهیم به بیماری مبتلا هست یا خیر.

رگرسیون:

- از رگرسیون می‌توان مثلا در پیش‌بینی دما و هوا با توجه به شرایط جوی استفاده کرد.
- بررسی ارتباط و تاثیر ورزش بر طول عمر و میزان سلامتی.
- تاثیر مصرف دخانیات بر احتمال ابتلا به بیماری‌های ریوی یا سرطان در افراد مصرف کننده.
- تاثیر کار با کامپیوتر و صفحه‌نمایش از یک تکنولوژی خاص بر چشمان افراد برنامه‌نویس.
- و یا برخی مثال‌های اقتصادی: بدست آوردن ارتباط نرخ اجاره با نرخ جمعیت شهر.
- بررسی تاثیر سال‌های تحصیل و سطح تحصیل خانوار بر میزان درآمد.

بخش اول:

۱ - Accuracy: یا صحت برابر است با مجموع تعداد پیش‌بینی‌های درست(چه مثبت و چه منفی) تقسیم بر تعداد کل پیش‌بینی‌ها. یعنی از کل پیش‌بینی‌هایی که کردیم چه قدر شان درست بودند.

$$accuracy = \frac{TP + TN}{All(TP + TN + FP + FN)}$$

۲ - Recall: (یا sensitivity یا پوشش) تعداد مثبت‌هایی که درست پیش‌بینی کردیم بر تعداد کل مثبت‌های واقعی (هم آن مثبت‌هایی که درست پیش‌بینی کردیم و هم آن مثبت‌هایی که اشتباه‌ها منفی پیش‌بینی کردیم - نرخ تشخیص مثبت‌های واقعی) و بیان می‌کند که چه قدر از داده‌های مثبت را توانستیم پوشش دهیم و پیش‌بینی کنیم.

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

۳ - Precision: یا دقّت؛ برابر است با تعداد مثبت‌هایی که درست پیش‌بینی کردیم بر کل پیش‌بینی‌های مثبت ما (هم آن‌هایی که به درستی مثبت پیش‌بینی کردیم و هم آن‌هایی که منفی هستند اما ما اشتباه‌ها مثبت درنظر گرفتیم) و بیان می‌کند چه قدر از داده‌هایی که مثبت پیش‌بینی کردیم، واقعاً مثبت هستند و درست پیش‌بینی شده‌اند.

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

۴ - F1-Score: (یا F1-measure) استفاده‌ی ترکیبی و یک نوع میانگین از معیارهای recall و precision است و می‌تواند به عنوان یک معیار برای امتیازدهی به تست‌ها مورد استفاده قرار بگیرد.

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

بخش دوم:

همانطور که متوجه شده‌اید همه معیارهای بالا برای اندازه‌گیری دقیق و میزان خوب بودن یک مدل است. ولی چرا فقط از همان مورد اول همه‌جا استفاده نمی‌کنیم؟ آیا کاربردهای آن‌ها در جاهای مختلف متفاوت است؟ ۳ سenario مطرح کنید و عدم کاربرد و یا خوبی معیارهای مختلف را با هم مقایسه کنید.

بله با توجه به فضای مورد بررسی و کاربرد مورد نظر می‌تواند استفاده از آن‌ها متفاوت باشد. چراکه مثلاً در برخی جاهای میزان عملکرد بالاست و میزان خطاهای بسیار ناچیز و حساس است و لازم است تا نرخ خطأ و دقیقت را بیشتر مورد توجه قرار دهیم به جای اینکه نرخ عملکرد درست و صحبت را در نظر بگیریم. گاهی هم بر عکس، حجم داده‌های مثبت و منفی بالانس است و صحبت به کار می‌آید پس میزان عملکرد صحیح مد نظر است و در نظر گرفتن خطأ معنای خاصی ندارد.

مثلاً بخواهیم بیماران مبتلا به کرونا را تشخیص دهیم و قرنطینه کنیم. در این حالت پیدا کردن افراد بیمار تا حد ممکن خیلی مهم‌تر از این است که به اشتباه تعداد بسیار کمی از افراد سالم را بیمار تشخیص دهیم. چراکه یک فرد بیمار می‌تواند خیلی خطر بیشتری داشته باشد. پس accuracy که صحبت کل پیش‌بینی‌های ما را بیان می‌کند (و نمی‌تواند تفاوتی بین خطای False Negative و خطای False Positive داشته باشد و تمامی خطاهای را یکسان در نظر می‌گیرد) خیلی مناسب نیست و معیارهای دیگر مهم‌تر هستند. مخصوصاً پوشش که FN ها را در نظر می‌گیرد. و یا F1-score که از هر دو استفاده می‌کند. (یا مثلاً در بررسی عملکرد نیروهای موشکی نظامی هم همین‌طور است چراکه کوچک ترین خطاهای می‌توانند آسیب‌بزرگی به خود سیستم وارد کنند و برای خود ارتش مرگ‌افرین باشند)

یا مثلاً برای یک مدل و الگوریتم بازیابی اطلاعات ویژگی Recall معیار بهتری خواهد بود. چراکه تعداد در پایگاه داده حجم داده‌های غیر مرتبط با کوئری وارد خیلی بیشتر از مرتبط‌ها است. و برای کاربر خیلی مهم‌تر خواهد بود که داده‌های مرتبط بیشتری پیدا کند.

از طرفی در کاربردی مانند بررسی وضعیت ذرات اتمی و زیر اتمی، accuracy معیار خوبی است چراکه توزیع نتایج مثبت و منفی بالانس است و همین‌طور احتمال خطأ پایین نیست. اما همان تعداد کم تخمین‌های صحیح می‌توانند موفقیت بزرگی باشند و اطلاعات زیادی به دانشمندان بدهند.

یا مثلاً می‌خواهیم عملکرد یک مدل یادگیری ماشین را که برای دسته‌بندی داده‌های مغزی و پزشکی طراحی شده است بررسی کنیم. در اجرا و استفاده بر روی داده‌های critical و واقعی (که عملکرد و تعداد تخمین‌های درست بالاست، حجم مثبت و منفی بالانس نیست و هدف پیدا کردن خطاهای دقیقت و پوشش مدل است)، صحبت (accuracy) بیان خیلی خوبی از عملکرد مدل نخواهد بود و معیارهای دیگر بهتر هستند.

سوال (۳)

درد سینه	گردش خون مناسب	عروق خونی بسته	بیماری قلبی دارد
خیر	خیر	خیر	خیر
بله	بله	بله	بله
بله	بله	خیر	خیر
بله	خیر	بله	بله

برای رسم درخت تصمیم گیری لازم است تا Gain ویژگی‌های مختلف را محاسبه کنیم و آن که از همه بیشتر بود را در ریشه قرار دهیم. در ادامه همین کار را برای بقیه ویژگی‌ها و رسم بقیه گره‌های درخت انجام می‌دهیم. هرگاه بله یا خیر در یال یک ویژگی به ما آنتروپی Yes یا No می‌گذاریم و درخت را دیگر ادامه نمی‌دهیم (به برگ رسیده‌ایم).

$$\begin{aligned}
 & \text{درد سینه} = E(\text{درد سینه}) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = +1 \\
 & \text{Gain} (\text{درد سینه}) = E_{\text{درد سینه}} - \frac{1}{2} E(\text{خیر}) - \frac{1}{2} E(\text{بله}) \\
 & \Rightarrow E(\text{خیر}) = -0.5 \times 1 = 0, \quad E(\text{بله}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.918 \\
 & \Rightarrow \text{Gain} = 1 - 0 - \frac{1}{2} \times 0.918 = 0.081 \\
 \\
 & \text{عرق خونی بسته} = E(\text{عرق خونی بسته}) = -\frac{1}{2} E(\text{خیر}) - \frac{1}{2} E(\text{بله}) \\
 & \Rightarrow E(\text{خیر}) = -0.5 \times 1 = 0, \quad E(\text{بله}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = +1 = E(\text{بله}) \\
 & \Rightarrow \text{Gain} = 1 - \frac{1}{2} \times 1 - \frac{1}{2} \times 1 = 0 \\
 \\
 & \text{Gain} (\text{عرق خونی بسته}) = E(\text{عرق خونی بسته}) - \frac{1}{2} E(\text{خیر}) - \frac{1}{2} E(\text{بله}) \\
 & \Rightarrow E(\text{خیر}) = -0 - \frac{1}{2} \log_2 \frac{1}{2} = 0 = E(\text{خیر}) \\
 & \Rightarrow \text{Gain} = 1 - 0 - 0 = 1
 \end{aligned}$$

Gain عرق خونی بسته نموده بین ۰ تا ۱ است، آنکه بیشتر خونه اش (بله) باشد است. این ای ویژگی بعینان را بسیار نزدیک شود و نتیجه‌ی این با ادامه داشتن است. درخت نیاز نمی‌داشته است.

سوال ۴

پاپ گزندوزت دارد؟	آب گازدار دوست دارد؟	سن	سیوال کلاه قرمز را دوست دارد؟
بله	بله	۷	خیر
بله	خیر	۱۲	خیر
خیر	بله	۱۸	بله
خیر	بله	۳۵	بله
بله	بله	۳۸	بله
بله	خیر	۵۰	خیر
خیر	خیر	۸۳	خیر

$$\Rightarrow E_{JF} = -\frac{1}{V} \log \frac{1}{1} - \frac{4}{V} \log \frac{4}{2} = 0,980$$

[۱+, F-] سیوال کلاه قرمز دارد

$$Gain(\text{پاپ گزندوزت}) = E_{JF} - \frac{1}{V} E(\text{نیاز}) - \frac{4}{V} E(\text{بیضیر})$$

$$E(\text{نیاز}) = -\frac{1}{12} \log \frac{1}{12} - \frac{1}{12} \log \frac{1}{12}, E(\text{بیضیر}) = -\frac{1}{12} \log \frac{1}{12} - \frac{3}{12} \log \frac{3}{4}$$

$$\Rightarrow Gain = 0,980 - \frac{1}{V} \times 0,92 - \frac{4}{V} \times 0,1811 \approx 0,12$$

[۱+, F-] [۱+, C-] پاپ گزندوزت دارد

$$Gain(\text{آب گازدار}) = E_{JF} - \frac{1}{V} E(\text{بیضیر}) - \frac{4}{V} E(\text{بیضیر})$$

$$E(\text{بیضیر}) = 0, , E(\text{بیضیر}) = -\frac{1}{8} \log \frac{1}{8} - \frac{1}{8} \log \frac{1}{8} = 0,1811$$

$$\Rightarrow Gain = E_{JF} - \frac{1}{V} \times 0, - \frac{4}{V} \times 0,1811 \approx 0,1811$$

[۱+, F-] [۱+, C-] NO آب گازدار دارد

دیگر یک سیوال داشتیم که می‌دانستیم بینه بازیه لیست
بله نه باشیم (نماینده ۷ بازیه) $\Rightarrow V \quad ۱۲ \quad ۱۸ \quad ۳۵ \quad ۳۸ \quad ۲۰ \quad ۸۳$
نیازی طبق بازیه +، همچنان ۱۲ ۱۸ ۳۵ ۳۸ ۲۰ ۸۳
نایاب نیزی طبق بازیه -، همچنان ۱۲ ۱۸ ۳۵ ۳۸ ۲۰ ۸۳

نایاب نیزی طبق بازیه - (که می‌توانیم تحریر نماینده باشیم) :
تعداد نسبت دسته هایی که می‌باشد که در آنها نیزی طبق بازیه داشته باشند

حالاتی که در آنها نیزی طبق بازیه داشته باشند (که می‌توانیم این را نیزی طبق بازیه نامیدیم) :

$$E(\text{بیضیر}) = 0, , E(\text{بیضیر}) = -\frac{1}{8} \log \frac{1}{8} - \frac{1}{8} \log \frac{1}{8} = 0,1811$$

$$\Rightarrow Gain(\text{age} > 15) = E_{JF} - \frac{1}{V} E(\text{بیضیر}) - \frac{4}{V} E(\text{بیضیر}) = 0,1811$$

$$\Rightarrow Gain(\text{age} > 15) = 0,1811$$

age > 15
[۱+, F-] [۱+, C-]

حالاتی که در آنها نیزی طبق بازیه داشته باشند :

age > ۱۵
[۱+, F-] [۱+, C-]

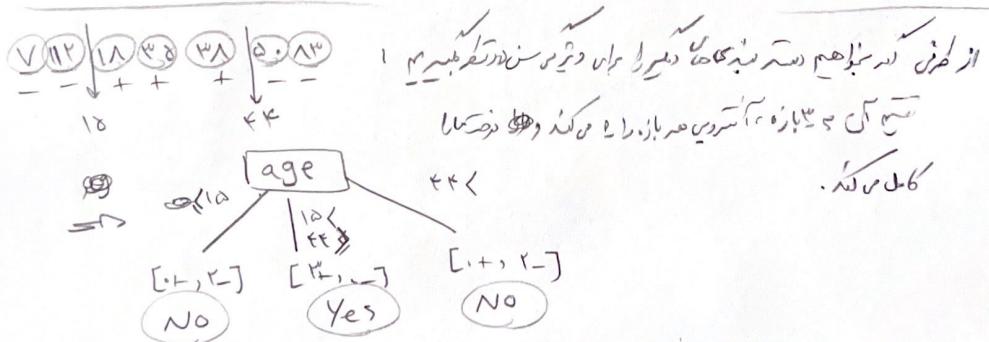
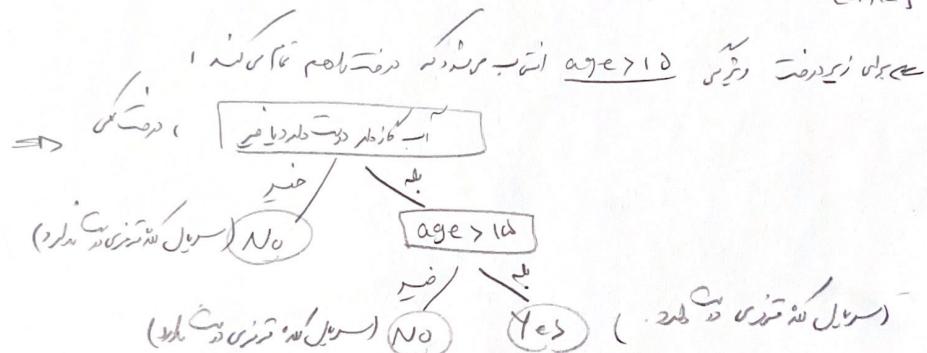
نه باشیم ۱۵ طبق بازیه آنها زیرا اولین ۱۵ نیزی طبق بازیه نداشتند.

حالاتی که در آنها نیزی طبق بازیه داشته باشند :

$$\text{Gain}(\text{age} > 10, \text{GDP}) = E(\text{GDP}) - \frac{1}{2} \times 1 - \frac{1}{2} \times 1 = 0.111$$

$$\text{gain} = E(\text{GDP}) - \dots = 0.111$$

$$\text{Gain} = 0.111 - \frac{1}{2} \times 0.111 = 0.055$$



سوال ۵) شاخص جینی چیست و برای چه از آن استفاده می‌کنیم؟ بالا و پایین بودن آن به چه معناست؟

شاخص جینی یک معیار اندازه‌گیری برای این است که یک عنصر انتخاب شده به صورت رندم از مجموعه، چه قدر احتمال دارد اشتباه انتخاب و برچسب‌گذاری شده باشد. پس اگر همه داده‌ها متعلق به یک دسته‌ی هدف باشند؛ مقدار شاخص جینی به حداقل (۰) می‌رسد. از این معیار در ساخت درخت‌های تصمیم و مقایسه‌ی ویژگی‌ها هم استفاده می‌شود. درخت تصمیم CART برای اینکه تصمیم بگیرد چگونه گره‌های درخت را انتخاب کند از معیار شاخص جینی (gini) استفاده می‌کند و برای هر ویژگی (بعد) هر چقدر شاخص جینی کمتر باشد، یعنی آن ویژگی اطلاعات بیشتری را به ما می‌دهد و می‌تواند در درخت ساخته شده، بالاتر (یعنی نزدیک به ریشه) قرار بگیرد و هرچه این مقدار بیشتر باشد یعنی ناخالصی بیشتری دارد و شанс انتخابش برای درخت تصمیم کمتر است و نزدیک به برگ خواهد بود. عموماً شاخص جینی برای داده‌هایی که دارای قسمت بزرگ‌تر هستند به درد می‌خورد این در حالی است که entropy (مانند درخت‌های ID3 و C4.5) به درد داده‌هایی می‌خورد که قسمت‌های کوچک زیادی دارند که مقادیر یکتا در آن‌ها بیشتر است.

$$1 - \sum_{i=1}^J p_i^2$$

آیا شاخص‌های دیگری نیز وجود دارند که کارایی مشابه داشته باشند؟

بله شاخص‌هایی مانند information gain که براساس مفهوم آتروری کار می‌کند (مانند الگوریتم‌های ID3 و C4.5) و یا معیار goodness و واریانس ریداکشن.

با اعداد دلخواه خود، چند نوب تعریف کنید و این شاخص را برای آن حساب کنید.

Chest pain	Proper Blood circulation	Clogged arteries	Heart Disease
Positive	Negative	Positive	Positive
Negative	Positive	Negative	Negative
Positive	Negative	Positive	Positive
Positive	Positive	Positive	Positive
Negative	Negative	Positive	Negative
Positive	Negative	Negative	Negative
Negative	Positive	Positive	Negative
Negative	Negative	Positive	Negative
Positive	Negative	Negative	Negative
Positive	Positive	Positive	Positive

Gini Index for Chest pain:

P(Chest pain=Positive): 6/10

P(Chest pain=Negative): 4/10

If (Chest pain = Positive & Heart Disease = Positive), probability = 4/6

If (Chest pain = Positive & Heart Disease = Negative), probability = 2/6

$$\text{Gini index} = 1 - ((4/6)^2 + (2/6)^2) = 0.45$$

If (Chest pain = Negative & Heart Disease = Positive), probability = 0

If (Chest pain = Negative & Heart Disease = Negative), probability = 4/4

$$\text{Gini index} = 1 - ((0)^2 + (4/4)^2) = 0$$

$$\text{Total Gini Index for Chest pain} = (6/10)0.45 + (4/10)0 = 0.27$$

Gini Index for Proper Blood circulation:

P(Proper Blood circulation=Positive): 4/10

P(Proper Blood circulation=Negative): 6/10

If (Proper Blood circulation = Positive & Heart Disease = Positive), probability = 2/4

If (Proper Blood circulation = Positive & Heart Disease = Negative), probability = 2/4

$$\text{Gini index} = 1 - ((2/4)^2 + (2/4)^2) = 0.5$$

If (Proper Blood circulation = Negative & Heart Disease = Positive), probability = 2/6

If (Proper Blood circulation = Negative & Heart Disease = Negative), probability = 4/6

$$\text{Gini index} = 1 - ((2/6)^2 + (4/6)^2) = 0.45$$

$$\text{Total Gini Index for Proper Blood circulation} = (4/10)0.5 + (6/10)0.45 = 0.47$$

Gini Index for Clogged arteries:

P(Clogged arteries=Positive): 7/10

P(Clogged arteries=Negative): 3/10

If (Clogged arteries = Positive & Heart Disease = Positive), probability = 4/7
 If (Clogged arteries = Positive & Heart Disease = Negative), probability = 3/7
 Gini index = $1 - ((4/7)^2 + (3/7)^2) = 0.49$

If (Clogged arteries = Negative & Heart Disease = Positive), probability = 0
 If (Clogged arteries = Negative & Heart Disease = Negative), probability = 3/3
 Gini index = $1 - ((0)^2 + (1)^2) = 0$

Total Gini Index for Clogged arteries = $(7/10)0.49 + (3/10)0 = 0.34$

→ Gini index:

Chest pain = 0.27

Proper Blood circulation = 0.47

Clogged arteries = 0.34

پس با مقایسه‌ی gini index این سه ویژگی می‌بینیم که مقدارش برای ویژگی Chest pain کمتر از بقیه است و می‌تواند در ریشه قرار بگیرد.

مراحل بالا را مجددا برای پیدا کردن گره‌های بعدی و زیر درخت انجام می‌دهیم با توجه به این موضوع که ریشه را داریم و گره بعدی فرزندش هستند و برای شاخه‌ی Positive اش باید گره پیدا کنیم:

Chest pain	Proper Blood circulation	Clogged arteries	Heart Disease
Positive	Negative	Positive	Positive
Positive	Negative	Positive	Positive
Positive	Positive	Positive	Positive
Positive	Negative	Negative	Negative
Positive	Negative	Negative	Negative
Positive	Positive	Positive	Positive

Gini Index of Proper Blood circulation for Positive Chest pain:

P(Proper Blood circulation=Positive): 2/6

P(Proper Blood circulation=Negative): 4/6

If (Proper Blood circulation = Positive & Heart Disease = Positive), probability = 2/2

If (Proper Blood circulation = Positive & Heart Disease = Negative), probability = 0
Gini index = $1 - ((2/2)^2 + (0)^2) = 0$

If (Proper Blood circulation = Negative & Heart Disease = Positive), probability = 2/4

If (Proper Blood circulation = Negative & Heart Disease = Negative), probability = 2/4

Gini index = $1 - ((0)^2 + (2/4)^2) = 0.75$

Total Gini Index for Proper Blood circulation = $(2/6)0 + (4/6)0.75 = 0.50$

Gini Index for Clogged arteries:

P(Clogged arteries=Positive): 4/6

P(Clogged arteries=Negative): 2/6

If (Clogged arteries = Positive & Heart Disease = Positive), probability = 4/4

If (Clogged arteries = Positive & Heart Disease = Negative), probability = 0

Gini index = $1 - ((4/4)^2 + (0)^2) = 0$

If (Clogged arteries = Negative & Heart Disease = Positive), probability = 0

If (Clogged arteries = Negative & Heart Disease = Negative), probability = 2/2

Gini index = $1 - ((0)^2 + (2/2)^2) = 0$

Total Gini Index for Clogged arteries = $(4/6)0 + (2/6)0 = 0$

→ Gini index:

Proper Blood circulation = 0.50

Clogged arteries = 0

شاخص جینی برای ویژگی Clogged arteries صفر و کمتر از ویژگی دیگر است. پس به عنوان ویژگی بعدی برای قرار گیری در درخت تصمیم گیری انتخاب می‌شود. همچنین شاخص جینی فرزندانش (بله و خیر) صفر است و به برگ‌ها می‌رسیم.

سوال ۶) مفهوم بیشبرازش چیست و کی اتفاق می‌افتد؟

بیشبرازش یا overfitting در یادگیری ماشین بدين معناست که مدل ما یکسری داده(دادههای آموزش) را بیش از اندازه یاد بگیرد و برای آن دقت بالایی کسب کند. بدین صورت هر نمونه از دادههای آموزشی دریافت کند با دقت بسیار بالایی پاسخ می‌دهد. اما مشکل اینجاست که این مدل حساسیت بالایی نسبت به این مجموعه داده پیدا می‌کند و پارامترهایش خیلی زیاد برای این مجموعه داده تنظیم می‌شوند و به همین دلیل با دریافت نمونه دادههای مشابه اما خارج از دیتاست استفاده شده در آموزش، خیلی ضعیف عمل می‌کند.

پس در بیشبرازش، مدل عملکرد مناسبی در برابر دادههای جدید نشان نمی‌دهد خطایش بالاست.

بیشبرازش به دلایل مختلفی می‌تواند رخ دهد از جمله اینکه تعداد دادههای آموزشی مان کم باشد، اندازه مدل (در صورتی که شبکه عصبی داریم؛ تعداد گرهها و اندازه شبکه عصبی مان) بیش از اندازه باشد و یا مدلمان زیادی پیچیده باشد، مرحله آموزش را بیش از اندازه تکرار کنیم، تعداد فیچرهای درنظر گرفته شده خیلی زیاد باشد، دادهها نویز داشته باشند و به درستی پیش پردازش نشده باشند.

برای رفع کردن آن چه کارهایی لازم است انجام دهیم؟ (به دلخواه چند مثال بزنید)

یکی از کارهایی که می‌توان انجام داد این است که اگر حجم دادههای آموزشی مان کم است آن را بیش تر کنیم (از طریق پیدا کردن دادههای جدید و یا حتی تولیدشان با برخی روش‌های یادگیری ماشین مانند GAN). با این کار مدل ما یادگیری‌اش گسترش‌های می‌شود و محدود به یک دیتاست کوچک نمی‌شود.

روش دیگر اعمال **Regularization** است که با یادگیری بیش از اندازه پارامترها مقابله می‌کند و با به کار بردن یکتابع مجازات(زیان-تبیه) یادگیری را کنترل می‌کند و تا حدودی ریسک بیشبرازش را کاهش می‌دهد.

در صورتی که مدل بزرگی استفاده کرده باشیم؛ کم کردن تعداد لایه‌ها و نودها روش دیگری است که برای جلوگیری از بیشبرازش می‌توانیم اتخاذ کنیم. با این کار مدل ساده‌تر شده و یادگیری سطحی‌تر اتفاق می‌افتد.

تکنیک دیگری است برای کنترل بیشبرازش که اگر تعداد ویژگی‌هایی که درنظر گرفتیم زیاد باشد، کمک کننده خواهد بود. در این روش از بین ویژگی‌های تعریف شده، آن‌هایی که مهم‌تر و موثرتر هستند را انتخاب می‌کنیم و بقیه را حذف می‌کنیم.

یک روش دیگر **Hold out data** است که درصورت داشتن دیتاست بزرگ می‌تواند مفید واقع شود. در این روش به جای اینکه همه‌ی دادههای دیتاست را برای آموزش استفاده کنیم، ۲۰ درصدش را برای تست درنظر می‌گیریم تا در تست ضعیف عمل نکند و با اعمال تست‌ها عملکردش دربرابر دادههای جدید ارتقا پیدا کند.

روش دیگر **Drop out** است به این صورت که با نادیده‌گرفتن لایه‌های مختلف در ایپاک‌ها مختلف، میزان یادگیری لایه‌ها و نودها را کاهش می‌دهیم. این لایه‌ها با احتمالاتی از پیش تعیین شده انتخاب می‌شوند.

یک مثال واقعی از مقابله با بیش‌برازش در تحلیل سیگنال‌های مغزی است. که با استفاده از GAN نمونه‌های جدید تولید می‌کنند و به این صورت از بیش‌برازش جلوگیری می‌کنند. به طور کلی افزایش حجم دیتابست با GAN در آزمایشاتی که داده‌ی نمونه کم است بسیار پرکاربرد است.

بخش برنامه‌نویسی

در google colab و با استفاده از jupyter notebook پیاده‌سازی انجام شده که فایل‌ها ضمیمه شده‌اند.