به نام خدا

درس یادگیری ماشین

تمرین سری اول

مریم رضوانی ۹۹۲۱۱۶۰۰۱۹

1) می‌خواهیم طبقه‌بندی کننده‌ای را پیاده‌سازی کنیم که دو ورودی می‌گیرد و مقدار هر ورودی(0،1یا2) است. اگر حداقل یکی از دو ورودی دارای مقدار2 باشد، خروجی 1 را می‌دهد. در غیراینصورت خروجی0 می‌دهد. آیا این مساله توسط یک پرسپترون قابل یادگیری است؟ پاسخ خودرا توضیح دهید.

با توجه به داده مسأله داریم :

if (x1=2) or (x2=2) or (x1=2 and x2=2) → f(x1,x2)=1

otherwise → f(x1,x2)=0

بنابراین برای اینکه بفهمیم توسط پرسپترون قابل یادگیری است یا خیر. با توجه به تابع پرسپترون داریم:

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

حال مقادیر فرض شده را به تابع بالا می دهیم: (bمقدار بایاس)

فرض:

x1' = 1 if x1 = 2

x1' = 0 otherwise

x2' = 1 if x2 = 2

x2' = 0 otherwise

فرض: x=2, x=0, x'1=2, x'2=0, w=1, w2=1, b=−0.5

f(x)= w.x'1 + w2.x'2 + b = 1*1 + 1*0 +(−0.5) = 0.5 > 0 , f(x)>0 → 1 با خروجی پرسپترون مطابقت دارد

فرض : x1=2 , x2=2 , x'1=1 , x'2=1, w1=1, w2=1, b=−0.5

f(x)= w.x1 + w2.x2 + b = 1*1 + 1*1 +(−0.5) = 1.5 > 0 , f(x)>0 → 1 با خروجی پرسپترون مطابقت دارد

در نتیجه می‌توانیم بگوییم پرسپترون یاد می‌گیرد و طبقه بندی صحیح انجام می دهد.

2) به سوالات‌زیر پاسخ کوتاه دهید.

الف)«مجموعه آموزش»و «مجموعه تست» در یک‌مدل یادگیری ماشینی‌چیست؟چه مقدار داده برای مجموعه‌های آموزشی، اعتبارسنجی و تست خود اختصاص‌خواهیدداد؟(5نمره)

مجموعه آموزش(Training Example):داده‌های دنیای واقعی یک توزیع (distribution) دارند که از آن‌ها نمونه(sample) بگیریم داده آموزشی را تولید می‌کنند. بطور کلی مجموعه آموزشی شامل داده‌هایی است که برای آموزش مدل بکار می‌روند.

مجموعه تست: شامل داده‌هایی است که برای اندازه‌گیری عمل‌کرد مدل وارزیابی درستی پیش‌بینی مدل بکار می‌روند.

بعنوان مثال: در دبیرستان سه پایه دهم یازدهم و دوازدهم داریم که برای شرکت در کنکور یکسری آموزش‌ها مثل کلاس کنکور و قلم چی شامل مجموعه آموزشی ما هستند و آزمون کنکور مجموعه تست ما برای دانش آموزان است که صحت و عمل‌کرد آن‌ها را در آزمون بررسی می‌کند.

ب)چه زمانی‌باید ازطبقه بندی به جای رگرسیون استفاده کرد؟

طبقه بندی یا classification برای داده‌هایی که خروجیشان مقداری گسسته دارد بکار می‌رود و رگرسیون برای زمانی است که داده‌هایی با مقداری پیوسته داشته باشیم. مثلاً برای تشخیص ایمیل اسپم و تعلق وام بانک به شخص که خروجی بله و خیر است از طبقه بندی و برای زمانی که برای داده‌های ورودی می‌خواهیم یک عدد واقعی پیش‌بینی کنیم مثلاً قیمت سهام. از رگرسیون استفاده می‌کنیم.

ج)ضرایب رگرسیون در رگرسیون خطی با چه تابعی صدا زده می‌شود؟ بطور مختصر توضیح دهید.

data loss function : هدف این تابع مینیمم کردن مجموع مربعات اختلاف بین مقدار واقعی با مقدار پیش‌بینی شده است.

$$MSE = \frac{1}{N}\sum_{i}^{N}(Yi - \hat{Y}i)^2$$

د)آیا می‌توان از رگرسیون‌خطی‌برای‌نمایش‌معادلات درجه دوم استفاده کرد؟ (5نمره)

خیر؛ نمی‌توان از رگرسیون خطی برای نمایش معادله درجه دوم استفاده کرد. چون رگرسیون خطی بر اساس یک متغیر وابسته و یک یا چند متغیر مستقل مدلی ارائه می‌دهد که بر اساس یک خط است و معادله درجه دوم غیر خطی می‌شود.

3) کدام گزینه سه سناریو زیر را به ترتیب به بهترین شکل نوع یادگیری توصیف می‌کند؟ دلیل خود را توضیح دهید.(۱۵نمره)

گزینه d

الف)یک سیستم طبقه بندی سکه برای یک دستگاه فروش خودکار ایجاد شده است. توسعه دهندگان مشخصات دقیق سکه را از ضرابخانه ایالت متحده به دست می‌آورند. و یک مدل آماری از اندازه؛ وزن و اسم آن استخراج می‌کنند. که سپس دستگاه فروش خودکار از آن برای طبقه بندی سکه ها استفاده می‌کند.

چون از مدل آماری استفاده می‌کنیم و از یک مدل ریاضی استفاده می‌کنیم یادگیری انجام نشده است.

ب) به‌جای تماس با ضرابخانه ایالت‌متحده برای به دست‌آوردن اطلاعات سکه،یک الگوریتم با مجموعه بزرگی ازسکه‌های برچسب‌دار ارائه می‌شود. الگوریتم از این داده‌ها برای استنباط‌مرزهای تصمیم‌گیری استفاده می‌کند که سپس ماشین‌فروش برای طبقه بندی سکه های خود استفاده می‌کند.

چون داده‌های برچسب داری داریم که دارای خروجی مشخص و طبقه بندی شده است یادگیری با نظارت است.

ج)یک کامپیوتر با بازی مکرر و تنظیم استراتژی خود با جریمه کردن حرکاتی که در نهایت منجر به باخت می شود، یک استراتژی برای بازی Tic-Tac-Toe ایجاد می‌کند.

در یک محیط عامل خودش با بازی مکرر یاد می‌گیرد و تجربه کسب می‌کند و منجر به یادگیری می‌شود. یادگیری تقویتی است.

a- الف)یادگیری تحت نظارت ب)یادگیری بدون نظارت ج)یادگیری تقویتی

b- الف) یادگیری تحت نظارت ب) یادگیری نکردن ج) یادگیری بدون نظارت

c- الف) یادنگرفتن ب) یادگیری تقویتی ج) یادگیری تحت نظارت

d- الف) یادنگرفتن ب) یادگیری تحت نظارت ج) یادگیری تقویتی (گزینه صحیح)

e- الف) یادگیری تحت نظارت ب) یادگیری تقویتی ج) یادگیری بدون نظارت

4) با در نظر گرفتن نمودار پراکندگی «scatter plot» بطور مختصر در این مورد توضیح دهید که در رگرسیون چگونه فرض خطی بودن بررسی می شود؟(۵ نمره)

فرض خطی بودن در رگرسیون: می‌توانیم از ۳ راه فرض خطی بودن را بررسی کنیم:

۱ ـ ببینیم Visualization (بصری کردن داده ها) در فضای دو بعدی چگونه است.

۲ ـ کارهای آماری:همبستگی(Correlation) دارند یانه؟ مثال: قد و وزن (correlation) همبستگی را بررسی می‌کنم ببینم همزمان با افزایش قد وزنم زیاد میشه یانه؟ اگر ۱+ باشد correlation مثبت است و رابطه مستقیم بین دو متغیر وجود دارد. (بصورت خطی). اگر ضریب همبستگی 1 ـ باشد رابطه عکس بین دو متغیر وجود دارد.و اگر ضریب همبستگی 0 باشد یعنی رابطه خطی بین دو متغیر وجود ندارد.

ضریب تعیین یک عدد بین 0و1 است که نشان می‌دهد چقدر تغییرات یک متغیر توسط تغییرات متغیر دیگر توضیح داده می‌شود. اگر ضریب تعیین نزدیک به ۱ باشد، یعنی رابطه خطی قوی بین دو متغیر وجود دارد. اگر ضریب تعیین نزدیک به صفر باشد، یعنی رابطه خطی ضعیف بین دو متغیر وجود دارد.

۳ ـ آزمون و خطا: ابتدا از رگرسیون خطیlinear regression استفاده می کنیم ببینیم loss function چقدر است. اگر خطا کم باشد(قابل قبول باشد) خطی استفاده می کنیم اگر خیلی داده‌ها پرت است درجه دو استفاده می‌کنیم مدام تست می‌کنیم و بررسی می‌کنیم.

آزمون فرض آماری چند متغیره چند روش برای بررسی اینکه آیا رابطه خطي بین دو متغیر به صورت تصادفي است یا خیر. برای انجام این آزمون، لازم است فرض صفر و فرض دیدگاه را تعریف کنیم و سپس از یك شاخص آماري مانند t-test یا F-test استفاده کنیم. در نهایت، با توجه به سطح اطمینان و مقدار p-value، فرض صفر را قبول یا رد مي‌کنیم.

مبحث tunnig کردن در یادگیری ماشین: راه‌های مختلف را تست کرد و باهم مقایسه کرد ببینیم کدام راه بهتر است آن را استفاده می‌کنیم.

5) یک پرسپترون منفرد با تابع فعال سازی sign را در نظر بگیرید. پرسپترون با وزن بردار $[۰.۴ ، -۰.۳ ، ۰.۱]^T$ بایاس نشان داده می شود. اگر بردار ورودی پرسپترون $X=[0.2,0.6,0.5]$ باشد.خروجی پرسپترون را بدست آورید.(۱۰ نمره) $\theta = 0$

با توجه به فرمول پرسپترون داریم:

$$Y=\text{sign}(W^T.X)+b$$

$$W^T.X = (0.4*0.2)+(-0.3*0.6)+(0.1*0.5)+0 = -0.05$$

$$Y= \text{sign}(-0.05)=-1$$

پس خروجی پرسپترون 1ـ است.

6) فرض کنید برای موارد زیر می‌خواهیم از تکنیک یادگیری ماشین استفاده کنیم. کدام روش یادگیری (بانظارت,بدون نظارت و تقویتی) را انتخاب می‌کنید؟در صورتی که یادگیری با ناظر است فضای ورودی و خروجی کدام است؟(۴۰نمره)

- شناسایی داده‌های خطرناک در یک شبکه بر مبنای تهدیدهای قبلی : یادگیری بانظارت. با استفاده از داده‌های آموزشی که شامل داده‌های خطرناک و بدون خطر هستند یک مدل طبقه‌بندی را آموزش می‌بیند و برای داده‌های جدید استفاده می کند.

  داده‌های ورودی: ویژگی‌های مربوط به داده‌های شبکه مانند:{IP Address,Port,Protocol, حجم ترافیک و ...}

  خروجی: یک لیبل که نشان می‌دهد داده خطرناک است یا خیر.

- شناسایی مشتریان مشابه یک فروشگاه اینترنتی : یادگیری بدون نظارت؛ با روش‌های داده کاوی و یادگیری ماشین است که با استفاده از داده‌های بدون برچسب سعی می‌کند مشتریان مشابه مشتریان عضوفروشگاه را با استفاده از ویژگی‌های آن‌ها مانند جنس، سن، علایق، رفتار و خرید دسته‌بندی (خوشه‌بندی) کند و افراد مشابه مشتریان فروشگاه را در دسته‌های مشابه قرار دهد.

- تشخیص زبان محتوایی یک متن که متن حداقل طول ۱۰ کاراکتر دارد : یادگیری با نظارت؛ با استفاده از داده‌های آموزشی که شامل متن‌های با طول حداقل ۱۰ کاراکتر به زبان‌های مختلف و لیبل آن ها هستند یک مدل را آموزش می‌بیند و با این مدل می‌تواند متون جدید را شناسایی کند.

  داده‌های ورودی: مجموعه‌ای از رشته‌های متنی با طول حداقل ۱۰ کاراکتر.

  داده‌های خروجی: یک برچسب که تشخیص می‌دهد متن مربوط به چه زبانی است.

- در بیمارستانی تعداد زیادی بیمار داریم اما امکان بررسی تک تک آن‌ها وجود ندارد می‌خواهیم تعدادی از آن‌ها که نماینده خوبی از بقیه می‌باشند انتخاب کنیم و آن‌ها را بطور دقیق بررسی کنیم: یادگیری بدون نظارت؛ داده‌ها بدون لیبل هستند و ما از کل بیماران نمونه‌هایی به تصادف انتخاب می‌کنیم و سپس آن‌ها را خوشه بندی می کنیم

- می‌خواهیم برنامه‌ای داشته باشیم که توانایی انجام بازی snake که چهار عمل بالا,پایین و راست و چپ را در هر لحظه می‌تواند انجام دهد داشته باشد: یادگیری تقویتی؛

در بازی snake، عامل مار را کنترل می‌کند و محیط صفحه‌ای است که در آن مار حرکت می‌کند. عامل برای هر حرکت خود، چهار انتخاب دارد: بالا، پایین، راست و چپ. اگر عامل بتواند غذای موجود در صفحه را بخورد، پاداش مثبت دریافت می‌کند. اگر عامل به دیوار بخورد یا خودش را بگزد، جریمه منفی دریافت می‌کند و بازی تمام می‌شود. عامل با تجربه کسب کردن از حالات مختلف بازی، سعی می‌کند یک استراتژی بهینه برای افزایش طول خود و جلوگیری از برخورد با دستاوردهای منفی پیدا کند.

- برنامه فروشگاهی می‌خواهد نظرات مشتریان در مورد کالاهای فروخته شده ارزیابی کند بنابراین در قسمتی امکان درج نظر متنی وجود دارد. با توجه به این نظرات متنی نظرات کاربران را از نظر خوب, بد و خنثی بودن بررسی می‌نماید. : یادگیری با نظارت؛ مجموعه داده‌های آموزشی شامل نظرات متنی و برچسب آن‌ها هستند.و با این مدل یاد می‌گیرد نظرات جدید را در دسته خوب، بد یا خنثی قرار دهد.

داده‌های ورودی: مجموعه‌ای از رشته‌های حرفی که نظرات مشتریان را نمایش می‌دهد.

داده‌های خروجی: یک لیبل است که نظرات خوب، بد یا خنثی را تشخیص می‌دهد.

- در چهارراهی می‌خواهیم ترافیک را به خوبی کنترل نماییم یعنی با توجه به هر ورودی چهارراه یکی را از سه رنگ موجود سبز زرد و قرمز انتخاب کنیم که باعث شود ترافیک به خوبی کنترل شود و یک چهارراه ازدحام زیاد و دیگری بدون ازدحام نباشد.) امکان اعمال اقدام های موجود و دیدن نتیجه آن وجود دارد.): یادگیری تقویتی؛

مثال، در کنترل ترافیک چهارراه، عامل سامانه‌ای است که مسئول تغییر رنگ چراغ‌های راهنمایی است. محیط صفحه‌ای است که در آن خودروها حرکت می‌کنند. عامل برای هر وضعیت ترافیک، سه انتخاب دارد: سبز، زرد و قرمز. اگر عامل بتواند جریان ترافیک را به خوبی هماهنگ کند و تعداد خودروهای منتظر را کمینه کند، پاداش مثبت دریافت می‌کند. اگر عامل باعث شود تصادفات رخ دهد یا تعداد خودروهای منتظر زیاد شود، جریمه منفی دریافت می‌کند. عامل با تجربه کسب کردن از حالات مختلف ترافیك، سعي مي‌كند يك استراتژي بهينه براي هماهنگ کردن جریان ترافیك پیدا کند.

- شرکت هواپیمایی مشخصاتی از هر پرواز شامل مبدأ مقصد و... دارد. در صورتی که تأخیر را برای پرواز ها اعلام کند از ضرر و دادن خسارت به مشتریان جلوگیری خواهد کرد. با توجه به این مورد کدام روش یادگیری را پیشنهاد می کنید؟ یادگیری بانظارت؛ مجموعه‌ای از داده‌های آموزشی که شامل مشخصات پروازهای گذشته و مقدار تأخیر آن‌ها هستند، یک مدل پیشبینی را آموزش می‌بینند. سپس، با استفاده از این مدل، می‌توانند تأخیر پروازهای جدید را بر اساس مشخصات آن‌ها برآورد کنند.

داده‌های ورودی: مجموعه‌ای از ویژگی‌های مربوط به پرواز است. مانند مبدأ، مقصد، زمان حرکت، زمان ورود، نوع هواپیما و غیره.

فضای خروجی : یک عدد حقیقی است که تأخیر پرواز به دقیقه را نشان می‌دهد چقدر است.