به نام خدا

درس هوش مصنوعی و سیستم های خبره

تمرین سیزدهم-تئوری

مدرس : دکتر محمدی

دانشجو : سارا سادات یونسی / ۹۸۵۳۳۰۵۳

سوالات تئوری)

سوال اول

درمورد روش Regression Logistic تحقیق کنید. تفاوت آن را با Regreesion Linear توضیح دهید.

توضیح سوال اول

اغلب برای بیان شدت رابطه خطی بین دو متغیر کمی از ضریب همبستگی استفاده می‌کنیم. همچنین برای نمایش مدل رابطه بین آن دو نیز از مدل رگرسیونی کمک می‌گیریم. در این میان یک الگو برای پیش‌بینی متغیر وابسته (Y) براساس متغیر مستقل (X) ایجاد می‌شود. ولی باید توجه داشت که در مدل ایجاد شده، هر دو متغیر مستقل و وابسته، کمی هستند. همچنین شرط پیوسته بودن این مقدارها نیز در روش رگرسیون نهفته است. ولی ممکن است بخواهیم رابطه بین یک متغیر مستقل (با مقدارهای پیوسته) را با یک متغیر وابسته با مقدارهای کیفی بسنجیم. در این حالت روش عادی رگرسیون خطی جوابگو نخواهد بود و باید از «رگرسیون لجستیک (Logistic Regression) «استفاده کرد.

از رگرسیون لجستیک برای تحلیل رابطه بین متغیرها بخصوص در زمینه‌های پزشکی، روانشناسی و علوم اجتماعی بسیار کمک گرفته می‌شود. برای مثال بررسی و ایجاد مدل رابطه بین میزان فعالیت روزانه و ابتلا به بیماری قند یک نمونه از تحلیل‌هایی است که در آن از مدل رگرسیون لجستیک کمک می‌گیرند. در این حالت متغیر مستقل، فعالیت روزانه با مقدارهای کمی است و متغیر وابسته کیفی نیز ابتلا یا عدم ابتلا به بیماری قند است که دارای دو مقدار ۰ و یا ۱ خواهد بود. همچنین در تحلیل حافظه انسان و رابطه آن با میزان خواب، روانشناسان آزمایشی را انجام می‌دهند که براساس مقدار ساعات متفاوت خواب افراد، یادآوری یا فراموشی کلمه‌ای را می‌سنجند. در این حالت میزان خواب متغیر مستقل با مقدارهای کمی پیوسته و متغیر وابسته کیفی با دو مقدار ۰ به معنی فراموشی و ۱ به معنی یادآوری صحیح است.

رگرسیون لجستیک و رگرسیون خطی

حال به تعریف رگرسیون برمی‌گردیم. می‌دانیم که منظور از رگرسیون خطی، ایجاد رابطه‌ای خطی برحسب پارامتر برای نمایش ارتباط بین متغیر مستقل و وابسته است. فرم مدل رگرسیون خطی ساده به صورت زیر است

$$Y=\beta 0+\beta 1X+\epsilon:$$

همانطور که دیده می‌شود این رابطه، معادله یک خط است که جمله خطا یا همان $\epsilon\epsilon$ به آن اضافه شده. پارامترهای این مدل خطی، عرض از مبدا ($\beta 0\beta 0$) و شیب خط ($\beta 1\beta 1$) هستند. در این حالت اگر $\hat{y}y\hat{}$ مقدار برآورد برای متغیر وابسته باشد، می‌توان آن را میانگین مشاهدات برای متغیر وابسته به ازای مقدار ثابت متغیر مستقل در نظر گرفت. پس اگر میانگین را با امید ریاضی جایگزین کنیم با فرض اینکه میانگین جمله خطا نیز صفر است، خواهیم داشت:

$$\hat{y}=E(Y|X=x)=\hat{\beta}0+\hat{\beta}1x$$

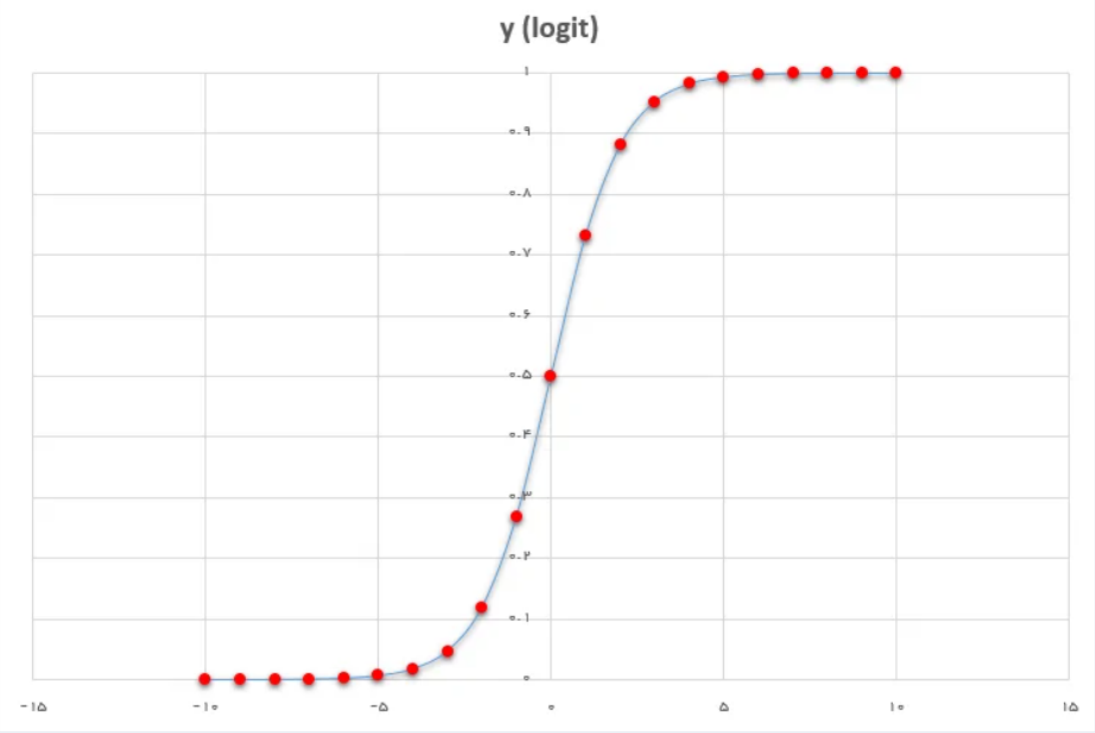
که در آن E(Y|X=x)E(Y|X=x)نشان‌دهنده امید ریاضی (متوسط) شرطی است و همچنین ^β0β0^ و ^β1β1^برآوردهای مربوط به هر یک از پارامترها هستند. اگر مقدار متغیر وابسته(Y) ، باینری (دو وضعیتی) و شامل ۰ و ۱ باشد مشخص است که دارای توزیع برنولی است و امید ریاضی آن به صورت زیر محاسبه می‌شود

y=E(Y|X=x)=P(Y=1|X=x)=p(x):

یک الگو در نظر گرفت، آنگاه مدل رگرسیون برای متغیر وابسته برنولی، مشخص p(x)p(x)به این ترتیب اگر بتوان برای تابع می‌شود. با توجه به این تعریف، برآورد پارامترهای رگرسیون لجستیک را مشخص می‌کنیم

از آنجایی که در بخش قبلی مقدار پیش‌بینی برای متغیر وابسته، با احتمال p(x)p(x)انجام شد، برای مشخص کردن مدل رابطه بین متغیر وابسته و مستقل به جای رابطه خطی، به تابعی احتیاج داریم که در حدود ۰ تا ۱ تغییر کند. در روش رگرسیون لجستیک از تابعی به نام «تابع لجستیک (Logistic Function) «استفاده می‌شود. به همین علت این روش رگرسیونی را رگرسیون لجستیک می‌نامند. در ادامه این تابع معرفی و نمودار مربوط به آن براساس پارامترهای b0=0b0=0 , b1=1b1=1 در تصویر دیده می‌شود.

$$f(x) = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}$$



همانطور که دیده می‌شود با افزایش مقدار x (x→∞x→∞) تابع لجستیک به ۱ نزدیک خواهد شد. همچنین با کاهش مقدار x (x→−∞x→−∞)مقدار تابع به سمت صفر میل می‌کند. حال فرض کنید برای رگرسیون لجستیک از این تابع برای بیان احتمال متغیر وابسته استفاده شود. پس خواهیم داشت:

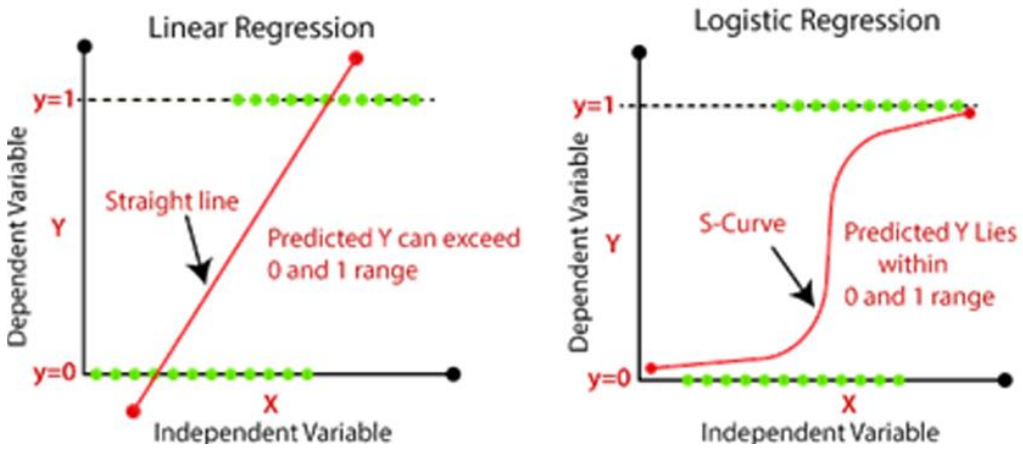
$$p(x) = \widehat{Y} = E(Y = 1|X = x) = \frac{e^{b_0+b_1x}}{1 + e^{b_0+b_1x}}$$

به منظور برآورد پارامترهای این مدل، می‌توان از «تبدیل لوجیت (Logit Transformation) »استفاده کرد. این تبدیل را روی بخت p(x)1−P(x)p(x)1−P(x)که قبلا بیان شده، اجرا می‌کنیم. در این صورت رابطه را می‌توان به شکل زیر نوشت:

$$g(x) = ln(\frac{p(x)}{1 - P(x)}) = \frac{\frac{e^{b_0+b_1x}}{1 + e^{b_0+b_1x}}}{1 - \frac{e^{b_0+b_1x}}{1 + e^{b_0+b_1x}}} = ln(e^{b_0+b_1x}) = b_0 + b_1x$$

با استفاده از تابع درستنمایی و حداکثر سازی آن می‌توان مدل را براساس برآورد پارامترها بدست آورد. با این کار به یک دستگاه معادلات می‌رسیم که متاسفانه برای حل آن روش تحلیلی وجود ندارد و باید به کمک روش‌های عددی برآورد را انجام داد.
خوشبختانه نرم‌افزارهای زیادی از جمله SPSS قادر هستند که محاسبات و برآوردهای مربوط به رگرسیون لجستیک را انجام دهند و پارامترهای b0b0و b1b1را محاسبه کنند. در ادامه به بررسی یک مثال به کمک نرم‌افزار SPSS می‌پردازیم.

تفاوت رگرسیون خطی و رگرسیون لجستیک : رگرسیون خطی و رگرسیون لجستیک دو الگوریتم مشهور یادگیری ماشین هستند که دسته ی تکنیک‌های یادگیری تحت نظارت قرار می‌گیرند. از آنجا که هر دو الگوریتم ماهیت نظارت شده دارند، بنابراین این الگوریتم ها از مجموعه داده های دارای برچسب برای پیش‌بینی استفاده می‌کنند. اما تفاوت اصلی بین آنها نحوه استفاده از آنهاست .رگرسیون خطی برای حل مشکلات رگرسیون استفاده می‌شود در حالی که رگرسیون لجستیک برای حل مشکلات طبقه بندی استفاده می‌شود .در شکل زیر نحوه طبقه بندی در هر دو روش به صورت نمودار مشخص است



Linear Regression	Logistic Regression
Used to predict a dependent output variable based on independent input variable	Used to classify a dependent output variable based on independent input variable
Accuracy is measured using Least squares estimation	Accuracy is measured using Maximum Likelihood estimation
The best fit line is a straight line	The best fit is given by a curve
The output is a predicted integer value	The output is a binary value between 0 and 1 value
Used in business domain, forecasting stocks	Used for classification, image processing

Linear Regression	**Logistic Regression**
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
In Linear regression, we predict the value of continuous variables.	In logistic Regression, we predict the values of categorical variables.
In linear regression, we find the best fit line, by which we can easily predict the output.	In Logistic Regression, we find the S-curve by which we can classify the samples.
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.
The output for Linear Regression must be a continuous value, such as price, age, etc.	The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc.
In Linear regression, it is required that relationship between dependent variable and independent variable must be linear.	In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable.
In linear regression, there may be collinearity between the independent variables.	In logistic regression, there should not be collinearity between the independent variable.

رگرسیون خطی:

رگرسیون خطی یکی از ساده‌ترین الگوریتم‌های یادگیری ماشینی است که تحت تکنیک یادگیری نظارتی قرار می‌گیرد و برای حل مسائل رگرسیون استفاده می‌شود.

برای پیش بینی متغیر وابسته پیوسته با کمک متغیرهای مستقل استفاده می شود.

هدف رگرسیون خطی یافتن بهترین خط مناسب است که بتواند خروجی متغیر وابسته پیوسته را به دقت پیش بینی کند.

اگر از یک متغیر مستقل برای پیش‌بینی استفاده شود، آن را رگرسیون خطی ساده و اگر بیش از دو متغیر مستقل وجود داشته باشد، به این رگرسیون رگرسیون خطی چندگانه می‌گویند.

با یافتن بهترین خط برازش، الگوریتم رابطه بین متغیر وابسته و متغیر مستقل را برقرار می کند. و رابطه باید ماهیت خطی داشته باشد.

خروجی رگرسیون خطی فقط باید مقادیر پیوسته مانند قیمت، سن، حقوق و غیره باشد.

رگرسیون لجستیک:

رگرسیون لجستیک یکی از محبوب‌ترین الگوریتم‌های یادگیری ماشینی است که تحت تکنیک‌های یادگیری نظارتی قرار می‌گیرد.

این می تواند برای طبقه بندی و همچنین برای مسائل رگرسیون استفاده شود، اما عمدتا برای مسائل طبقه بندی استفاده می شود.

از رگرسیون لجستیک برای پیش‌بینی متغیر وابسته طبقه‌ای با کمک متغیرهای مستقل استفاده می‌شود.

خروجی مسئله رگرسیون لجستیک فقط می تواند بین ۰ و ۱ باشد.

در مواردی که احتمالات بین دو کلاس مورد نیاز است، می توان از رگرسیون لجستیک استفاده کرد. از جمله اینکه آیا امروز باران خواهد بارید یا خیر، ۰ یا ۱، درست یا نادرست و غیره.

رگرسیون لجستیک مبتنی بر مفهوم برآورد حداکثر درستنمایی است. بر اساس این تخمین، داده های مشاهده شده باید محتمل ترین باشد.

در رگرسیون لجستیک، مجموع وزنی ورودی‌ها را از یک تابع فعال‌سازی عبور می‌دهیم که می‌تواند مقادیر بین ۰ و ۱ را ترسیم کند. این تابع فعال‌سازی به عنوان تابع سیگموئید و منحنی به‌دست‌آمده منحنی سیگموئید یا منحنی S نامیده می‌شود.

منابع : فرادرس

The difference between logistic regression and linear regression - Google Search

Linear Regression vs Logistic Regression - Javatpoint

سوال دوم

همانظور که در جدول زیر مشاهده م rکنید تعدادی داکیومنت وجود دارد که متعلق به ی ٔاز دو کلاس jیا c هستند. می خواهیم با استفاده از داده های آموزش rی Classifier Bayes Naive را آموزش داده و درنهایت مشخص کنید که داده ی آزمایش متعلق به کدوم کلاس می باشد. در این سوال شما باید از روش Smoothing Laplacian برای طبقه بند خود استفاده کنید. دقت کنید این تمرین نیازی به پیاده سازی ندارد اما باید تمام rمراحل به طور کامل نوشته شود. پارامتر Laplacian را برابر ۱ در نظر بگیرید.

توضیح سوال دوم

Laplace Smoothing

- Laplace's estimate:
  - Pretend you saw every outcome k extra times

$$P_{LAP,k}(x) = \frac{c(x) + k}{N + k|X|}$$

  - What's Laplace with k = 0?
  - k is the strength of the prior
- Laplace for conditionals:
  - Smooth each condition independently:

$$P_{LAP,k}(x|y) = \frac{c(x,y) + k}{c(y) + k|X|}$$

r r b

$P_{LAP,0}(X) =$

$P_{LAP,1}(X) =$

$P_{LAP,100}(X) =$

ابتدا فرمول های مرتبط با هر یک از بخش ها را می نویسیم

$P(c)=N_c/N$

P(w|c)=numOf(w,c)+1/numOf(c)+|V|

حال به محاسبه Prior ها می پردازیم

P(c) =3/4

P(j)=1/4

محاسبه ی احتمالات شرطی

P(chiniese|c)= 5+1/6+8=3/7

P(chiniese|j)=1+1/3+6=2/9

P(Tokyo|c)=0+1/8+6=1/14

P(Tokyo|j)=1+1/3+6=2/9

P(japan|c)=0+1/8+6=1/14

P(japan|j)=1+1/3+6=2/9

انتخاب کلاس با احتمال بالاتر

$P(c|d5) \rightarrow 3/4 * (3/7)^3 * 1/14 * 1/14 = 0.0003$

$P(j|d5) \rightarrow 1/4 * (2/9)^3 * 2/9 * 2/9 = 0.0001$

انتخاب ضریب برای نرمالایز کردن تامجموع احتمال ها برابر یک بشود.

$\beta = 1/0.0003 + 0.0001 = 2500$

احتمال بعد از اعمال ضریب نرمال شده

$P(c|d5) \rightarrow 0.75$

$P(j|d5) \rightarrow 0.25$

در نتیجه داده ی ما مربوط به کلاس" C" است که احتمال بیش تری هم دارد.