یادگیری ماشین

دکتر اعرابی- دکتر ابوالقاسمی

امیرحسین رکنی لموکی

۸۱۰۱۹۸۳۰۳

پاییز ۹۹

فهرست مطالب

۱ مقدمه

در این گزارش به بررسی و تحلیل سوالات تمرین سری اول درس یادگیری ماشین می‌پردازیم.

۲ سوال ۱

۱.۲ الف

در این مسئله برای یک مجموعه داده یکسان دو سایز متفاوت برای آموزش و تست در نظر گرفته‌ایم. در حالت اول داده‌های تست و آموزش را به دو قسمت مساوی تقسیم کرده اما در حالت دوم 95% داده را به آموزش اختصاص داده‌ایم. همانطور که مشاهده می‌شود به نظر می‌رسد طبقه بند دوم از طبقه بند اول دقیق‌تر عمل می‌نماید. اما در ایین بخش به طور دقیق تر جنبه‌های مختلف را مورد بررسی قرار می‌دهیم.

۱.۱.۲ مدل اول

به طور کلی این موضوع می‌تواند درست باشد که استفاده از تعداد دیتای کم برای آموزش می‌تواند منجر به Overfitting به داده‌های کم بشود، به طوریکه مدل به دلیل تعداد داده‌ی آموزش کم، به جواب زیر بهینه [1] همگرا بشود و نتواند قدرت تعمیم مناسبی داشته باشد.

۲.۱.۲ مدل دوم

در این مدل سعی شده است تا از بیشتر بخش‌های داده در فرایند آموزش استفاده بشود. اشکال این موضوع آن است که به دلیل تعداد پایین داده‌های تست، نمی‌توان چندان از صحت آزمون تست اطمینان داشت. در واقع ممکن است این ۲۰ داده نمونه مناسبی از کل جمعیت تمامی کلاس‌ها نباشد. به همین دلیل به نظر می‌رسد این مدل از نظر تعمیم دهی باید بیشتر مورد ارزیابی قرار بگیرد.

۳.۱.۲ مقایسه و نتیجه گیری

همانطور که اشاره شد مدل اول با مسئله کمبود داده‌ی آموزشی مواجه است و همچنین مدل دوم نیز(احتمالا)با کمبود داده‌ی تست مواجه است. برای حل این مسئله می‌توان از روش‌هایی مانند k-fold و یا Leave one out استفاده نمود.

[1] Suboptimal

۲.۲ ب

به طور کلی ما در مورد پدیده‌ای دارای یک نوع باور [۲] و یا دانش قبلی [۳] هستیم، سپس با مشاهده شواهد [۴] یک احتمال بر مبنای این مشاهدات در نظر گرفته می‌شود، سپس با تخمین این دو مرحله تصمیم گیری انجام می‌شود. برای مثال، فردی را در نظر بگیرید که دچار تب و سر درد است. این علائم می‌تواند هم نشان دهنده‌ی بیماری طاعون باشد، هم با علائم کووید-۱۹ مطابقت دارد. از طرفی این علائم نشان دهنده‌ی حساسیت پوستی باشد [۵]. بنابراین باید بین دو بیماری از سه بیماری تصمیم بگیریم. از طرفی، با توجه به دانش گذشته می‌دانیم که طاعون ریشه کن شده است و از طرفی در اوج پاندمی کووید-۱۹ می‌باشیم. بنابراین با ترکیب این دو مرحله تشخیص می‌دهیم که فرد احتمالا به کووید-۱۹ مبتلا است. توجه بفرمایید که در این مثال Inference و یا Reasoning در تشخیص بیماری بر اساس شواهد موجود مورد استفاده قرار گرفته است.

در مثالی دیگر می‌توان تشخیص سالم بودن یک میوه بر اساس رنگ، بو و مزه اشاره کرد. همچنین تشخیص یک object بر مبنای اطلاعات سنسورهای مختلف همگی از مثال‌های Inference هستند.

[2] Belief

[3] Prior Knowledge

[4] Evidents

[5] توجه بفرمایید تمامی این استدلال‌های پزشکی و بالینی مثال و فرض است و هیچ مبنای علمی برای گفته‌های اینجانب در این گزارش نمی‌توان قائل بود.

۳ سوال ۲

۱.۳ الف

۱.۱.۳ الگوریتم گرادیان کاهشی

به طور کلی یک تابع خوش فرم(مشتق پذیر) را می‌توان با استفاده از بسط تیلور به فرم زیر نوشت:

$$f(x+\Delta x) = f(x) + \nabla f(x)\Delta x + \frac{1}{2}(\Delta x)^T H(x)\Delta x + HOT. \tag{۱}$$

همچنین می‌توان تابع $f(x+\Delta x)$ را با استفاده از تابع ϕ معادلا باز نویسی کرد به طوریکه:

$$\phi(\alpha) = f(x+\alpha p_k); \qquad \alpha \in \mathbb{R}; x, p \in \mathbb{R}^n \tag{۲}$$

اصطلاحا به α طول گام و به P_k جهت گام گفته می‌شود. می‌توان نشان داد که به صورت کلی تابع در جهت عکس گرادیان بیشترین کاهش را دارد. از طرفی با توجه به خوش فرم بودن تابع، می‌توان تابع را در هر نقطه به صورت محلی با خط مماس بر آن تقریب زد. الگوریتم گرادیان کاهشی بر این مبنا ابتدا یک نقطه را به عنوان نقطه‌ی اولیه در نظر می‌گیرد، سپس با انتخاب جهت در جهت عکس گرادیان و انتخاب طول پله‌ی مناسب سعی در همگرایی پله‌ای به سمت نقطه‌ی کمینه را دارد.

۲.۱.۳ پیاده سازی الگوریتم گرادیان کاهشی

ابتدا گرادیان تابع هزینه را محاسبه می‌نماییم:

$$\nabla(J) = [\frac{\partial J}{\partial \theta_i}]^T; \qquad i \in \{1, ..., n\}, \nabla J \in \mathbb{R}^n$$

توجه بفرمایید، به نظر می‌رسد تابع مورد تابع هزینه ماشین‌های ساپورت برداری است، به همین دلیل از آن‌جا که کمی این تابع هزینه نگارش شده در صورت سوال دچار ابهام است، به نظر می‌رسد مقصود همون تابع هزینه معروف است و در این گزارش نیز همان تابع مدنظر قرار داده شده است از طرفی طبق مشتق زنجیره‌ای تنها برای داده i ام می‌توان نوشت:

$$f = 2 \times h(z(\theta) - y) \times \frac{\partial((h(z(\theta)) - y)}{\partial z} \times \frac{\partial z}{\partial \theta}; \qquad \theta = [w_1, ...w_n, b]^T, z(\theta) = w^T x + b, h = tanh$$

$$\rightarrow f = 2 \times (tanh(w^T x + b) - y) \times (1 - tanh(w^T x + b)^2) \times [x_1, x_2, ...x_n, 1]^T$$

که با ترکیب دو رابطه‌ی بالا به رابطه‌ی زیر می‌رسیم:

$$\nabla(J) = \Sigma_{i=1:q} \times (tanh(w^T x^i + b) - y^i) \times (1 - tanh(w^T x^i + b)^2) \times [x_1^i, x_2^i, ...x_n^i, 1]^T$$

حال با استفاده از مطالب گفته شده در بخش الگوریتم گرادیان، جهت حرکت را $P_k = -\nabla J(x_k)$ در نظر می‌گیریم و سپس با انتخاب طول پله α به طور گام به گام به صورت $f(x_{k+1}) = f(x_k + \alpha \times p_k)$ پیش می‌رویم. در قسمت ب در مورد انتخاب طول پله نیز صحبت خواهیم کرد.

۳.۱.۳ شرط توقف

این کار را تا جایی ادامه می‌دهیم که گرادیان صفر شود یا تعداد دو گام متوالی از حدی کوچکتر گردد و یا تعداد اجرای الگوریتم از حدی بیشتر شود.

۲.۳ ب

۱.۲.۳ انتخاب طول پله

به طور خلاصه الگوریتم را بازنویسی می‌نماییم:

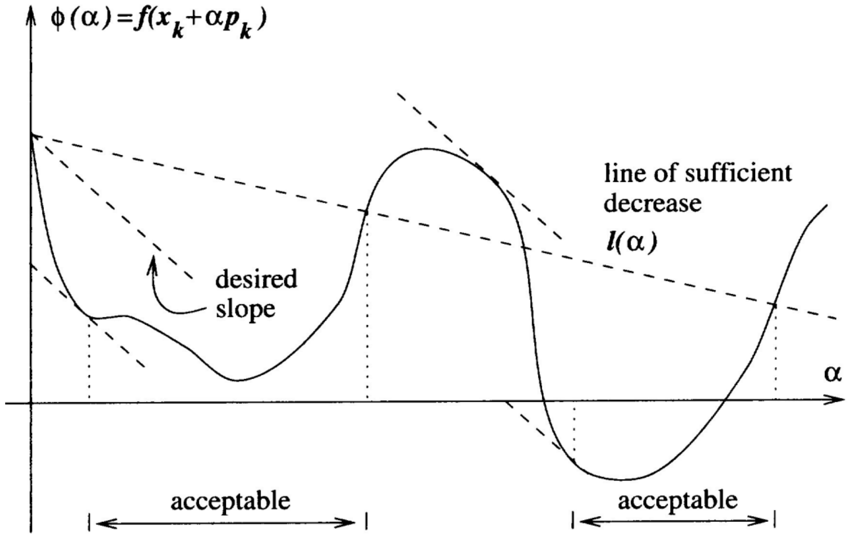
$$1.p_k = -\nabla J(x_k)$$
$$2.x_{k+1} = x_k + \alpha \times p_k$$
$$3.\text{Check Stopping Cond}$$
$$4.\text{Based on 3 stop or go 1}$$

حال تنها مورد انتخاب طول گام (α) می‌باشد. همانطور که اشاره شد، الگوریتم گرادیان کاهشی بر رفتار محلی تابع و هم ارزی آن با خط مماس مدنظر می‌باشد، بنابراین طول پله باید به قدری کم باشد که کماکان در حدود همسایگی نقطه مورد نظر باقی بمانیم(و بدین ترتیب همچنان در جهت کاهش تابع جرکت نماییم)، از طرفی نباید به قدری طول گام کوچک انتخاب بشود که الگوریتم طولانی بشود و عملا از نظر محاسباتی و زمانی به صرفه نباشد. بدین منظور از شروط ولفه [6] استفاده می‌گردد. ما در این بخش حالت قوی [7] شروط ولفه را بازنویسی می‌نماییم:

$$f(x_k + \alpha p_k) \leq f(x_k) + \mu_1 \alpha g_k^T p_k$$
$$|g(x_k + \alpha p_k)^T p_k| \leq \mu_2 |g_k^T p_k|,$$
$$0 \leq \mu_1 \leq \mu_2$$

که g_k همان بردار گرادیان است. برای درک بهتر تاثیر این شروط به شکل زیر توجه بشود:

[6] Wolfe Conditions
[7] Strong Wolfe Conditions

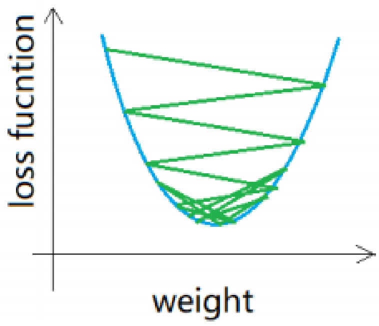


شکل ۱: تاثیر شرایط ولفه در برقراری طول پله‌ی مناسب.

۲.۲.۳ حالات حدی در انتخاب طول پله

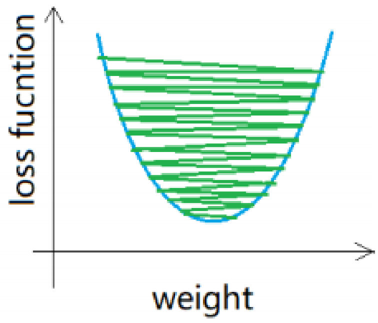
به طور کلی اگر قصد دو حالت حدی را داشته باشیم می‌توان به موارد زیر اشاره نمود:

۱. **انتخاب طول پله بزرگ** این کار موجب افزایش سرعت اجرای الگوریتم می‌شود اما از طرفی دقت لازم برای همگرایی به سمت نقطه‌ی بهنیه از دست می‌رود و پس از مدتی، الگوریتم در اطراف نقطه بهینه شروع به نوسان می‌نماید.



شکل ۲: انتخاب طول پله با اندازه بسیار بزرگ و رفتار نوسانی حاصل

۲. **انتخاب طول پله کوچک** انتخاب طول پله کوچک هر چند مشکل نوسان را از بین می‌برد، امامی‌تواند زمان رسیدن به نقطه‌ی بهینه را بسیار افزایش دهد و این از نظر زمانی و محاسباتی بهینه نخواهد بود.



شکل ۳: انتخاب طول پله با اندازه بسیار کوچک

بنابراین به صورت کلی می‌توان گفت انتخاب طول پله به صورت تطبیقی [8] می‌تواند موجب افزایش سرعت و دقت همگرایی شود و بنابراین استراتژی تطبیقی استراتژی مناسبی است.

[8] Adaptive

۴ سوال ۳

برای حل این مسئله از ابزار جبرخطی استفاده می‌نماییم. در نظر بگیرید که مجموعه تمامی نقاط را به وسیله‌ی زوج مرتب‌های $H = \{(x_i, y_i)\}_{i=1}^{k+1}$ نمایش بدهیم. همچنین از آن‌جا که تعداد $k+1$ نقطه در صفحه داریم، فضای همه‌ی چند جمله‌ای‌های کوچکتر از $k+1$ را در نظر می‌گیریم و می‌دانیم مجموعه $A = \{1, x, x^2, x^3, ..., x^k\}$ یک پایه برای این فضا با بعد $k+1$ است. حال مجموعه $B = \{f_1, ..., f_{k+1}\}; f_j \in \mathbb{P}_k(\mathbb{R})$ را در نظر بگیرید به طوریکه:

$$f_j(x) := \frac{\prod_{1 \le k \le n: k \ne j}(x - x_k)}{\prod_{1 \le k \le n: k \ne j}(x_j - x_k)},$$

باشد. به سادگی مشاهده می‌شود که:

$$f_j(x_i) = \begin{cases} 1 & i = j \\ 0 & i \ne j \end{cases} \tag{۳}$$

حال واضح است که f_i ها از یکدیگر مستقل خطی هستند. به عنوان اثبات در نظر داشته باشید:

$$f = a_1 f_1 + a_2 f_2 + ... + a_{k+1} f_{k+1} = 0$$

$$\to f(x_i) = a_1 f_1(x_i) + a_2 f_2(x_i) + ... a_{k+1} f_{k+1}(x_i) = 0$$

حال طبق ۳ می‌توان گفت:

$$f(x_i) = a_i = 0; \qquad i \in \{1, ..., k+1\}$$

بنابراین f_i ها مستقل خطی هستند و مجوعه B یک پایه برای $\mathbb{P}_k(\mathbb{R})$ است. حال با توجه به پایه بودن B می‌دانیم هر چند جمله‌ای از درجه $k+1$ را می‌توان دبا استفاده از این پایه به صورت یکتا ساخت. در واقع:

$$f = a_1 f_1 + a_2 f_2 + ... + a_{k+1} f_{k+1}$$

به طوریکه a_i ها به طور یکتا تعیین می‌گردند. برای ساختن چند جمله‌ای که بتواند نقاط مجموعه H را درونی یابی [9] نماید باید داشته باشیم:

$$f_(x_i) = y_i; \qquad i \in \{1, 2, ..., k+1\}$$

از طرفی طبق ۳ داریم:

$$f_(x_i) = a_1 f_1(x_i) + a_2 f_1(x_i) + ... + a_{k+1} f_{k+1}(x_i) = y_i \qquad i \in \{1, 2, ..., k+1\}$$

$$\to f_(x_i) = a_i = y_i \tag{۴}$$

[9] Interpolate

حال با توجه به معادلات ۴ و ۳ چند جمله‌ای مورد نظر به شرح زیر خواهد بود:

$$f := y_1 f_1 + y_2 f_2 + ... y_{k+1} f_{k+1} \quad (۵)$$

همچنین از آن‌جا که اشاره شد مجموعه B یک پایه برای چند جمله‌ای‌های کمتر از درجه $k+1$ می‌باشد و هرچند جمله‌ای در این فضا به صورت یکتا ساخته می‌شود، بنابراین با توجه به پایه بودن در این زیر فضا(از کل فضای چند جمله‌ای‌ها) می‌توان گفت که چند جمله‌ای معادله ۵ یکتا نیز می‌باشد.

۵ سوال ۴

۱.۵ مقدمه

در سوال مذکور بررسی خصوصیات آماری دو پارامتر β_0, β_1 مطرح شده‌اند. در این سوال فرض می‌کنیم x_i ها نمایانگر داده‌ی ورودی i ام و y_i نمایانگر داده خروجی متناظر اندازه‌گیری شده باشد. همچنین فرض می‌شود سیستم به صورت خطی است و داده‌های اندازه‌گیری شده با فرض نویز جمع شونده در خروجی از رابطه‌ی زیر پیروی نمایند:

$$y_i = \beta_1 x_i + b_2 + e \tag{۶}$$

که e یک نویز سفید با واریانس سیگما؛ $e \sim \mathcal{N}(0, \sigma)$ ؛ می‌باشد.

۲.۵ بررسی آماره‌های تخمین

در نظر بگیرید $\beta = [\beta_0, \beta_1]^T$ و $X_i = [1, x_i]^T$ در اینصورت:

$$Y_i = X_i^T \beta$$

. همچنین اگر تعداد مشاهدات(n) به اندازه کافی بزرگ باشد داریم:

$$\bar{y} = n \times (y_1 + ... + y_n); \qquad \bar{x} = n \times (x_1 + ... + x_n)$$

حال به راحتی می‌توان مشاهده کرد که روابط β_1, β_0 نگارش شده در صورت سوال با رابطه‌ی زیر معادل است:

$$\beta = (X^T X)^{-1} X^T Y \tag{۷}$$

که مقصود از X و Y به ترتیب ، ماتریس شامل تمامی ورودی‌ها(X_i) و خروجی‌ها(Y_i) می‌باشد. حال محاسبات زیر را خواهیم داشت:

$$\hat{\beta} = E\{\beta\} = E\{(X^T X)^{-1} X^T Y\} = E\{(X^T X)^{-1} X^T \beta\} + E\{(X^T X)^{-1} X^T e\} = (X^T X)^{-1} X^T \beta$$

$$\rightarrow cov\{\beta\} = E\{(\beta - \hat{\beta})(\beta - \hat{\beta})^T\} = E\{(X^T X)^{-1} X^T e e^T X (X^T X)^{-1}\}$$
$$\rightarrow (X^T X)^{-1} X^T E\{e e^T\} X (X^T X)^{-1}$$

بنابراین ماتریس کوواریانس پارامتر β به شرح زیر است:

$$\rightarrow cov\{\beta\} = (X^T X)^{-1} X^T E\{e e^T\} X (X^T X)^{-1} \tag{۸}$$

حال اگر فرض سفید بودن نویز را در نظر بگیریم :

$$\rightarrow (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

سپس با فرض اینکه تعداد نمونه‌ها(n) به اندازه کافی بزرگ باشد:

$$cov\beta = \sigma^2(X^TX)^{-1} = \frac{\sigma^2}{n}E\{[1,x]^T[1,x]\}$$

$$\rightarrow cov\beta = \frac{\sigma^2}{n}E\{\begin{pmatrix} x^2 & x \\ x & 1 \end{pmatrix}\}$$

$$\rightarrow cov\beta = \begin{pmatrix} E\{x^2\} & E\{x\} \\ E\{x\} & 1 \end{pmatrix} \tag{۹}$$

۳.۵ استقلال β_0, β_1

می‌دانیم برای آنکه دو پارامتر مستقل باشند کافی است که ماتریس کواریانس این دو پارامتر به صورت قطری باشد، بنابراین با توجه به معادله ۹ می‌توان معادلا گفت که $E\{x\} = x_1 + ...x_n = 0$ باشد که این بدان معناست نمونه‌ها به صورت تصادفی و به میانگین مبدا انتخاب بشوند(برای مثال انتخاب متقارن نسبت به مبدا می‌تواند یکی از حالات برآروده کننده باشد).

۶ سوال ۵

۱.۶ مقدمه

حال با استفاده از نتایج سوال قبل به حل این سوال می‌پردازیم.

۲.۶ الف

ابتدا واریانس نویز خروجی را حساب می‌نماییم:

$$Var\{e\} = \sigma^2 = Var\{y\} = \frac{\Sigma_{i=1:8}(y_i^2)}{8} - (\frac{\Sigma_{i=1:8}(y_i)}{8})^2 = 1.3594$$

همچنین طبق رابطه‌ی ۷ داریم:

$$\beta = [\beta_0, \beta_1]^T = (X^T X)^{-1} X^T Y;$$

$$X = \begin{pmatrix} 1 & 0.5 \\ 1 & 1 \\ 1 & 1.5 \\ 1 & 2 \\ 1 & 2.5 \\ 1 & 3 \\ 1 & 3.5 \\ 1 & 4 \end{pmatrix}; \qquad Y = \begin{pmatrix} 40 \\ 41 \\ 43 \\ 42 \\ 44 \\ 42 \\ 43 \\ 42 \end{pmatrix}$$

بنابراین بدست می‌آید:

$$\rightarrow \beta_0 = 40.8929, \qquad \beta_1 = 0.5476$$

۳.۶ ب

حال ماتریس کوواریانس دو پارامتر را با استفاده از ۹ محاسبه می‌نماییم:

$$cov\{\beta\} = \sigma^2 (X^T X)^{-1} = \begin{pmatrix} 0.12946429 & -0.29129464 \\ -0.29129464 & 0.82533482 \end{pmatrix} \tag{۱۰}$$

همچنین می‌دانیم:

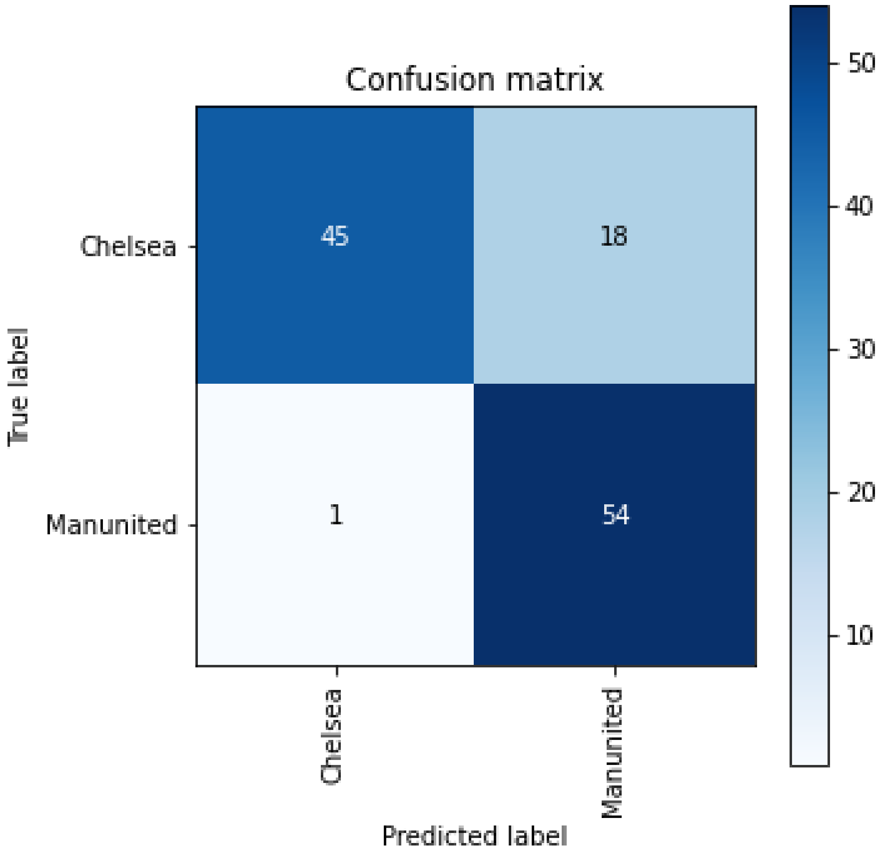
$$corr(\beta_0, \beta_1) = \frac{Cov(\beta_0, \beta_1)}{\sigma_{\beta_0} \sigma_{\beta_1}}$$

با توجه به رابطه ۱۰ داریم:

$$corr(\beta_0, \beta_1) = \frac{-0.291}{\sqrt{0.1295}\sqrt{0.8253}} = -0.89$$

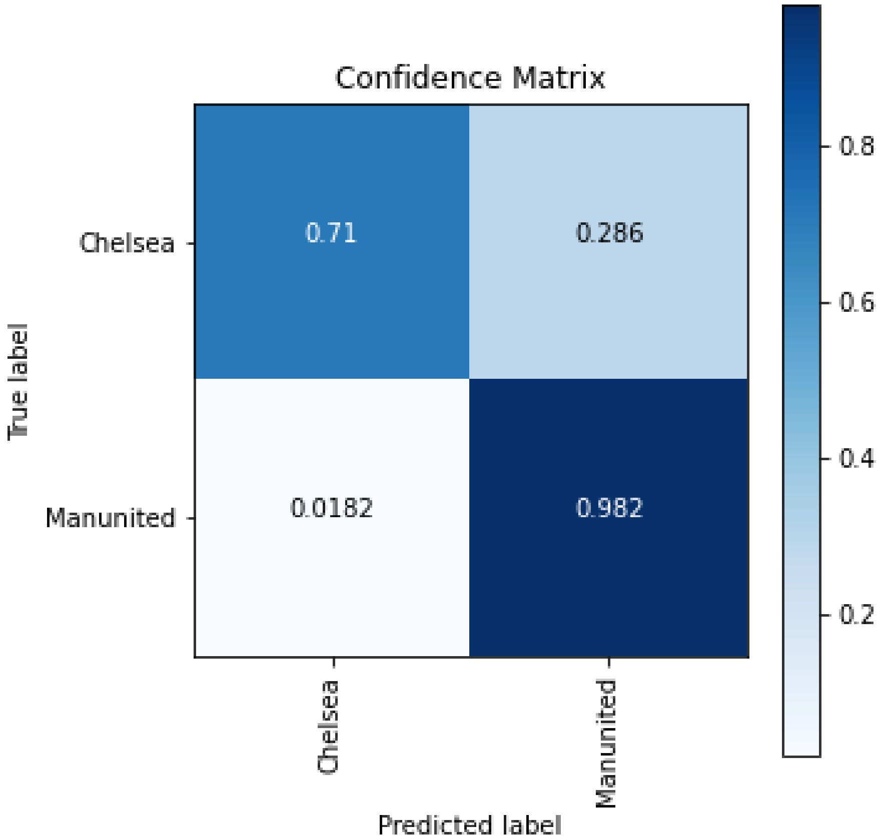
۷ سوال۶

در این مسئله با یک سوال طبقه بند مواجه هستیم و قرار است طبقه بندی با خاصیت تشخیص دو طبقه بند چلسی و منچستر یونایتد را طراحی نماییم. بدین منظور به مقایسه‌ی میانگین رنگی پیکسل‌های کل یک عکس و رنگ آبی و قرمز می‌پردازیم. پس از طراحی طبقه بند مشاهده می‌شود ماتریس درهم ریختگی [۱۰] به صورت زیر می‌باشد: همچنین ماتریس Confidence به شرح زیر می‌باشد:



شکل ۴: ماتریس درهم ریختگی طبقه بندی چلسی- منسچتر یونایتد

[10] Confusion Matrix



شکل ۵: ماتریس CONFIDENCE طبقه بندی چلسی- منسچتر یونایتد

همچنین با استفاده از این ماتریس می‌توان مقادیر accuracy ، precision و recall را بدست آورد: مطابق

جدول ۱: مقادیر گزارش شده از ماتریس درهم ریختگی

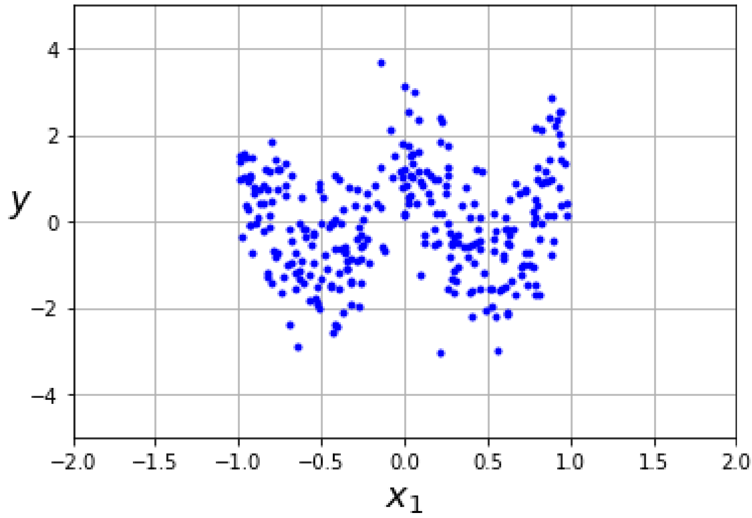
Accuracy	Recall	Precision	Club
0.84	0.71	0.98	Chelsea
	0.98	0.75	Man United

جدول ۱ با استفاده از مقادیر Precision می‌توان مشاهده کرد که در 98% مواقع وقتی طبقه بند خروجی چلسی تشخیص می‌دهد، تشخیصش درست است و در 75% مواقع وقتی خروجی منچستر را تشخیص می‌دهد، تشخیص درست بوده است. همچنین با توجه به مقدار Recall می‌توان مشاهده کرد که طبقه بندر 71% می‌تواند به درستی عکس‌های مربوط به چلسی را تشخیص دهد و در 98% مواقع می‌تواند عکس‌های مربوط به منچستر را به درستی تشخیص دهد. همچنین به صورت کلی دقت طبقه بند برابر با 84% می‌باشد.

۸ سوال ۷

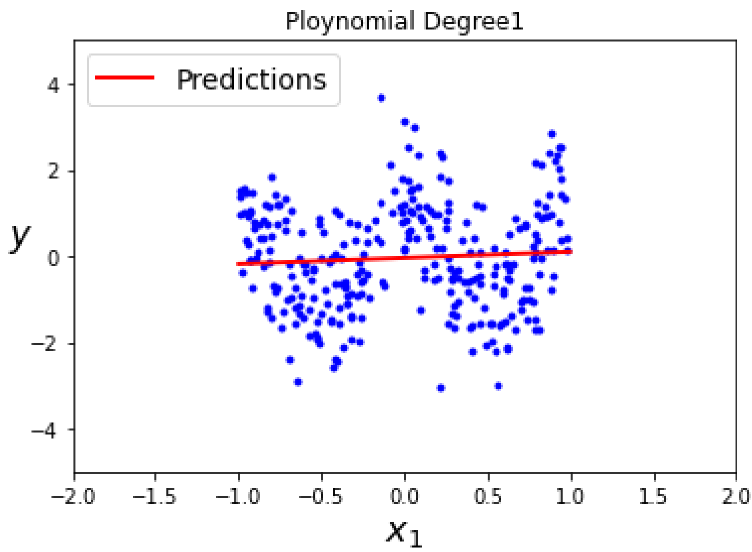
۱.۸ اجرای شبیه سازی

ابتدا داده‌ها را تحت اثر نویز با واریانس یک و میانگین صفر تولید می‌نماییم. شکل داده‌ها به شرح زیر است:
همانطور که مشاهده می‌شود، داده‌ها خروجی نویزی تابع کسینوسی در یک دوره‌ی تناوب می‌باشند. حال مدل‌ها

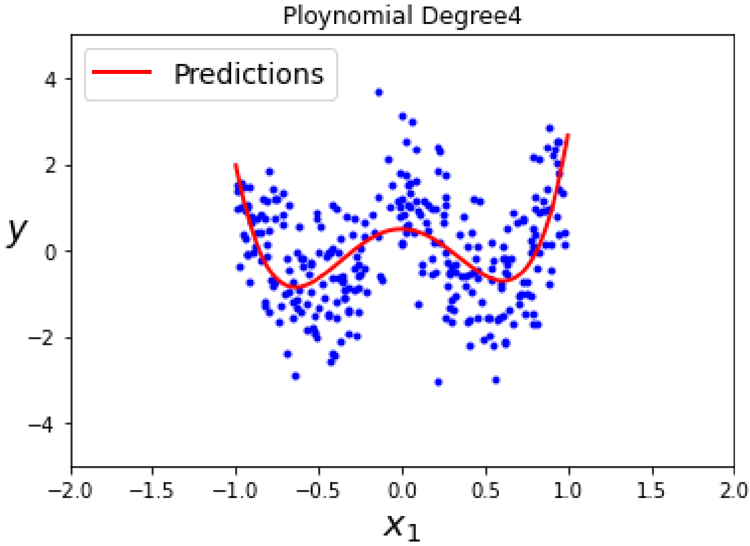


شکل ۶: تولید داده تحت اثر نویز

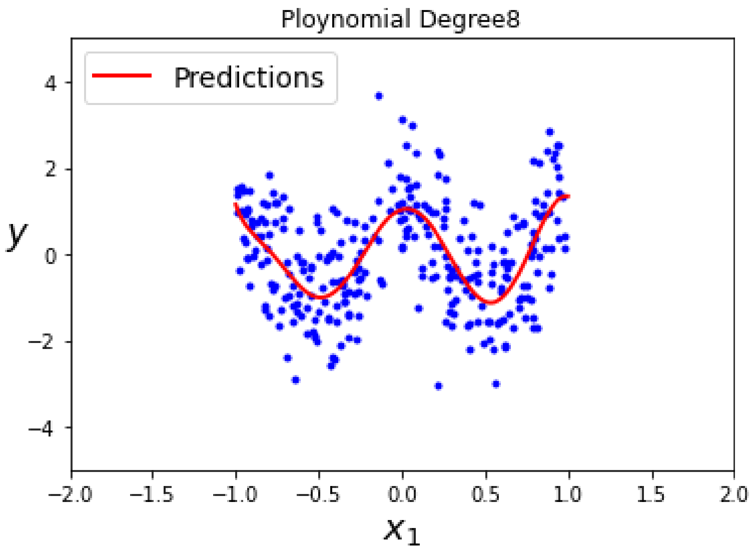
درجات مختلف را تحت این داده‌ها آموزش می‌دهیم، منحنی‌ها به شکل زیر خواهند بود:



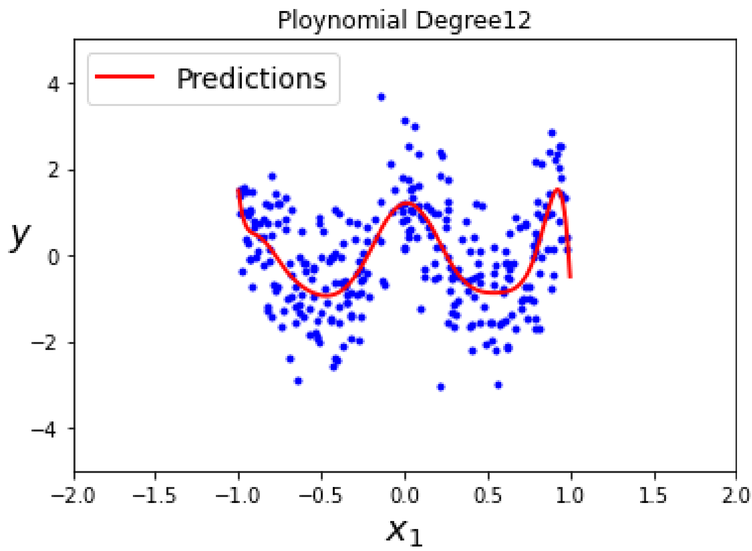
شکل ۷: تخمین با استفاده از مدل درجه ۱



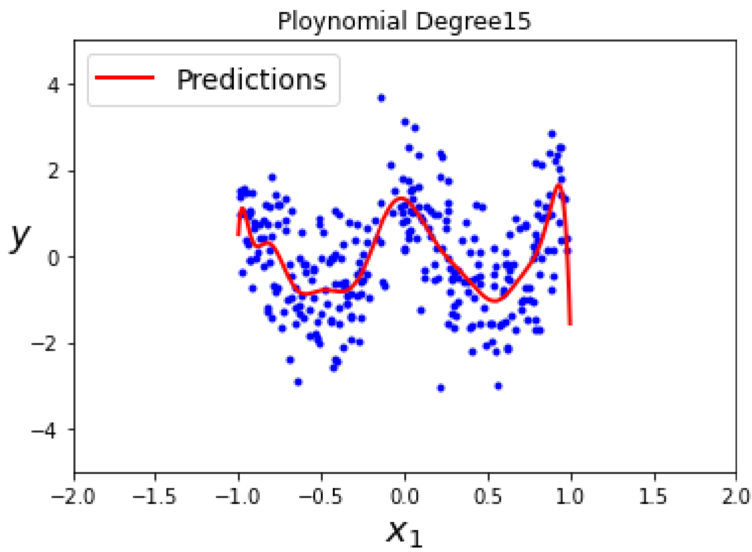
شکل ۸: تخمین با استفاده از مدل درجه ۴



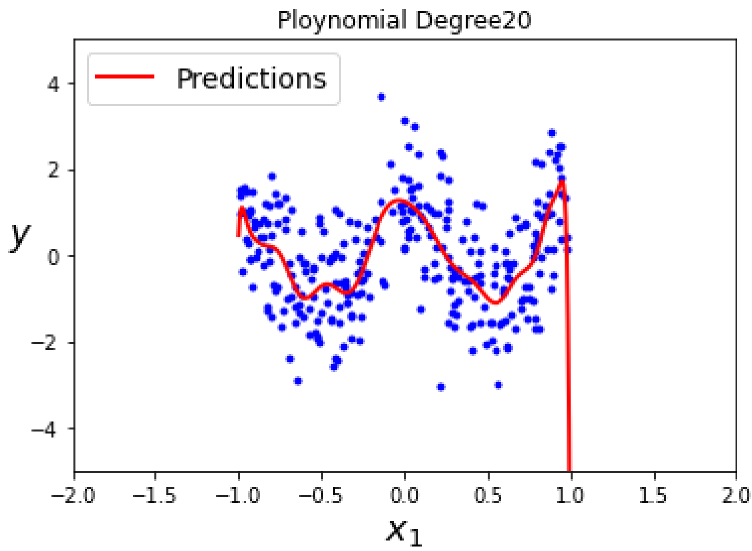
شکل ۹: تخمین با استفاده از مدل درجه ۸



شکل ۱۰: تخمین با استفاده از مدل درجه ۱۲

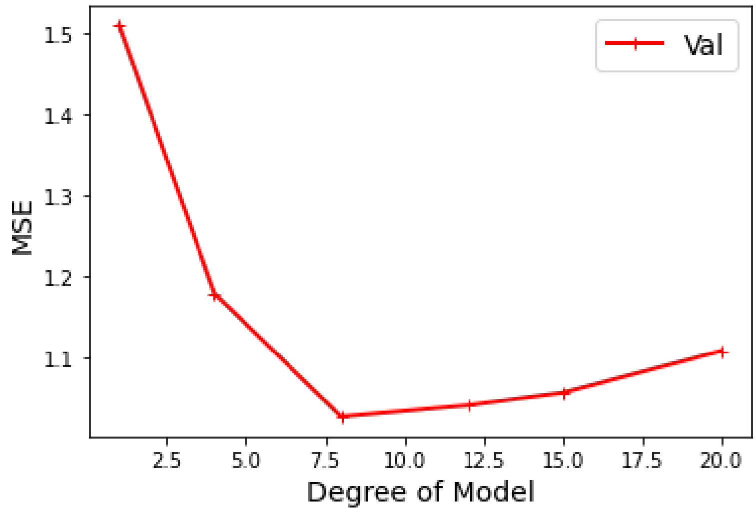


شکل ۱۱: تخمین با استفاده از مدل درجه ۱۵



شکل ۱۲: تخمین با استفاده از مدل درجه ۲۰

همچنین مشاهده می‌شود که مقادیر MSE روندی مطابق با نمودار زیر را طی خواهد کرد:



شکل ۱۳: نمودار MSEبرای مدل‌های مختلف

همچنین برای محاسبه بایاس و واریانس مطابق مطالب می‌توان گفت [11]

$$\text{Expected Loss} = \text{bias}^2 + \text{variance} + \text{noise}$$

که برای محاسبه‌ی آن‌ها می‌توان از روابط زیر استفاده کرد:

$$\bar{y}(x) = \frac{1}{L}\Sigma_{l=1}^{L} y^{(l)}(x)$$
$$\text{bias}^2 = \frac{1}{N}\Sigma_{n=1}^{N}\{\bar{y}(x_n) - h(x_n)\} \qquad (۱۱)$$
$$\text{variance} = \frac{1}{N}\Sigma_{n=1}^{N}\frac{1}{L}\Sigma_{l=1}^{L}\{y^{(l)}(x_n) - \bar{y}(x_n)\}^2$$

[11] Bishop, C.M., 2006. Pattern recognition and machine learning. springer, p.151

حال با پیاده سازی رابطه ۱۱ می‌توان جدول زیر را ارائه نمود: [۱۲]

جدول ۲: بررسی تفاوت دو بلوک FC و FILTERING

MSE	Noise	Variance	$Bias^2$	Model
1.4901207587744	0.9797	0.0113	0.4978	Linear
1.1745689486636965	0.9797	0.0199	0.1597	Degree 4
1.0189658887246866	0.9797	0.0289	0.0005	Degree 8
1.0493015615364785	0.9797	0.0424	0.0010	Degree 12
1.0802036249723488	0.9797	0.0574	0.0014	Degree 15
1.2540454181699006	0.9797	0.1081	0.0018	Degree 20

۲.۸ مقایسه و نتیجه گیری

مطابق جدول ۲ و نمودار ۱۲ مشاهده می‌شود که در ابتدای امر با افزایش درجه مدل، میزان بایاس و واریانس کاهش یافته است، سپس در ادامه، مدل شروع به یادگیری و تخمین نویز به عنوان هدف اصلی می‌نماید، در نتیجه این امر شاهد افزایش واریانس می‌باشیم. همچنین پس از مدتی عملا ضربه‌های متوالی که در فرم نمودار ایجاد می‌گردد (به خاطر یادگیری نویز توسط مدل)می‌تواند موجب افزایش خطای میانگین نیز بشود.

[۱۲] برای فهم بیشتر مطلب و همچنین پیاده سازی بخش بایاس و واریانس از کد پایتون در سایت به آدرس زیر استفاده شده است:
https://scikit-learn.org/stable/auto_examples/ensemble/plot_bias_variance.html لازم به ذکر است تنها در پیاده سازی باباس و واریانس از این کد کمک گرفته شده است و سایر بخش‌ها از این کد مستقل است.

۹ سوال۸

همانطور که در سوال ۶ مشاهده نمودیم، برای ارزیابی دقیق تر یک طبقه بند می‌توان از معیارهای مختلفی استفاده کرد که هر یک از این معیارها عملکرد به خصوصی از طبقه بند را مورد سنجش قرار می‌دهد. ذر این بخش مفاهیم مورد نظر را شرح می‌دهیم.

۱.۹ Accuracy

۱.۱.۹ تعریف - ذکر مثال

این معیار به بررسی عملکرد کلی یک طبقه بند می‌پردازد و در واقع بیان می‌کند که طبقه بندی در چند درصد مواقع به درستی عمل می‌کند و چقدر یک خروجی طبقه قابل اعتماد خواهد بود. برای مثال طبقه بند تشخیص فرد بیمار از سالم را در نظر بگیرید. تعداد کل تشخیص‌های درست طبقه‌بند به کل تعداد تشخیص‌ها بیانگر Accuracy می‌باشد. همچنین فرمول آن به شرح زیر می‌باشد:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

۲.۱.۹ نقطه ضعف

از آن‌جا که این طبقه بند به صورت کلی مورد بررسی قرار می‌گیرد، نمی‌تواند عملکرد تک تک کلاس‌ها را به درستی بیان کند. برای مثال ممکن است طبقه بند ممکن است در تشخیص فرد سالم به درستی عمل کند اما در تشخیص فرد بیمار ضعف داشته باشد اما به صورت کلی با توجه به عملکرد بالا در تشخیص فرد سالم این ناکارآمدی پنهان بماند. این ضعف می‌تواند پیامدهای جدی داشته باشد، برای مثال اگر فرد بیمار احتیاج به درمان فوری داشته باشد اما طبقه‌بند به اشتباه فرد را سالم تشخیص داده باشد می‌تواند عواقب جبران ناپذیری داشته باشد.

۲.۹ Precision

۱.۲.۹ تعریف - ذکر مثال

این معیار بیانگر قابل اتکار بودن طبقه بند در بیان تعلق به کلاس مشخیص است. در واقع تعداد تشخیص‌های درست به کلیه نمونه‌هایی است که به عنوان عضو کلاس مورد نظر تشخیص داده شدند.به عنوان نمونه، در مثال طبقه‌بند تشخیص فرد بیمار از سالم، تعداد افراد سالمی که به درستی سالم تشخیص داده شده‌اند، به جمع این افراد با افراد بیماری که به غلط سالم تشخیص داده شده‌اند بیانگر Precision طبقه‌بند، در تشخیص فرد سالم است. در واقع بالا بودن این شاخص به معنای آن است که احتمال آنکه فردی که سالم تشخیص داده شده است مریض باشد بسیار کم است. همچنین فرمول آن به شرح زیر است:

$$\text{Precision} = \frac{TP}{TP + FP}$$

۳.۹ نقطه ضعف

نقطه ضعف این معیار در این است که نمی‌تواند بیان کند که طبقه بند چقدر در تشخیص داده متعلق به کلاس موفق بوده است. در واقع در مثال طبقه بندی تشخیص فرد سالم و بیمار، تمی‌توان با این معیار متوجه شد که طبقه بند چند درصد از افراد سالم را توانسته است سالم تشخیص دهد.

۴.۹ Recall

۱.۴.۹ تعریف - ذکر مثال

این معیار بیانگر آن است که طبقه‌بند چند درصد داده‌های متعلق به یک کلاس را به درستی پیش بینی نماید. در واقع تعداد تشخیص‌های درست تعلق به کلاس به کلبه تمونه‌های متعلق به کلاس بیانگر این معیار است. برای نمونه در مثال طبقه‌بند تشخیص فرد سالم و بیمار، تعداد تشخیص‌های درست طبقه بند در تعیین فرد سالم به تعداد کل افراد سالم، بیانگر این Recall در کلاس افراد سالم است. در واقع این معیار دوگان Precision می‌باشد. همچنین فرمول آن به شرح زیر است:

$$\text{Recall} = \frac{TP}{TP + FN}$$

۵.۹ نقطه ضعف

این معیار به صورت دوگان معیار Precision عمل میکند و نقطه ضعف یکی، نقطه قوت دیگری است. در معیار Recall نمی‌توان متوجه شد که یک تشخیص طبقه بند به چه میزان قابل اتکا است. برای مثال، اگر طبقه بند فرد را سالم تشخیص داد، نمی‌توان مطمئن بود که آیا فرد بیماری هست که به اشتباه سالم تشخیص داده شده است یا خیر.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

شکل ۱۴: توپولوژی ماتریس درهم ریختگی و شاخص‌های مرتبط