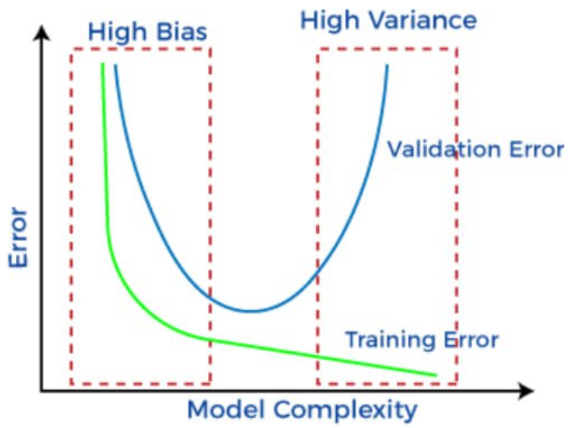
۱) الف) ۱) افزایش ویژگی ها ← منجر به بزرگ تر شدن hypothesis space می شود

۲) افزایش پیچیدگی مدل و توابع ← منجر به بزرگ تر شدن hypothesis space می شود

ب) چون بین داده ها همبستگی وجود دارد اور ها و خطا ها را بیشتر یاد می گیرد و باعث افزایش overfitting می شود یعنی variance بسیار زیاد می شود و در کل باعث پیچیدگی مدل می شود و اگر ویژگی های تکراری را حذف کنیم var کم می شود و bias زیاد تر می شود در کل مدل ساده تر شده و از overfit دوری شویم

ج) اولی: طبق تعریف داریم $Bias = E[\hat{f}(x)] - f(x)$ و حتی اگر تعداد داده ها را زیاد کنیم متوازن بایاس کم نمی شود و از آن مستقل است.
پس این گزاره غلط است



گزاره دوم نیز غلط است خطای train همیشه کم می شود و به overfit می رسیم ولی خطای test و validation تا جایی کم می شود و سپس زیاد می شود.

۲) الف) در درخت تصمیم خطا زمانی اتفاق می‌افتد که از X های یکسان Y های

مختلفی داشته باشیم در این جدول ۱ ← ۱۱۰ خواهیم رفت اما ۰ ← ۱۱۰ هم

در جدول داریم همینطور ? ← ۱۱۱ پیش بینی می‌شود (یکی ۱ و یکی ۲ داریم و رندوم عمل می‌کند)

که خطا رخ می‌دهد پس در مجموع ۲ خطا داریم که می‌شود:

$$\frac{۲}{۱۳} \sim ۱۵/۴\ \%$$

۱ ← ۱۱۰
۱ ← ۱۱۰
۰ ← ۱۱۰

۱ ← ۱۱۰

۰ ← ۱۱۱
۱ ← ۱۱۱

? ← ۱۱۱

ب) می‌خواهیم اثبات کنیم زمانی که نتیجه جدا کننده‌ای نداریم بیشترین خطا را داریم.

فرض کنید c تا کلاس و k تا داده داریم

زمانی که متوالی نتیجه‌های تمام داده‌ها یکسان است حداقل تعداد $\lceil \frac{k}{c} \rceil$

درست است در نتیجه خطا می‌شود $\frac{k-\lceil \frac{k}{c} \rceil}{k}$

اثبات (با استقرا): پایه $k=2$ فرض: برای $k<n$ برقرار است حکم: $k=n$ هم برقرار است.

برهان خلف: فرض می‌کنیم زمانی که نتیجه جدا کننده f را داریم که درخت برحسب آن جدا

می‌شود، حداکثر خطا را داریم در این صورت مثلاً اگر تعداد درست در دسته i ام $t_i<n$

باشد آنگاه تعداد درست $\ge \lceil \frac{t_i}{c} \rceil$ و در نتیجه تعداد کل درست $\ge \sum_i \lceil \frac{t_i}{c} \rceil$

همچنین $\sum_i \lceil \frac{t_i}{c} \rceil \ge \lceil \frac{\sum_i t_i}{c} \rceil$ و زمانی که نتیجه جدا کننده نداریم حداقل تعداد درست

$\lceil \frac{n}{c} \rceil = \lceil \frac{\sum_i t_i}{c} \rceil$ می‌شود که با فرض در تناقض است و ثابت می‌شود حکم برقرار است.

۳)

الف) فرض کنید k کلاس داریم که به شکل $D_y = \{1, 2, \ldots, k\}$ (w_1, w_2, …, w_k)
تعریف شده‌اند

و تبدیلی به شکل $X \to Y$ داریم که x یک بردار از فضای R^n است
با در نظر گرفتن BIAS به شکل $\begin{bmatrix} x \\ 1 \end{bmatrix}$ به فضای R^{n+1} می‌رود

(یعنی n تا feature داریم)

فضای R^n توسط k تا بردار $w_1 \ldots w_k$ به k دسته تقسیم شده است
که بردار w_i مربوط به کلاس i است.

با این توضیحات، خواهیم داشت:

(دقت کنید برای اضافه کردن بایاس باید $w = \begin{bmatrix} w \\ BIAS \end{bmatrix}$ و $x = \begin{bmatrix} x \\ 1 \end{bmatrix}$ نوشته شود.)

$$\Pr_{w_1, \ldots, w_k}(Y = i \mid X = x) = \frac{e^{w_i^T x}}{\sum_k e^{w_k^T x}}$$

$$e^{w_k^T x} = e^0 = 1$$

که طبق گفته سوال $w_k = 0$ قرار می‌دهیم که خواهیم داشت ↑

$$\Pr_{w_1, \ldots, w_k}(Y = i \mid X = x) = \frac{e^{w_i^T x}}{1 + \sum_{j=1}^{k-1} e^{w_j^T x}}$$

ب) $k-1$ بردار $w_1, w_2, \dots, w_{k-1}$ را باید Learn کنیم

بردار از فضای R^{n+1} است که n تعداد feature های

x است. اگر BIAS را صفر بگیریم در کل $(k-1)\times n$ مقدار

مجهول داریم اما اگر BIAS را در نظر بگیریم $(k-1)\times(n+1)$

مقدار مجهول خواهیم داشت که اگر از BIAS صرف نظر می کنیم

ج) n داده به شکل $(x_1, y_1), \dots, (x_n, y_n)$ داریم

که y_i به شکل $\{1, 2, \dots, k\}$ است

داده ها را به شکل iid در نظر بگیریم

$$P_{(y_{1:n} \mid x_{1:n})} = \prod_{i=1}^{n} P_{(Y = y_i \mid X = x_i ;\, w_1, \dots, w_k)}$$

$$\rightarrow L_{w_{1:k}} = \sum_{i=1}^{n} \ln P_{(Y = y_i \mid X = x_i)}$$

$$= \sum_{i=1}^{n} \ln \frac{e^{w_{y_i}^T x_i}}{1 + \sum_{l=1}^{k-1} e^{w_l^T x_i}}$$

$$= \boxed{\sum_{i=1}^{n} \left[w_{y_i}^T x_i - \ln\left(1 + \sum_{l=1}^{k-1} e^{w_l^T x_i}\right) \right]}$$

$$\nabla_{w_k} L = \sum_{i=1}^{n} \left[I_{(i,k)}\, x_i - \frac{x_i\, e^{w_k^T x_i}}{1 + \sum_{l=1}^{k-1} e^{w_l^T x_i}} \right]$$ ج

$$I_{(i,k)} = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases}$$ و

$$\rightarrow \nabla_{w_k} L = \sum_{i=1}^{n} x_i \left[I_{(i,k)} - P_{(Y=k \mid X = x_i)} \right]$$

$$f_{w_{1:k}} = L_{w_{1:k}} - \frac{\lambda}{2} \sum_{j=1}^{k-1} \|w_j\|^2$$ خ

$$\hookrightarrow \nabla_{w_k} f = \underbrace{\nabla_{w_k} L}_{} - \lambda w_k$$

از بخش ت

۴)

$$\underset{n\times 1}{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \qquad \underset{n\times(m+1)}{X} = \begin{bmatrix} 1 & - x_1^T - \\ 1 & - x_2^T - \\ & \vdots \\ 1 & - x_n^T - \end{bmatrix} \qquad \underset{(m+1)\times 1}{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_m \end{bmatrix}$$

اینجا x_n منظور داده‌ی n ام است ($x^{(n)}$)

ستون اول X بخاطر Bias و w_0 اضافه شده است.

که Loss به شکل: $\boxed{L = \frac{1}{2}\|y - Xw\|^2}$ تعریف می‌شود

که باید با حرکت روی w ها min شود

$$L = \frac{1}{2}(y - Xw)^T(y - Xw) = \frac{1}{2}\left(y^T y - y^T X w - w^T X^T y + w^T X^T X w\right) \Rightarrow \nabla_w L = \nabla_w \frac{1}{2}\left(y^T y - 2 w^T X^T y + w^T X^T X w\right)$$

$$= -X^T y + X^T X w \xrightarrow{=0} \boxed{\hat{w} = (X^T X)^{-1} X^T y}$$

حال به حل سوال می‌پردازیم:

الف) کافی است جای ماتریس X، x_j که ستون j ام X است را قرار دهیم:

$$w_j = (x_j^T x_j)^{-1} x_j^T y = \frac{x_j^T y}{x_j^T x_j}$$

یا از راه دیگر:

$$L(w_j) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - w_j x_{ji}\right)^2$$

x_{ji} عنصر ستون j ام سطر i ام ماتریس X

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^{n} (y_i - w_j x_{ji})(-x_{ji}) = 0$$

$$\rightarrow \sum_{i=1}^{n} y_i x_{ji} = w_j \sum_{i=1}^{n} x_{ji}^2$$

$$\rightarrow x_j^T y = w_j x_j^T x_j \rightarrow w_j = \frac{x_j^T y}{x_j^T x_j}$$

ب) چون ستون ها برهم عمود بودند پس ضرب داخلی آن ها صفر است یعنی :

$$X^T X = \mathrm{diag}\left(|x_1|^2, \dots, |x_m|^2\right)$$

$$\rightarrow (X^T X)^{-1} = \mathrm{diag}\left(\frac{1}{|x_1|^2}, \dots, \frac{1}{|x_m|^2}\right)$$

خونه های غیر قطر صفر می‌شوند

$$\rightarrow w = (X^T X)^{-1} X^T y \rightarrow w_j = \left((X^T X)^{-1} X^T y\right)_j = (X^T X)^{-1}_j \, (X^T y)_j$$

$$, (X^T X)^{-1}_j = \frac{1}{|x_j|^2} \rightarrow w_j = \frac{(X^T y)_j}{|x_j|^2} = \frac{x_j^T y}{|x_j|^2}$$

$$\rightarrow \boxed{w_j = \frac{x_j^T y}{x_j^T x_j}}$$

از بخش الف و ب نتیجه می‌گیریم محاسبه تک‌تک w_j ها و محاسبه ی همزمان w_j ها هر ۲ به جواب یکسان می‌رسند

ج) طبق فرض $\underset{n\times(m+1)}{X} = \begin{bmatrix} 1 & -x_1^T- \\ \vdots & \vdots \\ 1 & -x_n^T- \end{bmatrix}$ داریم:

$$X^T X = \begin{bmatrix} n & \mathrm{sum}(x_j) \\ \mathrm{sum}(x_j) & |x_j|^2 \end{bmatrix}$$

$\mathrm{sum}(x) =$ جمع تمام مقادیر بردار x یعنی $= \sum_i x^{(i)}$

$$X^T y = \begin{bmatrix} \mathrm{sum}(y) \\ x_j^T y \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{\begin{bmatrix} |x_j|^2 & -\mathrm{sum}(x_j) \\ -\mathrm{sum}(x_j) & n \end{bmatrix}}{n|x_j|^2 - \mathrm{sum}(x_j)^2}$$

name φ $\leftarrow$ $n|x_j|^2 - \mathrm{sum}(x_j)^2$

$$w = (X^T X)^{-1} X^T y = \frac{1}{\varphi} \begin{bmatrix} |x_j|^2 & -\mathrm{sum}(x_j) \\ -\mathrm{sum}(x_j) & n \end{bmatrix} \begin{bmatrix} \mathrm{sum}(y) \\ x_j^T y \end{bmatrix}$$

$$\rightarrow w_j = \frac{n x_j^T y - \mathrm{sum}(x_j)\,\mathrm{sum}(y)}{n|x_j|^2 - \mathrm{sum}(x_j)^2} = \frac{\frac{x_j^T y}{n} - \frac{\mathrm{sum}(x_j)}{n}\frac{\mathrm{sum}(y)}{n}}{\frac{|x_j|^2}{n} - \left(\frac{\mathrm{sum}(x_j)}{n}\right)^2}$$

$$= \frac{E[x_j y] - E[x_j]E[y]}{E[x_j^2] - E^2[x_j]} = \frac{\mathrm{Cor}(x_j, y)}{\mathrm{Var}(x_j)}$$

$\rightarrow \varphi = n^2 \mathrm{Var}(x_j)$

بخش دوم :

$$w_0 = \frac{sum(y)\,|x_j|^{2} - sum(x_j)\,(x_j^{T} y)}{n^{2}\, var(x_j)}$$

$$= \frac{\cancel{n^{2}}\, E[y]\, E[x_j^{2}] - \cancel{n^{2}}\, E[x_j]\, E[x_j y]}{\cancel{n^{2}}\, var(x_j)}$$

$$= \frac{E[y]\, E[x_j^{2}] - E[y]\, E^{2}[x_j] + E[y]\, E^{2}[x_j] - E[x_j]\, E[x_j y]}{var(x_j)}$$

$$= \frac{E[y]\, var(x_j) - E[x_j]\, cor(x_j, y)}{var(x_j)}$$ و $\frac{cor(x_j, y)}{var(x_j)} = w_j$ طبق الف

$$= E[y] - E[x_j]\, w_j$$

$$\rightsquigarrow \boxed{w_0 = E[y] - w_j\, E[x_j]}$$