پاسخنامه تمرین تئوری سوم

سوال ۱)

الف)

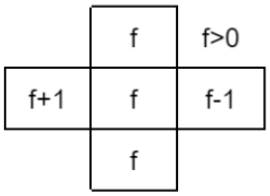
تابع $f(x,y)$ داده شده خروجی اش یک عدد است و یعنی بردار 1*1 است بنابراین بردار وزن w که در f ضرب میشود تا V را بدهد هم 1*1 است و صرفا 1 عدد است.

تابع $f(x,y)$ فاصله افقی تا گنج را میدهد یعنی هر چه به گنج نزدیکتر باشیم مقدار f کمتر خواهد شد.

برای هر دو خانه عمودی زیر هم، فاصله افقی تا گنج تفاوتی ندارد پس f یکسان دارند و چون w هم ثابت است V یکسان نیز دارند.

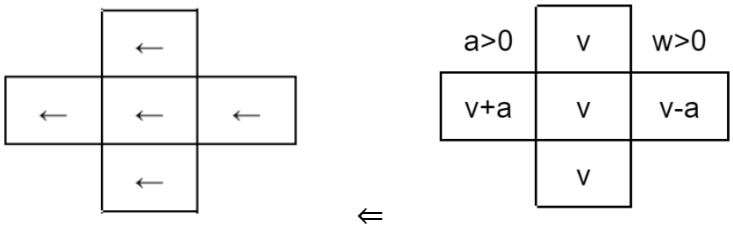
برای هر دو خانه افقی کنار هم، خانه سمت راست فاصله کمتری تا گنج دارد و f کمتری دارد (1 واحد کمتر).

میتوانیم به طور خلاصه وضعیت f های خانه های اطراف یک خانه را به شکل زیر در نظر بگیریم:



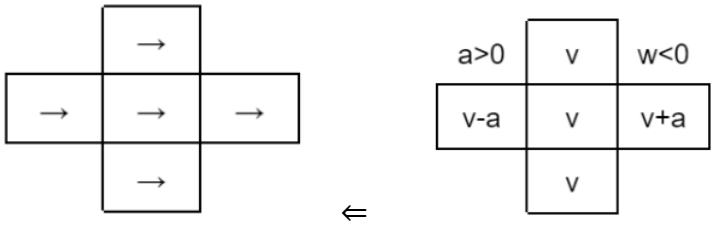
حال بسته به w ارزش های مختلفی به دست خواهیم آورد.

اگر w بزرگتر از 0 باشد، خانه سمت چپی ارزش بیشتری خواهد داشت بنابراین در تعیین سیاست به سمت آن اشاره میکنیم.



که سیاست iii را نتیجه میدهد.

اگر w منفی باشد خانه سمت راستی ارزش بیشتر خواهد داشت (عدد منفی با قدر مطلق کمتر، عددی بزرگتر است) بنابراین در سیاست به سمت راست اشاره میکنیم که سیاست i است.



در کل با تابع الف سیاست های i و iii را میتوانیم داشته باشیم.

ب)

خروجی این تابع نیز یک عدد است که فاصله منهتن تا گنج است. هرچه به گنج نزدیکتر باشیم f مقدار کمتر دارد و هرچه از گنج دورتر باشیم f مقدار بیشتر دارد. باز هم بردار وزن یک بردار 1*1 است و در واقع صرفا یک عدد است. اگر وزن مثبت باشد value هرچه به گنج نزدیک تر باشیم کمتر می شود بنابراین در سیاستمان به خانه ای با فاصله منهتن دورتر(چپ تر یا پایینتر) اشاره میکنیم. اگر وزن منفی باشد هر چه به گنج نزدیک تر باشیم value بیشتر میشود بنابراین در سیاستمان به خانه ای با فاصله منهتن نزدیکتر(بالاتر یا راست تر) اشاره میکنیم.

سیاست vi با این تابع f برای وزن منفی میتواند باشد.

ج)

خروجی تابع یک عدد است پس بردار وزن 1*1 است و وزن نیز 1 عدد است. اگر وزن مثبت باشد یعنی خانه های با فاصله واقعی نزدیکتر به گنج ارزش کمتری خواهند داشت(فاصله شان کمتر است پس fw نیز کمتر است) و خانه های دورتر ارزش بیشتر خواهند داشت. اگر w منفی باشد خانه های با فاصله واقعی کمتر ارزش بیشتر خواهند داشت و سیاست به شکلی به دست می آید که ما را به سمت گنج هدایت کند. سیاست v برای این تابع f با وزن منفی میتواند استفاده شود چون در بخش پایینی نقشه ما را به سمت چپ هدایت میکند که یک قدم در مسیر واقعی به سمت گنج نزدیکتر شویم و در بخش چپ نقشه ما را به بالا هدایت میکند و در بخش بالای نقشه ما را به راست هدایت میکند تا به گنج برسیم.

در کل برای این بخش می توان سیاست v را در نظر گرفت.

د)

دو سیاست باقی مانده ii و iv هستند. در هر دو، همه خانه ها یا به بالا یا به پایین اشاره میکنند. با در نظر گرفتن تابع الف که همه خانه ها فقط به چپ یا فقط به راست اشاره می کردند و در الف، ما فاصله افقی تا گنج را محاسبه می کردیم، میتوانیم برای ii و iv تابع فاصله عمودی تا گنج را در نظر بگیریم.

$$f(x, y) = |y - y^*|$$

این تابع هرچه از گنج در راستای عمودی دورتر باشیم مقدار بیشتر دارد. حال اگر وزن منفی باشد خانه های بالاتر ارزش بیشتر و خانه های پایینتر ارزش کمتر میگیرند بنابراین در سیاست به بالا اشاره میکنیم که همان سیاست ii است. اگر وزن مثبت باشد خانه های پایین تر ارزش بیشتر دارند بنابراین در سیاست به پایین اشاره می کنیم. قابل توجه هست که خانه های هر ردیف ، value های یکسان دارند (فاصله عمودی یکسان تا گنج).

سوال ۲)

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$

جواب نهایی:

V_2^*	9	1.5	9.75	36
V_1^*	6	0	5	24
V_0^*	0	0	0	0
	X_0	X_1	X_2	X_3

6	0	5	24
X_0	X_1	X_2	X_3

مراحل:

$$V_1(X_0) = \begin{cases} \text{right}: 0.5[0 + \gamma V_0(X_1)] + 0.5[6 + \gamma V_0(X_0)] = 3 \\ \text{stay}: 1[6 + \gamma V_0(X_0)] = 6 \leftarrow \max \end{cases}$$

$$V_1(X_1) = \begin{cases} \text{right}: 0.5[0 + \gamma V_0(X_2)] + 0.5[0 + \gamma V_0(X_1)] = 0 \\ \text{left}: 0.5[0 + \gamma V_0(X_0)] + 0.5[0 + \gamma V_0(X_1)] = 0 \\ \text{stay}: 1[0 + \gamma V_0(X_1)] = 0 \leftarrow \end{cases}$$

$$V_1(X_2) = \begin{cases} \text{right}: 0.5[0 + \gamma V_0(X_3)] + 0.5[5 + \gamma V_0(X_2)] = 2.5 \\ \text{left}: 0.5[0 + \gamma V_0(X_1)] + 0.5[5 + \gamma V_0(X_2)] = 2.5 \\ \text{stay}: 1[5 + \gamma V_0(X_2)] = 5 \leftarrow \max \end{cases}$$

$$V_1(X_3) = \begin{cases} \text{left}: 0.5[0 + \gamma V_0(X_2)] + 0.5[24 + \gamma V_0(X_3)] = 12 \\ \text{stay}: 1[24 + \gamma V_0(X_3)] = 24 \leftarrow \max \end{cases}$$

$$V_2(x_0) = \begin{cases} \text{right}: 0.5[0 + \gamma V_1(x_1)] + 0.5[6 + \gamma V_1(x_0)] = 4.5 \\ \text{stay}: 1[6 + \gamma V_1(x_0)] = 9 \leftarrow \max \end{cases}$$

$$V_2(x_1) = \begin{cases} \text{right}: 0.5[0 + \gamma V_1(x_2)] + 0.5[0 + \gamma V_1(x_1)] = 1.25 \\ \text{left}: 0.5[0 + \gamma V_1(x_0)] + 0.5[0 + \gamma V_1(x_1)] = 1.5 \leftarrow \max \\ \text{stay}: 1[0 + \gamma V_1(x_1)] = 0 \end{cases}$$

$$V_2(x_2) = \begin{cases} \text{right}: 0.5[0 + \gamma V_1(x_3)] + 0.5[5 + \gamma V_1(x_2)] = 9.75 \leftarrow \max \\ \text{left}: 0.5[0 + \gamma V_1(x_1)] + 0.5[5 + \gamma V_1(x_2)] = 3.75 \\ \text{Stay}: 1[5 + \gamma V_1(x_2)] = 7.5 \end{cases}$$

$$V_2(x_3) = \begin{cases} \text{left}: 0.5[0 + \gamma V_1(x_2)] + 0.5[24 + \gamma V_1(x_3)] = 19.25 \\ \text{stay}: 1[24 + \gamma V_1(x_3)] = 36 \leftarrow \max \end{cases}$$

سوال۳)

الف)برای policy evaluation داریم:

$$V_{k+1}^{\pi_i}(s) \leftarrow \sum_{s'} T(s, \pi_i(s), s') \left[R(s, \pi_i(s), s') + \gamma V_k^{\pi_i}(s') \right]$$

پس:

$$V_1^{\pi}(S_3) = 3, V_1^{\pi}(S_4) = 4, V_1^{\pi}(S_5) = 5, V_1^{\pi}(S_6) = 6$$

$$V_1^{\pi}(S_1) = V_1^{\pi}(S_2) = \frac{1}{6}(-1+0) + \frac{1}{6}(-1+0) + \frac{1}{6}(-1+0) + \frac{1}{6}(-1+0) + \frac{1}{6}(-1+0) + \frac{1}{6}(-1+0) = -1$$

$$V_2^{\pi}(S_1) = V_2^{\pi}(S_2) = \frac{1}{6}(-1-1) + \frac{1}{6}(-1-1) + \frac{1}{6}(-1+3) + \frac{1}{6}(-1+4) + \frac{1}{6}(-1+5) + \frac{1}{6}(-1+6) = 1.67$$

$$V_3^{\pi}(S_1) = V_3^{\pi}(S_2) = \frac{1}{6}(-1+1.67) + \frac{1}{6}(-1+1.67) + \frac{1}{6}(-1+3) + \frac{1}{6}(-1+4) + \frac{1}{6}(-1+5) + \frac{1}{6}(-1+6) = 2.56$$

$$V_4^{\pi}(S_1) = V_4^{\pi}(S_2) = \frac{1}{6}(-1+2.56) + \frac{1}{6}(-1+2.56) + \frac{1}{6}(-1+3) + \frac{1}{6}(-1+4) + \frac{1}{6}(-1+5) + \frac{1}{6}(-1+6) = 2.58$$

$$V_5^{\pi}(S_1) = V_5^{\pi}(S_2) = \frac{1}{6}(-1+2.58) + \frac{1}{6}(-1+2.58) + \frac{1}{6}(-1+3) + \frac{1}{6}(-1+4) + \frac{1}{6}(-1+5) + \frac{1}{6}(-1+6) = 2.95 \approx 3$$

برای حالت های 3 تا 6، مقدار ارزش آنها با یکبار محاسبه همگرا می شود. اما برای حالت های 1 و 2 باید تا 5 مرحله ادامه دهیم و میبینیم که ارزش آن ها به 3 همگرا می شود. پس:

حالت	S1	S2	S3	S4	S5	S6
π(i)	ریختن تاس	ریختن تاس	اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی
V(π(i))	3	3	3	4	5	6

ب) طبق ارزش های بدست آمده در قسمت الف، ارزش ریختن تاس برابر است با:

$$\frac{1}{6}(-1+3)+\frac{1}{6}(-1+3)+\frac{1}{6}(-1+3)+\frac{1}{6}(-1+4)+\frac{1}{6}(-1+5)+\frac{1}{6}(-1+6)=3$$

اکنون policy improvement را انجام می دهیم:

$$\pi_{i+1}(s)=\arg\max_a \sum_{s'} T(s,a,s')\left[R(s,a,s')+\gamma V^{\pi_i}(s')\right]$$

در حالت 1، اگر عمل تاس ریختن را انتخاب کنیم، مطابق بالا ارزش 3 بدست می آید. اگر عمل اتمام را انتخاب کنیم، ارزش 1 بدست می آید. پس عمل تاس ریختن را انتخاب می کنیم. برای حالت 2 نیز مانند بالا عمل تاس ریختن انتخاب می شود. برای حالت 3 هر دو عمل ارزش 3 را دارند، پس هر دوی آنها را می نویسیم.

برای حالتهای 4 تا 6، عمل اتمام بازی انتخاب می شود. پس در نهایت داریم:

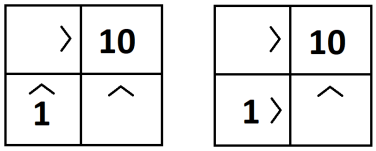
حالت	S1	S2	S3	S4	S5	S6
π(i)	ریختن تاس	ریختن تاس	اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی
π(i+1)	dice	dice	dice/finish	finish	finish	finish

ج) با مقایسه جدول بالا، می بینیم که سیاست ها تغییری نکردند، پس می توانیم نتیجه بگیریم که سیاست همگرا شده است، در حالیکه ارزش ها ممکن است بعد از سیاست ها همگرا شوند.

سوال ۴)

الف)

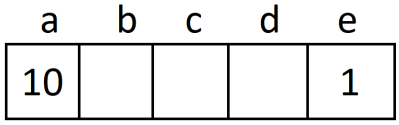
لزوما خیر، ماکزیمم پاداش یکتاست اما policy لزوما یکتا نیست و میتوان با مثال نشان داد. مثال ساده زیر را در نظر بگیرید.



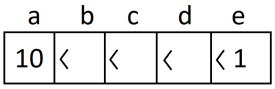
در هر دو حالت با اینکه در سیاست در خانه با امتیاز 1 تفاوت داریم اما در نهایت به یک maximum value خواهیم رسید.

ب)

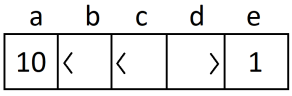
درست است و میتوان با مثال نشان داد. فرض کنید grid زیر را داریم و action های ما east و west و exit هستند.



اگر $\gamma = 1$ باشد optimal policy میشود :



اگر $\gamma = 0.1$ باشدoptimal policy میشود :



واضحا در خانه d تفاوت داریم. در حالت $\gamma = 1$ گذر زمان از ارزش 10 برای ما کم نمی کند. اما اگر $\gamma = 0.1$ و در d باشیم وقتی به 10 برسیم ارزش آن $10 * \gamma^3$ شده که 0.01 است پس بهتر است به سمت راست حرکت کنیم چون امتیاز 1 با $1 * \gamma$ یعنی 0.1 قابل دستیابی است. در نتیجه با لاندا های مختلف optimal policy های مختلفی می توانیم داشته باشیم.

ج)

به طور کلی بله اما نکته اینجاست شرایطی باید برقرار باشد تا به جواب نهایی برسیم 1-تعدا استیت ها محدود باشد 2-اندازه پاداش ها محدود باشد 3- discount factor کمتر از 1 باشد

د)

۱) مسئله: رانندگی خودکار

عامل: ماشین خودران محیط: خیابان شامل ماشین های دیگر و عابرین پیاده

2) مسئله: شطرنج

عامل: موتور شطرنج محیط: بازی شطرنج شامل حریف مقابل

3) مسئله: معاملات سهام

عامل: معامله کننده سهام محیط: قیمت ها و شاخص های سهام، اخبار مرتبط با سهامها و ...

یادگیری مبتنی بر مدل و بدون مدل هر دو معایب و مزیتهای خود را دارند. روش های مبتنی بر مدل به عامل اجازه برنامه ریزی و استنتاج بر اساس مدل را می دهند اما روش های بدون مد ل تنها بر یادگیری تکیه می کنند. طور کلی با وجود آخرین پیشرفتهای الگوریتم های یادگیری تقویتی، اگر همراه با تعریف مسئله)در دنیای واقعی(یک مدل دقیق ارائه نشده باشد، روش های بدون مدل برتری دارند زیرا کوچکترین در خطا در مدل می تواند باعث ناپایداری عامل شود.

1)رانندگی خودکار: ساختن یک مدل دقیق دشوار است بنابراین یادگیری بدون مدل برتری دارد.

2)شطرنج: ساخت یک مدل دقیق آسان است پس یادگیری با مدل برتری دارد.

3)معاملات سهام: ساخت یک مدل دقیق دشوار است پس یادگیری بدون مدل برتری دارد.

سوال ۵)

الف)

برای حل از رابطه زیر استفاده می‌کنیم:

$$V(s) : (1 - a) * V(s) + a(R(s, a, s') + gamma * V(s'))$$

مقدار استیت‌های اولیه را برابر صفر در نظر می‌گیریم.

درنتیجه داریم:

$$V(A) : 0 + 0.5 * (-1) + 0 = -0.5$$

$$V(B) : 0 + 0.5 * (-1) + 0 = -0.5$$

$$V(C) : 0 + 0.5 * 32 + 0 = 16$$

ب)

$$V(A) : 0.5 * (-0.5) + (-0.5) * (-1) - 0.5 = -1$$

$$V(B) : 0.5 * (-0.5) + (-0.5) * (-99) + 0 = -49.77$$

به این دلیل که در این قسمت برای C دیتاپوینت جدیدی نداریم، تغییر در مقدار آن ایجاد نمی‌شود.

ج)

در این قسمت از رابطه زیر استفاده می‌کنیم:

$$Q(s, a) : (1 - a) * Q(s, a) + a * (R(s, a, s') + gamma * max(s', a'))$$

$$Q(A, East) : 0 + 0.5 * (-1) + 0 = (-0.5)$$

$$Q(B, East) : 0 + 0.5 * (-1) + 0 = (-0.5)$$

$$Q(C, East) : 0 + 0.5 * 32 + 0 = 16$$

Q(A, East) : 0.5 * (-0.5) + 0.5 * (-1) + 0 = (-0.75)

Q(B, East) : 0.5 * (-0.5) + 0.5 * (-99) = (-49.75)

Q(A, South) : 0

سوال ۶)

الف)

V * (3, 2) = 100, V * (2, 2) = 50, V * (1, 3) = 12.5

مسیر بهینه که از (2،2) شروع می شود، رفتن به مربع 100+ است که دارای پاداش تخفیف 0 + 50 = γ * 100 است.

برای (1،3)، رفتن به یکی از 25+ یا 100+ دارای همان پاداش با تخفیف 12.5 و برای (3و2) حرکت به 100+.

ب)

Q(s, a) ← (1 - α)Q(s, a) + α(R(s, a, s0) + γ max_a' Q(s', a'))

Q((1, 3), S) ← (1 - ½)Q((1, 3), S) + ½ (R((1, 3), S, (1,2)) + ½ max_a' Q((1,2), a')) = ½ * 0 + ½ (0 + ½ * max(0, 0, 0)) = 0

Q((1, 2), E) ← (1 - ½)Q((1, 2), E) + ½ (R((1, 2), E, (2,2)) + ½ max_a' Q((2,2), a')) = ½ * 0 + ½ (0 + ½ * max(0, 0, 0, 0)) = 0

Q((2, 2), S) ← (1 - ½)Q((2, 2), S) + ½ (R((2, 2), S, (2,1)) + ½ max_a' Q((2,1), a')) = ½ * 0 + ½ (-100 + ½ * max(0, 0, 0)) = -50

–

Q((1, 3), S) ← (1 - ½)Q((1, 3), S) + ½ (R((1, 3), S, (1, 2)) + ½ max_a' Q((1, 2), a')) = ½ * 0 + ½ (0 + ½ * max(0, 0, 0)) = 0

Q((1, 2), E) ← (1 - ½)Q((1, 2), E) + ½ (R((1, 2), E, (2, 2)) + ½ max_a' Q((2, 2), a')) = ½ * 0 + ½ (0 + ½ * max(0, 0, 0, 0)) = 0

$Q((2, 2), E) \leftarrow (1 - ½)Q((2, 2), E) + ½ (R((2, 2), E, (3,2)) + ½ \ max\_a' \ Q((3,2), a')) = ½ * 0 + ½ (-100 + ½ * max(0, 0, 0)) = 0$

$Q((3, 2), N) \leftarrow (1 - ½)Q((3, 2), N) + ½ (R((3, 2), N, (3,3)) + ½ \ max\_a' \ Q((3,3), a')) = ½ * 0 + ½ (100 + ½ * max(0, 0)) = 50$

–

$Q((1, 3), S) \leftarrow (1 - ½)Q((1, 3), S) + ½ (R((1, 3), S, (1, 2)) + ½ \ max\_a' \ Q((1, 2), a')) = ½ * 0 + ½ (0 + ½ * max(0, 0, 0)) = 0$

$Q((1, 2), E) \leftarrow (1 - ½)Q((1, 2), E) + ½ (R((1, 2), E, (2,2)) + ½ \ max\_a' \ Q((2,2), a')) = ½ * 0 + ½ (0 + ½ * max(0, 0, 0, 0)) = 0$

$Q((2, 2), E) \leftarrow (1 - ½)Q((2, 2), E) + ½ (R((2, 2), E, (3,2)) + ½ \ max\_a' \ Q((3,2), a')) = ½ * 0 + ½ (-100 + ½ * max(50, 0, 0)) = 12.5$

$Q((3,2),N) = 50 \ Q((1,2),S) = 0 \ Q((2, 2), E) = 12.5$

پ)

1- با استفاده از به‌روزرسانی‌های تقریبی وزن یادگیری

$wi \leftarrow wi+\alpha[(R(s, a, s')+\gamma \ max\_a' \ Q(s', a')) - Q(s, a)] * fi(s, a).$

تنها زمانی که پاداش در قسمت اول غیر صفر است

$W1 \leftarrow 0 + ½ [0 + ½ \ max\_a' \ (Q(1,2), a')) - Q((1,3), S)] * 2 = 0 + ½ [0 + ½ \ 0 - 0]*2=0$

$W1 \leftarrow 0 + ½ [0 + ½ \ max\_a' \ (Q(2,2), a')) - Q((1,2), S)] * 2 = 0 + ½ [0 + ½ \ 0 - 0]*2=0$

$W1 \leftarrow 0 + ½ [-100 + ½ \ max\_a' \ (Q(2,1), a')) - Q((2,2), S)] * 2 = 0 + ½ [-100 + ½ \ 0 - 0]*2 = -100$

مشاهده میکنیم تنها زمانی که پاداش در اپیزود اول غیر صفر است و باعث میشود مقدار wi عوض شود، آ]رین جابه جایی یعنی زمانی است که به حالت -100 میرویم پس برای محاسبه w2, w3 همین یک تبدیل که موثر است را مینویسیم:

$W2 \leftarrow 0 + ½\ [-100 + ½\ max\_a'\ (Q(2,1), a')) - Q((2,2), S)] * 2 = 0 + ½\ [-100 + ½\ 0 - 0] * 2 = -100$

$W3 \leftarrow 0 + ½\ [-100 + ½\ max\_a'\ (Q(2,1), a')) - Q((2,2), S)] * 2 = 0 + ½\ [-100 + ½\ 0 - 0] * 2 = -100$

1- w1 = -100 w2 = -100 w3 = -100

2- غرب، در واقع این عمل $\max_a Q((1,2), a)$ است که در آن Q(s, a) با استفاده از تابع داده شده محاسبه می شود.

در این حالت ، مقدار Q برای غرب حداکثر است .

$Qf(s, a) = w1f1(s) + w2f2(s) + w3f3(a)$

$Qf((1,2), N) = -1 * 1 + 1 * 2 + 2 * 1 = 3$

$Qf((1,2), S) = -1 * 1 + 1 * 2 + 2 * 2 = 5$

$Qf((1,2), E) = -1 * 1 + 1 * 2 + 2 * 3 = 7$

$Qf((1,2), W) = -1 * 1 + 1 * 2 + 2 * 5 = 9$

سوال ۷)

ویژگی هایی که انتخاب می شوند باید بتوانند تا حد امکان موقعیت های متفاوت را از یکدیگر تمیز دهند. یک حالت از انتخاب ویژگی ها می تواند به شکل زیر باشد: 1)موقعیت خودرو 2)سرعت خودرو 3)فاصله خودرو از نزدیکترین جسم 4)سرعت نزدیک شدن به نزدیکترین جسم 5)فاصله خودرو از حاشیه جاده 6)میزان ترافیک جاده 7)میزان بارندگی، رطوبت و یخ زدگی جاده و .....

حتی اگر تمام این ویژگی ها در تابع ارزش خطی مورد استفاده قرار بگیرند و ضرایب نیز به بهترین شکل آموخته شوند، هنوز عامل می تواند دچار خطاهای فاحش شود. برای مثال عامل نمی تواند یک خودرو که چراغ راهنمای خود را روشن کرده و قصد پیچیدن در جاده دارد را از یک خودروی معمولی تشخیص دهد و همین ممکن است باعث رخ دادن حادثه شود.