دانشکده‌ی مهندسی کامپیوتر

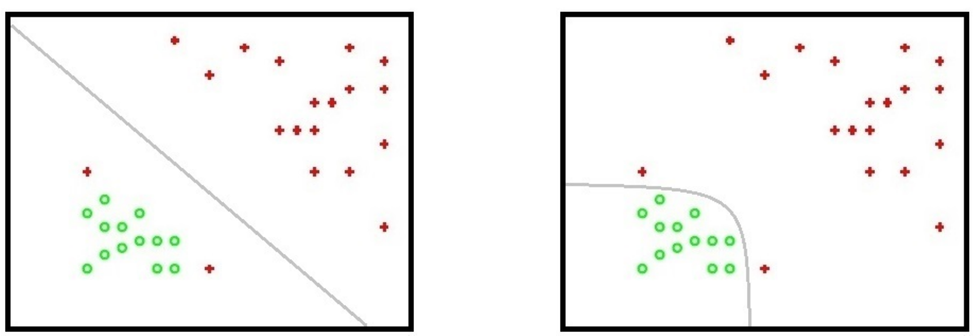
یادگیری ماشین	تابستان ۱۴۰۰

پاسخ امتحان پایان‌ترم

مدرس: دکتر محمّدحسین رهبان

پاسخ سوال ۱ *SVM & Feature Space*

آ-۱) در حالتی که $C \to \infty$، جریمه برای نقطه‌هایی که اشتباه دسته‌بندی می‌شوند زیاد است. بنابراین مرز تصمیم‌گیری به گونه‌ای است که به دقت نقطه‌ها را دسته‌بندی می‌کند. در حالتی که $C \to 0$، جریمه برای نقطه‌هایی که اشتباه دسته‌بندی می‌شوند زیاد نیست. در نتیجه دسته‌بند با وجود چند اشتباه هم می‌تواند حاشیه را بیشینه کند.

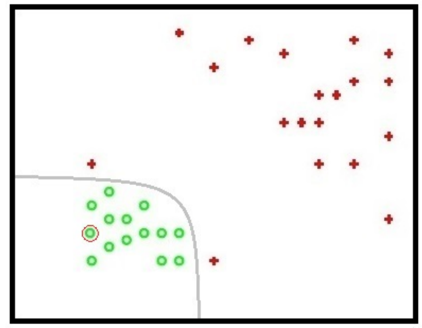


(آ) مرز تصمیم‌گیری در حالت $C \to \infty$

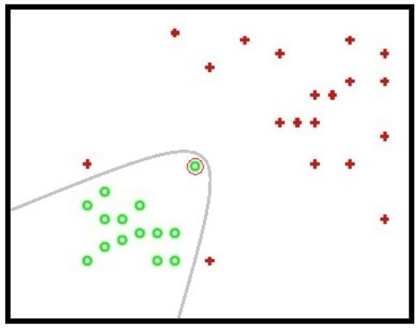
(ب) مرز تصمیم‌گیری در حالت $C \to 0$

آ-۲) با استدلال‌های گوناگونی می‌توان نشان داد که هر دو حالت بخش پیشین می‌تواند در شرایطی بهتر از دیگری باشند. در حالت $C \to \infty$ تلاش می‌شود داده‌ها درست دسته‌بندی شوند. بنابراین به دقت بالاتری می‌رسیم. با استدلال دیگر می‌توان در حالت $C \to 0$ دو داده موجود در کلاس دایره را داده خارج از محدوده[1] در نظر گرفت و با پذیرش این دو داده به عنوان خطا به نتیجه بهتری رسید. هر چند جواب دوم بهتر است و بیشتر دانشجویان به آن اشاره کرده‌اند ولی در صورت ارایه استدلالی درست برای حالت اول، نمره آن برای شما منظور شده است. نکته مهم اینکه اگر داده افزوده شده از مرز و حاشیه دور باشد (*Support Vector* نباشد) مرز را تغییر نمی‌دهد ولی اگر جز بردار پشتیبان‌ها باشد ممکن است تغییر مرز را به دنبال داشته باشد. چون در حالت $C \to \infty$ تلاش بر دسته‌بندی درست همه داده‌هاست، افزودن یک داده می‌تواند تاثیر زیادی روی مرز گذاشته و آن را خیلی تغییر دهد. از طرفی در حالت $C \to 0$ هم حتی افزودن یک داده خارج از محدوده هم ممکن است مرز را کمی جابه‌جا کند.

[1]Outlier



(ج) عدم تغییر مرز تصمیم‌گیری با افزودن داده جدید



(د) تغییر مرز تصمیم‌گیری با افزودن داده جدید

ب-۱)

$$k_1(\mathbf{a},\mathbf{b}) = \left(\mathbf{a}^\top \mathbf{b}\right)^2 = \left(a_1 b_1 + a_2 b_2\right)^2 = a_1^2 b_1^2 + 2\, a_1 b_1 a_2 b_2 + a_2^2 b_2^2$$

$$k_1(\mathbf{a},\mathbf{b}) = a_1^2 b_1^2 + 2\, a_1 a_2 b_1 b_2 + a_2^2 b_2^2 = \begin{pmatrix} a_1^2 \\ \sqrt{2}\, a_1 a_2 \\ a_2^2 \end{pmatrix}^\top \begin{pmatrix} b_1^2 \\ \sqrt{2}\, b_1 b_2 \\ b_2^2 \end{pmatrix}$$

ب-۲)

$$k_2(\mathbf{a},\mathbf{b}) = \left(\mathbf{a}^\top \mathbf{b} + 1\right)^2 = \left(\mathbf{a}^\top \mathbf{b}\right)^2 + 2\,\mathbf{a}^\top \mathbf{b} + 1 = \left(a_1^2 b_1^2 + 2\, a_1 a_2 b_1 b_2 + a_2^2 b_2^2\right) + 2\left(a_1 b_1 + a_2 b_2\right) + 1$$

$$k_2(\mathbf{a},\mathbf{b}) = a_1^2 b_1^2 + 2\, a_1 a_2 b_1 b_2 + a_2^2 b_2^2 + 2\, a_1 b_1 + 2\, a_2 b_2 + 1 = \begin{pmatrix} a_1^2 \\ \sqrt{2}\, a_1 a_2 \\ a_2^2 \\ \sqrt{2}\, a_1 \\ \sqrt{2}\, a_2 \\ 1 \end{pmatrix}^\top \begin{pmatrix} b_1^2 \\ \sqrt{2}\, b_1 b_2 \\ b_2^2 \\ \sqrt{2}\, b_1 \\ \sqrt{2}\, b_2 \\ 1 \end{pmatrix}$$

پاسخ سوال ۲ *Linear Regression - PCA*

آ) یکی از مهم‌ترین دلیل‌های بکار بردن توزیع گاوسی این است که انتگرال‌گیری $\int f(y \mid x,\theta)\, f(\theta \mid \mathcal{D})\, d_\theta$ در روش تمام بیزی [1] را برایمان آسان می‌کند؛ زیرا ضرب دو توزیع گاوسی نیز گاوسی می‌شود و انتگرال گرفتن از تابعی که شکل نمایی دارد سخت نیست. پاسخ‌های دیگری مانند ویژگی گاوسی بودن احتمال شرطی و حاشیه‌ای [2] دو توزیع گاوسی نیز قابل قبول است.

ب) با استفاده از رابطه‌های زیر می‌توان نشان داد که مقدار ویژه‌های ماتریس $XX^\top$ و $X^\top X$ برابر است و بردار ویژه‌های $XX^\top$ نیز با ضرب کردن ماتریس $X^\top$ در بردار ویژه‌های ماتریس X بدست می‌آیند.

$$XX^\top v_i = \lambda v_i \Longrightarrow X^\top X(X^\top v_i) = \lambda(X^\top v_i) \Longrightarrow X^\top X u_i = \lambda u_i$$

[1] Fully Bayesian

[2] Marginal Distribution

بنابراین با بدست آوردن مقدار و بردار ویژه‌های ماتریس $X^\top X$ در زمان $O(N^3)$، می‌توانیم روش PCA را اجرا کنیم.

پ) ۱. غلط است چون تنها مقدار ویژه‌های ماتریس‌اند که با دوران تغییر نمی‌کنند و بردارهای ویژه می‌توانند عوض شوند. به عنوان مثالی ساده می‌توان محور مختصات دو بعدی را در نظر گرفت که تمام داده‌ها روی یک خط قرار دارند و بعد از دوران، خط جابجا شده و در نتیجه بردار PCA جابجا می‌شود.

۲. درست است چون PCA در واقع L بزرگترین مقدار ویژه را در نظر گرفته و از بردار ویژه‌های نظیر آن‌ها استفاده می‌کند؛ پس با دوبار انجام دادنش، همچنان به همان بردارها با بزرگترین مقدار ویژه می‌رسیم.

۳. درست است چون ویژگی اضافه شده واریانس ۰ دارد و می‌دانیم PCA ویژگی‌های با بیشترین واریانس را انتخاب می‌کند. پس این ویژگی (با فرض ثابت نگه داشتن تعداد بعدهایی که به آن کاهش را انجام می‌دهیم) هیچوقت انتخاب نمی‌شود.

ت) برای h_0 سه حالت اعتبارسنجی متقابل [1] زیر را داریم:

اولین نمونه را داده آزمون در نظر بگیریم: در این حالت پارامتر مدل $b = \frac{1}{2}$ است و مقدار تابع هدف برای داده آزمون برابر $\left(\frac{1}{2} - 0\right)^2 = \frac{1}{4}$ می‌شود.

دومین نمونه را داده آزمون در نظر بگیریم: این حالت نیز مانند حالت قبل است.

سومین نمونه را داده آزمون در نظر بگیریم: در این حالت پارامتر مدل $b = 0$ است و مقدار تابع هدف برای داده آزمون برابر 1 می‌شود.

پس برای h_0، میانگین مستقل از v برابر با $\frac{1}{2}$ است.

برای h_1 سه حالت اعتبارسنجی متقابل زیر را داریم:

اولین نمونه را داده آزمون در نظر بگیریم: در این حالت پارامترهای مدل $a = \frac{1}{v-2}\,,\, b = \frac{-2}{v-2}$ هستند و مقدار تابع هدف برای داده آزمون برابر $\left(\frac{-4}{v-2}\right)^2$ می‌شود.

دومین نمونه را داده آزمون در نظر بگیریم: در این حالت پارامترهای مدل $a = \frac{1}{v+2}\,,\, b = \frac{2}{v+2}$ هستند و مقدار تابع هدف برای داده آزمون برابر $\left(\frac{4}{v+2}\right)^2$ می‌شود.

سومین نمونه را داده آزمون در نظر بگیریم: در این حالت پارامترهای مدل $a = b = 0$ هستند و مقدار تابع هدف برای داده آزمون برابر 1 می‌شود.

با مساوی قرار دادن میانگین سه حالت بالا با $\frac{1}{2}$ به معادله‌ای درجه دو می‌رسیم که با حل آن جواب مثبت $v = 2\sqrt{9 + 4\sqrt{6}}$ بدست می‌آید.

پاسخ سوال ۳ *Nearest Neighbour*

آ)

$$\begin{aligned}
\mathbb{E}_S\left[L_{\mathcal{D}}\left(h_S\right)\right] &= \mathbb{E}_{S_{\mathrm{x}} \sim D_{\mathcal{X}}^m\,,\, \mathrm{x} \sim D_{\mathcal{X}}\,,\, y \sim \eta(\mathrm{x})\,,\, y' \sim \eta(\pi_1(\mathrm{x}))}\left[\mathbb{1}_{[y \neq y']}\right] \\
&= \mathbb{E}_{S_{\mathrm{x}} \sim D_{\mathcal{X}}^m\,,\, \mathrm{x} \sim D_{\mathcal{X}}}\left[\mathbb{P}_{y \sim \eta(\mathrm{x})\,,\, y' \sim \eta(\pi_1(\mathrm{x}))}\left(y \neq y'\right)\right]
\end{aligned}$$

[1] Cross Validation

ب)

$$
\begin{aligned}
\mathbb{P}_{y\sim\eta(\mathbf{x})\,,\,y'\sim\eta(\pi_1(\mathbf{x}))}\left(y \neq y'\right) &= \eta(\mathbf{x}')\left(1-\eta(\mathbf{x})\right) + \left(1-\eta(\mathbf{x}')\right)\eta(\mathbf{x}) \\
&= \left(\eta(\mathbf{x})-\eta(\mathbf{x})+\eta(\mathbf{x}')\right)\left(1-\eta(\mathbf{x})\right) + \left(1-\eta(\mathbf{x})+\eta(\mathbf{x})-\eta(\mathbf{x}')\right)\eta(\mathbf{x}) \\
&= 2\,\eta(\mathbf{x})\left(1-\eta(\mathbf{x})\right) + \left(\eta(\mathbf{x})-\eta(\mathbf{x}')\right)\left(2\,\eta(\mathbf{x})-1\right)
\end{aligned}
$$

پ) حال به خاطر ویژگی لیپشیتز بودن تابع η و درستی نابرابری $|2\,\eta(\mathbf{x})-1| \leq 1$ داریم:

$$
\begin{aligned}
\mathbb{E}_S\left[L_{\mathcal{D}}\left(h_S\right)\right] &= \mathbb{E}_{S\sim D^m\,,\,\mathbf{x}\sim D}\left[2\,\eta(\mathbf{x})\left(1-\eta(\mathbf{x})\right) + \left(\eta(\mathbf{x})-\eta(\mathbf{x}')\right)\left(2\,\eta(\mathbf{x})-1\right)\right] \\
&= \mathbb{E}_{S\sim D^m\,,\,\mathbf{x}\sim D}\left[2\,\eta(\mathbf{x})\left(1-\eta(\mathbf{x})\right)\right] + \mathbb{E}_{S\sim D^m\,,\,\mathbf{x}\sim D}\left[\left(\eta(\mathbf{x})-\eta(\mathbf{x}')\right)\left(2\,\eta(\mathbf{x})-1\right)\right] \\
&\leq \mathbb{E}_{S\sim D^m\,,\,\mathbf{x}\sim D}\left[2\,\eta(\mathbf{x})\left(1-\eta(\mathbf{x})\right)\right] + \mathbb{E}_{S\sim D^m\,,\,\mathbf{x}\sim D}\left[\left(\eta(\mathbf{x})-\eta(\mathbf{x}')\right)\right] \\
&\leq 2\,L_{\mathcal{D}}\left(h^\star\right) + c\,\mathbb{E}_{S\sim D^m\,,\,\mathbf{x}\sim D}\left[\|\mathbf{x}-\mathbf{x}_{\pi_1(\mathbf{x})}\|\right]
\end{aligned}
$$

ت) با توجه به ویژگی‌های $h^\star$ داریم:

$$
L_{\mathcal{D}}\left(h^\star\right) = \mathbb{E}_{\mathbf{x}\sim D}\left[\boldsymbol{min}\{\eta(\mathbf{x}), 1-\eta(\mathbf{x})\}\right] \leq \mathbb{E}_{\mathbf{x}\sim D}\left[\eta(\mathbf{x})\left(1-\eta(\mathbf{x})\right)\right]
$$

به این ترتیب کرانی برای خطا بدست می‌آید.

پاسخ سوال ۴ *Semi-Supervised Learning*

آ) کمینه عبارت $M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})$ در $\mathbf{r}=0$ رخ می‌دهد و مقدار آن نیز صفر است. همچنین با توجه به فرض مشتق‌پذیری این عبارت می‌دانیم مقدار مشتق اول آن در $\mathbf{r}=0$ صفر است ($\nabla_{\mathbf{r}} M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})\big|_{\mathbf{r}=\mathbf{0}}=0$). بنابراین در بسط تیلور تا مرتبه ۲، دو جمله اول صفر هستند و فقط جمله سوم باقی می‌ماند:

$$
M(\mathbf{x},\mathbf{r},\boldsymbol{\theta}) \approx \frac{1}{2}\mathbf{r}^\top H(\mathbf{x},\boldsymbol{\theta})\,\mathbf{r}\,, \qquad H(\mathbf{x},\boldsymbol{\theta}) = \nabla_{\mathbf{r}}^2 M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})\big|_{\mathbf{r}=\mathbf{0}}
$$

ب) با نوشتن لاگرانژین عبارت بخش **آ** و مشتق آن نسبت به $\mathbf{r}$ داریم:

$$
H^\top = H
$$

$$
\|\mathbf{r}\|_2 \leq \varepsilon \implies \|\mathbf{r}\|_2^2 \leq \varepsilon^2 \implies 0 \leq \varepsilon^2 - \|\mathbf{r}\|_2^2
$$

$$
\boldsymbol{Lagrangian:}\ \mathcal{L}(\mathbf{r},\lambda) = \frac{1}{2}\mathbf{r}^\top H\,\mathbf{r} + \lambda\left(\varepsilon^2 - \|\mathbf{r}\|_2^2\right)
$$

$$
\begin{cases}
\dfrac{\partial\mathcal{L}(\mathbf{r},\lambda)}{\partial\mathbf{r}} = \mathbf{r}^\top H - 2\lambda\,\mathbf{r}^\top, \quad \dfrac{\partial\mathcal{L}(\mathbf{r},\lambda)}{\partial\mathbf{r}} = 0 \implies (H-2\lambda I)\mathbf{r} = 0 \quad * \\[2em]
\dfrac{\partial\mathcal{L}(\mathbf{r},\lambda)}{\partial\lambda} = \varepsilon^2 - \|\mathbf{r}\|_2^2, \qquad \dfrac{\partial\mathcal{L}(\mathbf{r},\lambda)}{\partial\lambda} = 0 \implies \|\mathbf{r}\|_2 = \varepsilon \qquad **
\end{cases}
$$

از رابطه * مشخص است که 2λ مقدار ویژه نظیر بردار ویژه $\mathbf{r}$ در ماتریس H است. حال کافی است با در نظر گرفتن رابطه ** بزرگترین مقدار ویژه این ماتریس را بدست آوریم:

$$\max_{\mathbf{r}\in\mathbb{R}^d} \frac{1}{2}\mathbf{r}^\top H\,\mathbf{r} = \max_{\mathbf{r}\in\mathbb{R}^d} \mathbf{r}^\top \lambda\,\mathbf{r} = \lambda_{max}\|\mathbf{u}\|_2^2 \implies \mathbf{r}^\star = \varepsilon\overline{\mathbf{u}}$$

پ) با در نظر گرفتن تجزیه طیفی ماتریس H داریم (ماتریس Q شامل بردار ویژه‌ها و ماتریس Λ شامل مقدار ویژه‌هاست):

$$H = Q\Lambda Q^\top$$

$$\mathbf{v}_n = \frac{H\mathbf{v}_{n-1}}{\|H\mathbf{v}_{n-1}\|} = \frac{H^n\mathbf{v}_0}{\|H^n\mathbf{v}_0\|} = \frac{\left(Q\Lambda Q^\top\right)^n\mathbf{v}_0}{\left\|\left(Q\Lambda Q^\top\right)^n\mathbf{v}_0\right\|} = \frac{Q\Lambda^n Q^\top\mathbf{v}_0}{\|Q\Lambda^n Q^\top\mathbf{v}_0\|} = \frac{Q\left(\frac{1}{\lambda_{max}}\Lambda\right)^n Q^\top\mathbf{v}_0}{\left\|Q\left(\frac{1}{\lambda_{max}}\Lambda\right)^n Q^\top\mathbf{v}_0\right\|}$$

و در حالت حدی داریم:

$$\lim_{n\to\infty}\mathbf{v}_n = \frac{\mathbf{v}_{max}\mathbf{v}_{max}^\top\mathbf{v}_0}{\|\mathbf{v}_{max}\mathbf{v}_{max}^\top\mathbf{v}_0\|} = \frac{\mathbf{v}_{max}^\top\mathbf{v}_0}{\|\mathbf{v}_{max}\mathbf{v}_{max}^\top\mathbf{v}_0\|}\mathbf{v}_{max} = \overline{\mathbf{v}_{max}}$$

ت) از تعریف مشتق برای نقطه‌ای اطراف $\mathbf{r}$ داریم:

$$\begin{aligned}
H &\approx \frac{\nabla_{\mathbf{r}}M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})\big|_{\mathbf{r}=\xi\mathbf{v_0}} - \nabla_{\mathbf{r}}M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})\big|_{\mathbf{r}=0}}{\xi\mathbf{v_0}} \implies\\
H\,\mathbf{v}_0 &\approx \frac{\nabla_{\mathbf{r}}M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})\big|_{\mathbf{r}=\xi\mathbf{v_0}} - \nabla_{\mathbf{r}}M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})\big|_{\mathbf{r}=0}}{\xi} = \frac{\nabla_{\mathbf{r}}M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})\big|_{\mathbf{r}=\xi\mathbf{v_0}}}{\xi} \implies\\
\mathbf{r}^\star &\approx \varepsilon\overline{\nabla_{\mathbf{r}}M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})\big|_{\mathbf{r}=\xi\mathbf{v_0}}}
\end{aligned}$$

پاسخ سوال ۵ RL

آ) برای هر خانه، استراتژی بهینه این است که کم هزینه‌ترین مسیر تا یکی از خانه‌های پایانی را پیدا کند و سپس آنجا بماند (مادامی که سود این کار از سود ماندن که منفی است بیشتر باشد). بنابراین بدیهی‌ست که در خانه ۹ بهترین کنش، ماندن است. بنابراین:

$$10 + 0.5\times V^\star(9) = V^\star(9) \implies V^\star(9) = 20$$

همچنین برای خانه ۷ نیز بهترین کنش یک حرکت رو به بالا و رسیدن به خانه ۹ است یعنی:

$$V^\star(7) = -1 + 0.5\times V^\star(9) = 9$$

با شروع از خانه ۵ دو سیاست را می‌توانیم دنبال کنیم. یکی اینکه با یک *Jump* به خانه ۹ برسیم که سود این حالت برابر با $-2 + 0.5\times 20 = 8$ است؛ یا اینکه در خانه ۵ بمانیم که در این صورت $V(5) = 2 + 0.5\times V(5) \implies V(5) = 6$. پس *Jump* به خانه ۹ بیشترین سود را دارد. در نتیجه برای خانه ۵ داریم $V^\star(5) = 8$. برای خانه‌های ۶ و ۸ داریم:

$Q(8, right) = -1 + 0.5 \times V^\star(7) = 3.5$

$Q(8, down) = -1 + 0.5 \times V^\star(5) = 3$

به طور مشابه (ولی با جهت کنش‌های متفاوت) برای خانه ۶، مقدارهای بالا بدست می‌آیند. بنابراین برای هر دو خانه ۶ و ۸ داریم:

$$V^\star(6) = V^\star(8) = 3.5$$

برای سایر خانه‌ها نیز داریم:

$$\left.\begin{array}{l} Q^\star(4, Jump\,to\,7) = -2 + 0.5 \times V^\star(7) = 2.5 \\ Q^\star(4, right) = -1 + 0.5 \times V^\star(5) = 3 \end{array}\right\} \Longrightarrow V^\star(4) = 3$$

$$\left.\begin{array}{l} Q^\star(3, Jump\,to\,8) = -2 + 0.5 \times V^\star(8) = -0.25 \\ Q^\star(3, right) = -1 + 0.5 \times 3 = 0.5 \end{array}\right\} \Longrightarrow V^\star(3) = 0.5$$

$$\left.\begin{array}{l} Q^\star(2, Jump\,to\,5) = -2 + 0.5 \times V^\star(5) = 2 \\ Q^\star(2, down) = -1 + 0.5 \times V^\star(3) = -0.75 \end{array}\right\} \Longrightarrow V^\star(2) = 2$$

$$\left.\begin{array}{l} Q^\star(1, Jump\,to\,4) = -2 + 0.5 \times V^\star(4) = -0.5 \\ Q^\star(1, down) = -1 + 0.5 \times V^\star(2) = 0 \end{array}\right\} \Longrightarrow V^\star(1) = 0$$

در نتیجه سیاست بهینه در هر حالت به شرح زیر است:

$$\pi^\star(1) = down, \quad \pi^\star(2) = Jump\,to\,5, \quad \pi^\star(3) = right, \quad \pi^\star(4) = right, \quad \pi^\star(5) = Jump\,to\,9,$$
$$\pi^\star(6) = up, \quad \pi^\star(7) = up, \quad \pi^\star(8) = right, \quad \pi^\star(9) = stay$$

ب) در این حالت داریم:

$$V^\star(9) = 10 + \gamma V^\star(9) \Longrightarrow V^\star(9) = \frac{10}{1-\gamma}$$
$$V^1(5) = 3 + \gamma V^1(5) \Longrightarrow V^1(5) = \frac{3}{1-\gamma}$$
$$V^2(5) = -2 + \gamma V^\star(9) \Longrightarrow V^2(5) = -2 + \frac{10\gamma}{1-\gamma}$$

بنابراین:

$$V^{1}(5) > V^{2}(5) \implies \frac{3}{1-\gamma} > -2 + \frac{10\gamma}{1-\gamma} \implies \gamma < \frac{5}{12}$$

پیروز باشید