به نام خدا

پاسخ سوالات میانترم – مقدمه‌ای بر یادگیری ماشین

شریفی – آذر خلیلی

پاییز ۱۴۰۱

سوال ۱

الف) از کلاس A در کل ۱۵ نمونه و از کلاس B در کل ۱۰ نمونه وجود دارد. در نتیجه آنتروپی در گره ۰ برابر خواهد بود با:

$$H(0) = -\frac{15}{25}\log_2\frac{15}{25} - \frac{10}{25}\log_2\frac{10}{25} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

ب)

$$H(1) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

$$H(2) = 0$$

$$IG(0) = H(0) - \frac{15}{25}H(1) - \frac{10}{25}H(2) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} + \frac{1}{5}\log_2\frac{1}{3} + \frac{2}{5}\log_2\frac{2}{3}$$

ج) دقت یادگیری برابر است با $\frac{21}{25}$. (84 درصد)

د) با ادامه دادن گره‌های ۳ و۴ می‌توانیم دقت یادگیری را افزایش دهیم. (یا به عبارتی با افزایش عمق درخت)

ه) ممکن است با ریسک overfit شدن مدل مواجه شویم که برای جلوگیری از آن، می‌توانیم درخت عمیق درست کرده و سپس آن را هرس کنیم. چرا که هرس کردن از overfit شدن جلوگیری می‌کند و باعث می‌شود دقت مدل روی داده‌ی تست افزایش یابد.

سوال ۲: مفاهیم

الف) بر خلاف مجموعه داده training دو مجموعه داده test و validation هر دو برای ارزیابی به کار می روند با این تفاوت که مجموعه داده validation عموما برای انتخاب ابرپارامتر بهینه استفاده می شود در حالی که مجموعه داده test برای ارزیابی الگوریتم یادگیری ماشین استفاده شده، مورد استفاده قرار می گیرد.

ب) میانگین خطای مطلق (MAE) ، میانگین مطلق فاصله بین داده های واقعی و داده های پیش بینی شده را اندازه گیری می کند، اما خطاهای بزرگ را در پیش بینی مجازات نمی کند. میانگین مربع خطا (MSE)میانگین مجذور فاصله بین داده های واقعی و داده های پیش بینی شده را اندازه گیری می کند و

به عبارتی خطاهای بزرگ را مجازات می کند. (البته MSE در مورد خطاهای کوچک (بین ۰ و ۱) با سختگیری کمتری نسبت به MAE عمل می‌کند.)

MAE بهترین تخمین‌گر میانه و MSE بهترین تخمین‌گر میانگین است.

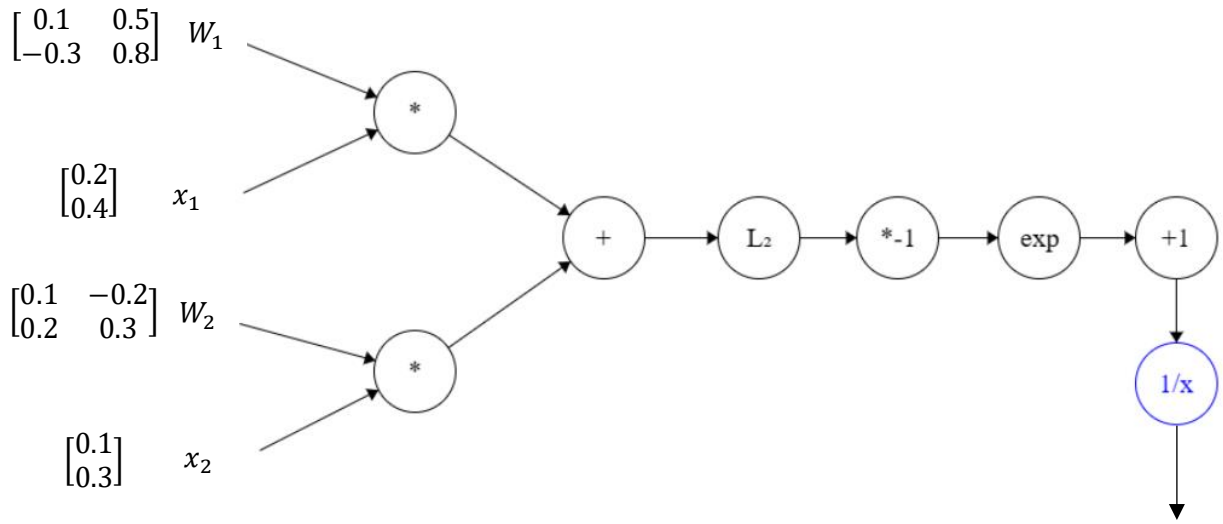
ج) زمانی که داده های آموزشی در یک مسئله یادگیری ماشین، نسبتاً کم باشد و یا اینکه نتیجه مربوط به داده های تست را خیلی دقیق ندانیم و بخواهیم چند بار آزمایش را تکرار کنیم و در نهایت میانگین این نتایج را برای ارزیابی نهایی در نظر بگیریم، از k-fold cross validation استفاده می‌کنیم. این روش نه تنها measurement بهتری از دقت مدل به ما می‌دهد، بلکه با دادن تعدادی عدد به ما کمک می‌کند میانگین و انحراف معیار حساب کرده تا در نهایت به کمک آن‌ها بتوانیم confidence interval بدست بیاوریم. در این روش داده ها را به K قسمت تقسیم می‌کنیم. سپس طی K مرحله مختلف هر بار یکی از K قسمت را به عنوان تست و K-1 قسمت دیگر را به عنوان داده آموزشی در نظر می‌گیریم که در نهایت میانگین نتایج ارزیابی به عنوان نتیجه نهایی ارزیابی در نظر گرفته می‌شود.

د) با توجه به خواصی که تابع ReLU دارد در اینجا از این تابع فعال ساز استفاده می شود زیرا که امکان تولید اعداد ۱ تا ۱۵ با وجود این تابع فعال‌ساز ممکن است. (تابع فعال‌سازی مانند sigmoid تنها اعداد بین ۰ تا ۱ را تولید می‌کند.) علاوه بر این تابع ReLU از نظر محاسباتی بسیار کارآمد است و به شبکه اجازه می‌دهد به سرعت همگرا شود و به شبکه خواص غیرخطی می‌دهد در عین اینکه محاسباتش همچنان خطی‌ست.

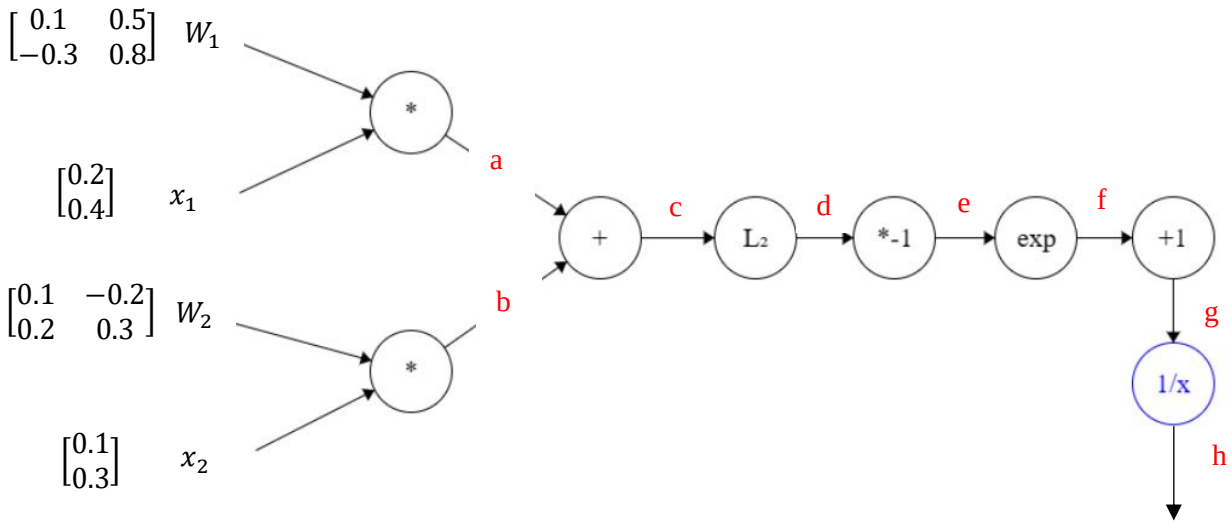
(هر جواب دیگری که منجر به تولید اعداد ۱ تا ۱۵ در خروجی شود مورد قبول است.)

سوال ۳

الف)



ب)



a = $W_1.x_1 = \begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}.\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix} = \begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$

b = $W_2.x_2 = \begin{bmatrix} 0.1 & -0.2 \\ 0.2 & 0.3 \end{bmatrix}.\begin{bmatrix} 0.1 \\ 0.3 \end{bmatrix} = \begin{bmatrix} -0.05 \\ 0.11 \end{bmatrix}$

c = $a + b = \begin{bmatrix} 0.17 \\ 0.37 \end{bmatrix}$

d = $\|c\|^2 = 0.17^2 + 0.37^2 = 0.166$

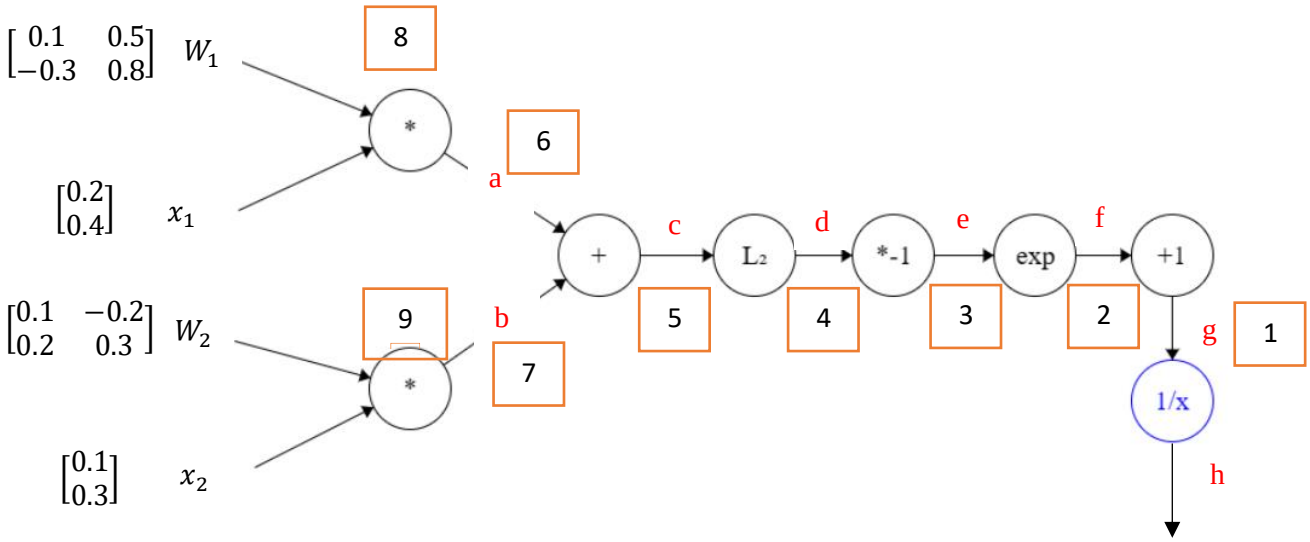
e = $-d = -0.166$

f = $e^{-0.166} = 0.85$

g = $f + 1 = 1.85$

h = $\frac{1}{g} = 0.54$

ج)



در شکل بالا مربع های کشیده شده مقدار مشتق در آن مکان را نشان میدهند. برای بدست آوردن مشتقات جزئی در هر یک از مربع ها ما باید با استفاده از قاعده زنجیری ابتدا در آن مرحله مشتق تابع مربوطه را با توجه به مقدار ورودی آن محاسبه کرده و سپس در مشتق مرحله قبلی بدست آوریم. فرضا برای مربع ۳، ما باید اول مشتق تابع نمایی را با توجه به ورودی e بدست آوریم و سپس آن را در مشتق مرحله ۲ ضرب کنیم.

در مربع ۱ : $F(g) = \frac{1}{g} \Longrightarrow \frac{dF}{dg} = -\frac{1}{g^2} = -\frac{1}{1.85^2} = -0.29$

در مربع ۲ : $g(f) = f + 1 \Longrightarrow \frac{dg}{df} = 1 \Longrightarrow \frac{dF}{df} = -0.29 \times 1 = -0.29$

در مربع ۳ : $f(e) = \exp(e) \Longrightarrow \frac{df}{de} = \exp(e) = 0.85 \Longrightarrow \frac{dF}{de} = -0.29 \times 0.85 = -0.25$

در مربع ۴ : $e(d) = -d \Longrightarrow \frac{de}{dd} = -1 \Longrightarrow \frac{dF}{dd} = -0.25 \times -1 = 0.25$

در مربع ۵ داریم:

$$d(c) = \|c\|^2 = {c_1}^2 + {c_2}^2 \Longrightarrow \frac{dd}{dc_i} = 2c_i \Longrightarrow \begin{cases} \frac{dd}{dc_1} = 2 \times 0.17 = 0.34 \\ \frac{dd}{dc_2} = 2 \times 0.34 = 0.68 \end{cases} \Longrightarrow$$

$$\begin{cases} \frac{dF}{dc_1} = 0.34 \times 0.25 = 0.085 \\ \frac{dF}{dc_2} = 0.68 \times 0.25 = 0.17 \end{cases} \Longrightarrow \frac{dF}{dc} = \begin{bmatrix} 0.085 \\ 0.17 \end{bmatrix}$$

در مربع ۶ داریم:

$$c(a,b) = a + b \Longrightarrow \frac{\partial c}{\partial a} = 1 \Longrightarrow \frac{dF}{da} = 1 \times \begin{bmatrix} 0.085 \\ 0.17 \end{bmatrix} = \begin{bmatrix} 0.085 \\ 0.17 \end{bmatrix}$$

در مربع ۷ داریم:

$$c(a,b) = a + b \Longrightarrow \frac{\partial c}{\partial b} = 1 \Longrightarrow \frac{dF}{db} = 1 \times \begin{bmatrix} 0.085 \\ 0.17 \end{bmatrix} = \begin{bmatrix} 0.085 \\ 0.17 \end{bmatrix}$$

در مربع ۸ داریم:

فرض میکنیم x_1 و W_1 را بصورت زیر داریم:

$$a(W_1, x_1) = W_1 . x_1 \Longrightarrow \frac{\partial a}{\partial W_1} = x_1^T = [0.2\ 0.4] \Longrightarrow \frac{dF}{dW_1} = [0.2\ 0.4] \times \begin{bmatrix} 0.085 \\ 0.17 \end{bmatrix}$$
$$= \begin{bmatrix} 0.017 & 0.034 \\ 0.034 & 0.068 \end{bmatrix}$$

در مربع ۹ داریم:

$$a(W_2, x_2) = W_2 . x_2 \Longrightarrow \frac{\partial a}{\partial W_2} = x_2^T = [0.1\ 0.3] \Longrightarrow \frac{dF}{dW_2} = [0.1\ 0.3] \times \begin{bmatrix} 0.085 \\ 0.17 \end{bmatrix}$$
$$= \begin{bmatrix} 0.0085 & 0.025 \\ 0.017 & 0.051 \end{bmatrix}$$

(محاسبه تا مربع‌های ۶ و ۷ کافی بوده و نمره‌ی کامل را دریافت می‌کند. محاسبه‌ی مربعات ۸ و ۹ می‌تواند تا ۵ نمره اضافه‌تر داشته باشد.)

سوال ۴

با توجه به اطلاعات موجود در سوال داریم:

	نتیجه آزمایش: تست مثبت (B)	نتیجه آزمایش: تست منفی (B^c)
نتیجه تست مثبت (A) در واقعیت	0.95	
نتیجه تست منفی (A^c) در واقعیت		0.96

همچنین داریم:

$$p(B|A) = 0.95, \quad p(B^c|A^c) = 0.96, \quad p(A) = 0.005, \quad p(A^c) = 0.995$$

لذا با استفاده از قاعده بیز می‌توان نوشت:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|A^c)p(A^c)}$$
$$= \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.04 \times 0.995} = 0.10$$

با توجه به اطلاعات داده شده کمیت های حساسیت و ویژگی برای این تست به ترتیب 0.95 و 0.96 است. کاربرد این معیارها در یادگیری ماشینن زمانی است که imbalance data داشته باشیم، زمانی که بخواهیم با نمودار ROC نتایج مقالات را با هم مقایسه کنیم یا استفاده در تست های غربالگری.

سوال ۵

ابتدا توجه به ماتریس واریانس-کوواریانس داده شده (A) و همچنین رابطه زیر میتوان مقادیر ویژه را به صورت زیر محاسبه کرد:

$$det(A - \lambda I) = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 3 - 4\lambda + \lambda^2 = 0$$

که $\lambda_1 = 3$ و $\lambda_2 = 1$ مقادیر ویژه و همچنین با استفاده از رابطه

$$Av = \lambda v$$

بردارهای ویژه نرمال شده مورد نظر نظیر به مقادیر ویژه به صورت زیر هستند

$$v_1 = \begin{bmatrix} 0.7 \\ 0.7 \end{bmatrix}, \qquad v_2 = \begin{bmatrix} -0.7 \\ 0.7 \end{bmatrix}$$

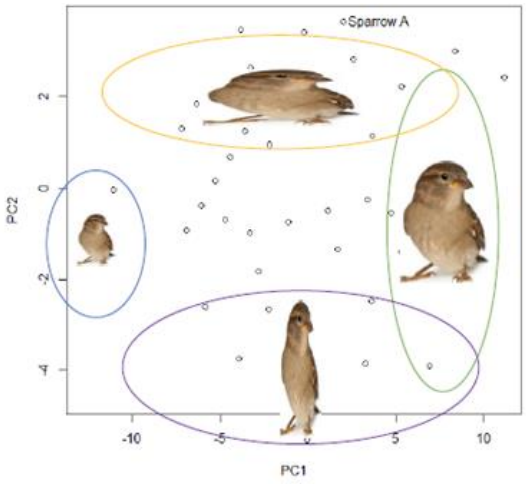
بنابراین مولفه های اصلی براساس متغیرهای طول (X_1) و وزن (X_2) گنجشک ها به صورت زیر تعریف می شوند

$$PC1 = 0.7\,X_1 + 0.7X_2; \qquad PC2 = -0.7\,X_1 + 0.7X_2$$

در ارتباط با میزان اطلاعاتی که هر متغییر اصلی شامل می شود، با توجه به مجموع مقادیر ویژه (۱+۳) ، به اولی میزان $\frac{۳}{۴}$ یا همان ۷۵ درصد و به دومی $\frac{۱}{۴}$ یا همان ۲۵ درصد اختصاص می یابد.

در این جا $PC1$ به صورت میانگین وزنی X_1 و X_2 است. بنابراین به نوعی معرف اندازه گنجشک ها است. پس مقدار $PC1$ بزرگ معرف گنجشک ها با اندازه بزرگ است (ناحیه 3 در شکل) و مقدار $PC1$ کوچک نشان دهنده گنجشک ها با اندازه کوچک است (ناحیه 1 در شکل). (شکل پس از نرمال شدن داده‌ها رسم شده است.)

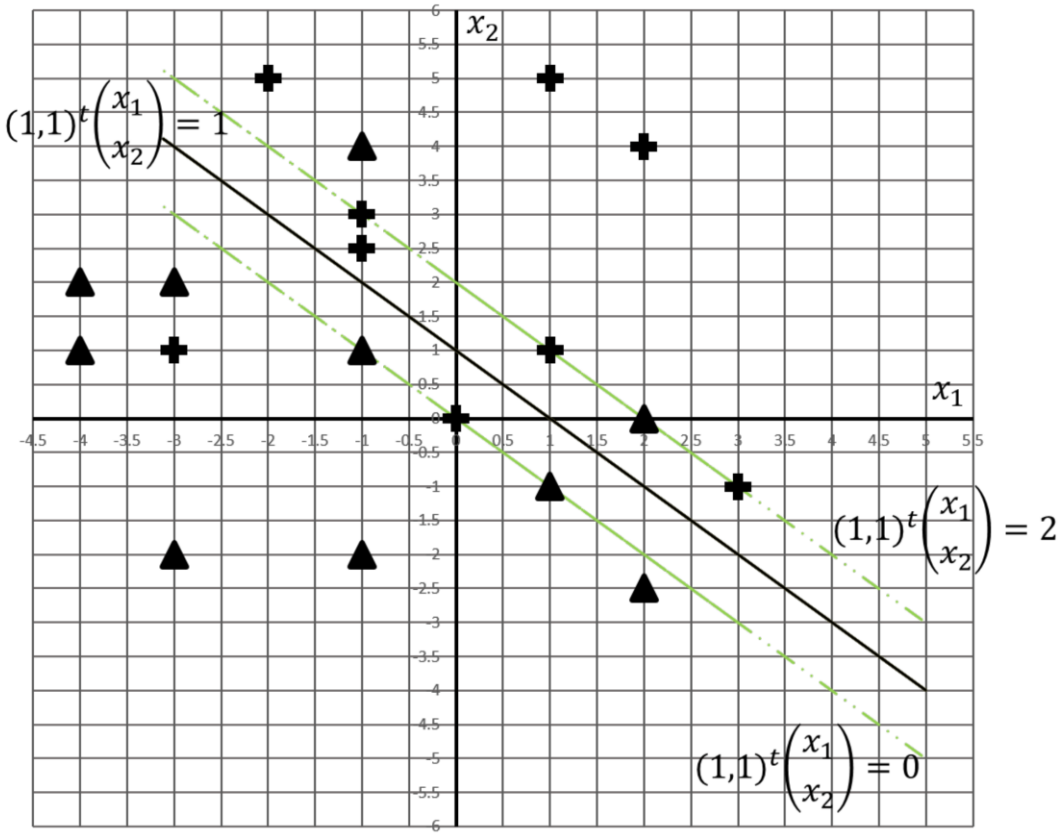
از طرفی در مورد $PC2$ ضریب یکی منفی و دیگری مثبت است پس مربوط به شکل گنجشک ها است که با توجه به ضرایب، مقادیر بزرگ $PC2$ یعنی گنجشک های چاق تر و کوهتاهتر (ناحیه 2 در شکل) و مقادیر کوچک $PC2$ یعنی گنجشک های لاغر و قد بلند (ناحیه 4 در شکل).



لازم به توضیح است با توجه به اینکه هدف بخش پایانی سوال، سنجش برداشت شما از PCهای بدست آمده بوده است، در نتیجه در صورتی که PCهای متفاوتی از مورد مطرح شده در پاسخنامه بدست آورده اید اما تفسیر درستی برای بخش پایانی از آن ذکر کردید، نمره این بخش از سوال برای شما منظور خواهد شد.

سوال ۶

الف)



ب)

$$\xi_{(2,4)} \lessdot \xi_{(2,0)} \qquad \xi_{(-1,1)} \eqq \xi_{(-1,-2)}$$
$$\xi_{(-3,1)} \gtrdot \xi_{(-1,2.5)} \qquad \xi_{(2,4)} \eqq \xi_{(-1,-2)}$$

ج)

نمونه‌ی $(2,4)$ مربوط به شرایط ۲،

نمونه‌ی $(1,1)$ مربوط به شرایط ۳ و

نمونه‌ی $(2,0)$ مربوط به شرایط ۱ است.