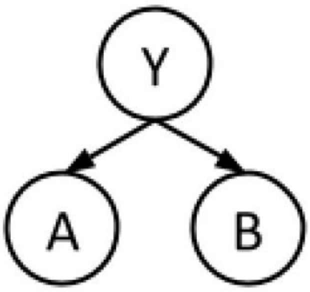
习题五：机器学习（共 75 分）

1、朴素贝叶斯（15 分）

在这个问题中，我们将训练一个朴素贝叶斯类来预测类标签 Y 作为输入特征的函数 A 和 B。Y、A 和 B 都是二进制变量，域为 0 和 1。我们有 10 条训练数据，用来估计我们的分布。我们的数据和模型如下图所示：

A	1	1	1	1	0	1	0	1	1	1
B	1	0	0	1	1	1	1	0	1	1
Y	1	1	0	0	0	1	1	0	0	0



(a) P（Y）、P（A|Y） 和 P（B|Y） 的最大似然估计是多少？（5 分）

Y	$P(Y)$
0	$3/5$
1	$2/5$

A	Y	$P(A\|Y)$
0	0	$1/6$
1	0	$5/6$
0	1	$1/4$
1	1	$3/4$

B	Y	$P(B\|Y)$
0	0	$1/3$
1	0	$2/3$
0	1	$1/4$
1	1	$3/4$

(b) 考虑一个新的数据点（A = 1，B = 1）。这个分类器将为此样本分配什么标签？（5 分）

$$\begin{aligned}
P(Y=0, A=1, B=1) &= P(Y=0)P(A=1|Y=0)P(B=1|Y=0) \tag{1}\\
&= (3/5)(5/6)(2/3) \tag{2}\\
&= 1/3 \tag{3}\\
P(Y=1, A=1, B=1) &= P(Y=1)P(A=1|Y=1)P(B=1|Y=1) \tag{4}\\
&= (2/5)(3/4)(3/4) \tag{5}\\
&= 9/40 \tag{6}\\
& \tag{7}
\end{aligned}$$

Our classifier will predict label 0.

(c) 让我们使用拉普拉斯平滑来平滑我们的分布。假如使用 k = 2 的拉普拉斯平滑，计算 P（A|Y） 的新分布。（5 分）

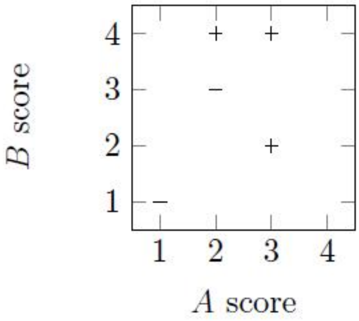
A	Y	$P(A\|Y)$
0	0	$3/10$
1	0	$7/10$
0	1	$3/8$
1	1	$5/8$

2、感知器（20 分）

您想根据电影的剧本预测电影是否会赢利。您聘请了两个评论家 A 和 B 来阅读您拥有的剧本，并以 1 到 4 的等级对其进行评分。批评者并不完美；以下是五个数据点，包括影评人的评分和电影的表现：

#	Movie Name	A	B	Profit?
1	Pellet Power	1	1	-
2	Ghosts!	3	2	+
3	Pac is Bac	2	4	+
4	Not a Pizza	3	4	+
5	Endless Maze	2	3	-

(a) 首先，您要检查数据的线性可分离性。在下面的 2D 平面上绘制数据；用 + 标记赢利的电影，用—标记不赢利的电影，并确定数据是否线性可分离。（5 分）



The data are linearly separable.

(b) 现在，您决定使用感知器对数据进行分类。假设您直接使用上面给出的分数作为特征，并使用一个偏置特征（bias）。即 $f_0=1$，$f_1=A$ 给出的分数，$f_2=B$ 给出的分数。使用感知器算法对数据进行一次遍历，将结果填入下表。按数据点的顺序，例如在步骤 1 中使用数据点 1，以此类推。（5 分）

step	Weights	Score	Correct?
1	[-1, 0, 0]	$-1\cdot1+0\cdot1+0\cdot1=-1$	yes
2	[-1, 0, 0]	$-1\cdot1+0\cdot3+0\cdot2=-1$	no
3	[0, 3, 2]	$0\cdot1+3\cdot2+2\cdot4=14$	yes
4	[0, 3, 2]	$0\cdot1+3\cdot3+2\cdot4=17$	yes
5	[0, 3, 2]	$0\cdot1+3\cdot2+2\cdot3=12$	no

最终的权重（Weights）： [-1, 1, -1]

(c) 你的算法是否学会了分离数据的权重（weights）？（4 分）

Have weights been learned that separate the data? With the current weights, points will be classified as positive if $-1\cdot1+1\cdot A+-1\cdot B\geq0$, or $A-B\geq1$. So we will have incorrect predictions for data points 3:

$$-1\cdot1+1\cdot2+-1\cdot4=-3<0$$

and 4:

$$-1\cdot1+1\cdot3+-1\cdot4=-2<0$$

Note that although point 2 has $w\cdot f=0$, it will be classified as positive (since we classify as positive if $w\cdot f\geq0$).

(d) 更一般地说，无论训练数据如何，您都想知道您的特征是否足够强大，能够允许你处理各种情形。圈出以下场景中，使用以上特征的感知器确实可以对能否赢利的电影作出分类的例子：（6 分）

1) 你的评论者很棒：如果他们的总分超过 8 分，那么电影会赢利，否则就不会。

   Can classify (consider weights [-8, 1, 1])

2) 您的评论家是艺术评论家：电影将赢利，当且仅当每个评论者给 2 或者 3 分。

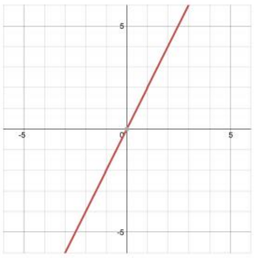
   Cannot classify

3) 你的审稿人品味奇特且迥异。您的电影将赢利当且仅当两者的评分相同。

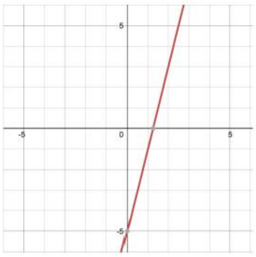
   Cannot classify

3、神经网络（40 分）

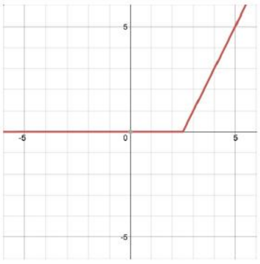
考虑单变量 x的多个函数 （a-h），如下图所示。后续页面上的神经元图开始简单，但会变得越来越复杂，逐渐建立起完整的神经网络。对每种图，指出它们能够表示以下的哪些函数。



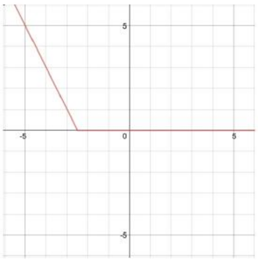
(a) $2x$



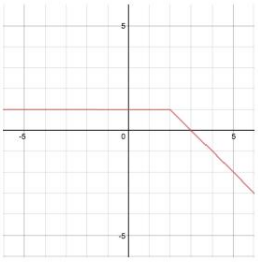
(b) $4x-5$



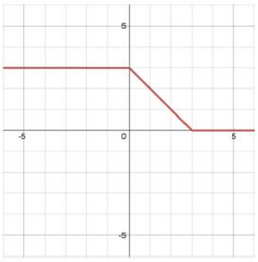
(c) $\begin{cases} 2x-5 & x \geq 2.5 \\ 0 & x < 2.5 \end{cases}$



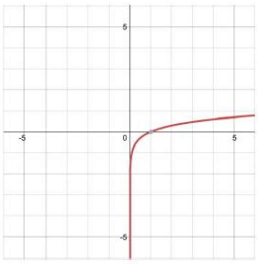
(d) $\begin{cases} -2x-5 & x \leq -2.5 \\ 0 & x > -2.5 \end{cases}$



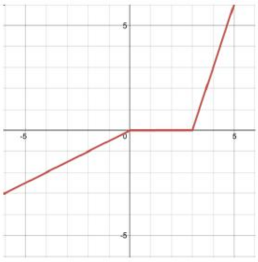
(e) $\begin{cases} -x+3 & x \geq 2 \\ 1 & x < 2 \end{cases}$



(f) $\begin{cases} 3 & x \leq 0 \\ 3-x & 0 < x \leq 3 \\ 0 & x > 3 \end{cases}$

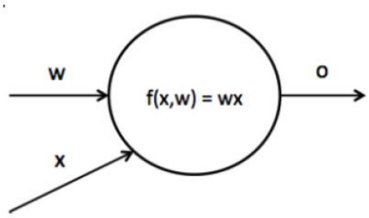


(g) $\log(x)$



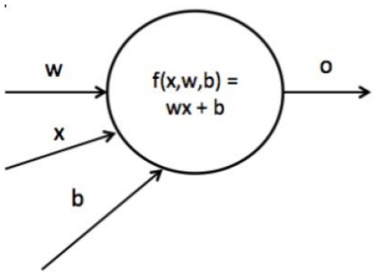
(h) $\begin{cases} 0.5x & x \leq 0 \\ 0 & 0 < x \leq 3 \\ 3x-9 & x > 3 \end{cases}$

(a) 考虑下图，计算一个线性变换，采用标量输入x 、权重w，输出 o，使得 $o = wx$。这个变换可以表示上图（a-h）种的哪些函数？对于可以表达的函数，写出适当值的w值。（5 分）



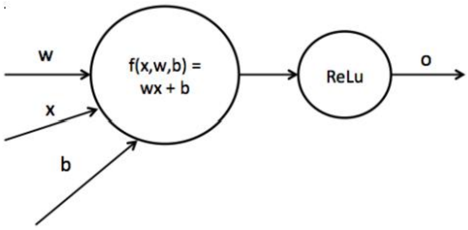
This graph can only represent (a), with $w = 2$. Since there is no bias term, the line must pass through the origin.

(b) 现在我们在图中引入一个偏置项 b，使得 $o = wx + b$（这被称为仿射函数）。这个变换可以代表上图（a-h）中的哪些函数？对于可以表达的函数，请给出适当的w和b值。（5 分）



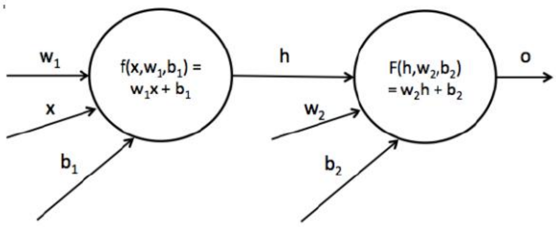
(a) with $w = 2$ and $b = 0$, and (b) with $w = 4$ and $b = -5$

(c) 我们可以引入非线性，如下所示。我们使用 ReLU 函数，$ReLU(x) = max(0; x)$。现在上图（a-h）中的哪些函数可以由这个神经网络来表示？对于可以的，给出适当的w, b值。（5 分）



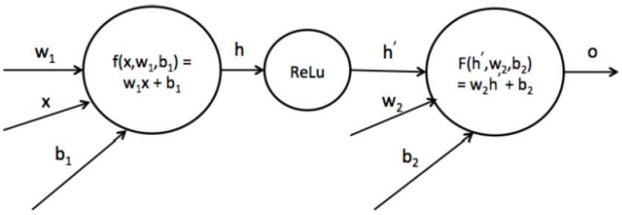
With the output coming directly from the ReLU, this cannot produce any values less than zero. It can produce (c) with $w = 2$ and $b = -5$, and (d) with $w = -2$ and $b = -5$

(d) 现在我们考虑具有多个仿射变换的神经网络，如下所示。我们现在有两组权重和偏差 w_1, b_1 和w_2, b_2。我们有第一层变换的结果 $h = w_1x + b_1$，和最终结果$o = w_2h + b_2$。这个网络可以表达（a-h）中的哪些函数？对于可以表达的函数，请写出适当的w_1, b_1 和w_2, b_2值。（5 分）



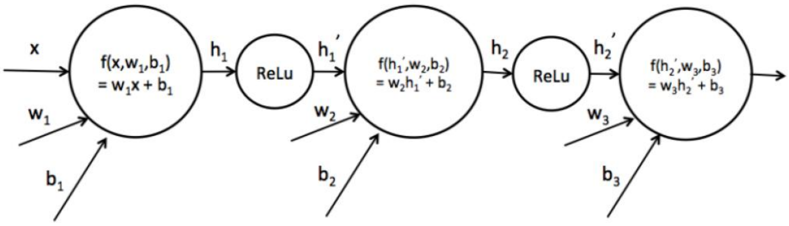
Applying multiple affine transformations (with no non-linearity in between) is not any more powerful than a single affine function: $w_2(w_1x + b_1) + b_2 = w_2w_1x + w_2b_1 + b_2$, so this is just a affine function with different coefficients. The functions we can represent are the same as in 1, if we choose $w_1 = w, w_2 = 0, b_1 = 0, b_2 = b$: (a) with $w_1 = 2, w_2 = 1, b_1 = 0, b_2 = 0$, and (b) with $w_1 = 4, w_2 = 1, b_1 = 0, b_2 = -5$.

(e) 接下来，我们在第一个仿射变换后向网络添加一个 ReLU 非线性，创建一个隐藏的层。这个网络可以代表哪些函数？对于可以表达的函数，请写出适当的w_1, b_1 和w_2, b_2值。（5 分）



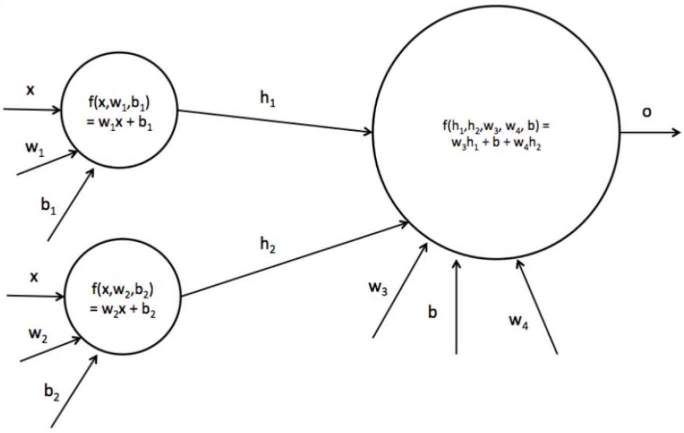
(c), (d), and (e). The affine transformation after the ReLU is capable of stretching (or flipping) and shifting the ReLU output in the vertical dimension. The parameters to produce these are: (c) with $w_1 = 2, b_1 = -5, w_2 = 1, b_2 = 0$, (d) with $w_1 = -2, b_1 = -5, w_2 = 1, b_2 = 0$, and (e) with $w_1 = 1, b_1 = -2, w_2 = -1, b_2 = 1$

(f) 现在我们向网络添加另一个隐藏层，如下所示。哪些函数可以是以这个网络来表达？（5 分）



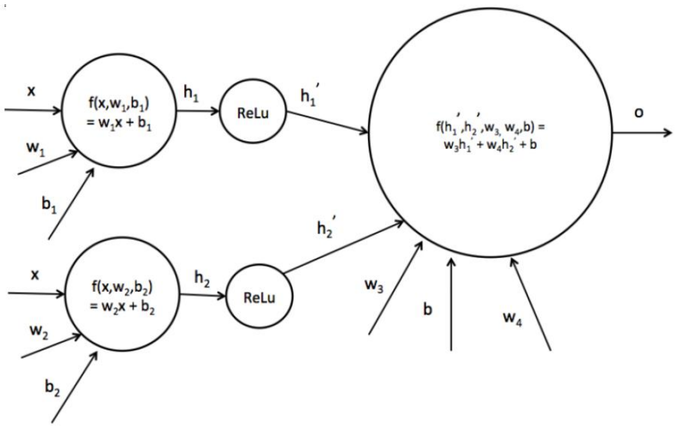
(c), (d), (e), and (f). The network can represent all the same functions as Q5 (because note that we could have $w_2 = 1$ and $b_2 = 0$). In addition it can represent (f): the first ReLU can produce the first flat segment, the affine transformation can flip and shift the resulting curve, and then the second ReLU can produce the second flat segment (with the final affine layer not doing anything). Note that (h) cannot be produced since its line has only one flat segment (and the affine layers can only scale, shift, and flip the graph in the vertical dimension; they can't rotate the graph).

(g) 我们想考虑使用只有一个隐藏层的神经网络，但让它更大——隐藏层的尺寸为 2。让我们考虑只使用两个仿射函数，两者之间没有非线性。这个网络可以代表那些函数？（5 分）



(a) and (b). With no non-linearity, this reduces to a single affine function (in the same way as Q4)

(h) 现在我们在两个仿射层之间添加一个非线性，产生下面的神经网络，其中包含尺寸为 2 的隐藏层。这个网络可以代表哪些函数？（5 分）



All functions except for (g). Note that we can recreate any network from (5) by setting w_4 to 0, so this allows us to produce (c), (d) and (e). To produce the rest of the functions, note that h'_1 and h'_2 will be two independent functions with a flat part lying on the x-axis, and a portion with positive slope. The final layer takes a weighted sum of these two functions. To produce (a) and (b), the flat portion of one ReLU should start at the point where the other ends ($x = 0$ for (a), or $x = 1$ for (b). The final layer then vertically flips the ReLU sloping down and adds it to the one sloping up, producing a single sloped line. To produce (h), the ReLU sloping down should have its flat portion end (at $x = 0$ before the other's flat portion begins (at $x = 3$). The down-sloping one is again flipped and added to the up-sloping. To produce (f), both ReLUs should have equal slope, which will cancel to produce the first flat portion above the x-axis.