Solutions to the Review Questions

1. Convert the following number from base 10 to base 16: 102.34

   SOLUTION: $102_{10} = 66_{16}$, since $102 = 6 \cdot 16^0 + 6 \cdot 16^1$. For the fractional part:

	Fractional part	Integer part
0.34×16	0.44	5
0.44×16	0.04	7
0.04×16	0.64	0
0.64×16	0.24	10
0.24×16	0.84	3
0.84×16	0.44	13
0.44×16	0.04	7

   And it starts to repeat, so all together we get: $66.5\ \overline{70A3D}$

2. Convert the following number from base 10 to base 5: 102.34.

   SOLUTION: The integer part is 402_5. Put it with the fractional part to get:

$$402.132222...$$

3. How big is machine ϵ, and how is it defined (for doubles)?

   SOLUTION: Machine ϵ is 2^{-52}, and is defined to be the distance between the number 1 and the next biggest number (we always use double precision, meaning we have 52 digits in the mantissa).

4. How would you round the following number using the IEEE rounding rule?

$$1.001001001\cdots \times 2^4$$

   SOLUTION: The following show where bits 51-54 are (space between 52 and 53):

$$1.001001\cdots 10\ \ 01001\cdots$$

   Therefore, bit 53 is zero and we truncate after bit 52.

5. Convert the following base 10 number to binary and express it as a floating point number using the IEEE rounding rule: 44/7

   SOLUTION: Probably best to write as a mixed fraction, $6\frac{2}{7}$. The conversion of 6 is 110, since $6 = 2^2 + 2^1$. Convert 2/7 as we showed in class: $0.01001001\cdots$. Therefore, put the two parts together. Additionally, we determined where bits 52 and 53 would be, and a space appears between them:

$$110.01001001\cdots = 1.1001001001\cdots \times 2^2 = 1.100100100\cdots 01\ \ 001001\cdots \times 2^2$$

   Since bit 53 is 0, we truncate again:

$$1.100100100\cdots 01 \times 2^2$$

6. Do the following sum by hand using IEEE rounding rule (and double precision):

   (a) $(1 + (2^{-51} + 2^{-52} + 2^{-59})) - 1$

   SOLUTION: Adding the three small numbers first, we get the following (the last number has 7 zeros)

   $$2^{-51}\,(1 + 0.1 + 0.00000001) = 1.10000001 \times 2^{-51}$$

   Now, by adding 1, bits 51, 52 and 59 are all ones, the rest zero. In particular, bit 53 is zero (again!), and so we truncate to get:

   $$1.00\cdots0011 \times 2^0$$

   Subtract 1 to get $2^{-51} + 2^{-52} = 2^{-52}(2 + 1) = 3 \times 2^{-52} \approx 6.66 \times 10^{-16}$, which we can verify in Matlab (otherwise, no need to do base 10).

7. What is a mathematical model?

   SOLUTION: A mathematical model is a set of mathematical statements that are designed to approximate some physical phenomena.

8. Was the $n-$armed bandit problem an example of *supervised* or *unsupervised* learning?

   SOLUTION: *Supervised* learning is learning by example- In the case of the $n-$armed bandit, we had no examples from which to learn. Therefore, this was unsupervised learning.

9. What is the greedy algorithm? How was it modified to get the $\epsilon-$greedy algorithm? In particular, how did our average reward depend on ϵ in the test trials (or in the figure that you produced)?

   SOLUTION: The greedy algorithm always plays the machine with the largest current payout. It was modified so that every once in a while (with probability ϵ), we choose a machine completely at random, without regard to the current payouts. We saw in the example that using $\epsilon = 0$ (the greedy algorithm) produced a lower average payout than the other values of ϵ used. In fact, we saw that as ϵ increased, our average payout also increased (although presumably that would not go ad infinitum).

10. What is the "softmax" action selection? In particular, how did we change a set of payouts Q_i to a set of probabilities, P_i?

    SOLUTION: In the softmax selection, we first convert the payouts to probabilities, then we play each machine according to its probability. Given n machines with current payout $Q_1, \ldots, Q_n$, and temperature τ, then (in Matlab notation)

    ```
    P=exp(Q/tau)./sum(exp(Q/tau));
    ```

    In mathematics notation,

    $$P_i = \frac{\mathrm{e}^{Q_i/\tau}}{\sum_{k=1}^{n} \mathrm{e}^{Q_k/\tau}}$$

11. Suppose $Q = [-0.5, 0, 0.5, 1.0]$. Use the softmax selection technique with $\tau = 0.1$ to compute the probabilities.

    SOLUTION: Using your calculator, you should find them to be (in order, truncated after two places):

    $$0.00, 0.00, 0.01, 0.99$$

12. If $Q_1 < Q_2 < Q_3 < Q_4$ for 4 machines, how do the probabilities change (under softmax) as $\tau \to 0$? As $\tau \to 1$?

    SOLUTION: We want to compute the limits, as we did in Homework 4.

13. What is the win-stay, lose-shift (or pursuit) strategy? What are the update rules?

    The win-stay, lose-shift strategy means just that- If we're winning, make the probability of using the machine we selected even more. If we're losing, make the probability of using that machine less.

    For the winning probability:

    $$P^{\text{new}} = P^{\text{old}} + \beta(1 - P^{\text{old}})$$

    For the losing probabilities (index not shown, but do this for all the other probabilities):

    $$P^{\text{new}} = P^{\text{old}} + \beta(0 - P^{\text{old}})$$

    where β is some fixed value between 0 and 1.

14. Suppose we play with three machines, and machine 3 is chosen and gives a big payout (enough to make $Q_t(3)$ the maximum). Update the probabilities for win-stay, lose-shift, if they are: $P_1 = 0.3, P_2 = 0.5, P_3 = 0.2$ and $\beta = 0.3$.

    SOLUTION: The new probabilities are:

    $$P_1 = 0.21 \qquad P_2 = 0.35 \qquad P_3 = 0.44$$

    Note that they do form a new set in that they are still all between 0 and 1, and they still sum to 1.

15. In the sample script on page 13, how are the initial probabilities set? (NOTE: I may give you a code "snippet" to interpret, like this question.).

    SOLUTION: The appropriate line is:

    ```
    Probs=(1/NumMachines)*ones(10,1);
    ```

    which sets all probabilities equal.

16. Find bases for the column, row and null space for the matrix A below. Also, find the null space of A^T.

    $$A = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}$$

    SOLUTION: The null space is the solution to $A\mathbf{x} = \mathbf{0}$, which is spanned by

    $[-2, 1, 0, 0]^T, [-3, 0, 1, 0]^T$, and $[-4, 0, 0, 1]$

(these are vectors in $\mathbb{R}^4$).

The column space is a subspace of $\mathbb{R}^1$, which is spanned by 1, and the row space is spanned by the row, $[1, 2, 3, 4]^T$. The null space of A^T is a subspace of $\mathbb{R}^1$ of dimension 0 (since the column space has dimension 1), so the null space of A^T contains only the zero "vector".

17. Show that, for the line of best fit, the normal equations produce the same equations as minimizing an appropriate error function:

    SOLUTION: Given data pairs $(x_1, y_1), \ldots, (x_n, y_n)$, the system of equations is given by:

$$\begin{array}{rl} y_1 &= mx_1 + b \\ y_2 &= mx_2 + b \\ \vdots & \vdots \\ y_n &= mx_n + b \end{array} \Rightarrow \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \Rightarrow A\mathbf{x} = \mathbf{b}$$

    Now we show that finding the least squares solution using linear algebra is the same as minimizing a certain error function:

    - Using linear algebra, we compute the normal equations $A^T A\mathbf{x} = A^T\mathbf{b}$, which results in the following system of equations:

$$\begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix}$$

    - Using Calculus, the error function we consider is the following (NOTE: The constant in the front of the sum is really irrelevant, but using the given value has a nice interpretation and cancels the 2's in the derivatives).

$$E(m, b) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - mx_i - b)^2$$

$$E_m = \frac{1}{n} \sum_{i=1}^{n} (y_i - mx_i - b)(-x_i) \qquad E_b = \frac{1}{n} \sum_{i=1}^{n} (y_i - mx_i - b)(-1)$$

    Setting the derivatives to zero and simplifying a bit, we see that we get the same system of equations.

18. Given data:

x	-1	0	1
y	2	1	1

    (a) Give the matrix equation for the *line of best fit.*
    (See the equation above; just substitute in the numerical values)

    (b) Compute the normal equations.

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \mathbf{x} = \begin{bmatrix} -1 \\ 4 \end{bmatrix}$$

(c) Solve the normal equations for the slope and intercept: $m = -1/2$ and $b = 4/3$.

19. Use the data in Exercise (18) to find the parabola of best fit: $y = ax^2 + bx + c$. (NOTE: Will you only get a least squares solution, or an actual solution to the appropriate matrix equation?)

    SOLUTION: In this case, the matrix equation is given by:

$$\begin{bmatrix} 1 & -1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

    The determinant of A is 2, so we get one unique solution. Solve by row reduction (about 4 row ops) to get: $a = 1/2, b = -1/2, c = 1$.

20. Use the data in Exercise (18) to find the equation of the median-median line.

    The median of one point is itself. First, find the equation for the line that runs through the medians of groups 1 and 3 (in this case, the first and third data point, $(-1, 2), (1, 1)$). The slope is $-1/2$ and using $(1, 1)$, we get:

$$y = \frac{3}{2} - \frac{1}{2}x$$

    The line is 1/2 unit higher than the middle data point- So we shift the line 1/3 of that distance down, or subtract 1/6 to get

$$y = \frac{4}{3} - \frac{1}{2}x$$

21. What is Hebb's rule (the biological version- you can paraphrase)?

    When cell A is near enough to cell B to repeatedly help in firing B, some change takes place so that A's efficiency in firing cell B is increased.

22. What is the Widrow-Hoff learning rule? How is it related to Hebb's rule?

    The original Hebb's rule did not use any targets:

$$W^{\text{new}} = W^{\text{old}} + \alpha \mathbf{y}\mathbf{x}^T$$

    where $\mathbf{y} = W\mathbf{x}$.

    Widrow-Hoff uses the targets $\mathbf{t}$ into account, and we update the weights on the k^{th} data point as:

$$W^{\text{new}} = W^{\text{old}} + \alpha(\mathbf{t}_k - \mathbf{y}_k)\mathbf{x}_k^T$$

    where $\mathbf{y}_k = W\mathbf{x}_k$. If we separated the bias terms $\mathbf{b}$ out, then $\mathbf{y}_k = W\mathbf{x}_k + \mathbf{b}$, and

$$\mathbf{b}^{\text{new}} = \mathbf{b}^{\text{old}} + \alpha(\mathbf{t}_k - \mathbf{y}_k)$$

23. Let $W = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. If $\mathbf{x} = [-1, 0, 1]^T$ and $\mathbf{t} = [2, 3]^T$, use Widrow-Hoff to update $W, \mathbf{b}$ one time using a learning rate of 1 (This is too big of a learning rate to actually use, but it will make your computations easier).

    SOLUTION: Compute $\mathbf{y} = W\mathbf{x} + \mathbf{b} = [1, 1]^T$, so

$$(\mathbf{t}_k - \mathbf{y}_k)\mathbf{x}^T = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} - & 0 & 1 \\ -2 & 0 & 2 \end{bmatrix}$$

    So the new $W, \mathbf{b}$ are given by:

$$W = \begin{bmatrix} 0 & 0 & 1 \\ -3 & 1 & 2 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

24. Let $\mathbf{x} = [1, 2]^T$. Find the matrix $\mathbf{x}\mathbf{x}^T$, its eigenvalues, and eigenvectors.

    SOLUTION:

$$\mathbf{x}\mathbf{x}^T = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

    You should find that $\lambda = \|\mathbf{x}\|^2 = 5$ and 0, with corresponding eigenvectors $\mathbf{x}$ and the other is something perpendicular to $\mathbf{x}$, like $[-2, 1]^T$.

25. Show that, if $A = \mathbf{x}\mathbf{x}^T$ for a non-zero vector $\mathbf{x}$, then A has one eigenvalue that is $\|\mathbf{x}\|^2$ and the eigenvector is $\mathbf{x}$. Show that all other eigenvalues are zero by finding the null space of A (think about it in terms of words).

    SOLUTION:

$$A\mathbf{x} = \mathbf{x}\mathbf{x}^T\mathbf{x} = \left(\|\mathbf{x}\|^2\right)\mathbf{x}$$

    which shows that $\lambda_1 = \|\mathbf{x}\|^2$ and $\mathbf{v}_1 = \mathbf{x}$. We know that $\lambda = 0$ is the other eigenvalue, whose eigenspace is the null space of $\mathbf{x}\mathbf{x}^T$. This is the space of all vectors orthogonal to $\mathbf{x}$, which (if A is $n \times n$) would be $n - 1$ dimensional, and that gives all possible eigenvalues.

26. Show that the affine mapping: $f(\mathbf{x}) = W\mathbf{x} + \mathbf{b}$ can be written as a linear mapping $\hat{W}\hat{\mathbf{x}}$ for an appropriate $\hat{W}$ and $\hat{\mathbf{x}}$

    SOLUTION: Take $\hat{W} = [W|\mathbf{b}]$ and we append a row of 1's to the bottom of the array (or vector) X.

27. What does "training" mean in terms of our mathematical model?

    "Training" means that we are attempting to find weights and biases that will fit our model the "best-" if we have an error function, that means that we are minimizing it.

28. If we use all the data we have at once, what kind of training are we doing? If we learn one data point at a time, what kind of training are we doing?

    SOLUTION: Using all the data at once is "batch" training. Updating the weights and biases one point at a time is "online" training.

29. Suppose I have some data in $\mathbb{R}^3$ that belongs to 4 different classes. Do I want my targets to be the real numbers $1, 2, 3, 4$, or are there better ways to build the target values? (Hint: Using $1, 2, 3, 4$ implies that class 1 and class 4 are very far apart- more so than 3 and 4. But this is probably not reflected in the data- Try targets that are equally spaced).

    SOLUTION: Better to make the classes into the four standard basis vectors of $\mathbb{R}^4$- That is, e_1 for Class 1, e_2 for Class 2, etc.

30. Given the function $z = f(x, y)$, show that the direction in which f decreases the fastest from a point (a, b) is given by the negative gradient (evaluated at (a, b)).

    SOLUTION: Consider the directional derivative at the point (a, b) in the direction of the unit vector $\mathbf{u}$:

$$D_u f(a, b) = \nabla f(a, b) \cdot \mathbf{u} = \|\nabla f\| \|\mathbf{u}\| \cos(\theta)$$

    Therefore, the right side of the equation is a maximum when $\cos(\theta) = 1$, or $\theta = 0$, and it is a minimum when $\cos(\theta) = -1$, or $\theta = \pi$. Therefore, we increase the most if we move in the direction of the gradient, and decrease the most by moving in the opposite direction (the negative of the gradient).

31. Illustrate the technique of gradient descent using

$$f(x, y) = x^2 + y^2 - 3xy + 2$$

    (a) Find the minimum.

    SOLUTION: Candidates for the minimum are where $\nabla f = 0$. If we set the gradient to zero, we see that the only candidate is at the origin, $x = 0, y = 0$. Since the x^2 and y^2 terms go to infinity as x, y get large, we expect the origin to be the minimizer.

    (b) Use the initial point $(1, 0)$ and $\alpha = 0.1$ to perform two steps of gradient descent (use your calculator).

    SOLUTION: Here is a table of values

$\mathbf{x}$	$f(\mathbf{x})$	$\nabla f(\mathbf{x})$	$\mathbf{x}^{\text{new}}$
$(1, 0)$	3	$[2, -3]^T$	$[0.8, 0.3]$
$(0.8, 0.3)$	2.01	$[0.7, -1.8]^T$	$[0.73, 0.48]^T$

32. If

$$f(t) = \begin{bmatrix} 3t - 1 \\ t^2 \end{bmatrix}$$

    find the tangent line to f at $t = 1$.

    SOLUTION: $f(1) = [2, 1]^T$ and $f'(1) = [3, 2]^T$ so the equation of the line (in parametric form) is:

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} + t \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

33. If $f(x, y) = x^2 + y^2 - 3xy + 2$, find the linearization of f at $(1, 0)$.

SOLUTION: We can use the information from the table we already constructed

$$L(x, y) = 3 + [2, -3] \begin{bmatrix} x - 1 \\ y - 0 \end{bmatrix} = 3 + 2(x - 1) - 3y$$

34. Given just one data point:

$$X = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \qquad T = [1]$$

Initializing W and $\mathbf{b}$ as an appropriately sized arrays of ones, perform three iterations of Widrow-Hoff using $\alpha = 0.1$ (by hand, you may use a calculator). You should verify that the the weights and biases are getting better.

Initially, $W = [1, 1]$ and $\mathbf{b} = 1$. Let's do the updates:

- $T = 1$, $y = W\mathbf{x} + \mathbf{b} = 2 - 1 + 1 = 2$, to $(t - y) = -1$ and the update is:

$$W = [1, 1] + (0.1)(-1)[2, -1] = [0.8, 1.1]$$

$$b = 1 + (0.1)(-1) = 0.9$$

- Now, $y = W\mathbf{x} + \mathbf{b} = 1.4$ (which is closer to the target of 1 than it was previously), and

$$W = [0.8, 1.1] + (0.1)(-0.4)[2, -1] = [0.72, 1.14]$$

$$b = 0.9 + (0.1)(-0.4) = 0.86$$

- This time, $y = W\mathbf{x} + \mathbf{b} = 1.16$ (which is even closer to the target of 1 than it was previously), and

$$W^{\text{new}} = [0.688 \quad 1.156] \qquad \mathbf{b}^{\text{new}} = 0.844$$

(and the new y after that is 1.064- Even closer to the target!

35. If a time series is given by:

$$x = \{1, 2, 0, 3, 4, 5, 2, 1, 0, 3, 4\}$$

Give the result of performing lag 2, and specify the domain-range pairing we would use in the novelty detection algorithm (same lag).

SOLUTION:

$$\begin{aligned}
(1, 2) &\to 0 \\
(2, 0) &\to 3 \\
(0, 3) &\to 4 \\
(3, 4) &\to 5 \\
(4, 5) &\to 2 \\
(5, 2) &\to 1 \\
(2, 1) &\to 0 \\
(1, 0) &\to 3 \\
(0, 3) &\to 4
\end{aligned}$$

36. Matlab Questions:

(a) What's the difference between a script file and a function?

A script file is a plain text file with Matlab commands. Matlab runs it as if the commands were being typed "live".

A function has certain inputs and certain outputs defined, as well as a rule.

(b) What does the following code fragment produce?

```
Q=[1 3 2 1 3];
idx=find(Q==max(Q));
```

SOLUTION: `find([0,1,0,0,1])` returns the indices $2, 5$.

(c) What will P be:

```
x=[0.3, 0.1, 0.2, 0.4];
P=cumsum(x);
```

P is the cumulative sum: $P = [0.3, 0.4, 0.6, 1.0]$