Part 1

1. The input value to the bias weight does not matter, as long as it is not zero!

2. The probability that classes are linearly separable increases when the features are nonlin-
early mapped to a higher dimensional feature space.

3. A scalar product in a, usually, high-dimensional feature space.

4. The direction of maximum variance in the subspace orthogonal to the first principal com-
ponent.

5. kNN

6. ICA can not solve the problem since the matrix

\[
\begin{pmatrix}
1 & 2 \\
2 & 4
\end{pmatrix}
\]

gives two mixtures that only differ by a scaling factor (2). It is also easy to see that the
mixture matrix is singular and therefore doesn’t have an inverse.

7. The correlation between projections of data is maximized.

8. The Q-function describes the value for each action in each state. The V-function only
describes the value of a state, assuming that the optimal action is taken in each state.

9. Problems where the cost function have many local optima and where it is sufficient to find
a good enough solution (not necessarily the global optimum). The solutions to the problem
should also be relatively easy/cheap to evaluate.

10. Dendrites
11. Suppose the data are stored as columns in a matrix $X$. The ordinary eigenvalue problem for PCA is then

$$XX^T e = \lambda e$$

Multiplying with the data matrix gives

$$X^T XX^T e = \lambda X^T e$$

which, by letting $f = X^T e$, can be written as

$$X^T X f = \lambda f.$$  

12. The discriminant function in SVM has the value $\pm 1$ for the support vectors. Hence, $x$ is a supportvector since $f(x) = 1$.

13. kNN with $k = 3$ will classify the cross to the class of squares since two of the three nearest samples are squares.

For SVM the cross is closer to the class of points.

14. a) We will find two correlations, since the smallest of the dimensionalities for $x$ and $y$ is 2.

b) The projections $w_i^T x$ and $w_j^T x$ are uncorrelated for all $i \neq j$.

15. Encode names and numbers binary using the symbols +/-1. If, for example, six bits are used and we reserve space for 25 characters/digits, each entry will be represented by a 150-dimensional vector. The outer products of these vectors are summed and this sum will be the weight matrix in the Hopfield memory. The advantage with such an approach is that only parts of the name or number have to be entered and then the memory can reconstruct the remaining information.
16. a) We can expand the cost function:
\[ e = E[(d - y)^T (d - y)] = E[dd - 2dy + yy] = E[dd - 2dw^T x + w^T xx^T w]. \]

Since we want to minimize the cost function we then differentiate with respect to \( w \) and set the derivative to 0.

\[ \frac{\partial e}{\partial w} = -2E[dx] + 2E[xx^T]w = 0 \]

This gives us:

\[ w_0 = E \{xx^T\}^{-1}E \{dx\} \quad \text{(The Wiener filter)} \]

b) The optimal weight vector is \( w = \begin{bmatrix} 1 \\ 4 \end{bmatrix} \)
17. a) Use the update rule for SOM:
\[ \Delta w_i = \gamma h(i, *) (x - w_i) \]

Since \( w_3 \) is closest to \( x \), \( h \) becomes

\[
\begin{align*}
    h(1, 1) &= e^{-\frac{4}{2.1^2}} = 0.1353 \\
    h(2, 1) &= e^{-\frac{1}{2.1^2}} = 0.6065 \\
    h(3, 1) &= e^0 = 1 \\
    h(4, 1) &= e^{-\frac{1}{2.1^2}} = 0.6065
\end{align*}
\]

The new weights become

\[
\begin{align*}
    w_1 &= \begin{bmatrix} 1.9323 \\ 2.1353 \end{bmatrix}, \\
    w_2 &= \begin{bmatrix} 2.3935 \\ 4.0000 \end{bmatrix}, \\
    w_3 &= \begin{bmatrix} 1.5000 \\ 3.5000 \end{bmatrix} \text{ and } w_4 &= \begin{bmatrix} 3.0902 \\ 4.6967 \end{bmatrix}
\end{align*}
\]

after the update.

b) With the update rule for SOM:
\[ \Delta w_i = \gamma h(i, *) (x - w_i) \]

and remembering that this is a gradient descent search
\[ \Delta w_i = -\gamma \frac{\partial E}{\partial w_i} \]

we get the cost function
\[ E = \frac{1}{2} \sum_i h(i, *) \| x - w_i \|^2 \]
18. a) $w_x = (1, 1)^T$, $w_y = (1, 0)^T$. Blending in the cosine-function can never increase the correlation to something we can achieve with through linear combinations of the $x$-variables. $w_x = (a, a)^T$ gives maximum variance for the sine-function in relation to the variance of the noise and therefore maximizes the variance for the part of the signal that is shared with the $y$-side.

b) $w_x = (1, 1)^T$ using the same reasoning as in (a). $w_y = (\cos(\phi), \sin(\phi))^T$ gives $y = \sin(\omega t + \phi)$, which of course maximizes the correlation with $x$.

c) $x = w_x^T x$ and $y = w_y^T y$ for the second correlation are orthogonal to their counterparts for the first correlation. The projection directions must therefore be $w_x = (1, -1)^T$ and $w_y = (\sin(\phi), -\cos(\phi))^T$ (possibly with the signs reversed). Since all of the signal that are shared between $x$ and $y$ are dealt with in the first correlation, we get $\rho = 0$ for the second correlation.
19. a) Model A: The only possible, i.e. forward. Model B: Shortcut at (3), otherwise forward.

b) $\epsilon$-greedy $\Rightarrow Q_{\epsilon} = V_{\epsilon}$. In this case, there are only one way, the optimal way. $\epsilon$ will therefore not matter.

\[
\begin{align*}
V(1) &= 3 + \gamma V(2) \\
V(2) &= \gamma (V3) \\
\vdots \\
V(6) &= \gamma V(1)
\end{align*}
\]

This leads to

\[
V(1) = 3 + \gamma^6 (V1) \Rightarrow V(1)(1 - \gamma^6) = 3 \Rightarrow V(1) = 3/(1 - \gamma^6)
\]

and

\[
\begin{align*}
V(6) &= 3\gamma/(1 - \gamma^6) \\
V(5) &= 3\gamma^2/(1 - \gamma^6) \\
V(4) &= 3\gamma^3/(1 - \gamma^6) \\
V(3) &= 3\gamma^4/(1 - \gamma^6) \\
V(2) &= 3\gamma^5/(1 - \gamma^6)
\end{align*}
\]

c)

\[
\begin{align*}
V_{\epsilon}(5) &= \gamma V_{\epsilon}(6) \\
V_{\epsilon}(4) &= \gamma^2 V_{\epsilon}(6) \\
V_{\epsilon}(3) &= (1 - \epsilon) \gamma V_{\epsilon}(6) + \epsilon \gamma^3 V_{\epsilon}(6) \\
&= (1 - \epsilon + \epsilon \gamma^2) \gamma V_{\epsilon}(6) \\
V_{\epsilon}(2) &= \gamma V_{\epsilon}(3) \\
V_{\epsilon}(1) &= \gamma^2 V_{\epsilon}(3) + 3 \\
V_{\epsilon}(6) &= \gamma^3 V_{\epsilon}(3) + 3\gamma
\end{align*}
\]

$\Rightarrow V(3) = (1 - \epsilon + \epsilon \gamma^2)(\gamma^3 V(3) + 3\gamma)\gamma$

$\Rightarrow V(3) = \frac{3(1 - \epsilon)\gamma^2 + \gamma^4 \epsilon}{1 - ((1 - \epsilon)\gamma^4 + \epsilon \gamma^6)}$