1. Expected Loss and Bayes Optimality.

a) The expression for the expected value of a function of a joint distribution:

$$\mathbb{E}[g(X,Y)] = \sum_{(x,y)} g(x,y)p(x,y)$$

Thus the expected loss $\mathbb{E}[\mathcal{L}(y,t)]$ for $y = \{\text{keep, remove}\}$ is:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(y=\text{keep},t)] &= \mathcal{L}(y=\text{keep},t=\text{NonSpam})\cdot P(t=\text{NonSpam}) + \mathcal{L}(y=\text{keep},t=\text{Spam})\cdot P(t=\text{Spam}) \\
&= (0)\cdot(1-0.1) + (1)\cdot(0.1) = 0.1 \\
\mathbb{E}[\mathcal{L}(y=\text{remove},t)] &= \mathcal{L}(y=\text{remove},t=\text{NonSpam})\cdot P(t=\text{NonSpam}) + \mathcal{L}(y=\text{remove},t=\text{Spam})\cdot P(t=\text{Spam}) \\
&= (100)\cdot(1-0.1) + (0)\cdot(0.1) = 90
\end{aligned}$$

b) Let $P(t = \text{spam}|\mathbf{x}) = p$, and $y = \{0,1\}$ correspond to {Keep, Remove}. To determine Bayes optimal decision, we need to find the y_* that minimizes the expected loss. The probability that the email is Spam given the feature $\mathbf{x}$ is p, and if we decide to keep it, then the expected loss is $(1)\cdot(p)$. Likewise, the probability that the email is NonSpam given the feature $\mathbf{x}$ is $(1-p)$, and if we decide to keep it, then expected loss will be $100\cdot(1-p)$. Thus, the classifier y_* can be represented in the following form:

$$y_*(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad P(t=\text{Spam}|\mathbf{x}) > 100\cdot(1-P(t=\text{Spam}|\mathbf{x})) \equiv P(t=\text{Spam}|\mathbf{x}) > 0.99 \\ 0 & \text{otherwise} \end{cases}$$

c) To determine the Bayes optimal decision y_*, we must derive $P(t = \text{spam}|\mathbf{x})$ for each value of $\mathbf{x}$.

$$\begin{aligned}
P(t=\text{Spam}|\mathbf{x}) &= \frac{P(\mathbf{x}|t=\text{Spam}\cdot P(t=\text{Spam})}{P(\mathbf{x})} \\
&= \frac{P(\mathbf{x}|t=\text{spam}\cdot P(t=\text{Spam})}{P(\mathbf{x}|t=\text{Spam}\cdot P(t=\text{Spam}) + P(\mathbf{x}|t=\text{NonSpam}\cdot P(t=\text{NonSpam})} \\
P(t=\text{Spam} \mid x_1=0,x_2=0) &= \frac{(0.4)\cdot(0.1)}{(0.4)\cdot(0.1)+(0.998)\cdot(0.9)} = 0.043 \\
P(t=\text{Spam} \mid x_1=0,x_2=1) &= \frac{(0.3)\cdot(0.1)}{(0.3)\cdot(0.1)+(0.001)\cdot(0.9)} = 0.971 \\
P(t=\text{Spam} \mid x_1=1,x_2=0) &= \frac{(0.2)\cdot(0.1)}{(0.2)\cdot(0.1)+(0.001)\cdot(0.9)} = 0.957 \\
P(t=\text{Spam} \mid x_1=1,x_2=1) &= \frac{(0.1)\cdot(0.1)}{(0.1)\cdot(0.1)+(0)\cdot(0.9)} = 1
\end{aligned}$$

Thus, the Bayes optimal decision y_* is represented in the following form:

$$y_*(\mathbf{x}) = \begin{cases} 1 & x_1=1, x_2=1 \\ 0 & \text{otherwise} \end{cases}$$

d)

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(y_*,t)] &= \sum_{y_*}\sum_{t} \mathcal{L}(y_*,t) \\
&= \mathcal{L}(\text{keep, spam})\cdot P(\text{keep, spam}) + \mathcal{L}(\text{remove, spam})\cdot P(\text{remove, spam}) + \\
&\mathcal{L}(\text{keep, NonSpam})\cdot P(\text{keep, NonSpam}) + \mathcal{L}(\text{remove, NonSpam})\cdot P(\text{remove, NonSpam}) \\
&= 1\cdot P(x_1=1,x_2=1,t=\text{Spam}) + 0 + 0 + 100(x_1=1,x_2=1,t=\text{NonSpam}) \\
&= 0.9 + 100\cdot 0 = 0.9
\end{aligned}$$

2. Feature Maps.

a)

Proof. Assume for contradiction that set is linearly separable i.e. we must linearly separate points {-1},{3} from {0}. However, since we are in one dimensional space, there must exist a line that separates points {-1} and {0}, and also {0} and {3} at the same time. A line that separates points in one dimensional space is equally represented by a point, but a point cannot be in two places at the same time and thus we have a contradiction. □

b) With the feature maps applied, we have the following table:

x	$\psi_1(x)$	$\psi_2(x)$	t
-1	-1	1	1
1	1	1	0
3	3	9	1

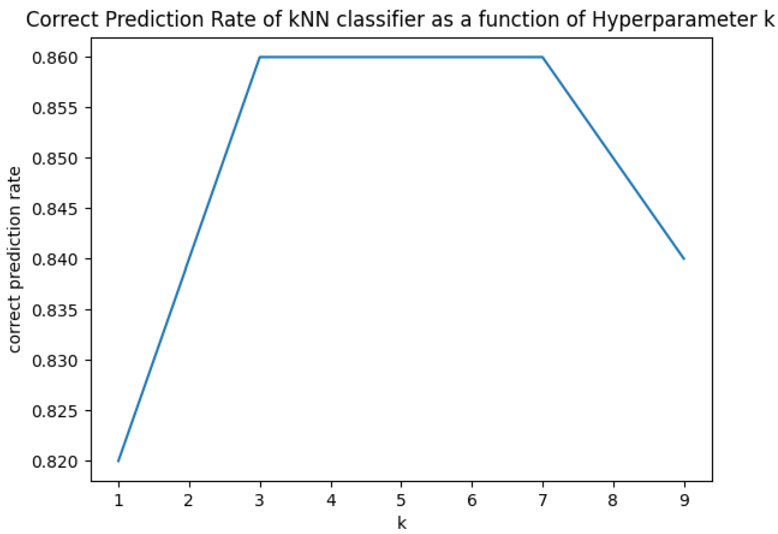
We can setup the following inequalities:

$$\begin{aligned} -w_1 + w_2 > 0 &\quad\Rightarrow\quad w_2 > w_1 \\ w_1 + w_2 < 0 &\quad\Rightarrow\quad w_1 < -w_2 \\ 3w_1 + w_2 > 0 &\quad\Rightarrow\quad w_2 > \frac{1}{3}w \end{aligned}$$

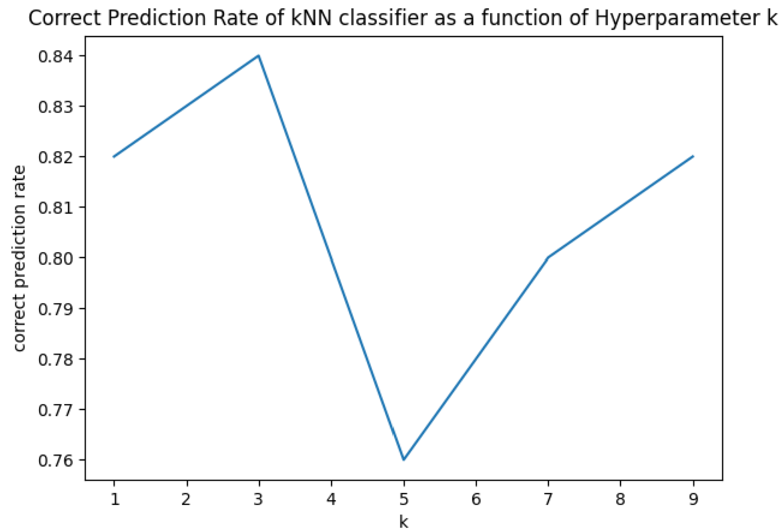
A possible set of solutions can be $w_1 = -2$ and $w_2 = 1$.

3.1. k-Nearest Neighbors.

a)



b) Based on the training set, the classifier performs best for $k = \{3, 5, 7\}$, and performs the worst for $k = \{1\}$. Since the the prediction rate is highest for $k = \{3, 5, 7\}$ (each of these hyperparameters produce the same classification rate), I would choose $k^* = 5$ as it is the median between them. As per the results of the test performance, $k = 3$ had the highest classification rate and $k = 5$ had the lowest classification rate of all the parameters and thus knowing this information, $k^* = 3$ would be the optimal choice.



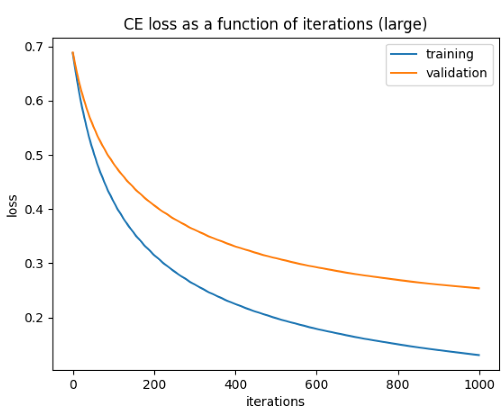
3.2. Logistic Regression.

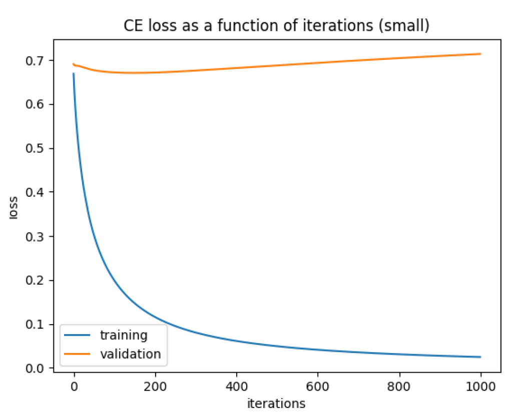
b) From testing various hyperparameters and weight initializations, I have found that initiliazing all weights to 0 produces, $\lambda = 0.01$ and 1000 iterations produces the best results (did not test for more than 1000 iterations due to slow computer performance). Below are results (first screenshot is for larger dataset).

```
Validation: Min CE: 0.2536710662674645 with index 999, error: 0.12
Training: Min Ce: 0.1306269841310717 error: 0.0
Testing: Min Ce: 0.23235634333047048 error: 0.07999999999999996
```

```
Validation: Min CE: 0.670792021497356 with index 149, error: 0.38
Training: Min Ce: 0.024430032402602796 error: 0.0
Testing: Min Ce: 0.6139844110276055 error: 0.21999999999999997
```

c)





From the results of these plots, it seems cross-entropy asymptotically decreases with the number of iterations. However, this is not the case for classification error. Based on this I would choose parameters where the classification error is smallest in the region where cross entropy starts to approach the asymptote.

4. Locally Weighted Regression.

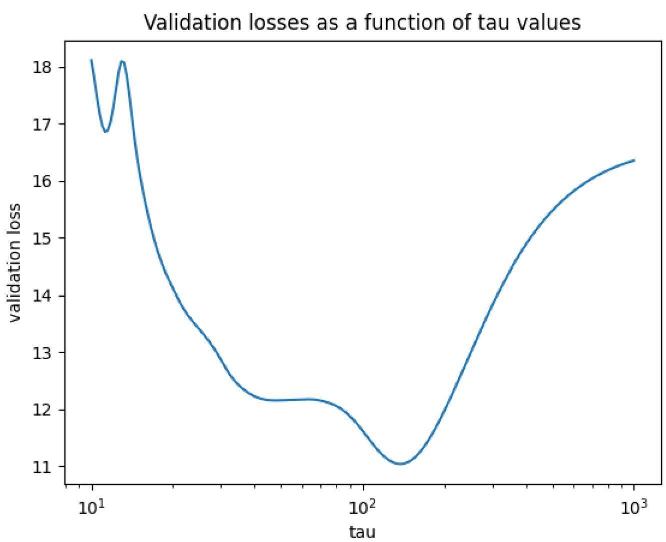
$$
\begin{aligned}
\mathbf{w}^* &= \arg\min \frac{1}{2}\sum_{i=1}^{N} a^{(i)} \left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \\
\frac{\partial}{\partial \mathbf{w}} & \frac{1}{2}\sum_{i=1}^{N} a^{(i)} \left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \\
&= \sum_{i=1}^{N} -x^{(i)}a^{(i)} \left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right) + \lambda \mathbf{w}
\end{aligned}
$$

Vectorizing we now get:

$$
\begin{aligned}
-\mathbf{X}^T\mathbf{A}\left(\mathbf{y} - \mathbf{W}\mathbf{x}\right) + \lambda\mathbf{w} &= 0 \\
-\mathbf{X}^T\mathbf{A}\mathbf{y} - \mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{x} + \lambda\mathbf{w} &= 0 \\
\left(\mathbf{X}^T\mathbf{A}\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w} &= \mathbf{X}^T\mathbf{A}\mathbf{y} \\
\mathbf{w}^* &= \left(\mathbf{X}^T\mathbf{A}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{A}\mathbf{y}
\end{aligned}
$$

Notice our loss function is a sum of convex functions and thus as a result is a convex function itself. Therefore the solution to the weighted least squares problem provides us with the global minimum as convex functions contain only one global minimum.

c)



d) The only term dependent on τ is $a^{(i)}$, and thus we can its behavior by applying limits to it. Observe that the norms in the numerator and denominator of $a^{(i)}$ stay constant w.r.t. to τ. Thus applying the limit:

$$
\begin{aligned}
&\lim_{\tau\to\infty} \frac{\exp\left(-\left\|\mathbf{x} - \mathbf{x}^{(i)}\right\|^2 / 2\tau^2\right)}{\sum_j \exp\left(-\left\|\mathbf{x} - \mathbf{x}^{(j)}\right\|^2 / 2\tau^2\right)} = \frac{1}{\sum_j 1} = \frac{1}{N} \\
&\lim_{\tau\to\infty} \frac{1}{2}\sum_{i=1}^{N} a^{(i)} \left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 = \frac{1}{2N}\sum_{i=1}^{N} \left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2
\end{aligned}
$$

So as $\tau \to \infty$, our validation loss asymptotically approaches some constant. Now for $\tau \to 0$:

$$
\lim_{\tau\to\infty} \frac{\exp\left(-\left\|\mathbf{x} - \mathbf{x}^{(i)}\right\|^2 / 2\tau^2\right)}{\sum_j \exp\left(-\left\|\mathbf{x} - \mathbf{x}^{(j)}\right\|^2 / 2\tau^2\right)} = \frac{0}{\sum_j 0} = \infty
$$

Hence, for $\tau \to 0$, the entire expression approaches infinity and is undefined at 0. These results also match up with the plot given in 4c.