1: Probability Decision Boundaries

(a)

$$Y \in \{T, F\}$$
$$P(Y = T) = 0.5$$
$$p_0 = P(Y = 0 \| X)$$
$$p_1 = P(Y = 1 \| X)$$
$$E(C \| \hat{Y} = 0) = p_0(0) + p_1(10)$$
$$E(C \| \hat{Y} = 1) = p_0(5) + p_1(0)$$
$$p_0 = 1 - p_1$$
$$E(C \| \hat{Y} = 0) = p_1(10)$$
$$E(C \| \hat{Y} = 1) = (1 - p_1)(5)$$

given

$$0 < \theta < 1$$

if

$$p_1 = 1$$

the cost function is minimized by

$$\hat{Y} = 1 \rightarrow (1 - p_1)(5)$$

and if

$$p_1 = 0$$

the cost function is minimized by

$$\hat{Y} = 0 \rightarrow p_1(10)$$

(b)

$$p_1(10) = (1 - p_1)(5)$$
$$p_1 = \theta = \frac{1}{3}$$

2: Double Counting the Evidence

(a) For X_1

$$P(Y = T \| X_1 = F) + P(Y = f \| X_1 = T)$$
$$= P(Y = T)(1 - P(X_1 = T \| Y = T) + P(Y = F)(1 - P(X_1 = F \| Y = F)$$
$$= 0.5x0.2 + 0.5x0.3$$
$$= 0.25$$

For X_2

$$P(Y = T \| X_2 = F) + P(Y = f \| X_2 = T)$$
$$= P(Y = T)(1 - P(X_2 = T \| Y = T) + P(Y = F)(1 - P(X_2 = F \| Y = F)$$
$$= 0.5x0.5 + 0.5x0.1$$
$$= 0.3$$

(b)

Table 1: Conditional Probabilities

X_1	X_2	Y	$H(x)$
T	T	T	0.2
T	T	F	0.015
T	F	T	0.2
T	F	F	0.135
F	F	T	0.05
F	F	F	0.315
F	T	T	0.05
F	T	F	0.035

$$P(Y = F)P(X_1 = T \| Y = F)P(X_2 = T \| Y = F) +$$
$$P(Y = F)P(X_1 = T \| Y = F)P(X_2 = F \| Y = F) +$$
$$P(Y = T)P(X_1 = F \| Y = F)P(X_2 = F \| Y = F) +$$
$$P(Y = F)P(X_1 = F \| Y = F)P(X_2 = T \| Y = F)$$
$$= 0.15 + 0.135 + 0.05 + 0.035$$
$$= 0.235$$

(c)

Table 2: Conditional Probabilities

X_1	X_2	X_3	Y	$H(x)$
T	T	T	T	0.1
T	T	T	F	0.0015
T	T	F	T	0.1
T	T	F	F	0.0135
T	F	T	T	0.1
T	F	T	F	0.0135
T	F	F	T	0.1
T	F	F	F	0.1215
F	F	F	T	0.025
F	F	F	F	0.2835
F	F	T	T	0.025
F	F	T	F	0.0315
F	T	F	T	0.025
F	T	F	F	0.0315
F	T	T	T	0.025
F	T	T	F	0.0035

$$
\begin{aligned}
&P(Y=F)P(X_1=T\|Y=F)P(X_2=T\|Y=F)P(X_3=T\|Y=F)+\\
&P(Y=F)P(X_1=T\|Y=F)P(X_2=T\|Y=F)P(X_3=F\|Y=F)+\\
&P(Y=F)P(X_1=T\|Y=F)P(X_2=F\|Y=F)P(X_3=T\|Y=F)+\\
&P(Y=T)P(X_1=T\|Y=F)P(X_2=F\|Y=F)P(X_3=F\|Y=F)+\\
&P(Y=T)P(X_1=F\|Y=F)P(X_2=F\|Y=F)P(X_3=F\|Y=F)+\\
&P(Y=T)P(X_1=F\|Y=F)P(X_2=F\|Y=F)P(X_3=T\|Y=F)+\\
&P(Y=T)P(X_1=F\|Y=F)P(X_2=T\|Y=F)P(X_3=F\|Y=F)+\\
&P(Y=F)P(X_1=F\|Y=F)P(X_2=T\|Y=F)P(X_3=T\|Y=F)+\\
&= 0.0015+0.0135+0.0135+0.1+0.025+0.025+0.025+0.0035\\
&= 0.207
\end{aligned}
$$

(d)

We are giving the classifier a non-independent feature which is causing the classifier to increase its confidence and essentially overfit the data.

(e)

Logistic regression does not require independent features and thus does not suffer from this issue. When given duplicate features or non-independent features, logistic regression simply lowers the weight assigned to each of these features - thus mitigating this problem.

3: Reject Option

(a)

$$P(y = 1 \| x) = p_1$$
$$E_cost(\hat{y} = 0 | p_1 = 0.2) = 0.8(0) + 0.2(10) = 2$$
$$E_cost(\hat{y} = 1 | p_1 = 0.2) = 0.8(10) + 0.2(0) = 8$$
$$E_cost(reject | p_1 = 0.2) = 0.8(3) + 0.2(3) = 3$$

Choosing a value of $\hat{y} = 0$ minimizes the cost

(b)

$$P(y = 1 \| x) = p_1$$
$$E_cost(\hat{y} = 0 | p_1 = 0.4) = 0.6(0) + 0.4(10) = 4$$
$$E_cost(\hat{y} = 1 | p_1 = 0.4) = 0.6(10) + 0.4(0) = 6$$
$$E_cost(reject | p_1 = 0.4) = 0.6(3) + 0.4(3) = 3$$

Choosing a value of $reject$ minimizes the cost

(c)

$$p_1 = P(Y = 1 \| X)$$
$$E(C \| \hat{Y} = 0) = p_0(0) + p_1(10)$$
$$E(C \| \hat{Y} = 1) = p_0(10) + p_1(0)$$
$$E(C \| \hat{Y} = 0) = p_1(10)$$
$$E(C \| \hat{Y} = 1) = (1 - p_1)(10)$$
$$E_cost(c \| reject) = 3$$

given

$$0 < \theta_0 < \theta_1 < 1$$

if

$$p_1 = 1$$

the cost function is minimized by

$$\hat{Y} = 1 \rightarrow (1 - p_1)(10)$$

and if

$$p_1 = 0$$

the cost function is minimized by

$$\hat{Y} = 0 \rightarrow p_1(10)$$

finally if

$$p_1 = 0.5, \hat{Y} = 1$$

the cost function is minimized by

$$reject \rightarrow 3$$

(d)

For θ_0

$$p_1(10) = 3$$
$$p_1 = \theta_0 = 0.3$$

For θ_0

$$(1 - p_1)(5) = 3$$
$$p_1 = \theta_1 = 0.4$$

1.2: Training the Best Classifier

For this challenge, I chose to use a Support Vector Classification tool from scipy's library with radial basis function kerel. I chose this ML classifier because according to the scikit documentation, it works well with multi-class prediction, learning from less than 100k samples of labeled data, and in cases where a linearSVC does not neccissarily cut it. I trained by simple partitioning the data into training features and labels and testing with crossvalidation over several out of the box ML algorithms. Once I settled on SVC, I trained and prediceted on the entire dataset.