Solution 1:

Logistic regression is a classification model, that estimates posterior probabilities $\pi(x)$ by linear functions in x. For a binary classification problem the model can be written as:

$$\hat{y} = 1 \quad \Leftrightarrow \quad \pi(x) = \frac{1}{1+\exp(-x^T\theta)} \geq a$$

For the decision boundary we have to set $\pi(x) = a$. Solving this equation for x yields:

$$\begin{aligned}
&\frac{1}{1+\exp(-x^T\theta)} = a \\
&\Leftrightarrow 1+\exp(-x^T\theta) = a^{-1} \\
&\Leftrightarrow \exp(-x^T\theta) = a^{-1} - 1 \\
&\Leftrightarrow \exp(-x^T\theta) = a^{-1} - 1 \\
&\Leftrightarrow -x^T\theta = \log(a^{-1} - 1) \\
&\Rightarrow x^T\theta = -\log(a^{-1} - 1)
\end{aligned}$$

For $a = 0.5$ we get:

$$x^T\theta = -\log(0.5^{-1} - 1) = -\log(2-1) = -\log(1) = 0$$

Solution 2:
See R code **sol_mlr_decision_boundaries.R**

Solution 3:

a) $\pi_1(x) = \frac{e^{\theta_1^T x}}{e^{\theta_1^T x} + e^{\theta_2^T x}}$

   $\pi_2(x) = \frac{e^{\theta_2^T x}}{e^{\theta_1^T x} + e^{\theta_2^T x}}$

   $\pi_1(x) = \frac{1}{(e^{\theta_1^T x} + e^{\theta_2^T x})/e^{\theta_1^T x}} = \frac{1}{1+e^{\theta^T x}}$ where $\theta = \theta_2 - \theta_1$ and $\pi_2(x) = 1 - \pi_1(x)$

b) Likelihood for one instance

$$L_i = P(y^{(i)} = k|\theta) = \prod_k \pi_k(x)^{I_k}$$

   where

$$\pi_k(x) = \frac{e^{\theta_k^T x}}{\sum_j e^{\theta_j^T x}}$$

$$\sum_{k=1}^{g} \pi_k = 1 \quad \text{and} \quad I_k = [y = k]$$

$$\sum_{k=1}^{g} I_k = 1$$

negative log likelihood for one instance:

$$-\log(L_i) = -\sum_{k=1}^{g} I_k \log(\pi_k(x))$$

negative log likelihood:

$$-\log(\mathcal{L}) = \sum_{i=1}^{n} -\log(L_i)$$

c) For g classes we get $g-1$ discriminant functions from the softmax $\pi_1(x), \dots, \pi_{g-1}(x)$ which can be interpreted as probability. The probability for class g can be calculated by using $1 - \sum_{k=1}^{g-1} \pi_k(x)$. To estimate the class we are using majority vote:

$$\hat{y} = \operatorname*{arg\,max}_k \pi_k(x)$$

The parameter of the softmax regression is defined as parameter matrix where each class has its own parameter vector θ_k, $k \in \{1, \dots, g-1\}$:

$$\theta = [\theta_1, \dots, \theta_{g-1}]$$

Solution 4:

a) To predict a new fruit as banana or not we have to calculate the posterior probabilities for “yes” and “no” regarding the Bayes’ theorem :

$$\begin{aligned} &P(\text{Banana} = \text{yes} \mid \text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported}) \\ &\quad = \frac{P(\text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported} \mid \text{Banana} = \text{yes}) P(\text{Banana} = \text{yes})}{P(\text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported})} \end{aligned}$$

$$\begin{aligned} &P(\text{Banana} = \text{no} \mid \text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported}) \\ &\quad = \frac{P(\text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported} \mid \text{Banana} = \text{no}) P(\text{Banana} = \text{no})}{P(\text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported})} \end{aligned}$$

- Due to the independence assumption we make for naive Bayes, we can rewrite the numerator as product of the individual probabilities (just done for “yes”):

$$\begin{aligned} &P(\text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported} \mid \text{Banana} = \text{yes}) \\ &\quad = P(\text{Color} = \text{yellow} \mid \text{Banana} = \text{yes}) P(\text{Form} = \text{round} \mid \text{Banana} = \text{yes}) P(\text{Origin} = \text{imported} \mid \text{Banana} = \text{yes}) \end{aligned}$$

- Since both posterior probabilities shares the same denominator it is sufficient to calculate just the numerator and standardize them afterwards (just done for “yes”):

$$\begin{aligned} &P(\text{Banana} = \text{yes} \mid \text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported}) \\ &\quad \propto P(\text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported} \mid \text{Banana} = \text{yes}) P(\text{Banana} = \text{yes}) \end{aligned}$$

To calculate the posterior probabilities we need:

$$\begin{aligned} &P(\text{Banana} = \text{yes}),\ P(\text{Color} = \text{yellow}|\text{Banana} = \text{yes}),\ P(\text{Form} = \text{round}|\text{Banana} = \text{yes}), \\ &P(\text{Origin} = \text{imported}|\text{Banana} = \text{yes}) \end{aligned}$$

and

$$\begin{aligned} &P(\text{Banana} = \text{no}),\ P(\text{Color} = \text{yellow}|\text{Banana} = \text{no}),\ P(\text{Form} = \text{round}|\text{Banana} = \text{no}), \\ &P(\text{Origin} = \text{imported}|\text{Banana} = \text{no}) \end{aligned}$$

Since the features are categorical we use as estimator the relative frequency:

$$\begin{aligned}
&P(\text{Banana} = \text{yes}) = 3/8 \\
&P(\text{Color} = \text{yellow}|\text{Banana} = \text{yes}) = 1/3 \\
&P(\text{Form} = \text{round}|\text{Banana} = \text{yes}) = 1/3 \\
&P(\text{Origin} = \text{imported}|\text{Banana} = \text{yes}) = 1
\end{aligned}$$

$$\begin{aligned}
&P(\text{Banana} = \text{no}) = 5/8 \\
&P(\text{Color} = \text{yellow}|\text{Banana} = \text{no}) = 2/5 \\
&P(\text{Form} = \text{round}|\text{Banana} = \text{no}) = 3/5 \\
&P(\text{Origin} = \text{imported}|\text{Banana} = \text{no}) = 2/5
\end{aligned}$$

With the independence assumption of naive Bayes, the final step is to calculate the product of the probabilities above:

$$\begin{aligned}
&P(\text{Banana} = \text{yes}|\text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported}) \\
&\quad \propto P(\text{Banana}) \cdot P(\text{Color} = \text{yellow}|\text{Banana} = \text{yes}) \cdot P(\text{Form} = \text{round}|\text{Banana} = \text{yes}) \\
&\quad \cdot P(\text{Origin} = \text{imported}|\text{Banana} = \text{yes}) \\
&\quad = 3/8 \cdot 1/3 \cdot 1/3 \cdot 1 \\
&\quad = 1/24 = 0.04167
\end{aligned}$$

$$\begin{aligned}
&P(\text{Banana} = \text{no}|\text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported}) \\
&\quad \propto P(\text{Banana}) \cdot P(\text{Color} = \text{yellow}|\text{Banana} = \text{no}) \cdot P(\text{Form} = \text{round}|\text{Banana} = \text{no}) \\
&\quad \cdot P(\text{Origin} = \text{imported}|\text{Banana} = \text{no}) \\
&\quad = 5/8 \cdot 2/5 \cdot 3/5 \cdot 2/5 \\
&\quad = 3/50 = 0.06
\end{aligned}$$

At this stage we can already see that the predicted label is "no" $0.06 > 0.04167$. To get the final posterior probabilities we have to calculate the constant c for standardization:

$$c = 0.04167 + 0.06 = 0.10167$$

This gives us the posterior probabilities:

$$\begin{aligned}
&P(\text{Banana} = \text{yes}|\text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported}) \\
&\quad = 0.04167/c \\
&\quad = 0.04167/0.10167 \\
&\quad = 0.4109
\end{aligned}$$

$$\begin{aligned}
&P(\text{Banana} = \text{no}|\text{Color} = \text{yellow},\ \text{Form} = \text{round},\ \text{Origin} = \text{imported}) \\
&\quad = 0.06/0.10167 \\
&\quad = 0.5901
\end{aligned}$$

Corresponding `R`-Code:

```r
df_banana = data.frame(
  Color = c("yellow", "yellow", "yellow", "brown", "brown", "green", "green", "red"),
  Form = c("oblong", "round", "oblong", "oblong", "round", "round", "oblong", "round"),
  Origin = c("imported", "domestic", "imported", "imported", "domestic", "imported",
    "domestic", "imported"),
  Banana = c("yes", "no", "no", "yes", "no", "yes", "no", "no")
)
```

```r
library(mlr)

nbayes_learner = makeLearner("classif.naiveBayes", predict.type = "prob")
banana_task = makeClassifTask(data = df_banana, target = "Banana")
model = train(learner = nbayes_learner, task = banana_task)

predict(model, newdata = data.frame(Color = "yellow", Form = "round", Origin = "imported"))

## Prediction: 1 observations
## predict.type: prob
## threshold: no=0.50,yes=0.50
## time: 0.00
##   prob.no prob.yes response
## 1  0.5902   0.4098       no
```

b) For numerical features, we would use a different probability distribution, not the Bernoulli distribution. For example, for the information "Length" we could assume that $P(Length|yes)$ and $P(Length|no)$ both follow a normal distribution.