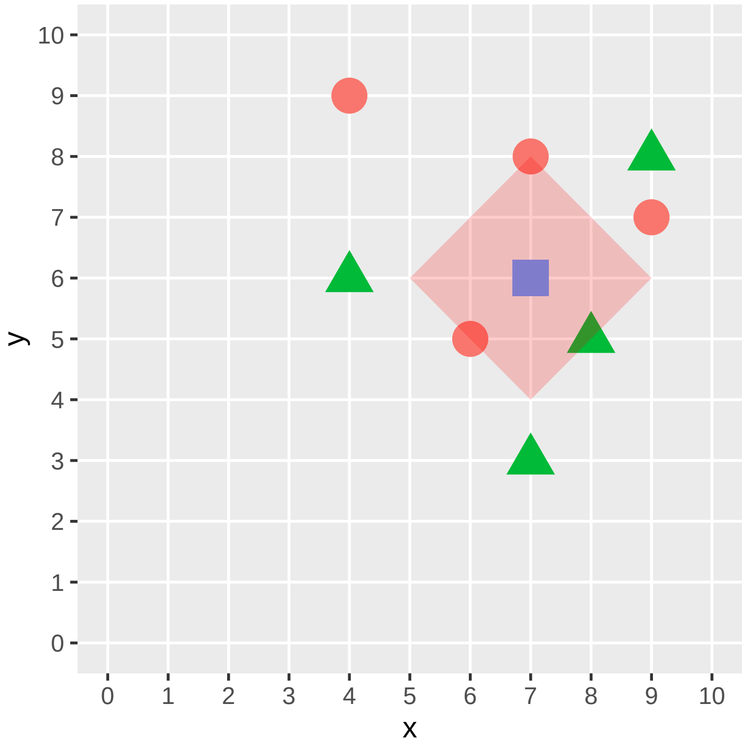
Solution 1:

See R code

Solution 2:

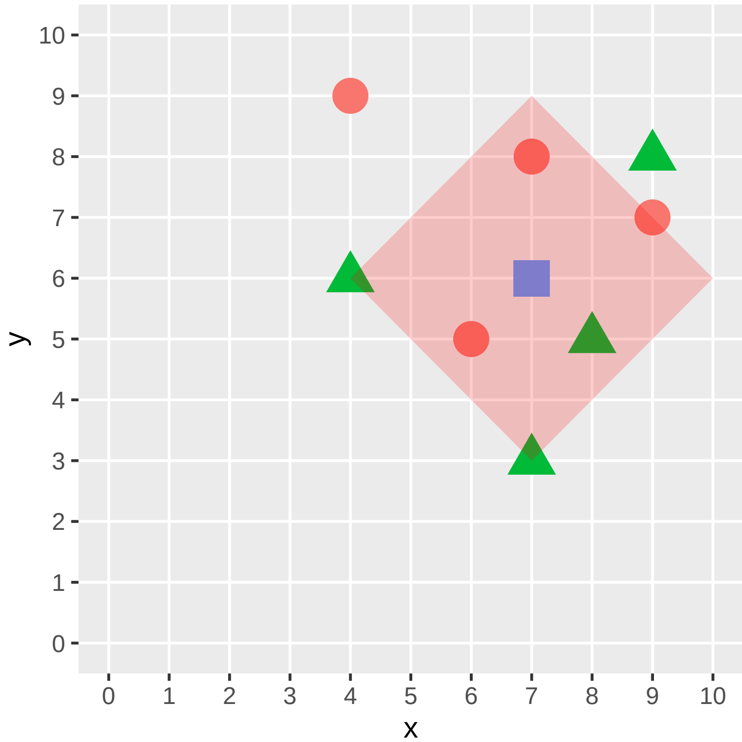
a) $k = 3$

2 circles and 1 triangle, so our point is also a circle



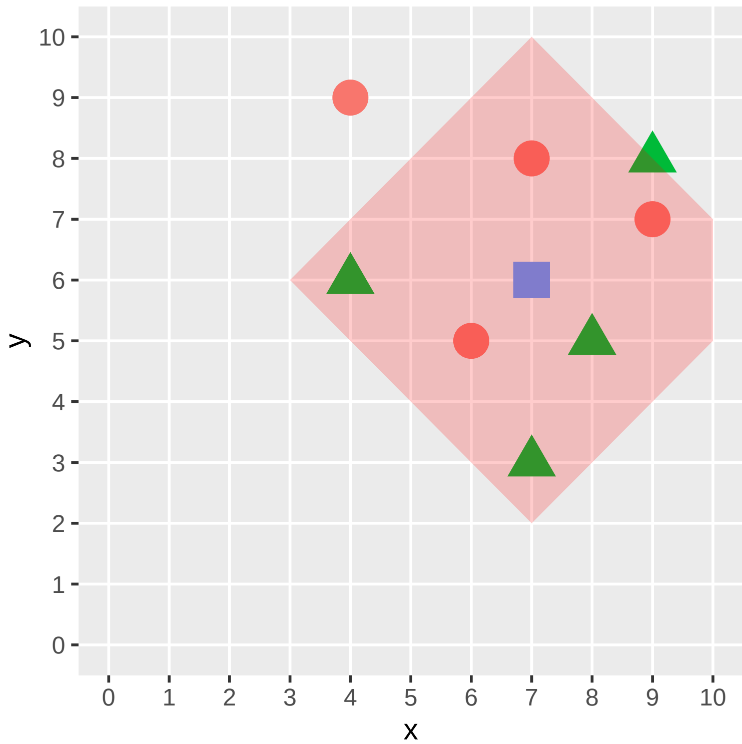
b) $k = 5$

3 circles and 3 triangles, we have to specify beforehand what to do in case of a tie



c) $k = 7$

3 circles and 4 triangles, so our point is also a triangle



Solution 3:

a) When using the naive Bayes classifier, the features $x := (x_{\text{Color}}, x_{\text{Form}}, x_{\text{Origin}})$ given the category $y \in \{\text{yes}, \text{no}\}$ are assumed to be conditionally independent of each other, s.t.

$$p((x_{\text{Color}}, x_{\text{Form}}, x_{\text{Origin}})|y = k) = p(x_{\text{Color}}|y = k) \cdot p(x_{\text{Form}}|y = k) \cdot p(x_{\text{Origin}}|y = k).$$

For the posterior probabilities $\pi_k(x)$ it holds that

$$\pi_k(x) \propto \underbrace{\pi_k \cdot p(x_{\text{Color}}|y = k) \cdot p(x_{\text{Form}}|y = k) \cdot p(x_{\text{Origin}}|y = k)}_{=:\alpha_k(x)}$$
$$\iff \exists c \in \mathbb{R} : \pi_k(x) = c \cdot \alpha_k(x),$$

where π_k is the prior probability of class k. From this and since the posterior probabilities need to sum up to 1, it holds that

$$1 = c \cdot \alpha_{\text{yes}}(x) + c \cdot \alpha_{\text{no}}(x)$$
$$\iff c = \frac{1}{\alpha_{\text{yes}}(x) + \alpha_{\text{no}}(x)}.$$

This means in order to compute $\pi_{\text{yes}}(x)$ the scores $\alpha_{\text{yes}}(x)$ and $\alpha_{\text{no}}(x)$ are needed.
Now we want to compute for a new fruit the posterior probability $\hat{\pi}_{yes}((\text{yellow}, \text{round}, \text{imported}))$.
Note that we do not know the *true* prior probability and the *true* conditional densities. Here -since the target and the features are categorical- we can estimate them with the relative frequencies encountered in the data, s.t.

$$\begin{aligned}
\hat{\alpha}_{\text{yes}}(x) &= \hat{\pi}_{yes} \cdot \hat{p}(\text{yellow}|y = \text{yes}) \cdot \hat{p}(\text{round}|y = \text{yes}) \cdot \hat{p}(\text{imported}|y = \text{yes}) \\
&= \hat{\mathbb{P}}(y = \text{yes}) \cdot \hat{\mathbb{P}}(x_{\text{Color}} = \text{yellow}|y = \text{yes}) \cdot \hat{\mathbb{P}}(x_{\text{Form}} = \text{round}|y = \text{yes}) \cdot \hat{\mathbb{P}}(x_{\text{Origin}} = \text{imported}|y = \text{yes}) \\
&= \frac{3}{8} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot 1 = \frac{1}{24} \approx 0.042, \\
\hat{\alpha}_{\text{no}}(x) &= \hat{\pi}_{no} \cdot \hat{p}(\text{yellow}|y = \text{no}) \cdot \hat{p}(\text{round}|y = \text{no}) \cdot \hat{p}(\text{imported}|y = \text{no}) \\
&= \hat{\mathbb{P}}(y = \text{no}) \cdot \hat{\mathbb{P}}(x_{\text{Color}} = \text{yellow}|y = \text{no}) \cdot \hat{\mathbb{P}}(x_{\text{Form}} = \text{round}|y = \text{no}) \cdot \hat{\mathbb{P}}(x_{\text{Origin}} = \text{imported}|y = \text{no}) \\
&= \frac{5}{8} \cdot \frac{2}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{3}{50} = 0.06.
\end{aligned}$$

At this stage we can already see that the predicted label is "no", since $\hat{\alpha}_{\text{no}}(x) = 0.06 > \frac{1}{24} = \hat{\alpha}_{\text{yes}}(x)$.
With this we can calculate the posterior probability

$$\hat{\pi}_{\text{yes}}(x) = \frac{\hat{\alpha}_{\text{yes}}(x)}{\hat{\alpha}_{\text{yes}}(x) + \hat{\alpha}_{\text{no}}(x)} \approx 0.41.$$

Corresponding `R`-Code:

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

new_fruit <- data.frame(Color = "yellow", Form = "round", Origin = "imported", Banana = NA)
df_banana <- rbind(df_banana, new_fruit)
```

```r
library(mlr3)
library(mlr3learners)

nb_learner <- lrn("classif.naive_bayes",
                  predict_type = "prob")

banana_task <- TaskClassif$new(
  id = "banana",
  backend = df_banana,
  target = "Banana"
)

nb_learner$train(banana_task, row_ids=1:8)

nb_learner$predict(banana_task, row_ids = 9)

## <PredictionClassif> for 1 observations:
##  row_ids truth response   prob.no  prob.yes
##        9  <NA>       no 0.5901639 0.4098361
```

b) For the distribution of a numerical feature given the the category we need to specify a probability distribution with continuous support. For example, for the information x_{Length} we could assume that $p(x_{\text{Length}}|y = \text{yes}) \sim \mathcal{N}(\mu_{\text{yes}}, \sigma^2_{\text{yes}})$ and $p(x_{\text{Length}}|y = \text{no}) \sim \mathcal{N}(\mu_{\text{no}}, \sigma^2_{\text{no}})$. (To estimate these normal distributions one would need to estimate their parameters $\mu_{\text{yes}}, \mu_{\text{no}}, \sigma^2_{\text{yes}}, \sigma^2_{\text{no}}$ on the data respectively)