Solutions week 3

3.1

In 1-nearest neighbours you will always perfectly fit the training set, so $e_{train} = 0$. The average over all data will be the average of the errors for the individual datasets since they are equally large, and so we get

$$e = \frac{e_{train} + e_{test}}{2} = \frac{e_{test}}{2} = 18\% \quad \Rightarrow \quad e_{test} = 36\% \tag{1}$$

Given that the test error for the logistic regression was 30%, which is less that 36%, we choose it over 1-NN since it has better generalization error.

3.2

Programming task...

3.3

We get the distributions for the test given the persons status (healthy = 1 or sick = 2) as

$$p(x \mid y = 1) = \mathcal{N}(10, 4^2)$$
$$p(x \mid y = 2) = \mathcal{N}(20, 5^2)$$

and we also get that $p(y = 1) = 0.99 = 1 - p(y = 2)$.

(a)

We can calculate the expression for the cancer probability given a measurement using the given distributions and bayes theorem.

$$\begin{aligned} p(y = 2 \mid x) &= \frac{p(x \mid y = 2)p(y = 2)}{p(x \mid y' = 1)p(y' = 1) + p(x \mid y' = 2)p(y' = 2)} \\ &= \frac{\frac{1}{\sqrt{2\pi 5^2}} e^{-\frac{(x-20)^2}{2 \cdot 5^2}} 0.01}{\frac{1}{\sqrt{2\pi 4^2}} e^{-\frac{(x-10)^2}{2 \cdot 4^2}} 0.99 + \frac{1}{\sqrt{2\pi 5^2}} e^{-\frac{(x-20)^2}{2 \cdot 5^2}} 0.01} \end{aligned}$$

We now just insert the values for our patients and get

$$p_A(y = 2 \mid 15) = 0.01059$$
$$p_B(y = 2 \mid 20) = 0.1553$$
$$p_C(y = 2 \mid 25) = 0.8472$$

(b)

Assuming our assumptions about the distributions are correct, we want to use the most probable according to Bayes classifier, so A and B are healthy while B has cancer.

(c)

Test (maybe in some programming language) for which x we get $p(y = 2 \mid x) = 0.5$, and this turns out to be $x \approx 22.59$

(d)

Given the impact of misclassification in the different cases (miss someone who has cancer, or do a more accurate test on someone who didn't have) we might want to err on the side of predicting cancer.

3.4

(a)

Programming task...

(b)

Catching 99% of all true cancer cases requires that the ratio between true positive cases and actual positive cases is larger or equal to 99%. This fraction is called TPR or recall.

It can be found looking at the cumulative distribution function for the measurements given it is a cancer case. So we want to find for what t we get that $p(x > t \mid y = 2) = 0.99$, and since the cdf is the cumulative sum we get that $p(x < t) = cdf(t)$ and can thus write

$$\begin{aligned} p(x > t \mid y = 2) &= 1 - p(x < t \mid y = 2) \\ &= 1 - cdf(t \mid y = 2) = 0.99 \\ cdf(t \mid y = 2) &= 0.01 \end{aligned}$$

In some languages you have functions for the inverse normal cdf, but if you you can just test what t gives a good approximation of 0.01, I got it to around $t \approx 8.36826$.

(c)

Now we want to look at the ratio of true positive cases and predicted positive cases which is called precision. We predict cancer if $x > t$, so what is the probability we actually have cancer given that $x > t$?

$$\begin{aligned}
p(y = 2 \mid x > t) &= \frac{p(x > t \mid y = 2)p(y = 2)}{p(x > t \mid y = 1)p(y = 1) + p(x > t \mid y = 2)p(y = 2)} \\
&= \frac{(1 - p(x < t \mid y = 2))0.01}{(1 - p(x < t \mid y = 1))0.99 + (1 - p(x < t \mid y = 2))0.01} \\
&= \frac{(1 - \mathit{cdf}(t \mid y = 2))0.01}{(1 - \mathit{cdf}(t \mid y = 1))0.99 + (1 - \mathit{cdf}(t \mid y = 2))0.01} \\
&\approx 0.014962
\end{aligned}$$

3.5

Programming task...