

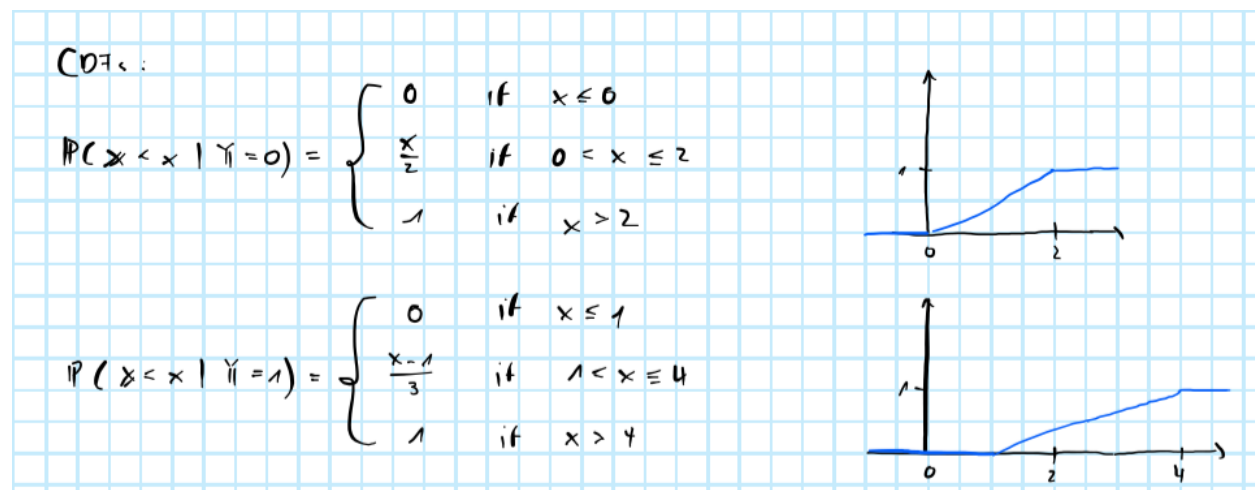
Problem Set 2

Patrick Altmeyer

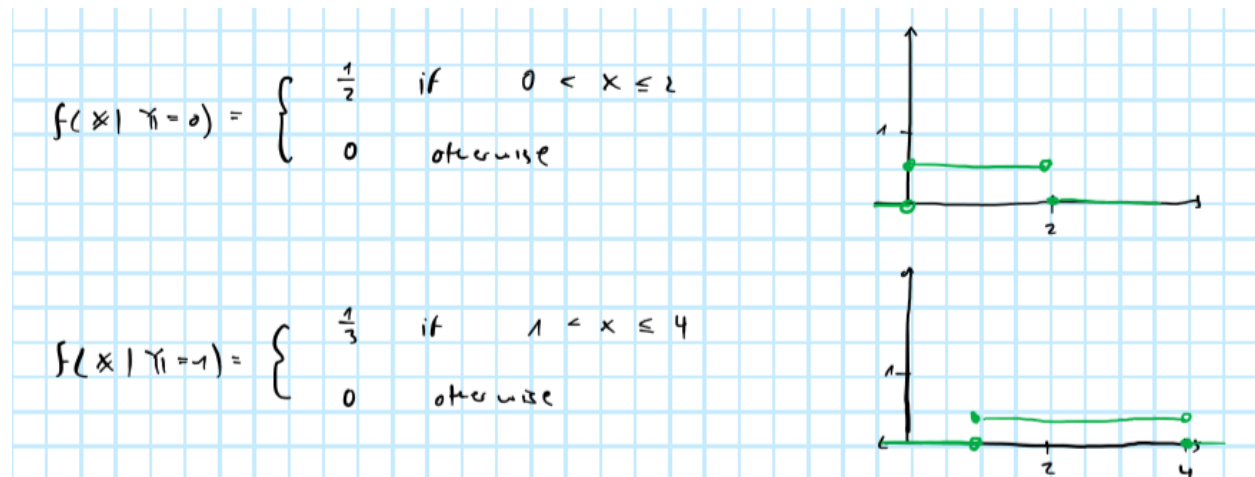
12 February, 2021

1 Piece-wise CDFs

We are given the conditional CDFs



from which the conditional PDFs can be derived:



Then we have for the joint PDF:

Taking it together:

$$f(x) = f(x|Y=0) P(Y=0) + f(x|Y=1) P(Y=1)$$

$$f(x) = \begin{cases} 0.5 \cdot \frac{1}{2} + 0.5 \cdot 0 = \frac{6}{24} & \text{if } 0 < x \leq 1 \\ 0.5 \cdot \frac{1}{2} + 0.5 \cdot \frac{1}{3} = \frac{10}{24} & \text{if } 1 < x \leq 2 \\ 0.5 \cdot 0 + 0.5 \cdot \frac{1}{3} = \frac{4}{24} & \text{if } 2 < x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

Using Bayes rule we can then determine a functional form for $\eta(X)$

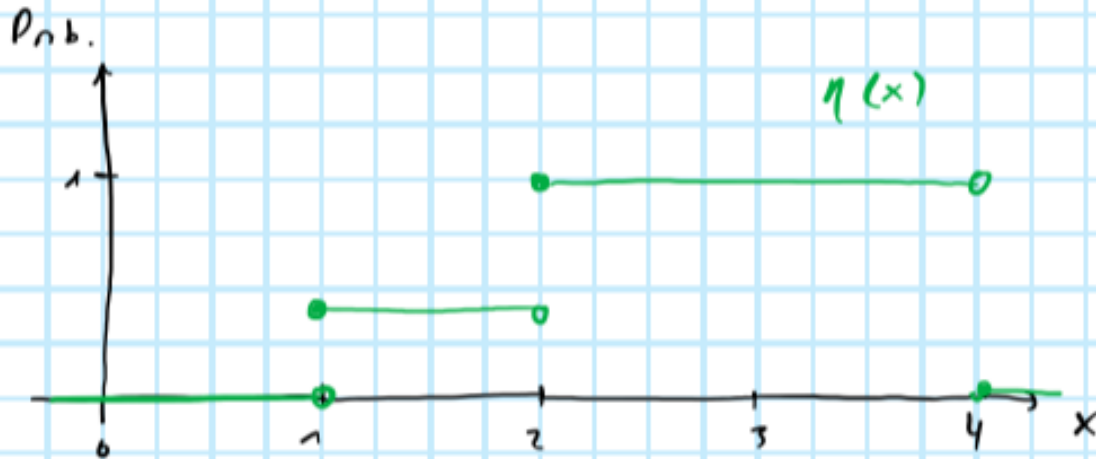
Bayes rule

$$\eta(x) = \frac{P(X=x|Y=1) P(Y=1)}{P(X=x)}$$

$$\eta(x) = \begin{cases} 0 \cdot \frac{1}{2} \cdot \frac{24}{6} = 0 & \text{if } 0 < x \leq 1 \\ \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{24}{10} = 0.4 & \text{if } 1 < x \leq 2 \\ \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{24}{4} = 1 & \text{if } 2 < x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

which can be illustrated as follows:

Bayes classifier and risk



1.1 Bayes classifier and risk

Then we have for the Bayes classifier and corresponding risk:

Classifier :

$$g^*(x) = \begin{cases} 1 & \text{if } 2 < x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

Bayes risk :

$$\begin{aligned} R^* &= \mathbb{E} \left[\mathbb{1}_{\eta(x) \leq \frac{1}{2}} \eta(x) + \mathbb{1}_{\eta(x) > \frac{1}{2}} (1 - \eta(x)) \right] \\ &= \mathbb{E} \min [\eta(x), (1 - \eta(x))] \end{aligned}$$

$$\begin{aligned} R^* &= \mathbb{E} \begin{cases} 0.4 & 1 < x \leq 2 \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{5}{12} \cdot \frac{4}{10} = \frac{20}{120} = \frac{1}{6} \end{aligned}$$

$\mathbb{P}(1 < x \leq 2)$

1.2 1-NN

For the 1-NN we can not that the following holds asymptotically:

Since $\eta(x)$ is (piece-wise) continuous we have
but

$$d(x, x_{(n)}(x)) \rightarrow 0$$

$$\Rightarrow x \approx x_{(n)}(x)$$

$$\Rightarrow \eta(x) \approx \eta(x_{(n)}(x))$$

for sufficiently large n (i.e. asymptotically).

So, the probability of error is

$$R(g_n) = P(Y_{(n)}(x) \neq Y \mid D_n) = E \left[P(Y_{(n)}(x) \neq Y \mid x, D_n) \mid D_n \right]$$

label of
nearest
neighbour

$$\approx P(Y' \neq Y \mid x)$$

where Y', Y have the same distribution given x
and Y' is conditionally independent of Y

We have $Y \sim \text{Bern}(\eta)$

$Y' \sim \text{Bern}(\eta)$

$$\begin{aligned} P(Y \neq Y') &= P(Y=1, Y'=0) + P(Y=0, Y'=1) \\ &= \eta(1-\eta) + \eta(1-\eta) \\ &= 2\eta(1-\eta) \end{aligned}$$

And consequently applying this here we get:

And hence

$$R^{1NN}(g_n) \approx E(2\eta(x)(1-\eta(x)))$$

Plugging in for η

$$R^{1NN}(g_n) = E \left\{ \begin{array}{ll} 2 \cdot \frac{0.48}{0.6} & \text{if } 1 < x \leq 2 \\ 0 & \text{otherwise} \end{array} \right. = \frac{5}{12} \cdot 0.48 = \frac{50}{120} \cdot \frac{48}{100} = \frac{2}{10}$$

1.3 3-NN

Similarly, we can show for the 3-NN rule:

3NN - classifier

By the same logic as for 1NN:

$$\begin{aligned}
 & \mathbb{P}(\text{majority}(Y_1', Y_2', Y_3') \neq Y) = \\
 &= \mathbb{P}(Y=1, \text{maj}(Y_1', Y_2', Y_3')=0) + \mathbb{P}(Y=0, \text{maj}(Y_1', Y_2', Y_3')=1) \\
 &= \eta \mathbb{P}(\text{maj}(Y_1' \dots) = 0) + (1-\eta) \mathbb{P}(\text{maj}(Y_1' \dots) = 1) \\
 &= \eta [(1-\eta)^3 + 3(1-\eta)^2 \eta] + (1-\eta) [\eta^3 + 3(1-\eta)\eta^2] \\
 &= \eta(1-\eta) [(1-\eta)^2 + 6\eta(1-\eta) + \eta^2]
 \end{aligned}$$

So, the asymptotic prob. of error of the 3 NN classifier

$$\begin{aligned}
 R^{3NN} &= \lim_{n \rightarrow \infty} \mathbb{E}(R(g_n)) = \mathbb{E}[\eta(x)(1-\eta(x))[(1-\eta(x))^2 + \eta(x)^2 + 6\eta(x)(1-\eta(x))] + \\
 &= \mathbb{E}[\eta(x)(1-\eta(x))] + 4 \mathbb{E}[\eta(x)^2(1-\eta(x))^2]
 \end{aligned}$$

$$R^{3NN}(g_n) = \mathbb{E} \begin{cases} 0.24 + 4[0.4^2 \cdot 0.6^2] = 0.4704 & \text{if } 1 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

$$= \frac{5}{12} \cdot 0.4704 = 0.196$$

2 Nearest neighbor regression

2.1 Maths

We can derive the optimal predictor as follows:

$$\begin{aligned} f^* &= \underset{f}{\operatorname{argmin}} R(f) \\ &= \underset{f}{\operatorname{argmin}} \mathbb{E} [f(x) - Y]^2 \\ &= \underset{f}{\operatorname{argmin}} \mathbb{E}_x \left[\mathbb{E}_{Y|X} [f(x) - Y]^2 | x \right] \\ f^* &= \underset{f}{\operatorname{argmin}} \mathbb{E}_{Y|X} [(f - Y)^2 | x] \\ \text{FOC: } \frac{d}{df} \mathbb{E}_{Y|X} [(f - Y)^2 | x] &= \mathbb{E}_{Y|X} [2f - 2Y | x] = 0 \\ \Rightarrow 2f - \mathbb{E}_{Y|X} [2Y | x] &= 0 \\ \Rightarrow f^* &= \mathbb{E}_{Y|X} [Y | x] \end{aligned}$$

In other words, the optimal predictor f^* of Y given X is just the conditional mean of Y given X .

Then the Bayes risk just corresponds to the irreducible error $R^* = \sigma_\varepsilon^2$:

$$\begin{aligned} R^* &= R(f^*) = \mathbb{E} [(f^*(x) - Y)^2 | x] \\ &= \mathbb{E} [(\mathbb{E}(Y | x) - Y)^2 | x] \\ Y &= \mathbb{E}(Y | x) + \varepsilon \\ \text{irreducible error} &= \mathbb{E} [(\mathbb{E}(Y | x) - \mathbb{E}(Y | x) - \varepsilon)^2 | x] \\ &= \mathbb{E} [\varepsilon^2 | x] \\ &= \sigma_\varepsilon^2 \end{aligned}$$

Now note that for any KNN regressor we can decompose its risk as follows:

$$\hat{f}_n^{KNN}(x) = \frac{1}{k} \sum_{\ell=1}^k Y_{(\ell)}(x)$$

Then

$$R(\hat{f}_n^{KNN}) = \mathbb{E}[(Y - \hat{f}_n^{KNN}(x))^2 | x = x] =$$

$$= \mathbb{E}[(f(x) + \varepsilon - \hat{f}_n^{KNN}(x))^2 | x = x]$$

$$\mathbb{E}(\varepsilon | x=x) = 0 \quad = \mathbb{E}[f(x) + \varepsilon)^2 - 2(f(x) + \varepsilon) \hat{f}_n^{KNN}(x) + \hat{f}_n^{KNN}(x)^2 | x = x]$$

$$= \mathbb{E}[f(x)^2 + \varepsilon^2 - 2f(x) \hat{f}_n^{KNN}(x) + \hat{f}_n^{KNN}(x)^2 | x = x]$$

$$\mathbb{E}(\varepsilon^2 | x=x) = \sigma_\varepsilon^2 + \mathbb{E}[(\hat{f}_n^{KNN}(x) - f(x))^2 | x = x]$$

$$= \sigma_\varepsilon^2 + \mathbb{E}[(\hat{f}_n^{KNN}(x) - \mathbb{E} \hat{f}_n^{KNN}(x) + \mathbb{E} \hat{f}_n^{KNN}(x) - f(x))^2 | x = x]$$

cross-term is 0

$$= \sigma_\varepsilon^2 + \mathbb{E}[(\hat{f}_n^{KNN}(x) - \mathbb{E} \hat{f}_n^{KNN}(x))^2 + (\mathbb{E} \hat{f}_n^{KNN}(x) - f(x))^2 | x = x]$$

$$= \sigma_\varepsilon^2 + \mathbb{E}[(\hat{f}_n^{KNN}(x) - \mathbb{E} \hat{f}_n^{KNN}(x))^2 | x = x] + \mathbb{E}[(\mathbb{E} \hat{f}_n^{KNN}(x) - f(x))^2 | x = x]$$

$$= \sigma_\varepsilon^2 + \underbrace{(\mathbb{E} \hat{f}_n^{KNN}(x) - f(x))^2}_{\text{sq. bias}} + \underbrace{\mathbb{E}[(\hat{f}_n^{KNN}(x) - \mathbb{E} \hat{f}_n^{KNN}(x))^2 | x = x]}_{\text{variance}}$$

↓
irreducible error

where for the variance term we can further simplify:

$$\text{Var}(\hat{f}_n^{KNN}(x)) = \text{Var}\left(\frac{1}{k} \sum_{\ell=1}^k Y_{(\ell)}(x)\right)$$

$$= \frac{1}{k^2} \text{Var}\left(\sum_{\ell=1}^k Y_{(\ell)}(x)\right)$$

||

$$= \frac{1}{k^2} \sum_{\ell=1}^k \text{Var}(Y_{(\ell)}(x))$$

ident. distr.

$$= \frac{1}{k} \sigma_\varepsilon$$

Then asymptotically we can show:

$$\begin{aligned}
\lim_{n \rightarrow \infty} \mathbb{E} R(f_n^{kNN}) &= \sigma^2_{\varepsilon} + \lim_{n \rightarrow \infty} \left(\mathbb{E} f_n^{kNN}(x) - f(x) \right)^2 + \frac{\sigma^2_{\varepsilon}}{k} \\
&= \sigma^2_{\varepsilon} \left(\frac{k+1}{k} \right) + \lim_{n \rightarrow \infty} \left(\mathbb{E} \left[\frac{1}{k} \sum_{\ell=1}^k Y_{(\ell)}(x) \right] - Y_i \right)^2 \\
&= \sigma^2_{\varepsilon} \left(\frac{k+1}{k} \right) + \lim_{n \rightarrow \infty} \left(d \left(\mathbb{E} \left[\frac{1}{k} \sum_{\ell=1}^k Y_{(\ell)}(x) \right], Y_i \right)^2 \right)
\end{aligned}$$

Note that $d \left(\mathbb{E} \left[\frac{1}{k} \sum_{\ell=1}^k Y_{(\ell)}(x) \right], Y_i \right) \rightarrow 0$ as $n \rightarrow \infty$

Then it is finally easy to see the asymptotic risks of the 1NN and KNN regressor:

For 1NN

$$\lim_{n \rightarrow \infty} \mathbb{E} R(f_n^{1NN}) = 2 \sigma^2_{\varepsilon} = 2 R^*$$

For kNN

$$\lim_{n \rightarrow \infty} \mathbb{E} R(f_n^{kNN}) = \left(\frac{k+1}{k} \right) \sigma^2_{\varepsilon} = \left(\frac{k+1}{k} \right) R^*$$

2.2 Program

The KNN regressor can be implemented in R as follows:

```

knn_regressor <- function(X,y,k, ...) {
  row_idx <- 1:nrow(X)
  distances <- data.table(t(combn(row_idx,2)))
  distances[,dist:=c(dist(X,...))]
  distances_rev <- copy(distances)
  setnames(distances_rev, c("V1", "V2"), c("V2", "V1"))
  distances <- rbind(distances, distances_rev)
  setorder(distances, V1, dist)
  setnames(distances, c("V1", "V2"), c("X", "neighbour"))
  distances[,y_neighbour:=y[neighbour]]
  fitted <- distances[,mean(y_neighbour[1:k]),by=X]$V1
  return(fitted)
}

```

Using this regressor I simulate data multiple times, fit and predict from the regressor and compute the mean squared error each time. I vary the dimensions d , the number of neighbours k and the sample size n .

Below in Figure 1 I plot the resulting estimated root mean squared error (RMSE) against (log) sample size for different values of dimension d . Each column represents a different choice of K increasing from left to right. We can make three important and intuitive observations:

1. The probability of error decreases with sample size. This is expected since asymptotically the distance between the label of x_i and its neighbours labels approaches zero.
2. The probability of error decreases somewhat with K (for large values of n). This reflects that the complexity of the KNN regressor is inversely related to K and hence its variance decreases as we increase the number of neighbours. Since asymptotically the bias of the KNN regressor is zero anyway, we can do better by decreasing the variance of the estimator through higher choices of K .
3. The probability of error for fixed n explodes as the dimensionality increases. This is driven by the fact that the bias – i.e. is the distance between x_i and its neighbours – takes longer to converge to zero as d grows. In particular, that distance is typically of the order of $n^{-\frac{1}{d}}$.

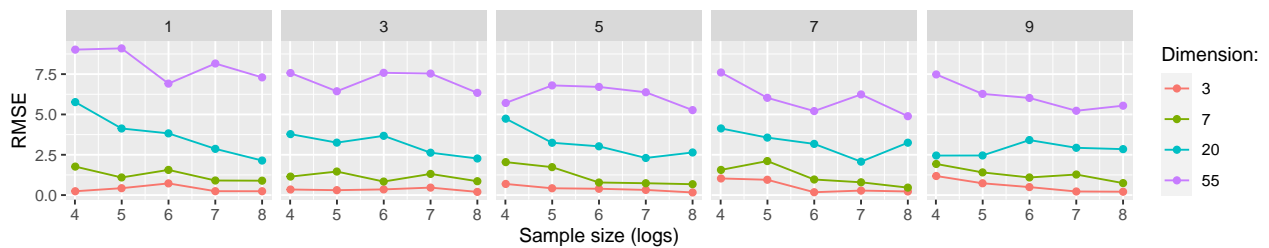


Figure 1: Root mean squared error (RMSE) of KNN regressor for different sample sizes and dimensions. The number of K nearest neighbours increases from left to right.

3 Bayes risk $R_k^* \leq e^{-ck}$

Disclaimer: I have spent quite some time on this one and thought of various approaches, but for none of them really managed to convince myself.

3.1 Using majority rule

An approaching using the majority rule is as follows:

Consider the Bayes classifier for observation pair (x_i, Y)

$$g^*(x_i) = \begin{cases} 1 & \text{if } \eta(x_i) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

We know that its Bayes risk is $R^* \in (0, \frac{1}{2})$.

Then since R_k^* is optimal by definition it must simply follow the majority rule with respect to x_1, \dots, x_k .

$$R_k^* = \mathbb{P}(g^*(x) \neq Y)$$

Consider $g(x_i) = \begin{cases} 1 & \eta(x_i) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$

$$= \mathbb{P}(\text{majority}(g(x_1), g(x_2), \dots, g(x_k)) \neq Y)$$

$$\leq \mathbb{P}(\text{Bin}(k, R^*) \geq \frac{k}{2})$$

$$= \mathbb{P}(\text{Bin}(k, R^*) - \mathbb{E} \text{Bin}(k, R^*) \geq \frac{k}{2} - \mathbb{E} \text{Bin}(k, R^*))$$

$$= \mathbb{P}(\text{Bin}(k, R^*) - kR^* \geq \frac{k}{2} - kR^*)$$

$$= \mathbb{P}(\frac{1}{k} \text{Bin}(k, R^*) - R^* \geq \frac{1}{2} - R^*)$$

$$\leq e^{-2(\frac{1}{2} - R^*)k}$$

And hence:

Since $R^* \in (0, \frac{1}{2})$ we have $c = 2(\frac{1}{2} - R^*) > 0$. Hence

$$R_k^* \leq e^{-ck}$$

A few intuitive results:

- R_k^* decreases as k increases and R^* decreases.
- For large k , $R_k^* < R^*$

3.2 Cases

Another approach I thought about was to look at the two cases of the minimum. By the conditional independence assumption we have:

$$\begin{aligned} \eta(x) &= P(Y=1 | \mathbf{x}=\mathbf{x}) = \frac{P(\mathbf{x}=\mathbf{x} | Y=1) P(Y=1)}{P(\mathbf{x}=\mathbf{x})} \\ &= \frac{\prod_{i=1}^n P(x_i=x_i | Y=1) P(Y=1)}{P(\mathbf{x}=\mathbf{x})} \\ \eta(x) &= \frac{(P(x_i=x_i | Y=1))^n P(Y=1)}{P(\mathbf{x}=\mathbf{x})} \end{aligned}$$

Now notice that $P(Y=1)$ is constant w.r.t. x_i and $P(\mathbf{x}=\mathbf{x})$ is a normalizing constant.

Hence, consider for a single i

$$\eta(x_i) = P(x_i=1 | Y=1) \cdot c$$

where $c = \frac{P(Y=1)}{P(\mathbf{x}=\mathbf{x})}$

By (1) we know that

$$c \eta(x_i) \stackrel{\text{iid}}{\sim} f(\eta(x_i))$$

Then we can look at the two cases as follows:

Case 1 $\operatorname{argmin}(\eta(x), 1 - \eta(x)) = 1 - \eta(x)$

Then
$$\begin{aligned} R_k^* &= c \cdot \mathbb{E} (1 - \eta(x))^k \\ &\leq e^{-\eta(x)k} \quad \text{by } (1-x) \leq e^{-x} \\ &= e^{-ck} \end{aligned}$$

where $c = \eta(x) > 0$.

Similarly we can argue for case 2 where $\operatorname{argmin} = \eta(x)$. \square

4 NN for binary classification

4.1 Maths

Bayes risk and the asymptotic risks of the different KNN classifiers involves an expectation with respect to $\eta(\mathbf{X}) = \frac{x^{(1)} + x^{(2)}}{2}$. To solve for that we first need to know the density of $z = x^{(1)} + x^{(2)}$. It turns out that z follows a *triangular* distribution with density

$$f(z) = \begin{cases} z & \text{if } 0 < z < 1 \\ 2 - z & \text{if } 1 \leq z < 2 \\ 0 & \text{otherwise} \end{cases}$$

This can be shown as follows:

$$\begin{aligned} f_z(z) &= \int_{-\infty}^{\infty} f_x(x) f_y(z-x) dx \\ &= \int_{-\infty}^{\infty} f(x) f(z-x) dx \end{aligned} \quad \text{Since } f_x = f_y$$

Since $x, y \sim \text{unif}(0,1)$, this implies that

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

so the integrand $f(x)f(z-x) \in \{0,1\}$

more specifically,

$$f(x)f(z-x) = \begin{cases} 1 & \text{if } 0 < x < 1 \text{ and } 0 < z-x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Clearly $f(z) = 0$ if $z < 0$ or $z \geq 2$. So we are dealing with

$$f_z(z) = \int_0^z f(x)f(z-x) dx$$

Then consider the following two cases: (1) $0 < z \leq 1$ and (2) $1 < z < 2$. For these cases we have the following:

(1) we need $x \leq z$ for $0 < z - x < 1$, $f(x) f(z - x) = 1$

Then :

$$f_z(z) = \int_0^z 1 \, dx = z$$

(2) we need $x \geq 1 - z$ for $z - x \leq 1$

Then :

$$f_z(z) = \int_{z-1}^1 1 \, dx = 2 - z$$

Now for our specific example we have

$$f(x^{(1)} + x^{(2)}) = \begin{cases} x^{(1)} + x^{(2)} & \text{if } 0 < x^{(1)} + x^{(2)} < 1 \\ 2 - (x^{(1)} + x^{(2)}) & \text{if } 1 \leq x^{(1)} + x^{(2)} < 2 \\ 0 & \text{otherwise} \end{cases}$$

Using this density function we can now solve for the expectation.

4.1.1 Bayes risk

The Bayes risk is $R^* = \frac{1}{3}$ which can be computed as follows:

$$R^* = \mathbb{E} \min \left(\frac{x^{(1)} + x^{(2)}}{2}, 1 - \frac{x^{(1)} + x^{(2)}}{2} \right)$$

$$\text{Let } x^{(1)} + x^{(2)} = z$$

$$= \int_{-\infty}^{\infty} \min \left(\frac{z}{2}, 1 - \frac{z}{2} \right) f(z) dz$$

$$= 0 + \int_0^2 \min \left(\frac{z}{2}, 1 - \frac{z}{2} \right) f(z) dz$$

$$= \int_0^1 \min \left(\frac{z}{2}, 1 - \frac{z}{2} \right) f(z) dz + \int_1^2 \min \left(\frac{z}{2}, 1 - \frac{z}{2} \right) f(z) dz$$

$$= \int_0^1 \frac{z}{2} dz + \int_1^2 \left(1 - \frac{z}{2} \right) dz$$

$$= \int_0^1 \frac{z^2}{2} dz + \frac{1}{2} \int_1^2 (z^2 - 4z + 4) dz$$

$$= \left[\frac{1}{6} z^3 \right]_0^1 + \frac{1}{2} \left[\frac{1}{3} z^3 - 2z^2 + 4z \right]_1^2$$

$$= \frac{1}{6} + \frac{1}{2} \left(\left[\frac{1}{3} \cdot 8 - 8 + 8 \right] - \left[\frac{1}{3} - 2 + 4 \right] \right)$$

$$= \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{3}$$

4.1.2 1NN

The risk of the 1NN classifier is $R^{1NN} = \frac{5}{12}$ which can be shown as follows:

$$\begin{aligned}
 R^{1NN} &= 2 \mathbb{E} \left[\frac{z}{2} \left(1 - \frac{z}{2} \right) \right] \\
 &= 2 \int_0^2 \frac{z}{2} \left(1 - \frac{z}{2} \right) f(z) dz \\
 &= 2 \left[\int_0^1 \frac{z}{2} \left(1 - \frac{z}{2} \right) z dz + \int_1^2 \frac{z}{2} \left(1 - \frac{z}{2} \right) (2-z) dz \right] \\
 &= 2 \left[\frac{1}{2} \int_0^1 z^2 - \frac{z^3}{2} dz + \frac{1}{2} \int_1^2 \left(\frac{z^3}{2} - z z^2 + z z \right) dz \right] \\
 &= \left[\frac{1}{3} z^3 - \frac{1}{8} z^4 \right]_0^1 + \left[\frac{1}{8} z^4 - \frac{2}{3} z^3 + z^2 \right]_1^2 \\
 &= \frac{5}{24} + \left[\left(2 - \frac{16}{3} + \frac{2}{1} \frac{16}{24} \right) - \left(\frac{1}{8} - \frac{2}{3} + 1 \right) \right] \\
 R^{1NN} &= \frac{10}{24} = \frac{5}{12}
 \end{aligned}$$

A quick sanity check shows that

$$R^{1NN} = \frac{5}{12} < \frac{4}{9} = 2R^*(1 - R^*)$$

4.1.3 3NN

The risk of the 3NN classifier is $R^{3NN} = \frac{47}{120}$ which can be shown as follows:

$$\begin{aligned}
R^{3NN} &= \mathbb{E} \left[\eta(x) (1 - \eta(x)) \right] + 4 \mathbb{E} \left[\eta(x)^2 (1 - \eta(x))^2 \right] \\
&= \frac{5}{24} + 4 \int_0^2 \frac{z^2}{4} \left(\frac{2-z}{2} \right)^2 f(z) dz \\
&= \frac{5}{24} + \frac{1}{4} \int_0^2 z^2 (2-z)^2 f(z) dz \\
&= \frac{5}{24} + \frac{1}{4} \left[\int_0^1 z^2 (2-z)^2 dz + \int_1^2 z^2 (2-z)^2 dz \right] \\
&= \frac{5}{24} + \frac{1}{4} \left[\int_0^1 z^5 - 4z^4 + 4z^3 dz + \int_1^2 z^2 (2^3 - 3 \cdot 2^2 \cdot z + 3 \cdot 2 \cdot z^2 - z^3) dz \right] \\
&= \frac{5}{24} + \frac{1}{4} \left[\left[\frac{1}{6} z^6 - \frac{4}{5} z^5 + z^4 \right]_0^1 + \int_1^2 (-z^5 + 6z^4 - 12z^3 + 8z^2) dz \right] \\
&= \frac{5}{24} + \frac{1}{4} \left[\left[\frac{1}{6} - \frac{4}{5} + 1 \right] + \left[-\frac{1}{6} z^6 + \frac{6}{5} z^5 - 3z^4 + \frac{8}{3} z^3 \right]_1^2 \right] \\
&= \frac{5}{24} + \frac{1}{4} \left[\frac{11}{30} + \left(-\frac{64}{6} + \frac{32}{5} - 3 \cdot 16 + \frac{64}{3} \right) - \left(-\frac{1}{6} + \frac{6}{5} - 3 + \frac{8}{3} \right) \right] \\
&= \frac{5}{24} + \frac{1}{4} \left[\frac{11}{30} + \frac{11}{30} \right] \\
&= \frac{5}{24} + \frac{1}{2} \cdot \frac{11}{30} = \frac{5}{24} + \frac{11}{60} = \frac{300}{1440} + \frac{264}{1440} = \frac{564}{1440} = \frac{47}{120} \\
R^{3NN} &= \frac{47}{120}
\end{aligned}$$

In conclusion we have that:

$$R^* = \frac{40}{120} < R^{3NN} = \frac{47}{120} < R^{1NN} = \frac{50}{120}$$

At the limit where $n \rightarrow \infty$ these asymptotic quantities do not depend on the dimensionality. This is because the bias of the KNN classifier – that is the distance between x_i and its neighbours – eventually converges to zero as n grows very large. As we saw earlier for the KNN regressor, this is the only point where the dimensionality comes into play, and evidently at the limit it has no effect on the probability of error. However, the dimensionality does affect the *rate* of convergence as we discussed in class. In other words, the higher d , the longer it takes to reach the asymptotic quantities, which will become more evident in the following section.

4.2 Program

The KNN classifier can be implemented in R as follows:

```

knn_classifier <- function(X,y,k,...) {
  row_idx <- 1:nrow(X)
  distances <- data.table(t(combn(row_idx,2)))
  distances[,dist:=c(dist(X,...))]
  distances_rev <- copy(distances)
  setnames(distances_rev, c("V1", "V2"), c("V2", "V1"))
  distances <- rbind(distances, distances_rev)
  setorder(distances, V1, dist)
  setnames(distances, c("V1", "V2"), c("X", "neighbour"))
  distances[,y_neighbour:=y[neighbour]]
  fitted <- distances[,median(y_neighbour[1:k]),by=X]$V1
  return(fitted)
}

```

To simulate the data I use the following helper function:

```

sim_data <- function(n,d) {
  X <- matrix(runif(n*d),n) # uniform 0,1
  p_y <- rowSums(X[,1:2])/2 # probabilities of each Bernoulli trial
  y <- rbinom(n, 1, p_y)
  return(list(X=X,y=y))
}

```

I then run the program multiple times varying the dimension d , the number of neighbours k and the sample size n . Each time I compute the frequency of error and finally average over those frequencies to obtain an estimate of the probability of error.

Below in Figure 2 I plot the resulting estimated probability of error against (log) sample size for different values of dimension d . Each column represents a different choice of K increasing from left to right. We can make three important and intuitive observations, very similar to what we have seen earlier for the KNN regressor.

1. The probability of error decreases with sample size. This is expected since asymptotically the distance between the label of x_i and its neighbours labels approaches zero.
2. The probability of error decreases with K (for large values of n). This reflects that the complexity of the KNN classifier is inversely related to K and hence its variance decreases as we increase the number of neighbours. Since asymptotically the bias of the KNN classifier is zero anyway, we can do better by decreasing the variance of the estimator through higher choices of K .
3. The probability of error for fixed n increases as the dimensionality increases. This is driven by the fact that the bias – i.e. is the distance between x_i and its neighbours – takes longer to converge to zero as d grows. In particular, that distance is typically of the order of $n^{-\frac{1}{d}}$. Contrary to the KNN regressor the distance between labels of x_i and its neighbours is of course bounded in this case, so we do not observe the same explosive increase in the error.

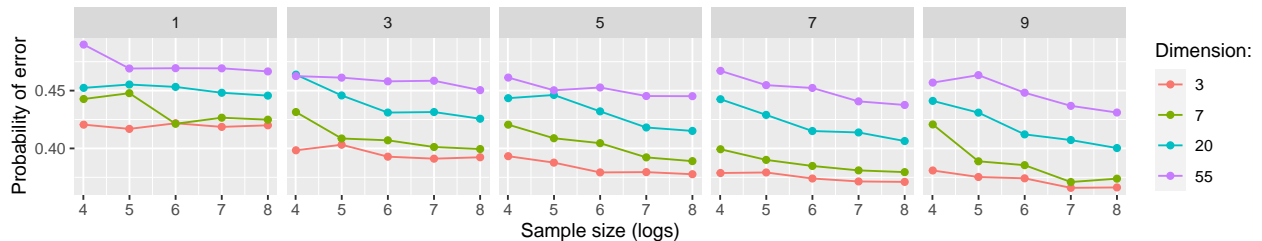


Figure 2: Probability of error of KNN classifier for different sample sizes and dimensions. The number of K nearest neighbours increases from left to right.