Question 8

Consider a binary classification problem in which X is uniformly distributed in $[0,1]^d$ and $\eta(x) = (1/d)\sum_{i=1}^{d} x_i$ (where $x_1, ..., x_d \in [0,1]$ are the components of x). Compute the asymptotic risk of the 1-nearest neighbor rule.

Write a program that generates training data of n i.i.d. pairs (X1, Y1), . . . , (Xn, Yn) according to the distribution above. Classify X using the 1, 3, 5, 7, 9-nearest neighbor rules. Re-draw (X, Y) many times so that you can estimate the risk of these rules.

Try this for various values of n and d and plot the estimated risk. Compare the estimated risk to the asymptotic risk of the 1-nearest neighbor rule.

Solution:

We know that the asymptotic risk of the nearest neighbor rule is $2\mathbb{E}[\eta(x)(1-\eta(x))]$. Then,

$$\begin{aligned} 2\mathbb{E}[\eta(x)(1-\eta(x))] &= 2\mathbb{E}[(\frac{1}{d}\sum_{i=1}^{d} X_i)(1-\frac{1}{d}\sum_{i=1}^{d} X_i)] \\ &= \frac{2}{d}\sum_{i=1}^{d}\mathbb{E}[X_i] - \frac{2}{d^2}\mathbb{E}[(\sum_{i=1}^{d} X_i)^2] \end{aligned}$$

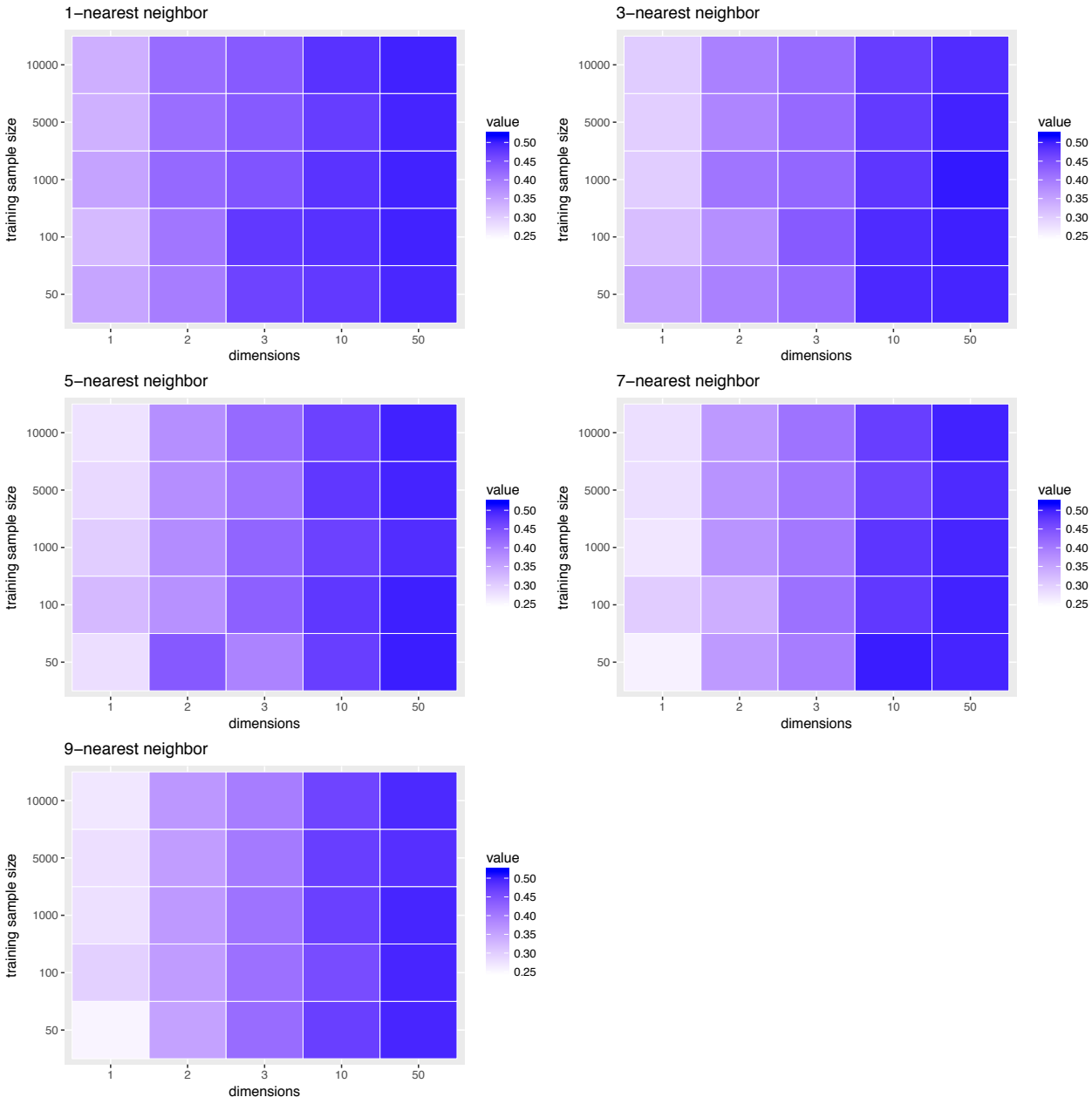
Since $\mathbb{E}[(\sum_{i=1}^{d} X_i)^2] - \mathbb{E}[(\sum_{i=1}^{d} X_i)]^2 = var(\sum_{i=1}^{d} X_i)$ with i.i.d. X, $\mathbb{E}[(\sum_{i=1}^{d} X_i)^2] = \frac{d}{12} + \frac{d^2}{4}$ since the mean and variance of a sum of independent uniform random variables over [0,1] is $\frac{d}{2}$ and $\frac{d}{12}$ respectively. Continuing,

$$\begin{aligned} R^{1-NN} = \frac{2}{d}\sum_{i=1}^{d}\mathbb{E}]X_i] - \frac{2}{d^2}\mathbb{E}[(\sum_{i=1}^{d} X_i)^2] &= 1 - \frac{2}{d^2}(\frac{d}{12} + \frac{d^2}{4}) \\ &= \frac{1}{2}(1 - \frac{1}{3d}) \end{aligned}$$

We find the risk of the 1,3,5,7,9-nearest neighbor rules initially for sample sizes 50, 100, 1000, 5000, 10000 and for various dimensions. In particular, we know that in this case, for the 1-nearest neighbor rule, the risk depends on the number of dimensions (although this is true for other nearest neighbor rules as well). Given the formula for the risk, we see that as d increases, R^{1-NN} clearly approaches 0.5 very quickly. It might be more interesting to look more at the behavior of lower dimension d for more variation in estimated risk values. I choose $d \in \{1, 2, 3, 10, 50\}$.

$$\begin{aligned} R_{d=1}^{1-NN} &= \frac{1}{3} = 0.3333 \\ R_{d=2}^{1-NN} &= \frac{5}{12} = 0.4166 \\ R_{d=3}^{1-NN} &= \frac{4}{9} = 0.4444 \\ R_{d=10}^{1-NN} &= \frac{29}{60} = 0.4833 \\ R_{d=50}^{1-NN} &= \frac{149}{300} = 0.4966 \end{aligned}$$

We compare these values with the empirical values obtained (with the largest n value since the empirical value will be more accurate with a larger sample). I use a test sample of 10000 for each combination of n and d.



1. For a fixed training sample size, for every nearest-neigbor rule, the estimated risk increases with the number of dimensions

In general, the pairwise distance of a fixed number of randomly sampled points is likely to be larger in higher dimensions, as we've shown in problem set 1. For a continuous distribution, the nearest-neighbor rule is more likely to work when the neighbors are close to the point whose label is being predicted (if a point has a label 1, another point which is very close to the first point is more likely to have a label 1 as well). However, for a fixed number of points, the points are farther apart the higher the dimension. Thus, our training data is less likely to be a good predictor in higher dimensions.

It is also worth noting in this case that the conditional distribution we're drawing from also tells us that the asymptotic risk increases with the number of dimensions up till 0.5. Note that this trend is true regardless of which nearest-neighbor rule we're using. This is because the aforementioned reasoning still applies: in higher dimensional spaces, all k neighbors are far away from the test point and are bad predictors of the test point's label.

2. Lower estimated risk when more neighbors are used to predict at low dimensions

We know that the distance between (X, k-nearest neighbors of X) $\leq \epsilon$ with high probability even at higher dimensions if $n\epsilon^d$ is large or when $n >> \frac{1}{e}^d$. This means that we need an exponentially large sample size for this statement to be true. Otherwise we experience the curse of dimensionality (points being far apart and being bad predictions of nearest neighbors).

However, the equation we found at the start of this question tells us that regardless of the value of n, the risk approaches 0.5 at high dimensions. This says that the provied distribution of data is such that regardless of the sample size of the training set, the nearest neighbor to a point tells us no information about the label of the point in question at high dimensions (we might as well flip a coin to predict the label of the point). This is a property of the particular space/distribution of the data we consider in this question. The implication is that there is something about higher dimensions in this example which removes the predictive capabilities of even neighboring points which are arbitrarily close to the test point.

The heat maps look mostly similar, but if one inspects the array of probabilities and looks closer at the heat maps, we see that the estimated risk decreases for all dimensions when more neighbors are used to predict the label of the test data up till dimension 10 (when looking at the largest training size). For dimension 10, the estimated risk decreases from 1-NN to 7-NN and then increases from 7-NN to 9-NN. For dimension 50, the estimated risk decreases from 1-NN to 5-NN and increases from 5-NN to 9-NN. This is probably because for with a (maximum) training sample size of 10000, at dimensions approx. 10 and higher, the space is so large that the the k number of nearest neighbors are simply not close enough to be good predictors of the test data. The higher the dimension, the less relevant each marginal neighbor is.

Comparing asymptotic and estimated risk for 1-NN rule

We use the estimated risk for the largest sample size and compare it with the pre-calculated asymptotic risk for the 1-Nearest Neighbors rule. It turns out that the estimated risk and the asymptotic risk are very similar to each other (even for lower sample sizes except at the lowest sample size of n=50, which had more fluctuation). This makes sense since we're sampling from the same conditional distribution which tells us the value of the asymptotic risk.

It also makes sense that there is a bigger fluctuation away from the asymptotic risk at lower training sizes since there's a greater chance that the drawn training sample is not representative of the true distribution, thus they would not be good predictors for the test data.

Table 1: Comparing empirical and asymptotic risk

d	asymptotic risk	n=50	n=100	n=1000	5000	10000
1	0.3333	0.3486	0.3223	0.3510	0.3331	0.3355
2	0.4166	0.3977	0.4063	0.4222	0.4160	0.4180
3	0.4444	0.4666	0.4782	0.4471	0.4401	0.4398
10	0.4833	0.4774	0.4864	0.4842	0.4738	0.4851
50	0.4966	0.4956	0.5003	0.5004	0.4982	0.5010