

Risk Analysis for Nearest-Neighbor

We investigate the problem of bounding the zero-one loss risk of the 1-NN binary classifier averaged with respect to the random draw of the training set. Under some assumptions on the data distribution \mathcal{D} , we prove a bound of the form

$$\mathbb{E}[\ell_{\mathcal{D}}(h_S)] \leq 2\ell_{\mathcal{D}}(f^*) + \varepsilon_m \quad (1)$$

where h_S denotes the 1-NN classifier generated on the training set S of size m , $\ell_{\mathcal{D}}(f^*)$ is the Bayes risk, and ε_m is a quantity which depends on m . Note that we are able to compare $\mathbb{E}[\ell_{\mathcal{D}}(h_S)]$ directly to the Bayes risk, showing that 1-NN is in some sense a powerful learning algorithm.

Recall that in binary classification we denote the joint distribution of (\mathbf{X}, Y) with the pair (\mathcal{D}_X, η) , where \mathcal{D}_X is the marginal of \mathcal{D} on \mathbf{X} and $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[\mathbb{I}\{Y = 1\} \mid \mathbf{X} = \mathbf{x}]$. Given a training set $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, we define the map $\pi(S, \cdot) : \mathbb{R}^d \rightarrow \{1, \dots, m\}$ by

$$\pi(S, \mathbf{x}) = \operatorname{argmin}_{t=1, \dots, m} \|\mathbf{x} - \mathbf{x}_t\| .$$

If there is more than one point \mathbf{x}_t achieving the minimum in the above expression, then $\pi(S, \mathbf{x})$ selects one of them using any deterministic tie-breaking rule; our analysis does not depend on the specific rule being used. The 1-NN classifier generated by S is defined by $h_S(\mathbf{x}) = y_{\pi(S, \mathbf{x})}$.

From now on, the training set $S = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_m, Y_m)\}$ is a sample drawn i.i.d. from \mathcal{D} . The expected risk is defined by

$$\mathbb{E}[\ell_{\mathcal{D}}(h_S)] = \mathbb{P}(Y_{\pi(S, \mathbf{X})} \neq Y)$$

Where probabilities and expectations are understood with respect to the random draw of training set S (used to define h_S) and of the example (\mathbf{X}, Y) with respect to which the risk of h_S is computed.

Given any two random variables X and Z , the deterministic quantity $\mathbb{E}[Z \mid X = x]$ denotes the conditional expectation of Z given $X = x$. We use $\mathbb{E}[Z \mid X]$ to denote the conditional expectation of Z as a function of the random variable X . We often use the following facts.

Fact 1. *For any random variables X, Z and for any functions f, g, h ,*

1. $\mathbb{E}[f(X)\mathbb{E}[Z \mid X]] = \mathbb{E}[\mathbb{E}[f(X)Z \mid X]] = \mathbb{E}[f(X)Z]$
2. *If $\mathbb{E}[g(Z, x)] = \mathbb{E}[h(Z, x)]$ for all x , then $\mathbb{E}[g(Z, X) \mid X]$ and $\mathbb{E}[h(Z, X) \mid X]$ denote the same random variable.*

A special case of part 1 is $\mathbb{E}[\mathbb{E}[Z \mid X]] = \mathbb{E}[Z]$.

We now state and prove a crucial lemma.

Lemma 2. *The expected risk of the 1-NN classifier can be written as follows*

$$\mathbb{E}[\ell_{\mathcal{D}}(h_S)] = \mathbb{E}\left[\eta(\mathbf{X}_{\pi(S,\mathbf{X})}) (1 - \eta(\mathbf{X}))\right] + \mathbb{E}\left[\left(1 - \eta(\mathbf{X}_{\pi(S,\mathbf{X})})\right)\eta(\mathbf{X})\right]$$

It is important to understand that the expectations above are with respect to the random draw of both S and (\mathbf{X}, Y) . **PROOF.** Note that

$$\mathbb{E}[\ell_{\mathcal{D}}(h_S)] = \mathbb{P}(Y_{\pi(S,\mathbf{X})} \neq Y) = \mathbb{P}(Y_{\pi(S,\mathbf{X})} = 1, Y = -1) + \mathbb{P}(Y_{\pi(S,\mathbf{X})} = -1, Y = 1)$$

Therefore, we need to prove that

$$\mathbb{P}(Y_{\pi(S,\mathbf{X})} = 1, Y = -1) = \mathbb{E}\left[\eta(\mathbf{X}_{\pi(S,\mathbf{X})}) (1 - \eta(\mathbf{X}))\right] \quad (2)$$

$$\mathbb{P}(Y_{\pi(S,\mathbf{X})} = -1, Y = 1) = \mathbb{E}\left[\left(1 - \eta(\mathbf{X}_{\pi(S,\mathbf{X})})\right)\eta(\mathbf{X})\right] \quad (3)$$

In order to prove (2), observe that for every $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned} \mathbb{E}[\mathbb{I}\{Y_{\pi(S,\mathbf{x})} = 1\}] &= \sum_{t=1}^m \mathbb{E}[\mathbb{I}\{Y_t = 1\} \mathbb{I}\{\pi(S, \mathbf{x}) = t\}] && \text{(by the law of total expectation)} \\ &= \sum_{t=1}^m \mathbb{E}\left[\mathbb{E}[\mathbb{I}\{Y_t = 1\} \mathbb{I}\{\pi(S, \mathbf{x}) = t\} \mid \mathbf{X}_t]\right] && (\mathbf{X}_t \text{ is the } t\text{-th datapoint in } S) \\ &= \sum_{t=1}^m \mathbb{E}\left[\mathbb{E}[\mathbb{I}\{Y_t = 1\} \mid \mathbf{X}_t] \mathbb{E}[\mathbb{I}\{\pi(S, \mathbf{x}) = t\} \mid \mathbf{X}_t]\right] && \text{(by independence conditioned on } \mathbf{X}_t) \\ &= \sum_{t=1}^m \mathbb{E}\left[\eta(\mathbf{X}_t) \mathbb{E}[\mathbb{I}\{\pi(S, \mathbf{x}) = t\} \mid \mathbf{X}_t]\right] && \text{(by definition of } \eta(\mathbf{X}_t)) \\ &= \sum_{t=1}^m \mathbb{E}\left[\eta(\mathbf{X}_t) \mathbb{I}\{\pi(S, \mathbf{x}) = t\}\right] && \text{(using part 1 of Fact 1)} \\ &= \mathbb{E}\left[\eta(\mathbf{X}_{\pi(S,\mathbf{x})})\right] && \text{(by the law of total expectation)} \end{aligned}$$

By part 2 of Fact 1, $\mathbb{E}[\mathbb{I}\{Y_{\pi(S,\mathbf{X})} = 1\} \mid \mathbf{X}]$ and $\mathbb{E}[\eta(\mathbf{X}_{\pi(S,\mathbf{X})}) \mid \mathbf{X}]$ denote the same random variable. Therefore,

$$\begin{aligned} \mathbb{P}(Y_{\pi(S,\mathbf{X})} = 1, Y = -1) &= \mathbb{E}[\mathbb{I}\{Y_{\pi(S,\mathbf{X})} = 1\} \mathbb{I}\{Y = -1\}] \\ &= \mathbb{E}\left[\mathbb{E}[\mathbb{I}\{Y_{\pi(S,\mathbf{X})} = 1\} \mathbb{I}\{Y = -1\} \mid \mathbf{X}]\right] && \text{(special case of Fact 1)} \\ &= \mathbb{E}\left[\mathbb{E}[\mathbb{I}\{Y_{\pi(S,\mathbf{X})} = 1\} \mid \mathbf{X}] \mathbb{E}[\mathbb{I}\{Y = -1\} \mid \mathbf{X}]\right] && \text{(see below)} \\ &= \mathbb{E}\left[\mathbb{E}[\eta(\mathbf{X}_{\pi(S,\mathbf{X})}) \mid \mathbf{X}] \mathbb{E}[\mathbb{I}\{Y = -1\} \mid \mathbf{X}]\right] && \text{(part 2 of Fact 1)} \\ &= \mathbb{E}\left[\eta(\mathbf{X}_{\pi(S,\mathbf{X})}) (1 - \eta(\mathbf{X}))\right] && \text{(definition of } \eta) \end{aligned}$$

In the third step we used the fact that when conditioned on \mathbf{X} , Y is independent of everything else, including $Y_{\pi(S,\mathbf{X})}$. The proof of (3) is very similar. \square

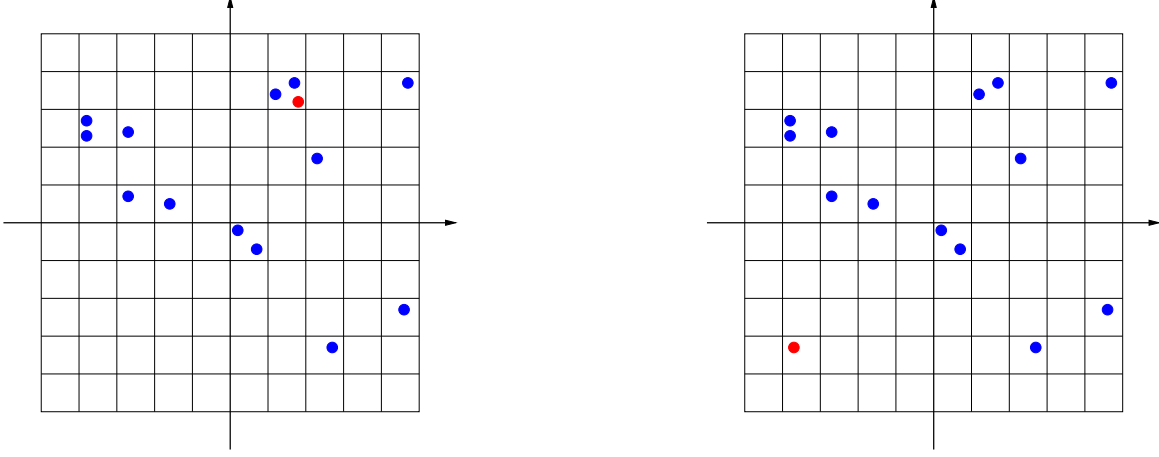


Figure 1: Bidimensional example of the construction used in the analysis of 1-NN. Left pane: \mathbf{X} (the red point) is in the same square C_i as its closest training point $\mathbf{X}_{\pi(S,\mathbf{X})}$. Hence, $\|\mathbf{X} - \mathbf{X}_{\pi(S,\mathbf{X})}\|$ is bounded by the length of the diagonal of this square. Right pane: here there are no training points in the square where \mathbf{X} lies. Hence, $\|\mathbf{X} - \mathbf{X}_{\pi(S,\mathbf{X})}\|$ can only be bounded by the length of the entire data space (the large square).

To proceed with the analysis, we now need some assumptions on D_X and η . First, all data points \mathbf{X} drawn from D_X satisfy $\max_i |X_i| \leq 1$ with probability one. In other words, $\mathbf{X} \in [-1, 1]^d$ with probability 1. Let $\mathcal{X} \equiv [-1, 1]^d$ the subsets of data points with this property. Second we assume that η is such that there exists $0 < c < \infty$ such that

$$|\eta(\mathbf{x}) - \eta(\mathbf{x}')| \leq c \|\mathbf{x} - \mathbf{x}'\| \quad \text{for all } \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

Technically, this conditions implies that η is Lipschitz in \mathcal{X} . This is a restriction on the set of all allowed η as $c < \infty$ implies continuity (but the opposite is not true). We can thus write

$$\eta(\mathbf{x}') \leq \eta(\mathbf{x}) + c \|\mathbf{x} - \mathbf{x}'\| \tag{4}$$

$$1 - \eta(\mathbf{x}') \leq 1 - \eta(\mathbf{x}) + c \|\mathbf{x} - \mathbf{x}'\| \tag{5}$$

Using (4) and (5), for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ we have

$$\begin{aligned} & \eta(\mathbf{x})(1 - \eta(\mathbf{x}')) + (1 - \eta(\mathbf{x}))\eta(\mathbf{x}') \\ & \leq \eta(\mathbf{x})(1 - \eta(\mathbf{x})) + \eta(\mathbf{x})c \|\mathbf{x} - \mathbf{x}'\| + (1 - \eta(\mathbf{x}))\eta(\mathbf{x}) + (1 - \eta(\mathbf{x}))c \|\mathbf{x} - \mathbf{x}'\| \\ & = 2\eta(\mathbf{x})(1 - \eta(\mathbf{x})) + c \|\mathbf{x} - \mathbf{x}'\| \\ & \leq 2 \min \{ \eta(\mathbf{x}), 1 - \eta(\mathbf{x}) \} + c \|\mathbf{x} - \mathbf{x}'\| \end{aligned}$$

where the last inequality holds because $z(1 - z) \leq \min\{z, 1 - z\}$ for all $z \in [0, 1]$. Therefore

$$\mathbb{E}[\ell_{\mathcal{D}}(h_S)] \leq 2 \mathbb{E} \left[\min \{ \eta(\mathbf{X}), 1 - \eta(\mathbf{X}) \} \right] + c \mathbb{E} \left[\|\mathbf{X} - \mathbf{X}_{\pi(S,\mathbf{X})}\| \right].$$

Recalling that the Bayes risk for the zero-one loss is $\ell_{\mathcal{D}}(f^*) = \mathbb{E} \left[\min \{ \eta(\mathbf{X}), 1 - \eta(\mathbf{X}) \} \right]$ we have

$$\mathbb{E}[\ell_{\mathcal{D}}(h_S)] \leq 2 \ell_{\mathcal{D}}(f^*) + c \mathbb{E} \left[\|\mathbf{X} - \mathbf{X}_{\pi(S,\mathbf{X})}\| \right].$$

In order to bound the term containing the expected value of $\|\mathbf{X} - \mathbf{X}_{\pi(S, \mathbf{X})}\|$ we partition the data space \mathcal{X} in d -dimensional hypercubes with side $\varepsilon > 0$ —see Figure 1 for a bidimensional example. Let C_1, \dots, C_r the resulting hypercubes. We can now bound $\|\mathbf{X} - \mathbf{X}_{\pi(S, \mathbf{X})}\|$ using a case analysis. Assume first that \mathbf{X} belongs to a C_i in which there is at least a training point \mathbf{X}_t . Then $\|\mathbf{X} - \mathbf{X}_{\pi(S, \mathbf{X})}\|$ is at most the length of the diagonal of the hypercube, which is $\varepsilon\sqrt{d}$ —see the left pane in Figure 1. Now assume that \mathbf{X} belongs to a C_i in which there are no training points. Then $\|\mathbf{X} - \mathbf{X}_{\pi(S, \mathbf{X})}\|$ is only bounded by the length of the diagonal of \mathcal{X} , which is $2\sqrt{d}$ —see the right pane in Figure 1. Hence, we may write

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{X} - \mathbf{X}_{\pi(S, \mathbf{X})}\| \right] &\leq \mathbb{E} \left[\varepsilon\sqrt{d} \sum_{i=1}^r \mathbb{I}\{C_i \cap S \neq \emptyset\} \mathbb{I}\{\mathbf{X} \in C_i\} + 2\sqrt{d} \sum_{i=1}^r \mathbb{I}\{C_i \cap S = \emptyset\} \mathbb{I}\{\mathbf{X} \in C_i\} \right] \\ &= \varepsilon\sqrt{d} \mathbb{E} \left[\sum_{i=1}^r \mathbb{I}\{C_i \cap S \neq \emptyset\} \mathbb{I}\{\mathbf{X} \in C_i\} \right] + 2\sqrt{d} \sum_{i=1}^r \mathbb{E} [\mathbb{I}\{C_i \cap S = \emptyset\} \mathbb{I}\{\mathbf{X} \in C_i\}] \end{aligned}$$

where in the last step we used linearity of the expected value. Now observe that, for all S and \mathbf{X} ,

$$\sum_{i=1}^r \mathbb{I}\{C_i \cap S \neq \emptyset\} \mathbb{I}\{\mathbf{X} \in C_i\} \in \{0, 1\}$$

because $\mathbf{X} \in C_i$ for only one $i = 1, \dots, d$. Therefore,

$$\mathbb{E} \left[\sum_{i=1}^r \mathbb{I}\{C_i \cap S \neq \emptyset\} \mathbb{I}\{\mathbf{X} \in C_i\} \right] \leq 1 .$$

To bound the remaining term, we use the independence between \mathbf{X} and the training set S ,

$$\mathbb{E} [\mathbb{I}\{C_i \cap S = \emptyset\} \mathbb{I}\{\mathbf{X} \in C_i\}] = \mathbb{E} [\mathbb{I}\{C_i \cap S = \emptyset\}] \mathbb{E} [\mathbb{I}\{\mathbf{X} \in C_i\}] = \mathbb{P}(C_i \cap S = \emptyset) \mathbb{P}(\mathbf{X} \in C_i) .$$

Since S contains m data points independently drawn, for a generic data point \mathbf{X}' we have that

$$\mathbb{P}(C_i \cap S = \emptyset) = (1 - \mathbb{P}(\mathbf{X}' \in C_i))^m \leq \exp(-m\mathbb{P}(\mathbf{X}' \in C_i))$$

where in the last step we used the inequality $(1 - p)^m \leq e^{-pm}$. Setting $p_i = \mathbb{P}(\mathbf{X}' \in C_i)$ we have

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{X} - \mathbf{X}_{\pi(S, \mathbf{X})}\| \right] &\leq \varepsilon\sqrt{d} + (2\sqrt{d}) \sum_{i=1}^r e^{-p_i m} p_i \\ &\leq \varepsilon\sqrt{d} + (2\sqrt{d}) \sum_{i=1}^r \max_{0 \leq p \leq 1} e^{-pm} p \\ &= \varepsilon\sqrt{d} + (2\sqrt{d}) r \max_{0 \leq p \leq 1} e^{-pm} p . \end{aligned}$$

The concave function $g(p) = e^{-pm} p$ is maximized for $p = \frac{1}{m}$. Therefore,

$$\mathbb{E} \left[\|\mathbf{X} - \mathbf{X}_{\pi(S, \mathbf{X})}\| \right] \leq \varepsilon\sqrt{d} + (2\sqrt{d}) \frac{r}{em} = \sqrt{d} \left(\varepsilon + \frac{2}{em} \left(\frac{2}{\varepsilon} \right)^d \right)$$

where we used the fact that the number r of hypercubes is equal to $\left(\frac{2}{\varepsilon}\right)^d$. Putting everything together we find that

$$\mathbb{E}[\ell_{\mathcal{D}}(h_S)] \leq 2\ell_{\mathcal{D}}(f^*) + c\sqrt{d} \left(\varepsilon + \frac{2}{em} \left(\frac{2}{\varepsilon}\right)^d \right)$$

Since this holds for all $0 < \varepsilon < 1$, we can set $\varepsilon = 2m^{-1/(d+1)}$. This gives

$$\varepsilon + \frac{2}{em} \left(\frac{2}{\varepsilon}\right)^d = 2m^{-1/(d+1)} + \frac{2^{d+1}2^{-d}m^{d/(d+1)}}{em} = 2m^{-1/(d+1)} \left(1 + \frac{1}{e}\right) \leq 4m^{-1/(d+1)}. \quad (6)$$

Substituting this bound in (6), we finally obtain

$$\mathbb{E}[\ell_{\mathcal{D}}(h_S)]t \leq 2\ell_{\mathcal{D}}(f^*) + c4m^{-1/(d+1)}\sqrt{d}.$$

This analysis allows us to draw two conclusions.

1. When $m \rightarrow \infty$, $\ell_{\mathcal{D}}(f^*) \leq \mathbb{E}[\ell_{\mathcal{D}}(h_S)] \leq 2\ell_{\mathcal{D}}(f^*)$. Namely, the asymptotic risk of 1-NN lies between the Bayes risk and twice the Bayes risk.
2. For $\mathbb{E}[\ell_{\mathcal{D}}(h_S)]$ to be at most $2\ell_{\mathcal{D}}(f^*) + \alpha$, the training set size must be at least $\left(\frac{4c}{\alpha}\sqrt{d}\right)^{d+1}$, which is exponential in the number d of features. This exponential dependence on the number of features of the training set size is known as **curse of dimensionality** and refers to the difficulty of learning when datapoints live in a high-dimensional space.

Acknowledgments

Thanks to Roberto Colomboni for helping with the proof of Lemma 2.