Machine Learning — Statistical Methods for Machine Learning

Risk Analysys for Nearest-Neighbor

Instructor: *Nicolò Cesa-Bianchi* version of April 7, 2021

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

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] \le 2\,\ell_{\mathcal{D}}(f^*) + \varepsilon_m \tag{1}$$

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 $\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big]$ 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 $(\boldsymbol{X}, Y)$ with the pair $(\mathcal{D}_X, \eta)$, where $\mathcal{D}_X$ is the marginal of $\mathcal{D}$ on $\boldsymbol{X}$ and $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}\big[\mathbb{I}\{Y = 1\} \mid \boldsymbol{X} = \boldsymbol{x}\big]$. Given a training set $S = \big\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)\big\}$, we define the map $\pi(S, \cdot) : \mathbb{R}^d \to \{1, \dots, m\}$ by

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

If there is more than one point $\boldsymbol{x}_t$ achieving the minimum in the above expression, then $\pi(S, \boldsymbol{x})$ 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 $h_S(\boldsymbol{x}) = y_{\pi(S,\boldsymbol{x})}$.

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

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

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

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

We now state and prove a crucial lemma.

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

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

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

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

Therefore, we need to prove that

$$\mathbb{P}\big(Y_{\pi(S,\boldsymbol{X})} = 1,\, Y = -1\big) = \mathbb{E}\Big[\eta\big(\boldsymbol{X}_{\pi(S,\boldsymbol{X})}\big)\,\big(1-\eta(\boldsymbol{X})\big)\Big] \tag{2}$$

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

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

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

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

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

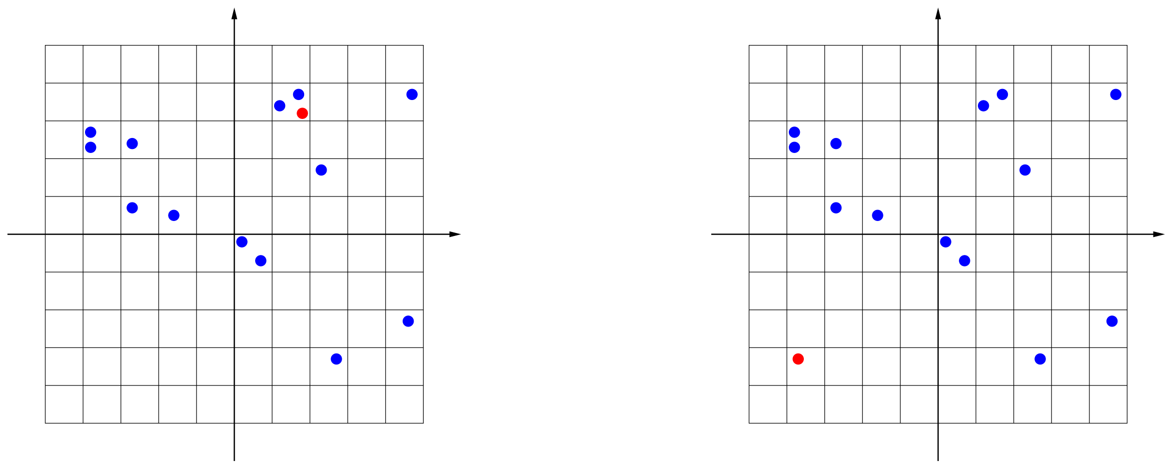

Figure 1: Bidimensional example of the construction used in the analysis of 1-NN. Left pane: $\boldsymbol{X}$ (the red point) is in the same square C_i as its closest training point $\boldsymbol{X}_{\pi(S,\boldsymbol{X})}$. Hence, $\|\boldsymbol{X} - \boldsymbol{X}_{\pi(S,\boldsymbol{X})}\|$ is bounded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

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

To proceed with the analysis, we now need some assumptions on D_X and η. First, all data points $\boldsymbol{X}$ drawn from D_X satisfy $\max_i |X_i| \le 1$ with probability one. In other words, $\boldsymbol{X} \in [-1,1]^d$ 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

$$\big|\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}')\big| \le c\,\|\boldsymbol{x} - \boldsymbol{x}'\| \qquad \text{for all } \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X} \ .$$

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

$$\eta(\boldsymbol{x}') \le \eta(\boldsymbol{x}) + c\,\|\boldsymbol{x} - \boldsymbol{x}'\| \tag{4}$$

$$1 - \eta(\boldsymbol{x}') \le 1 - \eta(\boldsymbol{x}) + c\,\|\boldsymbol{x} - \boldsymbol{x}'\| \tag{5}$$

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

$$\begin{aligned}
&\eta(\boldsymbol{x})\big(1 - \eta(\boldsymbol{x}')\big) + \big(1 - \eta(\boldsymbol{x})\big)\eta(\boldsymbol{x}') \\
&\qquad \le \eta(\boldsymbol{x})\big(1 - \eta(\boldsymbol{x})\big) + \eta(\boldsymbol{x})c\,\|\boldsymbol{x} - \boldsymbol{x}'\| + \big(1 - \eta(\boldsymbol{x})\big)\eta(\boldsymbol{x}) + \big(1 - \eta(\boldsymbol{x})\big)c\,\|\boldsymbol{x} - \boldsymbol{x}'\| \\
&\qquad = 2\eta(\boldsymbol{x})\big(1 - \eta(\boldsymbol{x})\big) + c\,\|\boldsymbol{x} - \boldsymbol{x}'\| \\
&\qquad \le 2\min\big\{\eta(\boldsymbol{x}), 1 - \eta(\boldsymbol{x})\big\} + c\,\|\boldsymbol{x} - \boldsymbol{x}'\|
\end{aligned}$$

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

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] \le 2\,\mathbb{E}\Big[\min\big\{\eta(\boldsymbol{X}), 1 - \eta(\boldsymbol{X})\big\}\Big] + c\,\mathbb{E}\big[\,\|\boldsymbol{X} - \boldsymbol{X}_{\pi(S,\boldsymbol{X})}\|\,\big] \ .$$

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

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] \le 2\,\ell_{\mathcal{D}}(f^*) + c\,\mathbb{E}\Big[\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi(S,\boldsymbol{X})}\big\|\Big] \ .$$

In order to bound the term containing the expected value of $\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi(S,\boldsymbol{X})}\big\|$ we partition the data space $\mathcal{X}$ in d-dimensional hypercubes with side $\varepsilon > 0$ —see Figure 1 for a bidimensional example. Let $C_1, \ldots, C_r$ the resulting hypercubes. We can now bound $\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi(S,\boldsymbol{X})}\big\|$ using a case analysis. Assume first that $\boldsymbol{X}$ belongs to a C_i in which there is at least a training point $\boldsymbol{X}_t$. Then $\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi(S,\boldsymbol{X})}\big\|$ 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 $\boldsymbol{X}$ belongs to a C_i in which there are no training points. Then $\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi(S,\boldsymbol{X})}\big\|$ 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

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

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

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

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

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

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

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

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

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

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

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

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

$$\mathbb{E}\Big[\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\big\|\Big] \le \varepsilon\sqrt{d} + \Big(2\sqrt{d}\Big)\,\frac{r}{em} = \sqrt{d}\left(\varepsilon + \frac{2}{em}\left(\frac{2}{\varepsilon}\right)^d\right)$$

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

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

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

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

Substituting this bound in (6), we finally obtain

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big]t \le 2\,\ell_{\mathcal{D}}(f^*) + c\,4m^{-1/(d+1)}\sqrt{d}\ .$$

This analysis allows us to draw two conclusions.

1. When $m \to \infty$, $\ell_{\mathcal{D}}(f^*) \le \mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] \le 2\,\ell_{\mathcal{D}}(f^*)$. Namely, the asymptotic risk of 1-NN lies between the Bayes risk and twice the Bayes risk.
2. For $\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big]$ to be at most $2\,\ell_{\mathcal{D}}(f^*) + \alpha$, the training set size must be at least $\big(\frac{4c}{\alpha}\sqrt{d}\big)^{d+1}$, 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.

The above risk analysis for 1-NN can be generalized to study, under the same Lipschitz assumptions, the risk of the classifier h_S generated by k-NN. In particular, the following result can be proven

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] \le \left(1 + \sqrt{\frac{8}{k}}\right)\ell_{\mathcal{D}}(f^*) + \mathcal{O}\big(k\,m^{-1/(d+1)}\big)\ .$$

Acknowledgments

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