Machine Learning — Statistical Methods for Machine Learning

Risk Analysis for Nearest-Neighbor

Instructor: *Nicolò Cesa-Bianchi* version of May 31, 2023

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

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

where A denotes the 1-NN algorithm, S_m the training set of size m, $\ell_{\mathcal{D}}(f^*)$ is the Bayes risk, and ε_m is a quantity that vanishes for $m \to \infty$. Note that we are able to compare $\mathbb{E}\big[\ell_{\mathcal{D}}\big(A(S_m)\big)\big]$ directly to the Bayes risk, showing that 1-NN is—in some sense—a powerful learning algorithm.

$(\mathcal{D}, \ell)$

Recall that in binary classification we denote the joint distribution of $(\boldsymbol{X}, Y)$ with the pair $\boxed{(\mathcal{D}_X, \eta)}$, where $\mathcal{D}_X$ is the marginal of $\mathcal{D}$ on $\boldsymbol{X}$ and $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$. Fix m and let $S = \big\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)\big\}$ be a training set of size m. we define the map $\pi(S, \cdot) : \mathbb{R}^d \to \{1, \dots, m\}$ $\in \mathbb{N}$ by

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

give me the index of closest data point

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 $h_S = A(S)$ is defined by $h_S(\boldsymbol{x}) = y_{\pi(S,\boldsymbol{x})}$. label of closest point

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

classifier

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

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

We now state a crucial lemma.

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

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] = \mathbb{E}\Big[\eta\big(\boldsymbol{X}_{\pi(S,\boldsymbol{X})}\big)\overset{P(Y=-1|X)}{\boxed{\big(1 - \eta(\boldsymbol{X})\big)}}\Big] + \mathbb{E}\Big[\Big(1 - \eta\big(\boldsymbol{X}_{\pi(S,\boldsymbol{X})}\big)\Big)\overset{P(Y=1|X)}{\boxed{\eta(\boldsymbol{X})}}\Big]$$

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 $\boxed{\max_i |X_i| \le 1}$ 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 Lipschitz on $\mathcal{X}$ with constant $c > 0$. We can thus write

$$\eta(\boldsymbol{x}') \le \eta(\boldsymbol{x}) + c\,\|\boldsymbol{x} - \boldsymbol{x}'\| \tag{2}$$

$$1 - \eta(\boldsymbol{x}') \le 1 - \eta(\boldsymbol{x}) + c\,\|\boldsymbol{x} - \boldsymbol{x}'\| \tag{3}$$

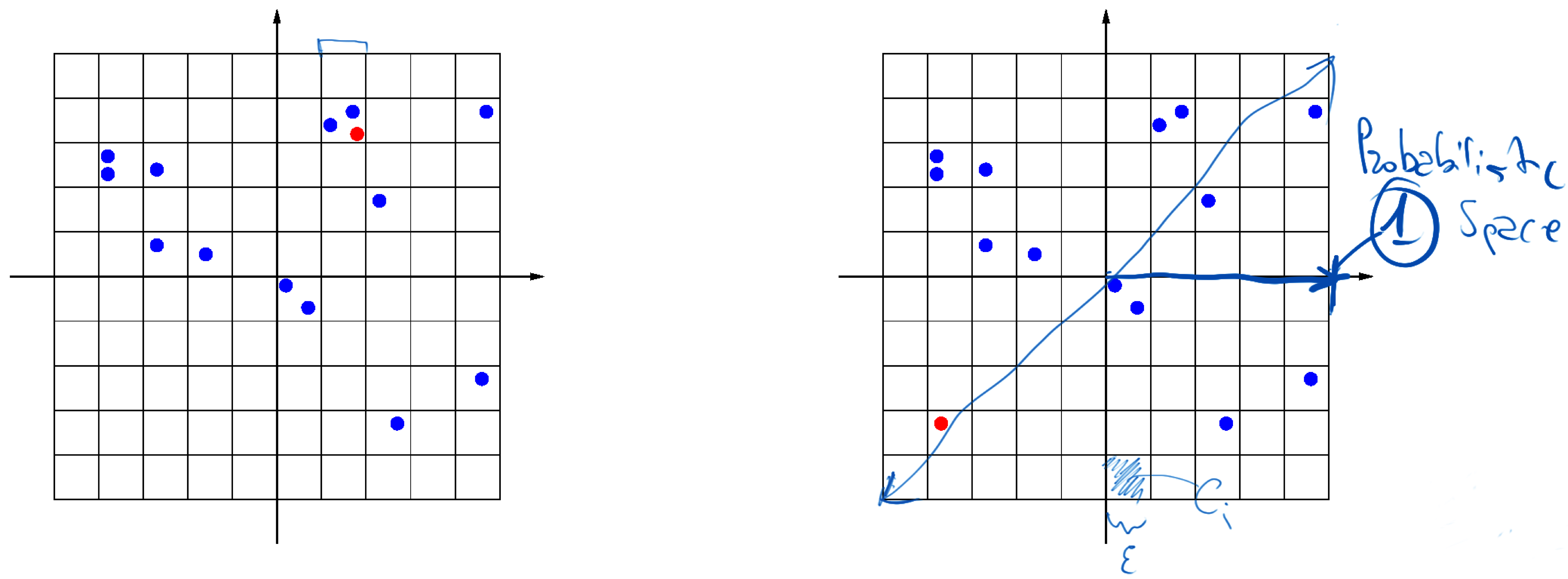

Figure 1: Bidimensional example of the construction used in the analysis of 1-NN. Left pane: $\boldsymbol{X}$ (the red point) is in the same square C_i as its closest training point $\boldsymbol{X}_{\pi(S,\boldsymbol{X})}$. Hence, $\left\|\boldsymbol{X}-\boldsymbol{X}_{\pi(S,\boldsymbol{X})}\right\|$ 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, $\left\|\boldsymbol{X}-\boldsymbol{X}_{\pi(S,\boldsymbol{X})}\right\|$ can only be bounded by the length of the entire data space (the large square).

Using (2) and (3), for all $\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}$ we have $\eta(x') \le \eta(x) + c\|x - x'\|$

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

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

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

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

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

In order to bound the term containing the expected value of $\left\|\boldsymbol{X}-\boldsymbol{X}_{\pi(S,\boldsymbol{X})}\right\|$ 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 $\left\|\boldsymbol{X}-\boldsymbol{X}_{\pi(S,\boldsymbol{X})}\right\|$ 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 $\left\|\boldsymbol{X}-\boldsymbol{X}_{\pi(S,\boldsymbol{X})}\right\|$ 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 $\left\|\boldsymbol{X}-\boldsymbol{X}_{\pi(S,\boldsymbol{X})}\right\|$ is only bounded by the length of the diagonal of $\mathcal{X}$, which is $2\sqrt{d}$, see the

$C = (C_1 \dots C_r)$ $\quad r$: # of HyperCubes

right pane in Figure 1. Hence, we may write

best case $\quad$ worst case $\quad$ true only one time on r

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

$$=\varepsilon\sqrt{d}\,\mathbb{E}\left[\sum_{i=1}^{r}\mathbb{I}\{C_i\cap S\neq\emptyset\}\mathbb{I}\{\boldsymbol{X}\in C_i\}\right]+2\sqrt{d}\sum_{i=1}^{r}\mathbb{E}\left[\mathbb{I}\{C_i\cap S=\emptyset\}\mathbb{I}\{\boldsymbol{X}\in C_i\}\right]$$

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 $i=1,\dots,d$. 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

$$\mathbb{E}\Big[\big\|\boldsymbol{X}-\boldsymbol{X}_{\pi(S,\boldsymbol{X})}\big\|\Big]\le\varepsilon\sqrt{d}+\big(2\sqrt{d}\big)\sum_{i=1}^{r}e^{-p_i m}p_i$$

$\ell_D(f^*)$

$$\le\varepsilon\sqrt{d}+\big(2\sqrt{d}\big)\sum_{i=1}^{r}\max_{0\le p\le 1}e^{-pm}p$$

worst case

$$=\varepsilon\sqrt{d}+\big(2\sqrt{d}\big)\,r\max_{0\le p\le 1}e^{-pm}p\,.$$

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

Probability that no data points lives in

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

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

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

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

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

Substituting this bound in (4), we finally obtain

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

Note that for $m \to \infty$, $\ell_{\mathcal{D}}(f^*) \leq \mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] \leq 2\,\ell_{\mathcal{D}}(f^*)$. Namely, the asymptotic risk of 1-NN lies between the Bayes risk and twice the Bayes risk.