Nearest Neighbor Classifiers

Matthieu R. Bloch

1 Nearest-neighbor classifier

In this section, we go back to our usual notation for our training dataset $\mathcal{D} \triangleq \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$, in which we do not explicitly distinguish random variables with capital letters.

We now investigate what is perhaps the simplest classifier of all. The *nearest-neighbor* (NN) classifier is $h^{\text{NN}}(\mathbf{x}) \triangleq y_{\text{NN}(\mathbf{x})}$ where $\text{NN}(\mathbf{x}) \triangleq \text{argmin}_i \|\mathbf{x}_i - \mathbf{x}\|$. The risk of NN classifier conditioned on fixed values of $\mathbf{x}$ and $\mathbf{x}_{\text{NN}(\mathbf{x})}$ is then

$$\begin{aligned} R_{\text{NN}}(\mathbf{x}, \mathbf{x}_{\text{NN}(\mathbf{x})}) &= \mathbb{P}\left(y_{\text{NN}(\mathbf{x})} \neq y | \mathbf{x}, \mathbf{x}_{\text{NN}(\mathbf{x})}\right) && (1)\\ &= \sum_k \mathbb{P}(y = k|\mathbf{x})\mathbb{P}\left(y_{\text{NN}(\mathbf{x})} \neq k | \mathbf{x}_{\text{NN}(\mathbf{x})}\right) && (2)\\ &= \sum_k \eta_k(\mathbf{x})(1 - \eta_k(\mathbf{x}_{\text{NN}(\mathbf{x})})) && (3)\\ &= \sum_k \eta_k(\mathbf{x}_{\text{NN}(\mathbf{x})})(1 - \eta_k(\mathbf{x})), && (4) \end{aligned}$$

where in (2) we have used the assumption that the dataset is generated independent and identically distributed (i.i.d.). How well does the risk $R_{\text{NN}} \triangleq \mathbb{E}_{\mathbf{x}, \mathbf{x}_{\text{NN}(\mathbf{x})}}\left(R_{\text{NN}}(\mathbf{x}, \mathbf{x}_{\text{NN}(\mathbf{x})})\right)$ compare to the Bayes risk for large N? The result is actually quite surprising as we show next.

Lemma 1.1. *Let* $\mathbf{x}$, $\{\mathbf{x}_i\}_{i=1}^N$ *be i.i.d.* $\sim P_{\mathbf{x}}$ *in a separable metric space* $\mathcal{X}$. *Let* $\mathbf{x}_{NN(\mathbf{x})}$ *be the nearest neighbor of* $\mathbf{x}$. *Then* $\mathbf{x}_{NN(\mathbf{x})} \to \mathbf{x}$ *with probability one as* $N \to \infty$

Before getting in the proof, let us analyze the assumptions of the lemma. Intuitively, the lemma is saying that if we sample enough data, we will eventually either sample any data point $\mathbf{x}$ or at least get close to it. Stating that $\mathcal{X}$ is a separable metric space is just a very mathematical of saying that there will not be points that we would not be able to approach by sampling sufficiently many points. The term separable is a bit ill-chosen, X separable means that it contains a countable and dense subset. The good news is that many metric spaces are separable, including compact metric spaces.

Proof. Let $d(\cdot, \cdot)$ be the metric associate to $\mathcal{X}$. For any $\mathbf{x} \in \mathcal{X}$ and $r > 0$ define the ball centered at $\mathbf{x}$ with radius r

$$\mathcal{S}_{\mathbf{x}}(r) \triangleq \{\mathbf{x}' \in \mathcal{X} : d(\mathbf{x}, \mathbf{x}') \leqslant r\}. \tag{5}$$

Consider now $\mathbf{x}_0 \in \mathcal{X}$ such that $\forall r > 0$, $\mathbb{P}(\mathcal{S}_{\mathbf{x}_0}(r)) > 0$; in other words, $\mathbf{x}_0$ is not a "lost" point that you would not be able to sample exactly or close to with $P_{\mathbf{x}}$. Then for any $r > 0$, the probability that the closest point to $\mathbf{x}_0$ in the dataset $\{\mathbf{x}_i\}_{i=1}^N$ is more than r away is

$$\mathbb{P}\left(\min_{k \in [\![1, N]\!]} d(\mathbf{x}_k, \mathbf{x}_0) \geqslant r\right) = (1 - \mathbb{P}(\mathcal{S}_{\mathbf{x}_0}(r)))^N \underset{N \to \infty}{\longrightarrow} 0. \tag{6}$$

We now show that $\mathbb{P}\big(\mathbf{x}_{\mathrm{NN}(\mathbf{x}_0)} \nrightarrow \mathbf{x}_0\big) = 0$, which requires a bit more care. In fact,

$$
\begin{aligned}
\mathbb{P}\big(\mathbf{x}_{\mathrm{NN}(\mathbf{x}_0)} \nrightarrow \mathbf{x}_0\big) &= \mathbb{P}\left(\exists k \in \mathbb{N}^* : \forall N \in \mathbb{N}^* \exists n_0 \geqslant N : \min_{i \in [\![1, n_0]\!]} d(\mathbf{x}_i, \mathbf{x}_0) \geqslant \frac{1}{k}\right) && (7)\\
&= \mathbb{P}\left(\exists k \in \mathbb{N}^* : \forall i \in \mathbb{N}^* \, d(\mathbf{x}_i, \mathbf{x}_0) \geqslant \frac{1}{k}\right) && (8)\\
&\leqslant \sum_{k \geqslant 1} \mathbb{P}\left(\bigcap_{i \geqslant 1} \{\forall j \leqslant i \, d(\mathbf{x}_j, \mathbf{x}_0) \geqslant \frac{1}{k}\}\right) && (9)\\
&\leqslant \sum_{k \geqslant 1} \lim_{i \to \infty} \mathbb{P}\left(\forall j \leqslant i \, d(\mathbf{x}_j, \mathbf{x}_0) \geqslant \frac{1}{k}\right) && (10)\\
&= 0. && (11)
\end{aligned}
$$

Note that the justification of the above inequalities is as follows:

- (8) is by definition of the nearest neighbor;
- (9) is by the union bound;
- (10) is by continuity of probabilities and the fact that the events $\mathcal{A}_i \triangleq \{\forall j \leqslant i \quad d(\mathbf{x}_j, \mathbf{x}_0) \geqslant \frac{1}{k}\}$ form a decreasing sequence;
- (11) follows by (6).

We now need to prove that the probability of sampling a point $\mathbf{x}$ for which the above reasoning is not true vanishes. Specifically let $\mathcal{N}$ be the set of points for which $\exists r_{\mathbf{x}} > 0$ such that $\mathbb{P}(\mathcal{S}_{\mathbf{x}}(r_{\mathbf{x}})) = 0$. By definition of $\mathcal{X}$ being separable, there exists a countable dense subset $\mathcal{A}$ of $\mathcal{X}$. In particular, for any $\mathbf{x} \in \mathcal{N}$, there exists $\mathbf{a}_{\mathbf{x}} \in \mathcal{A}$ such that $\mathbf{a}_{\mathbf{x}} \in \mathcal{S}_{\mathbf{x}}(\frac{r_{\mathbf{x}}}{3})$. Note that $\mathbf{x} \in \mathcal{S}_{\mathbf{a}_{\mathbf{x}}}(\frac{r_{\mathbf{x}}}{2}) \subset \mathcal{S}_{\mathbf{x}}(r_{\mathbf{x}})$. Consequently,

$$
0 \leqslant \mathbb{P}\big(\mathcal{S}_{\mathbf{a}_{\mathbf{x}}}(\tfrac{r_{\mathbf{x}}}{2})\big) \leqslant \mathbb{P}(\mathcal{S}_{\mathbf{x}}(r_{\mathbf{x}})) = 0. \tag{12}
$$

Therefore, $\mathcal{N}$ is included in $\bigcup_{\mathbf{a}_{\mathbf{x}}} \mathcal{S}_{\mathbf{a}_{\mathbf{x}}}(\frac{r_{\mathbf{x}}}{2})$, which is countable. Therefore,

$$
\mathbb{P}(\mathcal{N}) \leqslant \sum_{\mathbf{a}_{\mathbf{x}}} \mathbb{P}\big(\mathcal{S}_{\mathbf{a}_{\mathbf{x}}}(\tfrac{r_{\mathbf{x}}}{2})\big) = 0, \tag{13}
$$

which is what we needed. ■

Using Lemma 1.1, we now establish the following.

Lemma 1.2. *Let $\mathcal{X}$ be a separable metric space and consider a binary classifier with $K = 2$ and $\mathcal{Y} = \{0, 1\}$. Let $p(\mathbf{x}|y = 0)$, $p(\mathbf{x}|y = 1)$ be such that, $\mathbf{x}$ is either a continuity point of $p(\mathbf{x}|y = 0)$ and $p(\mathbf{x}|y = 1)$ with probability one, or a point of non-zero probability measure. Then, as $N \to \infty$,*

$$
R(h^B) \leqslant R(h^{NN}) \leqslant 2R(h^B)(1 - R(h^B))
$$

This result states that the risk of the NN classifier cannot be too far from the Bayes risk. Specifically, since we are looking here at binary loss functions we have $0 \leqslant 1 - R(h^{\mathrm{B}}) \leqslant 1$ and the risk of the NN classifier is at most twice the Bayes risk! In other words, simply assigning the label of the nearest point gives us close to the best performance when we have enough data points. The big catch behind the result is that there are two assumptions that might fail to hold in practice:

- We may not have enough data points, so that the nearest neighbor is actually very far from the point we are trying to predict. This is particularly true in high dimensions.
- Efficiently computing the nearest neighbor can be computationally tricky.

Proof. We need to consider two possible situations for $\mathbf{x}^* \in \mathcal{X}$.

- If $\mathbf{x}^*$ is a point of non-zero probability measure, i.e., such that $\mathbb{P}(\mathbf{x}^*) = p > 0$, then $\mathbb{P}\big(\mathbf{x}^* \neq \mathbf{x}_{\text{NN}(\mathbf{x}^*)}\big) = (1-p)^N \to_{N\to\infty} 0$. Consequently, as N gets large, one eventually generates the point $\mathbf{x}^*$ in the data set so that $\mathbf{x}_{\text{NN}(\mathbf{x}^*)} = \mathbf{x}^*$, at which point
$$R(\mathbf{x}^*, \mathbf{x}_{\text{NN}(\mathbf{x}^*)}) = R(\mathbf{x}^*, \mathbf{x}^*) = 2\eta_0(\mathbf{x}^*)(1-\eta_0(\mathbf{x}^*)). \tag{14}$$
- If $\mathbf{x}^*$ a continuity point of $p(\mathbf{x}|y=0)$ and $p(\mathbf{x}|y=1)$ with probability one, then as $N \to \infty$ we must have $\eta_0(\mathbf{x}_{\text{NN}(\mathbf{x}^*)}) \to \eta_0(\mathbf{x}^*)$ and $\eta_1(\mathbf{x}_{\text{NN}(\mathbf{x}^*)}) \to \eta_1(\mathbf{x}^*)$ with probability one by assumption. Consequently, with probability one
$$\begin{aligned}\lim_{N\to\infty} R(\mathbf{x}^*, \mathbf{x}_{\text{NN}(\mathbf{x}^*)}) &= \eta_0(\mathbf{x}^*)(1-\eta_0(\mathbf{x}^*)) + \eta_1(\mathbf{x}^*)(1-\eta_1(\mathbf{x}^*)) \qquad (15)\\ &= 2\eta_0(\mathbf{x}^*)(1-\eta_0(\mathbf{x}^*)). \qquad (16)\end{aligned}$$

Now recall the definition of the Bayes classifier as $h^{\text{B}}(\mathbf{x}^*) = \text{argmax}_k\, \eta_k(\mathbf{x}^*)$. This means that
$$R^B(\mathbf{x}^*) = \mathbb{E}_{y|\mathbf{x}^*}\left(\mathbb{1}\left\{\underset{k}{\text{argmin}}\, \eta_k(\mathbf{x}^*) = y\right\}\right) = \min_k \eta_k(\mathbf{x}^*) = \min\left(\eta_0(\mathbf{x}^*), 1-\eta_0(\mathbf{x}^*)\right). \tag{17}$$

Consequently,
$$R^B(\mathbf{x}^*)(1-R^B(\mathbf{x}^*)) = \eta_0(\mathbf{x}^*)(1-\eta_0(\mathbf{x}^*)). \tag{18}$$

Therefore, we have proved that
$$\lim_{N\to\infty} R(\mathbf{x}^*, \mathbf{x}_{\text{NN}(\mathbf{x}^*)}) = 2R^B(\mathbf{x}^*)(1-R^B(\mathbf{x}^*)). \tag{19}$$

It remains to look at the risk averaged over $\mathbf{x}^*$. Since the risk is bounded, by the dominated convergence theorem,
$$\lim_{N\to\infty} \mathbb{E}_{\mathbf{x}^*}\left(R(\mathbf{x}^*, \mathbf{x}_{\text{NN}(\mathbf{x}^*)})\right) = \mathbb{E}_{\mathbf{x}^*}\left(\lim_{N\to\infty} R(\mathbf{x}^*, \mathbf{x}_{\text{NN}(\mathbf{x}^*)})\right) = \mathbb{E}_{\mathbf{x}^*}(R(\mathbf{x}^*, \mathbf{x}^*)) \tag{20}$$

Now,
$$\begin{aligned}\mathbb{E}_{\mathbf{x}^*}(R(\mathbf{x}^*, \mathbf{x}^*)) &= 2\mathbb{E}_{\mathbf{x}^*}\left(R^B(\mathbf{x}^*)(1-R^B(\mathbf{x}^*))\right) \qquad (21)\\ &= 2\left(\mathbb{E}_{\mathbf{x}^*}\left(R^B(\mathbf{x}^*)\right) - \mathbb{E}_{\mathbf{x}^*}\left(R^B(\mathbf{x}^*)^2\right)\right) \qquad (22)\\ &= 2\left(\mathbb{E}_{\mathbf{x}^*}\left(R^B(\mathbf{x}^*)\right) - \mathbb{E}_{\mathbf{x}^*}\left(R^B(\mathbf{x}^*)\right)^2 + \underbrace{\mathbb{E}_{\mathbf{x}^*}\left(R^B(\mathbf{x}^*)\right)^2 - \mathbb{E}_{\mathbf{x}^*}\left(R^B(\mathbf{x}^*)^2\right)}_{\triangleq -\text{Var}(R^B(\mathbf{x}^*))}\right) \qquad (23)\\ &\leqslant 2\mathbb{E}_{\mathbf{x}^*}\left(R^B(\mathbf{x}^*)\right)\left(1-\mathbb{E}_{\mathbf{x}^*}\left(R^B(\mathbf{x}^*)\right)\right). \qquad (24)\end{aligned}$$

Since $R(h^B) \triangleq \mathbb{E}_{\mathbf{x}^*}\left(R^B(\mathbf{x}^*)\right)$, the result follows. ■

There are several extensions of this result. If there are $M > 2$ classes, one can show

$$R(h^{\mathrm{B}}) \leqslant R(h^{\mathrm{NN}}) \leqslant 2R(h^{\mathrm{B}})\left(1 - \frac{M}{M-1}R(h^{\mathrm{B}})\right). \tag{25}$$

One can also improve the bound by considering a K-NN classifier instead of the 1-NN classifier. In words, the K-NN classifier pools the labels of its K nearest neighbors and takes a majority vote to output the label. Formally, the K-NN classifier outputs

$$h^{K-\mathrm{NN}}(\mathbf{x}) = \operatorname*{argmax}_{\ell} \frac{1}{K} \sum_{i:\mathbf{x}_i \in \mathcal{N}_{K,\mathcal{D}}(\mathbf{x})} \mathbb{1}\{y_i = \ell\}, \tag{26}$$

where $\mathcal{N}_{K,\mathcal{D}}(\mathbf{x})$ indicates the set of K nearest neighbors of $\mathbf{x}$ in the dataset $\mathcal{D}$. One can show the following.

Theorem 1.3. *For all distributions and $K \geqslant 1$, as $N \to \infty$ we have*

$$R(h^{\mathrm{B}}) \leqslant R(h^{K-\mathrm{NN}}) \leqslant R(h^{\mathrm{B}})\left(1 + \sqrt{\frac{2}{K}}\right). \tag{27}$$

The larger K, the closer $R(h^{K-\mathrm{NN}})$ is to $R(h^{\mathrm{B}})$. This property is called *consistency* of the classifier.

Definition 1.4. *A classifier h_N on a dataset of size N is consistent (or asymptotically Bayes-risk efficient) for a given distribution $P_{\mathbf{x}y}$ if*[1]

$$\lim_{N\to\infty} R(h_N) \triangleq \lim_{N\to\infty} \mathbb{E}(\mathbb{P}(h_N(\mathbf{x}) \neq y)) = R(h_{\mathrm{B}}). \tag{28}$$

If the same property holds independently of $P_{\mathbf{x}y}$, the classifier is called universally *consistent.*

Intuitively, from Theorem 1.3, the K-NN classifier should be universally consistent. However, one has to state this carefully because the proof of Theorem 1.3 assumes that K is fixed and $N \to \infty$. Taking $K \to \infty$ in (27) might not make much sense. The following is true.

Proposition 1.5. *If $N \to \infty$, $K \to \infty$ while $\frac{K}{N} \to 0$, then*

$$\lim_{N\to\infty} R(h^{K-\mathrm{NN}}) \triangleq \lim_{N\to\infty} \mathbb{E}\big(\mathbb{P}\big(h^{K-\mathrm{NN}}(\mathbf{x}) \neq y\big)\big) = R(h^{\mathrm{B}}). \tag{29}$$

This result is saying that, given enough data, the NN classifier will perform just as well as the optimal Bayes classifier without knowing the underlying distribution. Unfortunately, in practice N is fixed and one must carefully choose K to obtain good performance. This is a problem of *model selection* that we will revisit later on. For now, let us just point out what is a *bad* idea to select K. If you were to select K by minimizing the empirical risk (as maybe suggested by what we discussed in Lecture 2), you would find that

$$\widehat{R}_N(h^{1\text{-NN}}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{h_1(\mathbf{x}_i) = y_i\} = 0. \tag{30}$$

[1] The expectation is over the dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, upon which h_N is dependent, as well as the test point $(\mathbf{x}, y)$.

2 To go further

Discussions of the Nearest Neighbor classifier can be found in [1, Section 2.3.2] and [2, Section 1.4.2]. The analysis of the risk of Nearest Neighbor classifiers is adapted from [3].

References

[1] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ser. Springer series in statistics. Springer, 2009.

[2] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[3] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.