

5 k -Nearest Neighbors (kNN)

Goal

Understand k -nearest neighbors for classification and regression. Relation to Bayes error.

Alert 5.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.

[This note is likely to be updated again soon.](#)

Definition 5.2: Distance

Given a domain $X \subseteq \mathbb{R}^d$, we define a **distance metric** $\text{dist} : X \times X \rightarrow \mathbb{R}_+$ as any function that satisfies the following axioms:

- nonnegative: $\text{dist}(\mathbf{x}, \mathbf{z}) \geq 0$;
- identity: $\text{dist}(\mathbf{x}, \mathbf{z}) = 0$ iff $\mathbf{x} = \mathbf{z}$;
- symmetric: $\text{dist}(\mathbf{x}, \mathbf{z}) = \text{dist}(\mathbf{z}, \mathbf{x})$;
- triangle inequality: $\text{dist}(\mathbf{x}, \mathbf{z}) \leq \text{dist}(\mathbf{x}, \mathbf{y}) + \text{dist}(\mathbf{y}, \mathbf{z})$.

We call the space X equipped with a distance metric dist a **metric space**, with notation (X, dist) .

If we relax the “iff” part in identity to “if” then we obtain pseudo-metric; if we drop symmetry we obtain quasi-metric; and finally if we drop the triangle inequality we get semi-metric.

Exercise 5.3: Example distances

Given any norm $\|\cdot\|$ on a vector space V , it immediately induces a distance metric:

$$\text{dist}_{\|\cdot\|}(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|.$$

Verify by yourself $\text{dist}_{\|\cdot\|}$ is indeed a distance metric.

In particular, for the ℓ_p norm defined in Definition 1.23, we obtain the ℓ_p distance.

Another often used “distance” is the cosine similarity:

$$\angle(\mathbf{x}, \mathbf{z}) = \frac{\mathbf{x}^\top \mathbf{z}}{\|\mathbf{x}\|_2 \cdot \|\mathbf{z}\|_2}.$$

Is it a distance metric?

Remark 5.4: kNN in a nutshell

Given a metric space (X, dist) and a dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, where $x_i \in X$, upon receiving a new instance $\mathbf{x} \in X$, it is natural to find near neighbors (e.g. “friends”) in our dataset \mathcal{D} according to the metric dist and predict $\hat{y}(\mathbf{x})$ according to the y -values of the neighbors. The underlying assumption is

neighboring feature vectors tend to have similar or same y -values.

The subtlety of course lies on what do we mean by neighboring, i.e., how do we choose the metric dist .

Remark 5.5: The power of an appropriate metric

Suppose we have (\mathbf{X}, Y) following some distribution on $X \times Y$, where the target space Y is equipped with some

metric dist_y (acting as a measure of our prediction error). Then, we may define a (pseudo)metric on \mathbf{X} as:

$$\text{dist}_x(\mathbf{x}, \mathbf{x}') := \mathbb{E}[\text{dist}_y(Y, Y') | \mathbf{X} = \mathbf{x}, \mathbf{X}' = \mathbf{x}'],$$

where (\mathbf{X}', Y') is an independent copy of (\mathbf{X}, Y) . (Note that $\text{dist}_x(\mathbf{x}, \mathbf{x}) = 0$ may not hold.) Given a test instance $\mathbf{X} = \mathbf{x}$, if we can find a near neighbor $\mathbf{X}' = \mathbf{x}'$ so that $\text{dist}_x(\mathbf{x}, \mathbf{x}') \leq \epsilon$, then predicting $Y(\mathbf{x})$ according to $Y(\mathbf{x}')$ gives us at most ϵ error:

$$\mathbb{E}[\text{dist}_y(Y(\mathbf{X}), Y(\mathbf{X}'))] = \mathbb{E}[\text{dist}_x(\mathbf{X}, \mathbf{X}')] \leq \epsilon.$$

Of course, we would not be able to construct the distance metric dist_x in practice, as it depends on the unknown distribution of our data.

Algorithm 5.6: kNN

Given a dataset $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$, where $\mathbf{x}_i \in (\mathbf{X}, \text{dist})$ and $\mathbf{y}_i \in \mathbf{Y}$, and a test instance \mathbf{x} , we predict according to the knn algorithm:

Algorithm: kNN

Input: Dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \in \mathbf{X} \times \mathbf{Y} : i = 1, \dots, n\}$, new instance $\mathbf{x} \in \mathbf{X}$, hyperparameter k

Output: $\mathbf{y} = \mathbf{y}(\mathbf{x})$

```

1 for  $i = 1, 2, \dots, n$  do
2    $d_i \leftarrow \text{dist}(\mathbf{x}, \mathbf{x}_i)$  // avoid for-loop if possible
3 find indices  $i_1, \dots, i_k$  of the  $k$  smallest entries in  $\mathbf{d}$ 
4  $\mathbf{y} \leftarrow \text{aggregate}(\mathbf{y}_{i_1}, \dots, \mathbf{y}_{i_k})$ 

```

For different target space \mathbf{Y} , we may use different aggregations:

- multi-class classification $\mathbf{Y} = \{1, \dots, c\}$: we can perform majority voting

$$\mathbf{y} \leftarrow \underset{j=1, \dots, c}{\text{argmax}} \#\{\mathbf{y}_{i_l} = j : l = 1, \dots, k\}, \quad (5.1)$$

where ties can be broken arbitrarily.

- regression: $\mathbf{Y} = \mathbb{R}^m$: we can perform averaging

$$\mathbf{y} \leftarrow \frac{1}{k} \sum_{l=1}^k \mathbf{y}_{i_l}. \quad (5.2)$$

Strictly speaking, there is no training time in kNN as we need only store the dataset \mathcal{D} . For testing, it costs $O(nd)$ as we have to go through the entire dataset to compute all distances to the test instance. There is a large literature that aims to bring down this complexity in test time by pre-processing our dataset and often by contending with near (but not necessarily nearest) neighbors (see e.g. (Andoni and Indyk 2008)).

Andoni, Alexandr and Piotr Indyk (2008). “Near-optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions”. *Communications of the ACM*, vol. 51, no. 1, pp. 117–122.

Exercise 5.7: The power of weights

More generally, suppose we also have a distance metric dist_y on \mathbf{Y} , we may set

$$\pi \leftarrow \underset{\pi}{\text{argmin}} \sum_{i=1}^n w_i^\downarrow \cdot \text{dist}_x(\mathbf{x}, \mathbf{x}_{\pi(i)}) \quad (5.3)$$

$$\mathbf{y} \leftarrow \underset{\mathbf{y} \in \mathbf{Y}}{\text{argmin}} \sum_{i=1}^n v_i^\downarrow \cdot \text{dist}_y^2(\mathbf{y}, \mathbf{y}_{\pi(i)}), \quad (5.4)$$

where $\pi : [n] \rightarrow [n]$ is a permutation, and $w_1 \geq w_2 \geq \dots \geq w_n \geq 0$, $v_1 \geq v_2 \geq \dots \geq v_n \geq 0$ are weights (e.g. how much each training instance should contribute to the final result). We may also use dist_y in (5.4) (without squaring). A popular choice is to set $v_i \propto 1/d_{\pi(i)}$ so that nearer neighbors will contribute more to predicting \mathbf{y} .

Prove that with the following choices we recover (5.1) and (5.2) from (5.3)-(5.4), respectively:

- Let $\mathcal{Y} = \{1, \dots, c\}$ and $\text{dist}_y(\mathbf{y}, \mathbf{y}') = \begin{cases} 0, & \text{if } \mathbf{y} = \mathbf{y}' \\ 1, & \text{o.w.} \end{cases}$ be the discrete distance. Use the kNN weights $\mathbf{w} = \mathbf{v} = \underbrace{(1, \dots, 1)}_k, 0, \dots, 0$.
- Let $\mathcal{Y} = \mathbb{R}^m$ and $\text{dist}_y(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|_2$ be the ℓ_2 distance.

Remark 5.8: Effect of k

Intuitively, using a larger k would give us more stable predictions (if we vary the training dataset), as we are averaging over more neighbors, corresponding to smaller variance but potentially larger bias (see ??):

- If we use $k = n$, then we always predict the same target irrespective of the input \mathbf{x} , which is clearly not varied at all but may incur a large bias.
- Indeed, if we have a dataset where different classes are *well* separated, then using a large k can bring significant bias while 1NN achieves near 0 error.

In practice we may select k using cross-validation (see Line 7). For a moderately large dataset, typically $k = 3$ or 5 suffices. A rule of thumb is we use larger k for larger and more difficult datasets.

Theorem 5.9: kNN generalization error (Biau and Devroye 2015)

Let k be odd and fixed. Then, for all distributions of (\mathbf{X}, Y) , as $n \rightarrow \infty$,

$$\mathbb{L}_{k\text{NN}} := \Pr[h_n(\mathbf{X}) \neq Y] \rightarrow \mathbb{E} \left[\sum_{l=0}^k \binom{k}{l} r^l(\mathbf{X})(1-r(\mathbf{X}))^{k-l} \left(r(\mathbf{X}) \mathbb{I}[l < \frac{k}{2}] + (1-r(\mathbf{X})) \mathbb{I}[l \geq \frac{k}{2}] \right) \right],$$

where the knn classifier h_n is defined in (5.5) and $r(\mathbf{x}) := \Pr[Y = 1 | \mathbf{X} = \mathbf{x}]$ is the regression function.

Proof. Let $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{i.i.d.}{\sim} \mathbf{X}$ and let $Y_i = \mathbb{I}[U_i \leq r(\mathbf{X}_i)]$, where $U_i \stackrel{i.i.d.}{\sim} \text{Uniform}([0, 1])$. Clearly, (\mathbf{X}_i, Y_i, U_i) form an i.i.d. sequence where $(\mathbf{X}_i, Y_i) \sim (\mathbf{X}, Y)$. Let $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i, U_i), i = 1, \dots, n\}$. Fixing \mathbf{x} , define $\tilde{Y}_i(\mathbf{x}) = \mathbb{I}[U_i \leq r(\mathbf{x})]$. Order $\mathbf{X}_{(l)}(\mathbf{x})$, $Y_{(l)}(\mathbf{x})$, $\tilde{Y}_{(l)}(\mathbf{x})$ and $U_{(l)}(\mathbf{x})$ according to the distance $\text{dist}(\mathbf{X}_i, \mathbf{x})$. Consider the classifiers:

$$h_n(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{l=1}^k Y_{(l)}(\mathbf{x}) > k/2 \\ 0, & \text{o.w.} \end{cases}, \quad \tilde{h}_n(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{l=1}^k \tilde{Y}_{(l)}(\mathbf{x}) > k/2 \\ 0, & \text{o.w.} \end{cases}. \quad (5.5)$$

Then, we have

$$\begin{aligned} \Pr[h_n(\mathbf{X}) \neq \tilde{h}_n(\mathbf{X})] &\leq \Pr \left[\sum_{l=1}^k Y_{(l)}(\mathbf{X}) \neq \sum_{l=1}^k \tilde{Y}_{(l)}(\mathbf{X}) \right] \\ &\leq \Pr \left[(Y_{(1)}(\mathbf{X}), \dots, Y_{(k)}(\mathbf{X})) \neq (\tilde{Y}_{(1)}(\mathbf{X}), \dots, \tilde{Y}_{(k)}(\mathbf{X})) \right] \\ &\leq \Pr \left[\bigcup_{l=1}^k \left[r(\mathbf{X}_{(l)}(\mathbf{X})) \wedge r(\mathbf{X}) < U_{(l)}(\mathbf{X}) \leq r(\mathbf{X}_{(l)}(\mathbf{X})) \vee r(\mathbf{X}) \right] \right] \end{aligned}$$

$$\leq \sum_{l=1}^k \mathbb{E} |r(\mathbf{X}_{(l)}(\mathbf{X})) - r(\mathbf{X})| \xrightarrow{n \rightarrow \infty} 0, \text{ see Stone's Lemma 5.13 below.}$$

Recall that $\mathfrak{L}(h_n) := \Pr(h_n(\mathbf{X}) \neq Y | \mathcal{D})$ and similarly for $\mathfrak{L}(\tilde{h}_n)$. Thus,

$$\mathbb{E} \left| \mathfrak{L}(h_n) - \mathfrak{L}(\tilde{h}_n) \right| \leq \Pr[h_n(\mathbf{X}) \neq \tilde{h}_n(\mathbf{X})] = o(1),$$

whereas noting that given \mathbf{x} , $\tilde{Y}_i(\mathbf{x}) \stackrel{i.i.d.}{\sim} \text{Bernoulli}(r(\mathbf{x}))$, hence

$$\begin{aligned} \mathbb{E} \mathfrak{L}(\tilde{h}_n) &= \Pr \left[\text{Binomial}(k, r(\mathbf{X})) > \frac{k}{2}, Y = 0 \right] + \Pr \left[\text{Binomial}(k, r(\mathbf{X})) \leq \frac{k}{2}, Y = 1 \right] \\ &= \mathbb{E} \left[(1 - r(\mathbf{X})) \mathbb{I}[\text{Binomial}(k, r(\mathbf{X})) > \frac{k}{2}] + r(\mathbf{X}) \mathbb{I}[\text{Binomial}(k, r(\mathbf{X})) \leq \frac{k}{2}] \right]. \end{aligned}$$

Combining the above completes the proof. \square

The proof above exploits the beautiful **decoupling** idea: $Y_{(i)}$'s, which the kNN classifier g_n depends on, are coupled through the ordering induced by the \mathbf{X}_i 's. On the other hand, $\tilde{Y}_{(i)}$'s are independent (conditioned on $\mathbf{X} = \mathbf{x}$) hence allow us to analyze the closely related classifier \tilde{g}_n with ease. Stone's Lemma 5.13 adds the final piece that establishes the asymptotic equivalence of the two classifiers.

Biau, Gérard and Luc Devroye (2015). *Lectures on the Nearest Neighbor Method*. Springer.

Corollary 5.10: 1NN $\leq 2 \times$ Bayes (Cover and Hart 1967)

For $n \rightarrow \infty$, we have

$$\mathbb{L}_{\text{Bayes}} \leq \mathbb{L}_{1\text{NN}} \leq 2\mathbb{L}_{\text{Bayes}}(1 - \mathbb{L}_{\text{Bayes}}) \leq 2\mathbb{L}_{\text{Bayes}},$$

and $\mathbb{L}_{3\text{NN}} = \mathbb{E}[r(\mathbf{X})(1 - r(\mathbf{X}))] + 4\mathbb{E}[r^2(\mathbf{X})(1 - r(\mathbf{X}))^2]$.

Proof. For $k = 1$, it follows from Theorem 5.9 that

$$\mathbb{L}_{1\text{NN}} = 2\mathbb{E}[r(\mathbf{X})(1 - r(\mathbf{X}))]$$

whereas the Bayes error is

$$\mathbb{L}_{\text{Bayes}} = \mathbb{E}[r(\mathbf{X}) \wedge (1 - r(\mathbf{X}))].$$

Therefore, letting $\mathbf{s}(\mathbf{x}) = r(\mathbf{x}) \wedge (1 - r(\mathbf{x}))$, we have

$$\mathbb{L}_{1\text{NN}} = 2\mathbb{E}[\mathbf{s}(\mathbf{X})(1 - \mathbf{s}(\mathbf{X}))] = 2\mathbb{E}\mathbf{s}(\mathbf{X}) \cdot \mathbb{E}(1 - \mathbf{s}(\mathbf{X})) - 2 \cdot \text{Variance}(\mathbf{s}(\mathbf{X})) \leq 2\mathbb{L}_{\text{Bayes}}(1 - \mathbb{L}_{\text{Bayes}}).$$

The formula for $\mathbb{L}_{3\text{NN}}$ follows immediately from Theorem 5.9. \square

We note that for trivial problems where $\mathbb{L}_{\text{Bayes}} = 0$ or $\mathbb{L}_{\text{Bayes}} = \frac{1}{2}$, $\mathbb{L}_{1\text{NN}} = \mathbb{L}_{\text{Bayes}}$. On the other hand, when the Bayes error is small, $\mathbb{L}_{1\text{NN}} \sim 2\mathbb{L}_{\text{Bayes}}$ while $\mathbb{L}_{3\text{NN}} \sim \mathbb{L}_{\text{Bayes}}$.

Cover, T. M. and P. E. Hart (1967). "Nearest Neighbor Pattern Classification". *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27.

Proposition 5.11: Continuity

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be (Lebesgue) integrable. If $k/n \rightarrow 0$, then

$$\frac{1}{k} \sum_{l=1}^k \mathbb{E} |f(\mathbf{X}_l(\mathbf{X})) - f(\mathbf{X})| \rightarrow 0,$$

where $\mathbf{X}_{(i)}(\mathbf{X})$ is ordered by the distance $\|\mathbf{X}_i - \mathbf{X}\|_2$ and $\mathbf{X}_i \sim \mathbf{X}$ for $i = 1, \dots, n$.

Proof. Since \mathcal{C}_c is dense in \mathcal{L}_1 , we may approximate f by a (uniformly) continuous function f_ϵ with compact support. In particular, for $\epsilon > 0$ there exists $\delta > 0$ such that $\text{dist}(\mathbf{x}, \mathbf{z}) \leq \delta \implies |f_\epsilon(\mathbf{x}) - f_\epsilon(\mathbf{z})| \leq \epsilon$. Thus,

$$\begin{aligned} \frac{1}{k} \sum_{l=1}^k \mathbb{E} |f(\mathbf{X}_l(\mathbf{X})) - f(\mathbf{X})| &\leq \frac{1}{k} \sum_{l=1}^k \mathbb{E} |f(\mathbf{X}_l(\mathbf{X})) - f_\epsilon(\mathbf{X}_l(\mathbf{X}))| + \mathbb{E} |f_\epsilon(\mathbf{X}_l(\mathbf{X})) - f_\epsilon(\mathbf{X})| + \mathbb{E} |f_\epsilon(\mathbf{X}) - f(\mathbf{X})| \\ &\stackrel{\text{(Stone's Lemma 5.13)}}{\leq} (\gamma_d + 2) \mathbb{E} |f(\mathbf{X}) - f_\epsilon(\mathbf{X})| + 2 \|f_\epsilon\|_\infty \cdot \Pr[\text{dist}(\mathbf{X}_{(k)}, \mathbf{X}) > \delta] + \epsilon \\ &\leq (\gamma_d + 2) \epsilon + 2 \|f_\epsilon\|_\infty \cdot \Pr[\text{dist}(\mathbf{X}_{(k)}, \mathbf{X}) > \delta] \\ &\leq (\gamma_d + 3) \epsilon, \text{ thanks to Theorem 5.12 when } n \text{ is large.} \end{aligned}$$

The proof is complete by noting that ϵ is arbitrary. \square

Theorem 5.12: projection through kNN

Fix \mathbf{x} and define $\rho = \text{dist}(\mathbf{x}, \text{supp}\mu)$ where $\text{supp}\mu$ is the support of some measure μ . If $k/n \rightarrow 0$, then almost surely

$$\text{dist}(\mathbf{X}_{(k)}(\mathbf{x}), \mathbf{x}) \rightarrow \rho,$$

where $\mathbf{X}_i \stackrel{i.i.d.}{\sim} \mu$ and $\mathbf{X}_{(i)}$ is ordered by $\text{dist}(\mathbf{X}_i, \mathbf{x})$, $i = 1, \dots, n$.

Proof. Fix any $\epsilon > 0$ and let $p = \Pr(\text{dist}(\mathbf{X}, \mathbf{x}) \leq \epsilon + \rho) > 0$. Then, for large n ,

$$\begin{aligned} \Pr(\text{dist}(\mathbf{X}_{(k)}, \mathbf{x}) - \rho > \epsilon) &= \Pr\left(\sum_{i=1}^n B_i < k\right), \text{ where } B_i \stackrel{i.i.d.}{\sim} \text{Bernoulli}(p) \\ &= \Pr\left(\frac{1}{n} \sum_{i=1}^n (B_i - p) < k/n - p\right) \\ &\leq \exp(-2n(p - k/n)^2). \end{aligned}$$

Since $p > 0$ and $k/n \rightarrow 0$, the theorem follows. \square

Let $\mathbf{X} \sim \mu$ be another independent copy, then with $k/n \rightarrow 0$:

$$\text{dist}(\mathbf{X}_{(k)}, \mathbf{X}) \xrightarrow{a.s.} 0.$$

Indeed, for μ -almost all \mathbf{x} and large n , we have

$$\Pr\left[\sup_{m \geq n} \text{dist}(\mathbf{X}_{(k,m)}(\mathbf{x}), \mathbf{x}) \geq \epsilon\right] \leq \sum_{m \geq n} \exp(-mp^2) \xrightarrow{n \rightarrow \infty} 0.$$

Lemma 5.13: Stone's Lemma (Stone 1977)

Let $(w_1^{(n)}, \dots, w_n^{(n)})$ be a probability vector with $w_1^{(n)} \geq \dots \geq w_n^{(n)}$ for all n . Then, for any integrable function $f: \mathbb{R}^d \rightarrow \mathbb{R}$,

$$\mathbb{E} \left[\sum_{i=1}^n w_i^{(n)} |f(\mathbf{X}_{(i)}(\mathbf{X}))| \right] \leq (1 + \gamma_d) \mathbb{E} |f(\mathbf{X})|,$$

where \mathbf{X}_i 's are i.i.d. copies of \mathbf{X} , $\mathbf{X}_{(i)}$'s are ordered by $\|\mathbf{X}_i - \mathbf{X}\|_2$, and $\gamma_d < \infty$ only depends on d .

Proof. Define

$$W_i^{(n)}(\mathbf{x}) := W_i^{(n)}(\mathbf{x}; \mathbf{x}_1, \dots, \mathbf{x}_n) := w_k^{(n)}$$

if \mathbf{x}_i is the k -th nearest neighbor of \mathbf{x} (ties broken by index). We first prove

$$\sum_{i=1}^n W_i^{(n)}(\mathbf{x}; \mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n) \leq (1 + \gamma_d). \quad (5.6)$$

Cover \mathbb{R}^d with γ_d angular cones $K_t, t = 1, \dots, \gamma_d$, each with angle $\pi/12$. Let $A = \{i : \mathbf{x}_i = \mathbf{x}\}$ and $B_t = \{i : \mathbf{x}_i \in (K_t + \mathbf{x}) \setminus \{\mathbf{x}\}\}$. Choose any $a, b \in B_t$ such that $0 < \|\mathbf{x}_a - \mathbf{x}\| \leq \|\mathbf{x}_b - \mathbf{x}\|$, then

$$\|\mathbf{x}_a - \mathbf{x}_b\|^2 \leq \|\mathbf{x}_a - \mathbf{x}\|^2 + \|\mathbf{x}_b - \mathbf{x}\|^2 - 2\|\mathbf{x}_a - \mathbf{x}\|\|\mathbf{x}_b - \mathbf{x}\|\cos(\pi/6) < \|\mathbf{x}_b - \mathbf{x}\|^2. \quad (5.7)$$

Therefore, if \mathbf{x}_b is the k -th closest to \mathbf{x} among \mathbf{x}_{B_t} , then \mathbf{x} is at best the k -th closest to \mathbf{x}_b among $\mathbf{x}, \mathbf{x}_{B_t \setminus \{b\}}$. Since the weights $w_i^{(n)}$ are ordered, we have

$$\begin{aligned} \sum_{i \in B_t} W_i^{(n)}(\mathbf{x}; \mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n) &\leq \sum_{i=1}^{n-|A|} w_i^{(n)} \leq 1 \\ \sum_{i \in A} W_i^{(n)}(\mathbf{x}; \mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n) &= \sum_{i=1}^{|A|} w_i^{(n)} \leq 1. \end{aligned}$$

Taking unions over the γ_d angular cones proves (5.6).

Therefore,

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^n w_i^{(n)} |f(\mathbf{X}_{(i)}(\mathbf{X}))| \right] &= \mathbb{E} \left[\sum_{i=1}^n W_i^{(n)}(\mathbf{X}) |f(\mathbf{X}_i)| \right] \\ (\text{symmetrization}) &= \mathbb{E} \left[|f(\mathbf{X})| \sum_{i=1}^n W_i^{(n)}(\mathbf{X}_i; \mathbf{X}_1, \dots, \mathbf{X}_{i-1}, \mathbf{X}, \mathbf{X}_{i+1}, \dots, \mathbf{X}_n) \right] \\ &\leq (1 + \gamma_d) \mathbb{E} |f(\mathbf{X})|. \end{aligned}$$

□

Here γ_d is the covering number of \mathbb{R}^d by angular cones:

$$K(\mathbf{z}, \theta) := \{\mathbf{x} \in \mathbb{R}^d : \angle(\mathbf{x}, \mathbf{z}) \leq \theta\}.$$

The proof above relies on the ℓ_2 distance only in (5.7).

Stone, Charles J. (1977). “Consistent Nonparametric Regression”. *The Annals of Statistics*, vol. 5, no. 4, pp. 595–620.

Theorem 5.14: No free lunch (Shalev-Shwartz and Ben-David 2014)

Let h be **any** classifier learned from a training set \mathcal{D}_n with size $n \leq |\mathcal{X}|/2$. Then, there exists a distribution (\mathbf{X}, Y) over $\mathcal{X} \times \{0, 1\}$ such that the Bayes error is zero while

$$\Pr [h(\mathbf{X}; \mathcal{D}_n) \neq Y] \geq \frac{1}{4}.$$

In particular, with probability at least $\frac{1}{8}$ over the training set \mathcal{D}_n we have $\Pr [h(\mathbf{X}; \mathcal{D}_n) \neq Y | \mathcal{D}_n] \geq \frac{1}{8}$.

Proof. We may assume w.l.o.g. that $|\mathcal{X}| = 2n$. Enumerate all $T = 2^{2n}$ functions $h_t : \mathcal{X} \rightarrow \{0, 1\}$, each of which induces a distribution where $\mathbf{X} \in \mathcal{X}$ is uniformly random while $Y = h_t(\mathbf{X})$. For each labeling function h_t , we

have $S = (2n)^n$ possible training sets $\mathcal{D}_n(s, t)$. Thus,

$$\begin{aligned}
 \max_{t \in [T]} \frac{1}{S} \sum_{s=1}^S \Pr[h(\mathbf{X}; \mathcal{D}_n(s, t)) \neq h_t(\mathbf{X})] &\geq \frac{1}{T} \sum_{t=1}^T \frac{1}{S} \sum_{s=1}^S \Pr[h(\mathbf{X}; \mathcal{D}_n(s, t)) \neq h_t(\mathbf{X})] \\
 &\geq \min_{s \in [S]} \frac{1}{T} \sum_{t=1}^T \Pr[h(\mathbf{X}; \mathcal{D}_n(s, t)) \neq h_t(\mathbf{X})] \\
 &\geq \min_{s \in [S]} \frac{1}{T} \sum_{t=1}^T \frac{1}{2|\mathbf{X} \setminus \mathcal{D}_n(s, t)|} \sum_{\mathbf{x}_i \in \mathbf{X} \setminus \mathcal{D}_n(s, t)} \mathbb{I}[h(\mathbf{x}_i; \mathcal{D}_n(s, t)) \neq h_t(\mathbf{x}_i)] \\
 &= \min_{s \in [S]} \frac{1}{2|\mathbf{X} \setminus \mathcal{D}_n(s, t)|} \frac{1}{T} \sum_{t=1}^T \sum_{\mathbf{x}_i \in \mathbf{X} \setminus \mathcal{D}_n(s, t)} \mathbb{I}[h(\mathbf{x}_i; \mathcal{D}_n(s, t)) \neq h_t(\mathbf{x}_i)] \\
 &\geq \frac{1}{4},
 \end{aligned}$$

since we apparently have

$$\mathbb{I}[h(\mathbf{x}_i; \mathcal{D}_n(s, t)) \neq h_t(\mathbf{x}_i)] + \mathbb{I}[h(\mathbf{x}_i; \mathcal{D}_n(s, \tau)) \neq h_\tau(\mathbf{x}_i)] = 1,$$

for two labeling functions h_t and h_τ which agree on \mathbf{x} iff $\mathbf{x} \in \mathcal{D}_n$. □

Let $c > 1$ be arbitrary. Consider the uniform grid \mathbf{X} in the cube $[0, 1]^d$ with $1/c$ distance between neighbors. Clearly, there are $(c+1)^d$ points in \mathbf{X} . If our training set is smaller than $(c+1)^d/2$, then kNN suffers at least $1/4$ error while the Bayes error is 0! Thus, **the condition $n \rightarrow \infty$ in Theorem 5.9 can be very unrealistic in high dimensions!**

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.