11 k-Nearest Neighbors (kNN)

Goal

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

Alert 11.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.
This note is likely to be updated again soon.

Definition 11.2: Distance

Given a domain $\mathsf{X} \subseteq \mathbb{R}^d$, we define a distance metric $\text{dist} : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$ 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 $(\mathsf{X}, \text{dist})$.
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 11.3: 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 1.25, 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 11.4: kNN in a nutshell

Given a metric space $(\mathsf{X}, \text{dist})$ and a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1) \dots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathsf{X}$, upon receiving a new instance $\mathbf{x} \in \mathsf{X}$, 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 11.5: The power of an appropriate metric

Suppose we have $(\mathbf{X}, Y)$ following some distribution on $\mathsf{X} \times \mathsf{Y}$, 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 X as:

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

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:

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

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 11.6: kNN

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

Algorithm: kNN

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

```
for i = 1, 2, ..., n do
    d_i ← dist(x, x_i)                    // avoid for-loop if possible
find indices i_1, ..., i_k of the k smallest entries in d
y ← aggregate(y_{i_1}, ..., y_{i_k})
```

For different target space Y, we may use different aggregations:

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

$$\mathbf{y} \leftarrow \underset{j=1,\ldots,\mathsf{c}}{\text{argmax}} \#\{\mathbf{y}_{i_l} = j : l = 1, \ldots, k\}, \tag{11.1}$$

  where ties can be broken arbitrarily.

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

$$\mathbf{y} \leftarrow \frac{1}{k} \sum_{l=1}^{k} \mathbf{y}_{i_l}. \tag{11.2}$$

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 11.7: The power of weights

More generally, suppose we also have a distance metric dist_y on Y, we may set

$$\pi \leftarrow \operatorname*{argmin}_{\pi} \sum_{i=1}^{n} w_i^{\downarrow} \cdot \text{dist}_x(\mathbf{x}, \mathbf{x}_{\pi(i)}) \tag{11.3}$$

$$\mathbf{y} \leftarrow \operatorname*{argmin}_{\mathbf{y}\in\mathsf{Y}} \sum_{i=1}^{n} v_i^{\downarrow} \cdot \text{dist}_y^2(\mathbf{y}, \mathbf{y}_{\pi(i)}), \tag{11.4}$$

where $\pi : [n] \to [n]$ is a permutation, and $w_1 \geq w_2 \geq \cdots \geq w_n \geq 0$, $v_1 \geq v_2 \geq \cdots \geq v_n \geq 0$ are weights (e.g. how much each training instance should contribute to the final result). We may also use dist_y in (11.4) (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 (11.1) and (11.2) from (11.3)-(11.4), respectively:

- Let $\mathsf{Y} = \{1, \dots, \mathsf{c}\}$ 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 $\mathsf{Y} = \mathbb{R}^m$ and $\text{dist}_y(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|_2$ be the ℓ_2 distance.

Remark 11.8: 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 Algorithm 2.31). 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 11.9: kNN generalization error (Biau and Devroye 2015)

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

$$\mathbb{L}_{kNN} := \Pr[h_n(\mathbf{X}) \neq Y] \to \mathsf{E}\left[\sum_{l=0}^{k} \binom{k}{l} \mathsf{r}^l(\mathbf{X})(1 - \mathsf{r}(\mathbf{X}))^{k-l}\Big(\mathsf{r}(\mathbf{X})[\![l < \tfrac{l}{2}]\!] + (1 - \mathsf{r}(\mathbf{X}))[\![l \geq \tfrac{k}{2}]\!]\Big)\right],$$

where the knn classifier h_n is defined in (11.5) *and $\mathsf{r}(\mathbf{x}) := \Pr[Y = 1|\mathbf{X} = \mathbf{x}]$ is the regression function.*

Proof. Let $\mathbf{X}_1, \dots, \mathbf{X}_n \overset{i.i.d.}{\sim} \mathbf{X}$ and let $Y_i = [\![U_i \leq \mathsf{r}(\mathbf{X}_i)]\!]$, where $U_i \overset{i.i.d.}{\sim} \text{Uniform}([0, 1])$. 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 $\tilde{Y}_i(\mathbf{x}) = [\![U_i \leq \mathsf{r}(\mathbf{x})]\!]$. Order $\mathbf{X}_{(i)}(\mathbf{x})$, $Y_{(i)}(\mathbf{x})$, $\tilde{Y}_{(i)}(\mathbf{x})$ and $U_{(i)}(\mathbf{x})$ according to the distance $\text{dist}(\mathbf{X}_i, \mathbf{x})$. Consider the classifiers:

$$h_n(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{l=1}^{k} Y_{(l)}(\mathbf{x}) > k/2 \\ 0, & \text{o.w.} \end{cases}, \qquad \tilde{h}_n(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{l=1}^{k} \tilde{Y}_{(l)}(\mathbf{x}) > k/2 \\ 0, & \text{o.w.} \end{cases}. \tag{11.5}$$

Then, we have

$$\begin{aligned}
\Pr[h_n(\mathbf{X}) \neq \tilde{h}_n(\mathbf{X})] &\leq \Pr\left[\sum_{l=1}^k Y_{(l)}(\mathbf{X}) \neq \sum_{l=1}^k \tilde{Y}_{(l)}(\mathbf{X})\right] \\
&\leq \Pr\left[\left(Y_{(1)}(\mathbf{X}), \ldots, Y_{(k)}(\mathbf{X})\right) \neq \left(\tilde{Y}_{(1)}(\mathbf{X}), \ldots, \tilde{Y}_{(k)}(\mathbf{X})\right)\right] \\
&\leq \Pr\left[\bigcup_{l=1}^k [\![\mathsf{r}(\mathbf{X}_{(l)}(\mathbf{X})) \wedge \mathsf{r}(\mathbf{X}) < U_{(l)}(\mathbf{X}) \leq \mathsf{r}(\mathbf{X}_{(l)}(\mathbf{X})) \vee \mathsf{r}(\mathbf{X})]\!]\right] \\
&\leq \sum_{l=1}^k \mathsf{E}\left|\mathsf{r}(\mathbf{X}_{(l)}(\mathbf{X})) - \mathsf{r}(\mathbf{X})\right| \overset{n\to\infty}{\longrightarrow} 0, \text{ see Stone's Lemma 11.13 below.}
\end{aligned}$$

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

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

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

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

Combining the above completes the proof. □

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 11.13 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 11.10: 1NN $\leq 2\times$Bayes (Cover and Hart 1967)

For $n \to \infty$, we have

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

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

Proof. For $k = 1$, it follows from Theorem 11.9 that

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

whereas the Bayes error is

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

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

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

The formula for $\mathbb{L}_{3\text{NN}}$ follows immediately from Theorem 11.9. □

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 11.11: Continuity

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

$$\frac{1}{k}\sum_{l=1}^{k} \mathsf{E}\left|f\big(\mathbf{X}_l(\mathbf{X})\big) - f(\mathbf{X})\right| \to 0,$$

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 $\operatorname{dist}(\mathbf{x}, \mathbf{z}) \le \delta \implies |f_\epsilon(\mathbf{x}) - f_\epsilon(\mathbf{z})| \le \epsilon$. Thus,

$$\begin{aligned}
\frac{1}{k}\sum_{l=1}^{k} \mathsf{E}\left|f\big(\mathbf{X}_l(\mathbf{X})\big) - f(\mathbf{X})\right| &\le \frac{1}{k}\sum_{l=1}^{k} \mathsf{E}\left|f\big(\mathbf{X}_l(\mathbf{X})\big) - f_\epsilon(\mathbf{X}_l(\mathbf{X}))\right| + \mathsf{E}\left|f_\epsilon\big(\mathbf{X}_l(\mathbf{X})\big) - f_\epsilon(\mathbf{X})\right| + \mathsf{E}\left|f_\epsilon(\mathbf{X}) - f(\mathbf{X})\right| \\
\text{(Stone's Lemma 11.13)} \quad &\le (\gamma_d + 2)\mathsf{E}\left|f(\mathbf{X}) - f_\epsilon(\mathbf{X})\right| + 2\|f_\epsilon\|_\infty \cdot \Pr[\operatorname{dist}(\mathbf{X}_{(k)}, \mathbf{X}) > \delta] + \epsilon \\
&\le (\gamma_d + 2)\epsilon + 2\|f_\epsilon\|_\infty \cdot \Pr[\operatorname{dist}(\mathbf{X}_{(k)}, \mathbf{X}) > \delta] \\
&\le (\gamma_d + 3)\epsilon, \quad \text{thanks to Theorem 11.12 when } n \text{ is large.}
\end{aligned}$$

The proof is complete by noting that ϵ is arbitrary. □

Theorem 11.12: projection through kNN

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

$$\operatorname{dist}(\mathbf{X}_{(k)}(\mathbf{x}), \mathbf{x}) \to \rho,$$

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

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

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

Since $p > 0$ and $k/n \to 0$, the theorem follows. □

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

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

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

$$\Pr\left[\sup_{m \ge n} \operatorname{dist}(\mathbf{X}_{(k,m)}(\mathbf{x}), \mathbf{x}) \ge \epsilon\right] \le \sum_{m \ge n} \exp(-mp^2) \overset{n\to\infty}{\longrightarrow} 0.$$

Lemma 11.13: Stone's Lemma (Stone 1977)

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

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

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

$$\sum_{i=1}^n W_i^{(n)}(\mathbf{x}_i; \mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n) \leq (1+\gamma_d). \tag{11.6}$$

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

$$\|\mathbf{x}_a - \mathbf{x}_b\|^2 \leq \|\mathbf{x}_a - \mathbf{x}\|^2 + \|\mathbf{x}_b - \mathbf{x}\|^2 - 2\|\mathbf{x}_a - \mathbf{x}\|\|\mathbf{x}_b - \mathbf{x}\|\cos(\pi/6) < \|\mathbf{x}_b - \mathbf{x}\|^2. \tag{11.7}$$

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

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

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

Taking unions over the γ_d angular cones proves (11.6).

Therefore,

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

□

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

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

The proof above relies on the ℓ_2 distance only in (11.7).

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

Theorem 11.14: No free lunch (Shalev-Shwartz and Ben-David 2014)

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

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

In particular, with probability at least $\frac{1}{7}$ over the training set $\mathcal{D}_n$ we have $\Pr\left[h(\mathbf{X}; \mathcal{D}_n) \neq Y | \mathcal{D}_n\right] \geq \frac{1}{8}$.

Proof. We may assume w.l.o.g. that $|\mathsf{X}| = 2n$. Enumerate all $T = 2^{2n}$ functions $h_t : \mathsf{X} \to \{0,1\}$, each of which induces a distribution where $\mathbf{X} \in \mathsf{X}$ 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,

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

since we apparently have

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

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 X in the cube $[0,1]^d$ with $1/c$ distance between neighbors. Clearly, there are $(c+1)^d$ points in 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 $n \to \infty$ in Theorem 11.9 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.