Nearest Neighbor

Peng Lingwei

August 16, 2019

Contents

19 Nearest Neighbor

19.1 NEAREST NEIGHBOR

1. Instance domain $(\mathcal{X}, \mathcal{Y}) \sim \mathcal{D}$;
2. Metric function $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$;
3. Training examples $S = ((\vec{x}_1, y_1), \ldots, (\vec{x}_m, y_m))$;
4. For each $\vec{x} \in \mathcal{X}$, let $(\pi_1(\vec{x}), \ldots, \pi_m(\vec{x})) = \pi(\rho(\vec{x}, \vec{x}_i), \ldots, \rho(\vec{x}, \vec{x}_m))$
5. Rules of k-NN in classification: return the majority label among $\{y_i : \pi_i(\vec{x}) \leq k\}$
6. Rules of k-NN in regression: return $h_S(\vec{x}) = \frac{\sum_{\pi_i \leq k} \rho(\vec{x}, \vec{x}_i) y_i}{\sum_{\pi_j \leq k} \rho(\vec{x}, \vec{x}_j)}$

19.2 ANALYSIS 1-NN

1. $\mathcal{X} = [0,1]^d$, $\mathcal{Y} = \{0, 1\}$, $l(h, (\vec{x}, y)) = 1_{[h(\vec{x}) \neq y]}$, ρ is the Euclidean distance;
2. Define conditional probability: $\eta(\vec{x}) = \mathbb{P}_{\mathcal{D}}[y = 1 | \vec{x}]$;
3. Bayes optimal rule: $h^*(\vec{x}) = 1_{[\eta(\vec{x}) > 1/2]}$;
4. Assume that η is c-Lipschitz: $\forall \vec{x}, \vec{x}' \in \mathcal{X}, |\eta(\vec{x}) - \eta(\vec{x}')| \leq c \|\vec{x} - \vec{x}'\|$

Lemma 1. *In 1-NN:*

$$\mathbb{E}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S)] \leq 2L_{\mathcal{D}}(h^*) + c\mathbb{E}_{S \sim \mathcal{D}^m, \vec{x} \sim \mathcal{D}}\left[\|\vec{x} - \vec{x}_{i:\pi_i(\vec{x})=1}\|\right]$$

Proof.

$$\begin{aligned}
\mathbb{E}_{S \sim \mathcal{D}^m}\{L_{\mathcal{D}}(h_S)\} =& \mathbb{E}_{S \sim \mathcal{D}^m}\{\mathbb{E}_{(\vec{x}, y) \sim \mathcal{D}}[1_{[h_S(\vec{x}) \neq y]}]\} \\
=& \mathbb{E}_{S_x \sim \mathcal{D}_{\mathcal{X}}^m, \vec{x} \sim \mathcal{D}, y \sim \eta(\vec{x}), y' \sim \eta(x_{i:\pi_i(\vec{x})=1})}\left[1_{[y \neq y']}\right] \\
=& \mathbb{E}_{S_x \sim \mathcal{D}_{\mathcal{X}}^m, \vec{x} \sim \mathcal{D}}\left[\mathbb{P}_{y \sim \eta(\vec{x}), y' \sim \eta(x_{i:\pi_i(\vec{x})=1})}\left[y \neq y'\right]\right]
\end{aligned}$$

For any two domain points $\vec{x}, \vec{x}'$:

$$\begin{aligned}
\mathbb{P}_{y \sim \eta(\vec{x}), y' \sim \eta(\vec{x}')} =& \eta(\vec{x}')(1 - \eta(\vec{x})) + (1 - \eta(\vec{x}'))\eta(\vec{x}) \\
=& 2\eta(\vec{x})(1 - \eta(\vec{x})) + (\eta(\vec{x}) - \eta(\vec{x}'))(2\eta(\vec{x}) - 1).
\end{aligned}$$

Using $|2\eta(\vec{x}) - 1| \leq 1$ and the assumption that η is $c - Lipschitz$, then

$$\mathbb{P}_{y \sim \eta(\vec{x}), y' \sim \eta(\vec{x}')} = 2\eta(\vec{x})(1 - \eta(\vec{x})) + c\|\vec{x} - \vec{x}'\|.$$

$$\begin{aligned}
\mathbb{E}_{S \sim \mathcal{D}}\left[L_{\mathcal{D}}(h_S)\right] \leq \mathbb{E}_{\vec{x} \sim \mathcal{D}}\left[2\eta(\vec{x})(1 - \eta(\vec{x}))\right] + c\mathbb{E}_{S_x \sim \mathcal{D}, \vec{x} \sim \mathcal{D}}\left[\|\vec{x} - \vec{x}_{i:\pi_i(\vec{x})=1}\|\right] \\
L_{\mathcal{D}}(h^*) = \mathbb{E}_{\vec{x} \sim \mathcal{D}}\left[\min\{\eta(\vec{x}), 1 - \eta(\vec{x})\}\right] \geq \mathbb{E}_{\vec{x}}\left[\eta(\vec{x})(1 - \eta(\vec{x}))\right].
\end{aligned}$$

□

Then we bound the second part of preceeding inequation's right side.

Lemma 2. *Let $C_1, \ldots, C_r$ be a collection of subsets of some domain set $\mathcal{X}$. Then,*

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sum_{i:C_i\cap S=\emptyset}\mathbb{P}[C_i]\right] \le \frac{r}{me}$$

Proof.

$$\begin{aligned}
&\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sum_{i:C_i\cap S=\emptyset}\mathbb{P}[C_i]\right]\\
=&\sum_{i=1}^{r}\mathbb{P}\left[C_i\right]\mathbb{E}_{S\sim\mathcal{D}^m}\left[1_{[C_i\cap S=\emptyset]}\right] = \sum_{i=1}^{r}\mathbb{P}\left[C_i\right]\mathbb{P}_{S\sim\mathcal{D}}\left[C_i\cap S=\emptyset\right]\\
=&\sum_{i=1}^{r}\mathbb{P}\left[C_i\right]\left(1-\mathbb{P}\left[C_i\right]\right)^m \le \sum_{i=1}^{r}\mathbb{P}\left[C_i\right]e^{-\mathbb{P}[C_i]m}\\
\le& r\max_i \mathbb{P}\left[C_i\right]e^{-\mathbb{P}[C_i]m} \le \frac{r}{me}
\end{aligned}$$

□

Theorem 1. $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_S)\right] \le 2L_{\mathcal{D}}(h^*) + 2c\sqrt{d}m^{-\frac{1}{d+1}}$

Proof. We cut $\mathcal{X} = [0,1]^d$ into $N\times\cdots\times N$ hypertable, which divide sample space into $r = N^d$ pieces, $C_1, \ldots, C_r$.

$\forall \vec{x}, \vec{x}'$, if they are in the same box, we have $\|\vec{x}-\vec{x}'\| \le \frac{\sqrt{d}}{T}$. Otherwise, $\|\vec{x}-\vec{x}'\| \le \sqrt{d}$.

$$\begin{aligned}
\mathbb{E}_{\vec{x},S}\left[\|\vec{x}-\vec{x}_{i:\pi_i(\vec{x})=1}\|\right] \le& \mathbb{E}_S\left[\mathbb{P}\left[\cup_{i:C_i\cap S=\emptyset}C_i\right]\sqrt{d} + \mathbb{P}\left[\cup_{i:C_i\cap S\neq\emptyset}C_i\right]\sqrt{d}/T\right]\\
\le&\sqrt{d}\left(\frac{T^d}{me}+\frac{1}{T}\right) \le \sqrt{d}\Big(\frac{me}{d}\Big)^{-\frac{1}{d+1}}\left\{\frac{1}{d}+1\right\}\\
\le& 2\sqrt{d}m^{-1/(d+1)}
\end{aligned}$$

□

The theorem shows that if we want the error gap is smaller than ϵ, the sample size $m \ge (2c\sqrt{d}/\epsilon)^{d+1}$, we call it the "curse of dimensionality".

$\forall c > 1$, guarantees $\eta(\vec{x})$ is c-Lipschitz. If $m \le (c+1)^d/2$, the true error of the rule L is greater than 1/8 with probability greater than 1/7. (The proof is in the book.)

19.3 Chernoff Bound

Chebyshev's Inequality only requires the pairwise independence of the variables $\{X_i\}$. Donote $Z = \sum X_i$, so the bound

$$\forall a > 0, \mathbb{P}\left[|Z-\mathbb{E}\left[Z\right]| \ge a\right] = \mathbb{P}\left[(Z-\mathbb{E}\left[Z\right])^2 \ge a^2\right] \le \frac{Var\left[Z\right]}{a^2}$$

is not satisfying for i.i.d. variables X_i.

Theorem 2. *Let $X_1, \dots, X_m$ be independent Bernoulli variables where for every i, $\mathbb{P}[X_i = 1] = p_i$ and $\mathbb{P}[X_i = 0] = 1 - p_i$. Let $Z = \sum_{i=1}^m X_i$ and $p = \mathbb{E}[Z] = \sum_{i=1}^m p_i$.*

1. *Upper Tail: $\forall \delta > 0, \mathbb{P}(Z \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2}{2+\delta}\mu}$;*

2. *Lower Tail: $\forall \delta \in (0,1), \mathbb{P}(Z \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2}{2+\delta}\mu}$*

Proof. Step1: $\delta > 0$:

$$\mathbb{E}\left[e^{tZ}\right] = \mathbb{E}\left[e^{t\sum_i X_i}\right] = \prod_i \mathbb{E}\left[e^{tX_i}\right] = \prod_i \left(p_i e^t + (1-p_i)\right) \leq \prod_i e^{p_i(e^t-1)} = e^{p(e^t-1)}$$

$$\mathbb{P}[Z \geq (1+\delta)p] \leq \min_{t>0} \frac{\mathbb{E}\left[e^{tZ}\right]}{e^{(1+\delta)tp}} \leq \min_{t>0} e^{p(e^t-1)-(1+\delta)tp} = e^{-p[(1+\delta)\ln(1+\delta)-\delta]}$$

Let's take a break, and study the function $f(\delta) = \ln(1+\delta) - \frac{\delta}{1+k\delta}$: $f'(\delta) = \frac{k^2\delta^2+(2k-1)\delta}{(1+\delta)(1+k\delta)^2}$. If $k \geq \frac{1}{2}$, $\forall \delta > 0, f(\delta) \geq f(0) = 0 \Rightarrow \ln(1+\delta) \geq \frac{\delta}{1+k\delta}$.

$$\mathbb{P}[Z \geq (1+\delta)p] \leq e^{-p\cdot\frac{(1-k)\delta^2}{1+k\delta}} = e^{-p\frac{\delta^2}{2+\delta}}$$

Step2: $\delta \in (0,1)$:

$$\mathbb{P}[Z \leq (1-\delta)p] \leq \min_{t>0} \frac{\mathbb{E}\left[e^{-tZ}\right]}{e^{-tp(1-\delta)}} \leq \min_{t>0} e^{p(e^{-t}-1)+tp(1-\delta)} \leq e^{-p((1-\delta)\ln(1-\delta)+\delta)}$$

$$(1-\delta)\ln(1-\delta) + \delta = \sum_{i=1}^{\infty} \frac{\delta^{i+1}}{i(i+1)} \geq \sum_{i=1}^{\infty} \frac{(-\delta)^{i+1}}{i(i+1)} = ((1+\delta)\ln(1+\delta) - \delta)$$

Then, we can get the same bound:

$$\mathbb{P}[Z \leq (1-\delta)p] \leq e^{-p\cdot\frac{(1-k)\delta^2}{1+k\delta}} = e^{-p\frac{\delta^2}{2+\delta}}$$

□

19.4 Analysis k-NN

Lemma 3. *Let $C_1, \dots, C_r$ be a collection of subsets of some domain set, $\mathcal{X}$. Then $\forall k \geq 2$,*

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sum_{i:|C_i\cap S|<k} \mathbb{P}[C_i]\right] \leq \frac{2rk}{m}.$$

Proof.

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sum_{i:|C_i\cap S|<k} \mathbb{P}_{\mathcal{D}}[C_i]\right] &= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sum_{i=1}^r \mathbb{P}_{\mathcal{D}}[C_i]\, 1_{[|C_i\cap S|<k]}\right] \\ &= \sum_{i=1}^r \mathbb{P}_{\mathcal{D}}[C_i]\, \mathbb{P}_{S\sim\mathcal{D}}[|C_i\cap S| < k]\end{aligned}$$

If $k \geq \mathbb{P}\left[C_i\right] m/2$,

$$\mathbb{P}_{\mathcal{D}}\left[C_i\right] \mathbb{P}_{S\sim\mathcal{D}}\left[\left|C_i \cap S\right| < k\right] \leq \mathbb{P}_{\mathcal{D}}\left[C_i\right] \leq \frac{2k}{m}$$

If $k < \mathcal{P}_{\mathcal{D}}\left[C_i\right] m/2$, then

$$\mathbb{P}_{S\sim\mathcal{D}}\left[\left|C_i \cap S\right| < k\right] \leq \mathbb{P}_{S\sim\mathcal{D}}\left[\left|C_i \cap S\right| < \left(1 - \frac{1}{2}\right) \mathbb{P}_{\mathcal{D}}\left[C_i\right] m\right] \leq e^{-\mathbb{P}_{\mathcal{D}}[C_i]m/10}$$

$$\mathbb{P}_{\mathcal{D}}\left[C_i\right] \mathbb{P}_{S\sim\mathcal{D}}\left[\left|C_i \cap S\right| < k\right] \leq \mathbb{P}_{\mathcal{D}}\left[C_i\right] e^{-P_{D}[D_i]m/10} \leq \frac{10}{me} \leq \frac{4}{m} \leq \frac{2k}{m}$$

□

Lemma 4. *Let $p = \frac{1}{k}\sum_{i=1}^{k} p_i$, and $p' = \frac{1}{k}\sum_{i=1}^{k} X_i$. Then*

$$\mathbb{E}_{X_1,\ldots,Z_k} \mathbb{P}_{y\sim p}\left[y \neq 1_{[p'>1/2]}\right] \leq \left(1 + \sqrt{\frac{8}{k}}\right) \mathbb{P}_{y\sim p}\left[y \neq 1_{[p>1/2]}\right]$$

Proof.

$$\begin{aligned}\mathbb{E}_{X_1,\ldots,X_k}\mathbb{P}_{y\sim p}\left[y \neq 1_{[p'>1/2]}\right] =& p\left(1 - \mathbb{P}_{X_1,\ldots,X_k}\left[p' > 1/2\right]\right) + (1-p)\left(\mathbb{P}_{X_1,\ldots,X_k}\left[p' > 1/2\right]\right) \\ =& p + (1-2p)\left(\mathbb{P}_{X_1,\ldots,X_k}\left[p' > 1/2\right]\right)\end{aligned}$$

$$\mathbb{P}_{X_1,\ldots,X_k}\left[p' > 1/2\right] = \mathbb{P}_{X_1,\ldots,X_k}\left[\sum_{i=1}^{k} X_i \geq k/2\right] = \mathbb{P}_{X_1,\ldots,X_k}\left[\sum_{i=1}^{k} X_i \geq (1 + \frac{1}{2p} - 1)kp\right]$$

If $p \leq \frac{1}{2}$, $\mathbb{P}_{X_1,\ldots,X_k}\left[p' > 1/2\right] \leq e^{-kph\left(\frac{1}{2p}-1\right)} = e^{-kp+\frac{k}{2}(\log(2p)+1)}$
(If $p > \frac{1}{2}$, we study the random variables $1 - X_1, \ldots, 1 - X_k$, the error times keep unchanged.)

There is a inequation: $(1-2p)e^{-kp+\frac{k}{2}(\log(2p)-1)} \leq p\sqrt{\frac{8}{k}}$

$$\mathbb{E}_{X_1,\ldots,X_k}\mathbb{P}_{y\sim p}\left[y \neq 1_{[p'>1/2]}\right] \leq \left(1 + \sqrt{\frac{8}{k}}\right) p$$

□

Lemma 5. $\forall p, p' \in [0,1], y' \in \{y, y'\}, \mathbb{P}_{y\sim p}\left[y \neq y'\right] - \mathbb{P}_{y\sim p'}\left[y \neq y'\right] \leq |p - p'|$.

Proof. If $y' = 0$, $\mathbb{P}_{y\sim p}\left[y \neq 0\right] - \mathbb{P}_{y\sim p'}\left[y \neq 0\right] \leq p - p$;
If $y' = 1$, $\mathbb{P}_{y\sim p}\left[y \neq 1\right] - \mathbb{P}_{y\sim p'}\left[y \neq 1\right] \leq (1-p) - (1-p') = p' - p$. □

Theorem 3. *Let $C_1, \ldots, C_r$ be the cover of the set $\mathcal{X}$ using boxes of length ϵ.*

$$\mathbb{E}_S\left[L_{\mathcal{D}}(h_S)\right] \leq \left(1 + \sqrt{\frac{8}{k}}\right) L_{\mathcal{D}}(h^*) + (3c\sqrt{d} + 2k)m^{-1/(d+1)}.$$

Proof. First we get a loose bound:

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}}\left[L_{\mathcal{D}}(h_S)\right] \leq &\mathbb{E}_{S\sim\mathcal{D}}\left[\sum_{i:|C_i\cap S|<k} P_{\mathcal{D}}\left[C_i\right]\right] \\ &+ \max_i \mathbb{P}_{S,(\vec{x},y)}\left[h_S(\vec{x}) \neq y | \forall j \in [k], \|\vec{x}-\vec{x}_{j:\pi_j(\vec{x})\leq k}\| \leq \epsilon\sqrt{d}\right]\end{aligned}$$

If a cell doesn't contain k instances from the training set and test point $\vec{x}$ gets from this "bad cell", we think it's a kind of mistake.Only if test point $\vec{x}$ gets from a "good cell", there is probability for correct prediction.

Let $p = \frac{1}{k}\sum_{i=1}^{k}\eta(\vec{x}_i) < 1/2$.

$$\begin{aligned}&\mathbb{E}_{y_1,\ldots,y_m}\mathbb{P}_{y\sim\eta(\vec{x})}\left[h_S(\vec{x}) \neq y\right] \leq \mathbb{E}_{y_1,\ldots,y_m}\mathbb{E}_{y\sim p}\left[h_S(\vec{x}) \neq y\right] + |p-\eta(\vec{x})| \\ &\leq \left(1+\sqrt{\frac{8}{k}}\right)\mathbb{P}_{y\sim p}\left[1_{[p>1/2]} \neq y\right] + |p-\eta(\vec{x})| \\ &\leq \left(1+\sqrt{\frac{8}{k}}\right)\left(\min\{\eta(\vec{x}), 1-\eta(\vec{x})\} + |p-\eta(\vec{x})|\right) + |p-\eta(\vec{x})| \\ &\leq \left(1+\sqrt{\frac{8}{k}}\right)L_{\mathcal{D}}(h^*) + \left(2+\sqrt{\frac{8}{k}}\right)|p-\eta(\vec{x})| \\ &\leq \left(1+\sqrt{\frac{8}{k}}\right)L_{\mathcal{D}}(h^*) + 3c\epsilon\sqrt{d}\end{aligned}$$

$$\mathbb{E}_{S\sim\mathcal{D}}\left[L_{\mathcal{D}}\left(h_S\right)\right] \leq \left(1+\sqrt{\frac{8}{k}}\right)L_{\mathcal{D}}(h^*) + 3c\epsilon\sqrt{d} + \frac{2k}{m\epsilon^d}$$

If $\epsilon = m^{-1/(d+1)}$, $\mathbb{E}_{S\sim\mathcal{D}}\left[L_{\mathcal{D}}(h_S)\right] \leq \left(1+\sqrt{\frac{8}{k}}\right)L_{\mathcal{D}}(h^*) + (3c\sqrt{d}+2k)m^{-1/(d+1)}$ □