Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology

Technical University of Munich

k-Nearest Neighbour Classification

Outline

- k-nearest neighbour classification
  - Generalisation error of k-NN for finite k as $m \to \infty$
- Consistency / Universal consistency: Asymptotically achieving Bayes risk
- Plug-in classifiers
  - Stone's theorem: Universal consistency of plug-in classifiers
  - Universal consistency of k-NN rule
  - Proof of Stone's thoerem

Nearest neighbour rule

- Assume $\mathcal{X} \subset \mathbb{R}^p$ and we use Euclidean distance $\|x - x'\|$
- Given $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$
- Nearest neighbour and k-nearest neighbour rules:
  - For test data $x \in \mathcal{X}$, sort $x_1, \ldots, x_m$ according to $\|x - x_i\|$
  - $\pi_k(x) \in [m]$ such that $x_{\pi_k(x)}$ is k-th nearest neighbour of x

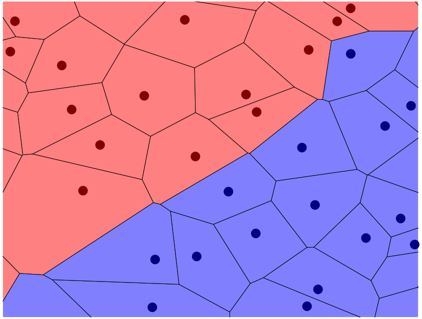
$$\|x - x_{\pi_1(x)}\| \leq \|x - x_{\pi_2(x)}\| \ldots \leq \|x - x_{\pi_m(x)}\|$$

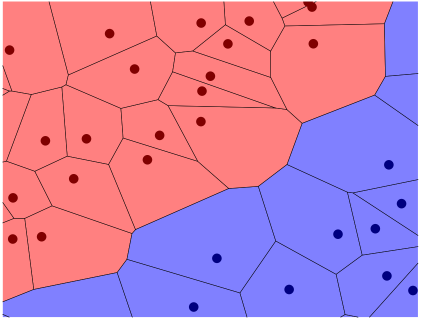
  - $h_S^{NN}(x) = y_{\pi_1(x)}$ and $h_S^{kNN}(x) =$ majority vote of $y_{\pi_1(x)}, \ldots, y_{\pi_k(x)}$

There can be ties (not discussed here)

Nearest neighbour rule

- Finite m: Decision boundary depends significantly on S
- Large m: Can learn very complex decision boundaries (more complex for $k > 1$)





Recap: Bayes risk

Learning problem characterised by:

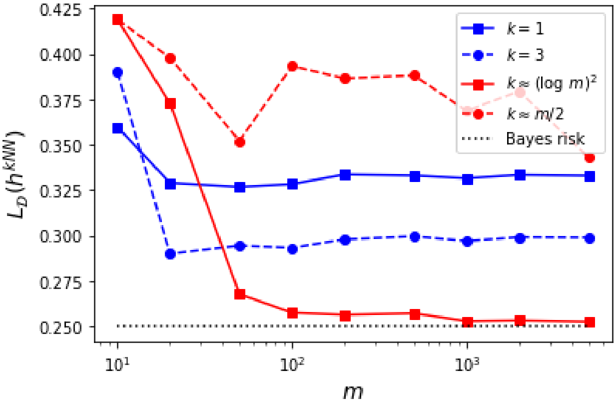
$$\mathcal{D} \quad = \quad \underbrace{\mathcal{D}_{\mathcal{X}}}_{\text{marginal of features}} \quad \times \quad \underbrace{\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)}_{\text{conditional probability of label}}$$

- $\eta(x) = \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1|x)$
- Bayes risk $\quad L^*_{\mathcal{D}} = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\min\{\eta(x), 1 - \eta(x)\}\right]$
- Bayes risk is smallest possible risk / generalisation error for a learning problem
- Bayes risk achieved by Bayes classifier (needs knowledge of η)

Example: Performance of k-NN rule

- Predicting software crash:

$$\mathcal{D}_{\mathcal{X}} = \text{Uniform}[0,1] \qquad \text{and} \qquad \eta(x) = |1 - 2x|$$



- Test error of k-NN for different values of m and k
  - Test data has 5000 samples
  - Errors averaged over 25 trials
- We will try to explain these results mathematically

Expected generalisation error for NN rule

Theorem kNN.1 (Asymptotic expected risk of NN rule)

- *Define* $L_{\mathcal{D}}^{NN} = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[2\eta(x)(1-\eta(x))\right]$
  As $m \to \infty$, $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S^{NN}\right)\right] \longrightarrow L_{\mathcal{D}}^{NN}$
- *Comparison with Bayes risk:* $L_{\mathcal{D}}^{*} \ \leq \ L_{\mathcal{D}}^{NN} \ \leq \ 2L_{\mathcal{D}}^{*}$

- Easy to verify 2^{nd} statement (exercise)
- We will prove the 1^{st} statement (under some assumptions)

Expected generalisation error of k-NN

Theorem kNN.2 (Asymptotic expected risk of k-NN rule)

- *Assume k is fixed*
- *Define* $L_{\mathcal{D}}^{kNN} = \lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S^{kNN}\right)\right]$ *(limit exists)*

$$L_{\mathcal{D}}^{*} \;\leq\; L_{\mathcal{D}}^{kNN} \;\leq\; \left(1 + \frac{2}{\sqrt{k}}\right) L_{\mathcal{D}}^{*}$$

- Proof skipped. Similar to proof for $k = 1$ (more involved)
- Result suggests that we need $k \to \infty$ to get close to Bayes risk

Intuition why NN (or k-NN) works

- Let $x_1, x_2, \ldots, x_m \sim_{iid} \mathcal{D}_{\mathcal{X}}$. Consider some x.
- Intuition: For large m, $x_{\pi_1(x)}, \ldots x_{\pi_k(x)}$ are arbitrarily close to x
  Hence, the label of x is likely to be same as $y_{\pi_1(x)}, \ldots, y_{\pi_k(x)}$
- Bit more formal:
  If $\dfrac{k}{m} \to 0$ as $m \to \infty$, then $x_{\pi_1(x)}, \ldots, x_{\pi_k(x)} \to x$ in probability
  - Convergence in probability: A sequence of random variables $z_1, z_2, \ldots$ is said to converge to random variable x in probability if, for every $\epsilon > 0$, $\lim_{m\to\infty} \mathbb{P}(\|z_m - x\| > \epsilon) = 0$.

Formal proof of Theorem kNN.1 (under assumptions)

- Goal: Show $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S^{\mathrm{NN}}\right)\right] \to \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[2\eta(x)(1-\eta(x))]$ as $m\to\infty$

- $\ell =$ 0-1 loss $\quad\Longrightarrow\quad L_{\mathcal{D}}(h_S^{NN}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}\left\{y \neq y_{\pi_1(x)}\right\}\right]$

- $S = \{(x_1,y_1),\ldots,(x_m,y_m)\} \sim \mathcal{D}^m$, and test data $(x,y)\sim\mathcal{D}$
  - View as $x, x_1, \ldots, x_m \sim_{iid} \mathcal{D}_{\mathcal{X}}$ generated first
  - Then labels generated according to $\eta(\cdot)$

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S^{\mathrm{NN}}\right)\right] &= \mathbb{E}_{S,(x,y)\sim\mathcal{D}^{m+1}}\left[\mathbf{1}\left\{y\neq y_{\pi_1(x)}\right\}\right]\\ &= \mathbb{E}_{x,x_1,\ldots,x_m}\left[\mathbb{E}_{y,y_1,\ldots,y_m}\left[\mathbf{1}\left\{y\neq y_{\pi_1(x)}\right\}\;\middle|\; x,x_1,\ldots,x_m\right]\right]\end{aligned}$$

Proof of Theorem kNN.1 (continued)

- Conditioned on $x, x_{\pi_1(x)}$, labels y and $y_{\pi_1(x)}$ are independent

$$\mathbb{E}_{y,y_1,\ldots,y_m}\left[\mathbf{1}\left\{y \neq y_{\pi_1(x)}\right\} \;\middle|\; x, x_1, \ldots, x_m\right] = \eta(x)(1-\eta(x_{\pi_1(x)})) + (1-\eta(x))\eta(x_{\pi_1(x)})$$

- Need to show: As $m \to \infty$,

$$\mathbb{E}_{x,x_1,\ldots,x_m}\left[\eta(x)(1-\eta(x_{\pi_1(x)})) + (1-\eta(x))\eta(x_{\pi_1(x)})\right] - \mathbb{E}_x\left[2\eta(x)(1-\eta(x))\right] \to 0$$

Equivalently,

$$\left|\mathbb{E}_{x,x_1,\ldots,x_m}\left[\eta(x)(1-\eta(x_{\pi_1(x)})) + (1-\eta(x))\eta(x_{\pi_1(x)}) - 2\eta(x)(1-\eta(x))\right]\right| \to 0$$

- Jensen's inequality: If $f(z)$ is a convex function, then $f(\mathbb{E}[z]) \leq \mathbb{E}[f(z)]$
  Example: $\left|\mathbb{E}[z]\right| \leq \mathbb{E}[|z|]$

Proof of Theorem kNN.1 (continued)

$$\begin{aligned}
&\left|\eta(x)(1-\eta(x_{\pi_1(x)})) + (1-\eta(x))\eta(x_{\pi_1(x)}) - 2\eta(x)(1-\eta(x))\right| \\
&\qquad = \left|(1-2\eta(x))(\eta(x_{\pi_1(x)}) - \eta(x))\right| \\
&\qquad = |1-2\eta(x)| \cdot |\eta(x_{\pi_1(x)}) - \eta(x)| \\
&\qquad \leq |\eta(x_{\pi_1(x)}) - \eta(x)| \qquad\qquad \text{since } |1-2\eta(x)| \leq 1 \text{ for all } x
\end{aligned}$$

- Hence, suffices to show that: $\mathbb{E}_{x,x_1,\ldots,x_m}\left[|\eta(x_{\pi_1(x)}) - \eta(x)|\right] \to 0$

Proof of Theorem kNN.1 (adding assumptions on $\mathcal{D}_\mathcal{X}, \eta$)

(A1) $\eta : \mathcal{X} \to [0,1]$ is uniformly continuous

- Uniform continuity:
  η is uniformly continuous if for every $\delta > 0$, there exists $\epsilon_\delta > 0$ such that for every $x, x' \in \mathcal{X}$ with $\|x - x'\| \leq \epsilon_\delta$, $|\eta(x) - \eta(x')| \leq \delta$

- Equivalently, there exists $\epsilon_\delta > 0$ such that $|\eta(x) - \eta(x')| > \delta \implies \|x - x'\| > \epsilon_\delta$

(A2) $\text{support}(\mathcal{D}_\mathcal{X}) = \mathcal{X}$

- Define $\mathcal{D}_\mathcal{X}(x;\epsilon) = \mathbb{P}_{x' \sim \mathcal{D}_\mathcal{X}}(\|x' - x\| \leq \epsilon)$, probability mass in ϵ-neighbourhood of x

- support$(\mathcal{D}_\mathcal{X})$ = set of all $x \in \mathcal{X}$ for which $\mathcal{D}_\mathcal{X}(x;\epsilon) > 0$ for every $\epsilon > 0$

- These assumptions are not necessary, but lead to a simpler proof

Proof of Theorem kNN.1 (continued)

- Choose any $\delta \in (0,1)$. We can write

$$\begin{aligned}\mathbb{E}\left[|\eta(x_{\pi_1(x)}) - \eta(x)|\right] &= \mathbb{E}\left[|\eta(x_{\pi_1(x)}) - \eta(x)| \cdot \mathbf{1}\left\{|\eta(x_{\pi_1(x)}) - \eta(x)| > \delta\right\}\right] \\ &\quad + \mathbb{E}\left[|\eta(x_{\pi_1(x)}) - \eta(x)| \cdot \mathbf{1}\left\{|\eta(x_{\pi_1(x)}) - \eta(x)| \leq \delta\right\}\right] \\ &\leq \mathbb{P}\left(|\eta(x_{\pi_1(x)}) - \eta(x)| > \delta\right) + \delta\end{aligned}$$

- From uniform continuity of η: $|\eta(x_{\pi_1(x)}) - \eta(x)| > \delta \implies \|x_{\pi_1(x)} - x\| > \epsilon_\delta$

$$\begin{aligned}\mathbb{P}_{x,x_1,\ldots,x_m}\left(|\eta(x_{\pi_1(x)}) - \eta(x)| > \delta\right) &\leq \mathbb{P}_{x,x_1,\ldots,x_m}\left(\|x_{\pi_1(x)} - x\| > \epsilon_\delta\right) \\ &= \mathbb{P}_{x,x_1,\ldots,x_m}\left(\min_{i\in\{1,\ldots,m\}} \|x_i - x\| > \epsilon_\delta\right) \\ &= \mathbb{E}_x\left[\mathbb{P}_{x_1,\ldots,x_m}\left(\min_{i\in\{1,\ldots,m\}} \|x_i - x\| > \epsilon_\delta \,\middle|\, x\right)\right]\end{aligned}$$

Proof of Theorem kNN.1 (continued)

- Recall $x_1, \ldots, x_m$ are independent

$$
\begin{aligned}
\mathbb{P}_{x_1,\ldots,x_m}\left(\min_{i\in\{1,\ldots,m\}} \|x_i - x\| > \epsilon_\delta \,\middle|\, x\right) &= \prod_{i=1}^{m} \mathbb{P}_{x_i}\left(\|x_i - x\| > \epsilon_\delta \,\middle|\, x\right) \\
&= \prod_{i=1}^{m} \left(1 - \mathbb{P}_{x_i}\left(\|x_i - x\| \leq \epsilon_\delta \,\middle|\, x\right)\right) \\
&= \left(1 - \mathcal{D}_{\mathcal{X}}(x; \epsilon_\delta)\right)^m
\end{aligned}
$$

- For every $x \in \text{support}(\mathcal{D}_{\mathcal{X}})$, $\mathcal{D}_{\mathcal{X}}(x; \epsilon_\delta) > 0$ and so

$$
\left(1 - \mathcal{D}_{\mathcal{X}}(x; \epsilon_\delta)\right)^m \to 0 \text{ as } m \to \infty
$$

- By assumption (A2), this is true for evert $x \in \mathcal{X}$. So for every $\epsilon_\delta > 0$

$$
\mathbb{E}_x\left[\mathbb{P}_{x_1,\ldots,x_m}\left(\min_{i\in\{1,\ldots,m\}} \|x_i - x\| > \epsilon_\delta \,\middle|\, x\right)\right] \to 0 \text{ as } m \to \infty
$$

Proof of Theorem kNN.1 (conclusion)

Combining everything

$$
\begin{aligned}
&\left|\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S^{\mathrm{NN}}\right)\right] - L_{\mathcal{D}}^*\right| \\
&= \left|\mathbb{E}_{x,x_1,\ldots,x_m}\left[\eta(x)(1-\eta(x_{\pi_1(x)})) + (1-\eta(x))\eta(x_{\pi_1(x)}) - 2\eta(x)(1-\eta(x))\right]\right| \\
&\leq \mathbb{E}_{x,x_1,\ldots,x_m}\left[|\eta(x_{\pi_1(x)}) - \eta(x)|\right] \\
&\leq \underbrace{\mathbb{P}_{x,x_1,\ldots,x_m}\left(|\eta(x_{\pi_1(x)}) - \eta(x)| > \delta\right)}_{\to 0 \text{ as } m\to\infty} + \delta \qquad \text{for any chosen } \delta > 0
\end{aligned}
$$

For any $\delta > 0$, $\lim\limits_{m\to\infty}\left|\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S^{\mathrm{NN}}\right)\right] - L_{\mathcal{D}}^*\right| \leq \delta$

Hence, limit must be zero. Concludes the proof.

Proof idea of Theorem kNN.1, kNN.2 (without assumptions)

- Use above proof till you get $\mathbb{E}_{x,x_1,\ldots,x_m}\left[|\eta(x_{\pi_1(x)}) - \eta(x)|\right]$
- Then one applies 2nd statement below with $k = 1$, which has no assumption on $\eta, \mathcal{D}_\mathcal{X}$
- Proof skipped. If interested, see Lemmas 5.3-5.4 in Devroye's book

Lemma kNN.3 (Convergence of function computed on k-NN)

Let $f : \mathcal{X} \to \mathbb{R}$ be an integrable function. If $\dfrac{k}{m} \to 0$, then

(i) $\displaystyle \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left[\left|f(x_{\pi_i(x)})\right|\right] \leq \left(\left(1 + \tfrac{2}{\sqrt{2-\sqrt{3}}}\right)^p - 1\right)\mathbb{E}[|f(x)|]$ *(Stone's lemma)*

(ii) $\displaystyle \mathbb{E}_{x,x_1,\ldots,x_m \sim \mathcal{D}_\mathcal{X}^{m+1}}\left[\frac{1}{k}\sum_{i=1}^{k}\left|f(x_{\pi_i(x)}) - f(x)\right|\right] \to 0 \quad$ *as* $m \to \infty$

Consistency and Universal consistency

- $\mathcal{D}$ = distribution on $\mathcal{X} \times \mathcal{Y}$
- h_S = predictor learned by algorithm $\mathcal{A}$ given sample $S \sim \mathcal{D}^m$
- $\mathcal{A}$ is consistent with respect to $\mathcal{D}$ and specified loss if

$$\mathbb{E}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S)] \to L^*_{\mathcal{D}} \qquad \text{as } m \to \infty$$

- $\mathcal{A}$ is universally consistent if it is consistent for every $\mathcal{D}$

Practical approach to Bayes classification

- Bayes binary classifier
  - $h^*(x) = \mathbf{1}\left\{\eta(x) \geq \frac{1}{2}\right\}$ or, $h^*(x) = \text{sign}\left(\eta(x) - \frac{1}{2}\right) \in \{-1, +1\}$
  - Main challenge: $\eta(\cdot)$ not known
- Plug-in classifier:
  - $\widehat{\eta}(\cdot) =$ estimate $\eta(\cdot)$ from labelled examples S
  - Predictor

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{\eta}(x) \geq \dfrac{1}{2} \\ 0 & \text{otherwise} \end{cases} \qquad \text{OR} \qquad \widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{\eta}(x) \geq \dfrac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

Example of plug-in classifier: Naïve Bayes

- From Bayes theorem:

$$\eta(x) = \mathbb{P}(y=1|x) = \frac{\mathbb{P}(y=1)\mathbb{P}(x|y=1)}{\mathbb{P}(x)} \qquad 1-\eta(x) = \frac{\mathbb{P}(y=0)\mathbb{P}(x|y=0)}{\mathbb{P}(x)}$$

- Rewriting Bayes classifier: $h^*(x) = \mathbf{1}\left\{\mathbb{P}(y=1)\mathbb{P}(x|y=1) > \mathbb{P}(y=0)\mathbb{P}(x|y=0)\right\}$

- A plug-in classifier: $\widehat{h}(x) = \mathbf{1}\left\{\widehat{p}_1\widehat{b}_1(x) > \widehat{p}_0\widehat{b}_0(x)\right\}$

  - Easy to estimate $\mathbb{P}(y=i)$; $\qquad \widehat{p}_i =$ fraction of training data with label-i

  - Difficult to estimate class-conditional probability/density $\mathbb{P}(x|y)$ if x is high dimensional

  - Naïve Bayes: For $x = \left(x^{(1)},\ldots,x^{(p)}\right) \in \mathbb{R}^p$, assume $\mathbb{P}(x|y=i) = \underbrace{\prod_{j=1}^{p} \mathbb{P}\left(x^{(j)}|y=i\right)}_{\widehat{b}_i(x) \text{ estimates this}}$

Example of plug-in classifier: NN rule

- $\widehat{\eta}(\cdot)$ as weighted average:
  - Given $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$
  - For test data x, define weights $w_1(x), \dots, w_m(x) \in [0, 1]$ with $\sum_{i=1}^{m} w_i(x) = 1$

$$\widehat{\eta}(x) \;=\; \sum_{i=1}^{m} \mathbf{1}\{y_i = 1\}\, w_i(x)$$

- NN rule: $w_i(x) = 1$ for $i = \pi_1(x)$, and 0 otherwise

$$\widehat{\eta}(x) \;=\; \mathbf{1}\left\{y_{\pi_1(x)} = 1\right\} \qquad \Longrightarrow \qquad \widehat{h}(x) = y_{\pi_1(x)}$$

- Questions: What are $w_1(\cdot), \dots, w_m(\cdot)$ for kNN?

  Can we write Naïve Bayes in terms of weighted average?

Universal consistency of plug-in classifiers

Theorem kNN.4 (Stone's consistency theorem)

$\widehat{h}$ *is universally consistent if weights for estimating* $\widehat{\eta}$ *satisfy:*

(i) $\exists c$ *such that, for every non-negative integrable function* f *with* $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)] < \infty$,

$$\mathbb{E}_{x,x_1,\ldots,x_m\sim\mathcal{D}^{m+1}}\left[\sum_{i=1}^{m} w_i(x)\cdot f(x_i)\right] \leq c\,\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)]$$

(ii) *For all* $a > 0$, $\displaystyle\lim_{m\to\infty}\mathbb{E}_{x,x_1,\ldots,x_m\sim\mathcal{D}^{m+1}}\left[\sum_{i=1}^{m} w_i(x)\cdot \mathbf{1}\left\{\|x_i - x\| > a\right\}\right] = 0$

(iii) $\displaystyle\lim_{m\to\infty}\mathbb{E}_{x,x_1,\ldots,x_m\sim\mathcal{D}^{m+1}}\left[\max_{i\in[m]} w_i(x)\right] = 0$

k-nearest neighbour rule

- Assume $\mathcal{X} \subset \mathbb{R}^p$ and we use Euclidean distance $\|x - x'\|$
- Given $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$
- k-nearest neighbour rule:
  - For test data $x \in \mathcal{X}$, sort $x_1, \ldots, x_m$ according to $\|x - x_i\|$
  - $\pi_k(x) =$ index for is k-th nearest neighbour of x
  - Predict $h_S^{kNN}(x) =$ majority vote of $y_{\pi_1(x)}, \ldots, y_{\pi_k(x)}$

OR for ± 1 labels, $h_S^{kNN}(x) = \text{sign}\left(\frac{1}{k}\sum_{i=1}^{k} y_{\pi_i(x)}\right)$

Universal consistency of k-NN

Theorem kNN.5 (Universal consistency of k-NN)

If $k \to \infty$ and $\dfrac{k}{m} \to 0$ as $m \to \infty$, then for all distributions $\mathcal{D}$,

$$\mathbb{E}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}} \left(h_S^{kNN} \right) \right] \to L_{\mathcal{D}}^* \qquad \textit{as } m \to \infty$$

Proved by verifying conditions of Stone's theorem

Proof: Universal consistency of k-NN

- k-NN as plug-in classifier

$$\widehat{\eta}(x) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\left\{y_{\pi_i(x)} = 1\right\} = \sum_{i=1}^{m} \mathbf{1}\left\{y_i = 1\right\} \underbrace{\frac{\mathbf{1}\left\{\pi_i(x) \leq k\right\}}{k}}_{=w_i(x)}$$

- $\displaystyle\sum_{i=1}^{m} w_i(x) \cdot f(x) = \frac{1}{k}\sum_{i=1}^{k} f\left(x_{\pi_i(x)}\right) \implies$ Condition (i) holds due to Stone's lemma

- Condition (ii) holds since $x_{\pi_k(x)} \to x$ in probability if $\dfrac{k}{m} \to 0$

- Condition (iii) holds since $\displaystyle\max_{i\in[m]} w_i(x) = \frac{1}{k} \to 0$ as $k \to \infty$

Recap Stone's theorem

Theorem kNN.6 (Stone's consistency theorem)

Let $\widehat{\eta}(x) = \sum_{i=1}^{m} \mathbf{1}\{y_i = 1\}\, w_i(x)$, *and* $\widehat{h}(x) = \mathbf{1}\left\{\widehat{\eta}(x) \geq \frac{1}{2}\right\}$.

$\widehat{h}$ *is universally consistent if weights* $w_i(x)$ *satisfy:*

(i) $\exists c$ *such that, for every non-negative integrable function* f *with* $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)] < \infty$,

$$\mathbb{E}_{x,x_1,\ldots,x_m\sim\mathcal{D}^{m+1}}\left[\sum_{i=1}^{m} w_i(x)\cdot f(x_i)\right] \leq c\,\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)]$$

(ii) *For all* $a > 0$, $\quad \lim_{m\to\infty} \mathbb{E}_{x,x_1,\ldots,x_m\sim\mathcal{D}^{m+1}}\left[\sum_{i=1}^{m} w_i(x)\cdot \mathbf{1}\{\|x_i - x\| > a\}\right] = 0$

(iii) $\lim_{m\to\infty} \mathbb{E}_{x,x_1,\ldots,x_m\sim\mathcal{D}^{m+1}}\left[\max_{i\in[m]} w_i(x)\right] = 0$

Proof of Stone's theorem: Main idea

Lemma kNN.7 (Risk bound for plug-in classifier)

Consider 0-1 loss, and let $\widehat{h}(x) = \mathbf{1}\left\{\widehat{\eta}(x) \geq \frac{1}{2}\right\}$ *be a plug-in classifier. Then*

$$L_{\mathcal{D}}(\widehat{h}) - L^*_{\mathcal{D}} \;\leq\; 2\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\,|\widehat{\eta}(x) - \eta(x)|\,\right] \;\leq\; 2\sqrt{\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[(\widehat{\eta}(x) - \eta(x))^2\right]}$$

Proof idea: (complete the steps)

- 2nd inequality uses Jensen's inequality: For a convex function f, $f(\mathbb{E}[z]) \leq \mathbb{E}[f(z)]$
- 1st inequality: First show that $L_{\mathcal{D}}(\widehat{h}) - L^*_{\mathcal{D}} = 2\mathbb{E}_x\left[\left|\eta(x) - \frac{1}{2}\right| \cdot \mathbf{1}\left\{h^*(x) \neq \widehat{h}(x)\right\}\right]$

  Then find upper bound for the term inside observing that, whenever $h^*(x) \neq \widehat{h}(x)$, $\left|\eta(x) - \frac{1}{2}\right| \leq |\eta(x) - \widehat{\eta}(x)|$

Proof of Stone's theorem: Main idea

- Taking expectation w.r.t S, we can write

$$\mathbb{E}_S[L_\mathcal{D}(\widehat{h}) - L^*_\mathcal{D}] \;\leq\; 2\mathbb{E}_{S,x}\Big[\,|\widehat{\eta}(x) - \eta(x)|\,\Big] \;\leq\; 2\sqrt{\mathbb{E}_{S,x}\left[\left(\widehat{\eta}(x) - \eta(x)\right)^2\right]}$$

- Due to above, suffices to show

$$\mathbb{E}_{S,x}\left[\left(\widehat{\eta}(x) - \eta(x)\right)^2\right] \to 0 \qquad \text{as } m \to \infty$$

- Assumption: $\eta : \mathcal{X} \to [0,1]$ is uniformly continuous
  - The assumption is not necessary, but simplifies parts of the proof

Proof of Stone's theorem: Main idea

- Recall $\widehat{\eta}(x) = \sum_{i=1}^{m} \mathbf{1}\{y_i = 1\} w_i(x)$ and define $\widetilde{\eta}(x) = \sum_{i=1}^{m} \eta(x_i) w_i(x)$

$$\text{Then} \quad \left(\widehat{\eta}(x) - \eta(x)\right)^2 = \left(\widehat{\eta}(x) - \widetilde{\eta}(x) + \widetilde{\eta}(x) - \eta(x)\right)^2$$
$$\leq 2\Big(\left(\widehat{\eta}(x) - \widetilde{\eta}(x)\right)^2 + \left(\widetilde{\eta}(x) - \eta(x)\right)^2\Big)$$

- Separately show expectation of each squared term goes to 0

- Note: $\widehat{\eta}(x) - \widetilde{\eta}(x) = \sum_{i=1}^{m} (\mathbf{1}\{y_i = 1\} - \eta(x_i)) w_i(x)$

  and $\quad \widetilde{\eta}(x) - \eta(x) = \sum_{i=1}^{m} (\eta(x_i) - \eta(x)) w_i(x)$

Proof: $\left(\widehat{\eta}(x) - \widetilde{\eta}(x)\right)^2 \to 0$ in expectation

$$\mathbb{E}_{S,x}\left[\left(\widehat{\eta}(x) - \widetilde{\eta}(x)\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{1}\{y_i = 1\} - \eta(x_i))w_i(x) \cdot (\mathbf{1}\{y_j = 1\} - \eta(x_j))w_j(x)\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{m}\underbrace{(\mathbf{1}\{y_i = 1\} - \eta(x_i))^2}_{\leq 1}(w_i(x))^2\right]$$

given x, x_i, x_j,
y_i independent of y_j
Expectation $= 0$ for $i \neq j$

$$\leq \mathbb{E}\left[\sum_{i=1}^{m}(w_i(x))^2\right]$$

$$\leq \mathbb{E}\left[\max_{i\in[m]} w_i(x) \cdot \underbrace{\sum_{i=1}^{m} w_i(x)}_{=1}\right] \quad = \quad \mathbb{E}\left[\max_{i\in[m]} w_i(x)\right] \quad \to \quad 0 \quad \text{(condition (iii))}$$

Proof: $(\widetilde{\eta}(x) - \eta(x))^2 \to 0$ in expectation

$$\mathbb{E}_{S,x}\left[(\widetilde{\eta}(x) - \eta(x))^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{m} w_i(x) \cdot (\eta(x_i) - \eta(x))\right)^2\right]$$

$$\leq \mathbb{E}\left[\sum_{i=1}^{m} w_i(x) \cdot (\eta(x_i) - \eta(x))^2\right] \qquad \text{(Jensen's inequality)}$$

- Jensen's inequality: For convex f and weights $w_1, \ldots, w_m$ such that $\sum_i w_i = 1$,

$$f\left(\sum_i w_i z_i\right) \leq \sum_i w_i f(z_i)$$

- Fix some $\epsilon > 0$. Since η is uniformly continuous,

$$\exists a_\epsilon > 0 \;\text{ such that }\; \|x_i - x\| \leq a_\epsilon \implies |\eta(x_i) - \eta(x)| \leq \epsilon$$

Proof: $(\widetilde{\eta}(x) - \eta(x))^2 \to 0$ in expectation

$$\mathbb{E}\left[\sum_{i=1}^{m} w_i(x) \cdot (\eta(x_i) - \eta(x))^2\right]$$
$$\leq \underbrace{\mathbb{E}\left[\sum_{i=1}^{m} w_i(x) \cdot \epsilon^2 \cdot \mathbf{1}\left\{\|x_i - x\| \leq a_\epsilon\right\}\right]}_{\leq \epsilon^2} + \underbrace{\mathbb{E}\left[\sum_{i=1}^{m} w_i(x) \cdot 1 \cdot \mathbf{1}\left\{\|x_i - x\| > a_\epsilon\right\}\right]}_{\to 0 \text{ due to condition (ii)}}$$

- From above, for any $\epsilon > 0$,

$$\lim_{m\to\infty} \mathbb{E}_{S,x}\left[(\widetilde{\eta}(x) - \eta(x))^2\right] \leq \epsilon^2$$

- Hence, limit is 0

Proof: Need for condition (i)

- Above proof, assuming η is uniformly continuous, does not need condition (i)
- If uniform continuity is not assumed
  - $\eta(\cdot)$ is bounded $\implies$ Can be approximated by a uniformly continuous function η^*

$$\text{for any } \epsilon > 0, \ \exists \text{ unif. cont. } \eta^* \text{ such that } \mathbb{E}_x\left[(\eta(x) - \eta^*(x))^2\right] < \epsilon$$

- Use previous slide to prove $\mathbb{E}_{S,x}\left[(\widetilde{\eta}^*(x) - \eta^*(x))^2\right] \to 0$
- Need condition (i) to bound $\mathbb{E}_{S,x}\left[(\widetilde{\eta}(x) - \widetilde{\eta}^*(x))^2\right]$ and $\mathbb{E}_{S,x}\left[(\eta^*(x) - \eta(x))^2\right]$

Conclusion / Up Next

- Given infinite training data $(m \to \infty)$, NN (or kNN) is good compared with optimal (Bayes) predictor, $L_{\mathcal{D}}^{NN} \leq 2L_{\mathcal{D}}^{*}$

- With $k \to \infty$ and $k/m \to 0$, kNN is universally consistent
  $\ldots$ optimal for any $\mathcal{D}$ if it has infinite training data, $m \to \infty$

- Next: What happens for finite m?
  - For finite m, complex models can easily overfit
  - We restrict ERM to certain classes of models (say linear classifier)
  - For ERM solution $\widehat{h}$, we bound $L_{\mathcal{D}}(\widehat{h})$ in terms of complexity of model class