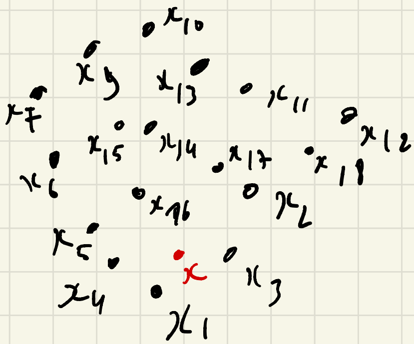
K-nearest Neighbor Classifier

K-nearest neighbor

- $(X, Y) \sim D$ $\quad S_{Train} = \{(x_i, y_i)\}_{i=1}^{N}$ iid $\sim D$
- Nearest neighbors function:

$$nbh_{S_{Train}, k}(x) = \{\text{set of } k \text{ elements of } S_{Train} \text{ closest to } x\}$$

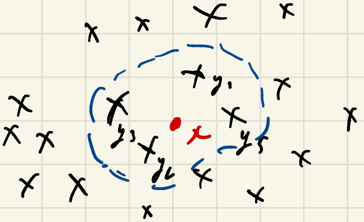
- ex:



$$nbh_{S_{Tr}, 4}(x) = \{x_1, x_3, x_4, x_{16}\}$$

- Regression: $y \in \mathbb{R}$

$$f_{S_{Train}, k}(x) = \frac{1}{k} \sum_{i \in nbh_{S_{Tr}, k}(x)} y_i$$

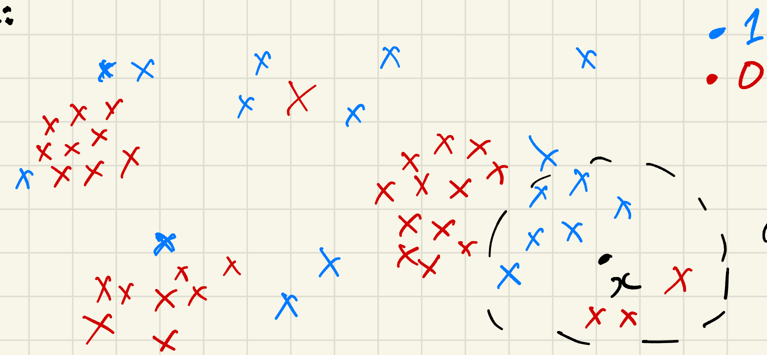


$$f(x) = \frac{1}{4}(y_1 + y_2 + y_3 + y_5)$$

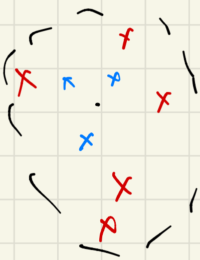
• **Classification**: $y \in \{0,1\}$ (Binary classification)

$$f_{S_{Tn},k}(x) = \underset{i \in \text{nbh}_{S_{Tn},k}(x)}{\text{majority}} (y_i)$$

Ex:

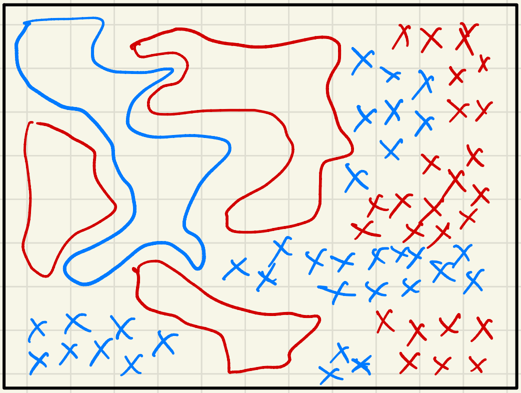


$f_{S_{Tn},9}(x) = 1$



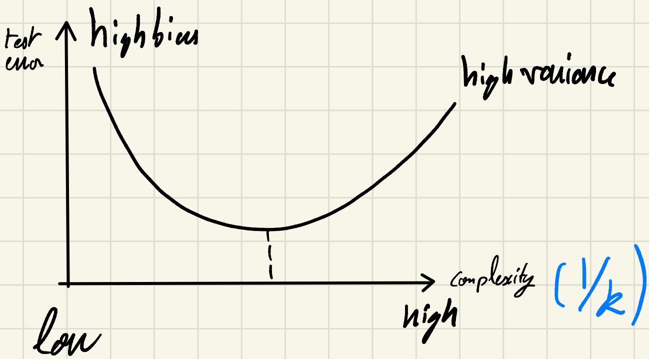
, weight vote ?

why is it meaningful ?



- meaningful when there is spatial correlation
- Implicitly learn very complex decision boundaries in low dimension!

• Bias variance in function of K.



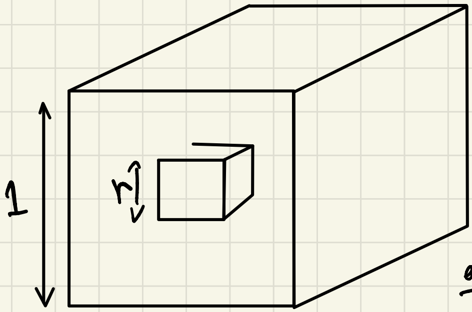
- small K, $K=1$ ⟶ complex model: low bias, high variance
- large K, $K = \#$ samples ⟶ simple model (high bias) (low variance)

• complexity scales as $1/k$

→ complexity ↗ when k ↘

Curse of dimensionality

$\mathcal{X} = [0,1]^d$



. what is the chance that an uniform point in the large box is in the small box

$\rightarrow r^d$ (very small)

. $(1-r^d)^N$: none of the samples is inside the small box.

ex: if $N = 500$, $d = 10$ then $r = 0{,}52$ to find a neighbor

- **Analysis:**

$(x,y) \sim D$; $y \in \{0,1\}$

. $\mathcal{L}(f) = \mathbb{P}(y \neq f(x))$

Bayes Predictor:

. $\eta(x) = \mathbb{P}(y = 1 \mid x) = 1 - \mathbb{P}(y = 0 \mid x)$

if I know η, how can I predict when given x_0

if $\eta(x_0) \geq 1/2 \rightarrow 1$

if $\eta(x_0) < 1/2 \rightarrow 0$

. $f^*(x) = \mathbb{1}_{\eta(x) \geq 1/2}$

$$\mathcal{L}(f^*) = \mathbb{P}(f^*(x) \neq y) = \mathbb{E}_{x \sim D_x}\left[\min\{\eta(x), 1 - \eta(x)\}\right]$$

• **Claim**

$$\mathbb{E}_{S_{Train}}[\mathcal{L}(f_{S_{Train},1})] \leq 2\,\mathcal{L}(f^*) + \text{geometric term}$$

. We assume that $\exists\, c \geq 0, \forall\, x, x' \in \mathcal{X}$

$$|\eta(x) - \eta(x')| \leq c\, \|x - x'\|_2$$

$\Rightarrow$ Geometric term $= c\, \mathbb{E}_{x \sim D_{S_{Train}}}[\|x - \text{nbh}_{S_{Train},1}(x)\|]$

$$\boxed{\mathbb{E}_{S_{Train}}[\mathcal{L}(f_{S_{Train},1})] \leq 2\,\mathcal{L}(f^*) + c\, \mathbb{E}_{x \sim D_{S_{Train}}}[\|x - \text{nbh}_{S_{Train},1}(x)\|]}$$

$$\boxed{\mathbb{E}_{S_{Tr}}[\mathcal{L}(f_{S_{Train},1})] \leq 2\,\mathcal{L}(f^*) + 4c\sqrt{d}\,(N)^{-\frac{1}{d+1}}}$$

$$\mathcal{L}(f_{S_{Tr}}) = \mathbb{P}(y \neq f_{S_{Tr}}(x))$$

• **Proof**:

given 2 samples x and x'

(x', y') is in the training set

$\overset{\times}{x}$ $\quad$ $\overset{\times}{x'}$ $\quad$ I want to predict the label of x

Question: $\mathbb{P}(y \neq y')$?

Simple case: $x = x'$

$$\mathbb{P}(y' \neq y) = 2\eta(x)(1-\eta(x))$$

$y' = 0 \quad y = 1$ and the inverse

$(1-\eta(x)) \quad \eta(x)$

$$\mathbb{P}(y' \neq y) = 2\eta(x)(1-\eta(x)) \leq 2\min\{\eta(x), 1-\eta(x)\}$$

• general case:

$$\mathbb{P}(y \neq y') = \eta(x)(1-\eta(x')) + \eta(x')(1-\eta(x))$$

$$= \eta(x)(1-\eta(x)) + \eta(x)[\eta(x') - \eta(x)]$$
$$+ \eta(x)(1-\eta(x)) + (\eta(x') - \eta(x))(1-\eta(x))$$
$$= 2\eta(x)(1-\eta(x)) + \eta(x) - \eta(x')$$

$$\mathbb{P}(y \neq y') \leq 2\eta(x)(1-\eta(x)) + |\eta(x) - \eta(x')|$$

$$\leq 2\eta(x)(1-\eta(x)) + C\|x - x'\|_2$$
$$\leq 2\min\{\eta(x), 1-\eta(x)\} + C\|x - x'\|_2$$

Lemma: Given x and x'

$$\mathbb{P}(y \neq y') \leq 2\min\{\eta(x), 1-\eta(x)\} + C\|x - x'\|$$

• Draw S_{Train} then we draw $x \sim D$, define $x' = nbh_{S_{Train}}(x)$

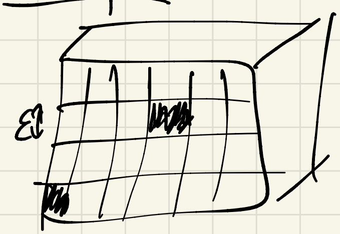
$$\mathbb{E}_{\substack{S_{Train}\\ x}}[\mathbb{P}(y \neq y')] \leq 2\mathbb{E}_{x\sim D}[\min(\eta(x), 1-\eta(x))] + C\mathbb{E}_{\substack{x\sim D\\ S_{Train}}}[\|x - x'\|]$$
$$\leq 2\,2(f^*) + \mathbb{E}[\|x - nbh(x)\|]$$

$$\mathbb{E}_{\substack{S_{Train} \\ x\sim D}}[\mathbb{P}(y\neq y')] = \mathbb{E}_{S_{Train}}\left[\mathbb{E}_{x\sim D}[\mathbb{P}(y\neq f_{S_{Tr},1}(x))]\right]$$

$$= \mathbb{E}_{S_{Tr}}\left[\mathbb{P}_{x,y\sim D}(y\neq f_{S_{Tr},1}(x))\right]$$

$$= \mathbb{E}_{S_{Tr}}\left[\mathcal{L}(f_{S_{Tr}})\right] \quad \blacksquare$$

Second part:



$[0,1]^d$

$\mathbb{P}(x \sim \text{box } k) = p_k$

- if we are lucky, this box contains an element of S_{Train}.

  → its distance will be at most $\sqrt{d}\cdot\varepsilon$ $(0<\varepsilon\ll 1)$

- if we are unlucky: no element of S_{Train} in the box.

  → the distance can be as large as $\sqrt{d}$

. The prob. that if we pick a box k we get no train. sample inside:

$$(1-p_k)^N$$

$$\sum p_k\left[(1-p_k)^N\sqrt{d} + \left(1-(1-p_k)^N\right)\sqrt{d}\,\varepsilon\right]$$

if $0 \ll p_k < 1 \to 1-p_k \ll 1$

$0 < p_k \ll 1 \to p_k \ll 1$

look at the variation of $p(1-p)^n$