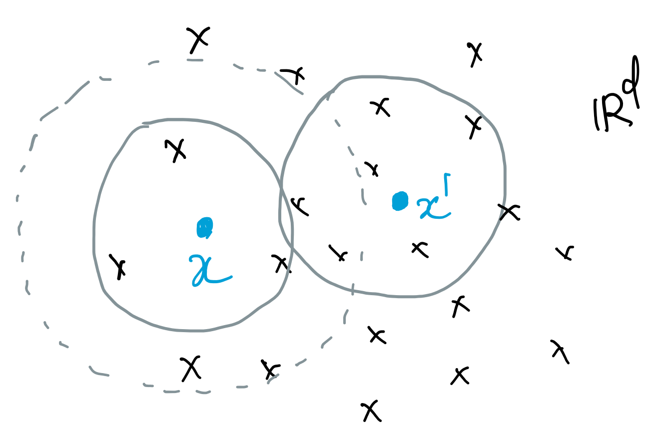
L1: Data analysis and ML basics

- ML bascis
  - Supervised and unsupervised learning
  - Data, model, and algorithm
  - No-free-lunch theorem
- Training/test split and generalization error
  - underfitting and overfitting
  - Bias-variance trade-off
  - Model selection and validation
- The break-down of three errors in machine learning
- Overview of kernel methods
  - Parametric and non-parametric model

• KDE bias-variance trade-off

$p(x)$ density function on $\mathbb{R}^d$,

$x_i \sim p(x)dx$, $i = 1, \cdots, n$ i.i.d.



$$\hat{p}_\sigma(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sigma^d} K\left(\frac{\|x_i - x\|}{\sigma}\right), \quad \sigma > 0$$

bandwidth para

$$K(z) = e^{-z^2/2}$$

other functions on $\mathbb{R}^d$, regular, decay

Question: $|\hat{p}_\sigma(x) - p(x)| \leq ?$

- Bias error

$$\mathbb{E}\hat{p}_\sigma(x) = \frac{1}{\sigma^d}\int_{\mathbb{R}^d} K\left(\frac{\|y-x\|}{\sigma}\right) p(y)\, dy := \bar{p}_\sigma(x)$$

If p and K are smooth, σ small,

$y = x + \sigma v$

$$\bar{p}_\sigma(x) = \int_{\mathbb{R}^d} \frac{1}{\sigma^d} K\left(\frac{\|y-x\|}{\sigma}\right) p(y)\, dy$$

$$= \int_{\mathbb{R}^d} K(\|v\|)\, \underbrace{p(x+\sigma v)}_{= p(x) + \sigma \nabla p(x)^T v + O(\sigma^2)}\, dv$$

suppose $K(\|v\|)$ is symmetric, s.t.

$$\int_{\mathbb{R}^d} v K(\|v\|)\, dv = 0,$$

also $K(\|v\|)$ is rescaled by a constant s.t.

$$\int_{\mathbb{R}^d} K(\|v\|)\, dv = 1$$

then

$$\bar{p}_\sigma(x) = \int_{\mathbb{R}^d} K(\|v\|)\left(p(x) + \sigma \nabla p(x)^T v + O(\sigma^2)\right) dv$$

$$= \underbrace{\left(\int_{\mathbb{R}^d} K(\|v\|)\, dv\right)}_{1} p(x) + \sigma \nabla p(x)^T \overbrace{\cancel{\left(\int v K(\|v\|)\, dv\right)}}^{0} + O(\sigma^2)$$

$$= p(x) + O(\sigma^2)$$

- Variance error

$$\hat{p}_\sigma(x) - \mathbb{E}\hat{p}_\sigma(x) = ?$$

For fixed $x \in \mathbb{R}^d$,

$$\hat{p}_\sigma(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{1}{\sigma^d} K\left(\frac{\|x_i - x\|}{\sigma}\right)}_{\xi_i \text{ r.v. i.i.d.}}$$

$$\mathrm{Var}(\xi_i) \leq \mathbb{E}\xi_i^2 = \frac{1}{\sigma^{2d}}\int_{\mathbb{R}^d} K\Big(\underbrace{\frac{\|y-x\|}{\sigma}}_{\|v\|}\Big)^2 p(y)\, dy$$

$$= \frac{1}{\sigma^d}\int_{\mathbb{R}^d} K(\|v\|)^2 p(x+\sigma v)\, dv$$

suppose $\int_{\mathbb{R}^d} K(\|v\|)^2 dv < \infty$,

p bdd. : $|p(x)| \leq C$, $\forall x \in \mathbb{R}^d$,

then

$$\mathbb{E}\xi_i^2 \leq \frac{1}{\sigma^d} \cdot C \cdot \int_{\mathbb{R}^d} K(\|v\|)^2 dv = O\left(\frac{1}{\sigma^d}\right)$$

Then, by concentration of independent sum,

$$|\hat{p}_\sigma(x) - \mathbb{E}\hat{p}_\sigma(x)| = \tilde{O}\left(\sqrt{\frac{1}{n\sigma^d}}\right)$$

- Trade-off.

$$O(\sigma^2) + O(n^{-1/2}\sigma^{-d/2})$$

if we match

$$\sigma^2 \sim n^{-1/2}\sigma^{-d/2}$$

$$\Rightarrow \sigma \sim n^{-\frac{1}{4+d}}$$

The overall error is $O(n^{-\frac{2}{d+4}})$.