

Linear regression vs. Nearest neighbors

- Different treatment of training examples
 - Nearest neighbor is memory- or instance-based: need to remember training set
 - Linear regression is parametric: can discard training set and retain learned parameters
- Different underlying assumptions:
 - Nearest neighbor makes few assumptions about data but choice of k is key
 - Linear regression makes strong assumptions about data

k Nearest Neighbors: Advantages

- retains all information in training instances
- can approximate complex target functions -- only using simple
- local approximations
- need not pre-classify entire input space
- learning can be on-line or batch
- key assumption : **smoothness** (property of input point \mathbf{x} likely to be similar to those of points in its neighborhood)
- but problems with dimensionality: **neighbors in high-dimensional spaces are far away**

k Nearest Neighbors: Disadvantages

- Problems with dimensionality
- Suppose we have a dataset of size N in d -dimensional unit hypercube, points uniformly distributed
 - Assume neighborhoods of side b ; so volume b^d
 - To contain k pts, average neighborhood must occupy k / N of entire volume (=1)
 - $b^d = k / N$
 - $b = (k / N)^{(1 / d)}$
 - $d=2 \quad k=10 \quad N=1\text{mill} \rightarrow b \approx .003$
 - $d=100 \quad k=10 \quad N=1\text{mill} \rightarrow b \approx .89$: almost entire input space
- **Why is the volume of the neighborhood relevant?**
 - What would $b \approx .89$ look like in 2D space?