Machine Learning Topic 2

The Nearest Neighbor Rule

Aimee Barciauskas

29 March 2016

Nearest Neighbors

- Definition of **distance**: *slide 1*
- Risk of Nearest Neighbor: *slide 3*
- Asymptotic Probability of Error Theorem and Proof: *slide 4*
- K-Nearest Neighbor Definition and formula for asymptotic risk: *slide 5-6*
  - Theorem of Universal Consistency of $k - NN$: *slide 7*
- Description of partitioning classifier: *bottom of slide 7-8*
  - Curse of dimensionality for partitioning classifier: *slide 8*

Notes

- The probability that the nearest neighbor "far away" is small when $\epsilon >> \frac{1}{n^{1/d}}$
  - As dimension increases, the number of points required grows exponentially
- If the distance to the nearest neighbors is small and $\eta(x)$ smooth and continuous, then the distribution of $Y^{(1)}$ is similar to $Y' \sim \eta(X)$

Definitions

(1-sided) Risk of the Nearest Neigbor classifier: $\mathbb{I}_{g_k(X)=0,Y=1} = \mathbb{P}\big\{Bin(k, \eta(x)) < \frac{k}{2}|X\big\}$

Nearest Neighbor Theorem

If $X'_n(X)$ are the k nearest neighbors of X from the training set X_n $(X_1, ..., X_n)$ are i.i.d in a seperable metrix space), then:

$$X'_n \to X$$

In other words, they are "close" to X in a sense that they are asymptotically co-located.

Nearest Neighbor Classifiers Notes (Vittorio, Columbia)

Proof of Nearest Neighbor Theorem:

To prove the theorem, we prove the probability that the converse happens goes to zero exponentially fast.

We define "good" points as those with positive probabaility that they fall in $S_x(\delta)$ centered at x with radius δ, e.g.: $\forall \delta > 0$, $\mathbb{P}\{S_x(\delta)\} > 0$

Since the training points are independent, the probabiltiy that all training points lie outside the $S_x(\delta)$ is the probability of each individual training point lies outside $S_x(\delta)$, which is the n-th power of the individual probability.

$$\mathbb{P}\big\{d(X'_n(x), x) > 0\big\} = \mathbb{P}\big\{X'_n(x) \notin S_x(r)\big\} = \big(1 - \mathbb{P}\{S_x(\delta)\}\big)^n \to 0$$

Rate of Convergence to R^*

Depends on distribution (*slide 4a*)

$\mathbb{P}\{Bin(n, \frac{1}{2} = 0 \text{ or } n\} = 2^{-n}+2^{-n},$ probability there is not at least one data point in each of two disjoint buckets
$\mathbb{E}R(g_n) = 2^{-n}$ goes to 0 exponentially fast