MLISP: Machine Learning in Signal Processing

Lecture 5

Prof. V. I. Morgenshtern

Scribe: M. Solomon

Illustrations: The elements of statistical learning, Hastie, Tibshirani, Friedman

Agenda:

1. Local methods in high dimensions
2. Curse of dimensionality
3. Bias and variance for nearest neighbors

1 Local methods in high dimensions

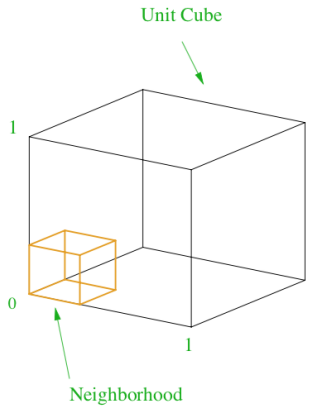
So far we have considered two learning methods:

- The *linear model* that is very rigid.
- The *k-nearest neighbors* algorithm that can flexibly adapt to the data.

It may seem that with a reasonably large training set we can always find many enough observations very close to any $\mathbf{x}$, and therefore we could approximate the theoretically optimal conditional expectation by the k-nearest neighbors algorithm. This approach breaks down in high dimensions as we will see next.

2 The curse of dimensionality

Edge-length of the k-nearest neighborhoods: The reason is that the k-nearest neighborhoods are extremely large in high dimensional spaces. Consider a p-dimensional unit hypercube:



Assume that we have 10000 examples that are uniformly distributed in the hypercube. Suppose we want to make predictions about the point $[0, 0, ..., 0]$. We create a cubic neighborhood around this point to capture the fraction r of the observations.

In our case r could be $1/1000$, so that we have about 10 points in the cube to have reasonable averaging.

What is the edge-length of the cube? It is:

$$l_p(r) = r^{1/p}.$$

Why? This follows from the uniformity of the distribution because the volume of this neighborhood is

$$(r^{1/p})^p = r.$$

- For $p = 2$:

  $$l_p(1/1000) = (1/1000)^{1/2} = 0.03$$

  So the neighborhood is indeed local and very small.

- For $p = 10$:

  $$l_p(1/1000) = (1/1000)^{1/10} = 0.5$$

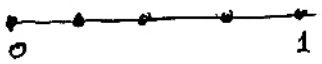
  So the neighborhood spans 50% of the input variable range is no longer local or small.

- For $p = 30$:

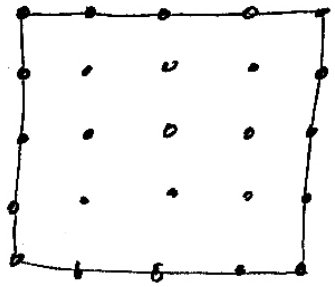
  $$l_p(1/1000) = (1/1000)^{1/30} = 0.75$$

  The neighborhood is very large.

Sampling density: The other side of the same coin is that the sampling density is always small in high dimensions. Let's define the sampling density as the approximate number of points on each line of length one. For example in 1d sampling density of 5 looks like this:



In 2d the sampling density of 5 looks like this:



Therefore to have the sampling density 5 in p dimensions, we need 5^p data points. For $p = 100$, this is more than the total number of atoms in the universe!

3 Bias and variance for 1-nearest neighbors algorithm

Suppose we have 100 training examples $\mathtt{x}^{(i)}$ generated uniformity in the hypercube $[-1, 1]^p$. Assume that the true relationship between $\mathtt{x}$ and $\mathtt{y}$ is nonrandom and given by:

$$y = f(\mathbf{x}) = e^{-8||\mathbf{x}||^2}.$$

There is no measurement error. We use the 1-nearest-neighbor rule to predict y_0 at the test point $\mathbf{x}_0 = \mathbf{0}$. This generates $\hat{\mathtt{y}}_0$. Denote the *random* training set by T.

Let's calculate the expected prediction error at $\mathbf{x}_0$:

$$\text{EPE}(\mathbf{x}_0) = \mathbb{E}_T[f(\mathbf{x}_0) - \hat{\mathtt{y}}_0]^2$$

where the expectation is over there random training set. This expression can be decomposed as

$$\text{EPE}(\mathbf{x}_0) = \mathbb{E}_T[\hat{\mathtt{y}}_0 - \mathbb{E}_T(\hat{\mathtt{y}}_0)]^2 + [\mathbb{E}_T(\hat{\mathtt{y}}_0) - f(\mathbf{x}_0)]^2 + 2\underbrace{\mathbb{E}_T(\hat{\mathtt{y}}_0 - \mathbb{E}_T(\hat{\mathtt{y}}_0))}_{0}\underbrace{(\mathbb{E}_T(\hat{\mathtt{y}}_0) - f(\mathbf{x}_0))}_{\text{nonrandom}}.$$

The first term above is called the *variance* of our estimator and the second is the *squared bias*:

$$\text{EPE}(\mathbf{x}_0) = \text{Var}_T(\hat{\mathtt{y}}_0) + (\text{Bias}(\hat{\mathtt{y}}_0))^2.$$

The variance term measures the variability of our estimate because of the randomness of the training data. The squared bias term measures how far the expected value of our estimator is from the truth. The squared bias term is non-zero as can be observed from Figure 1:

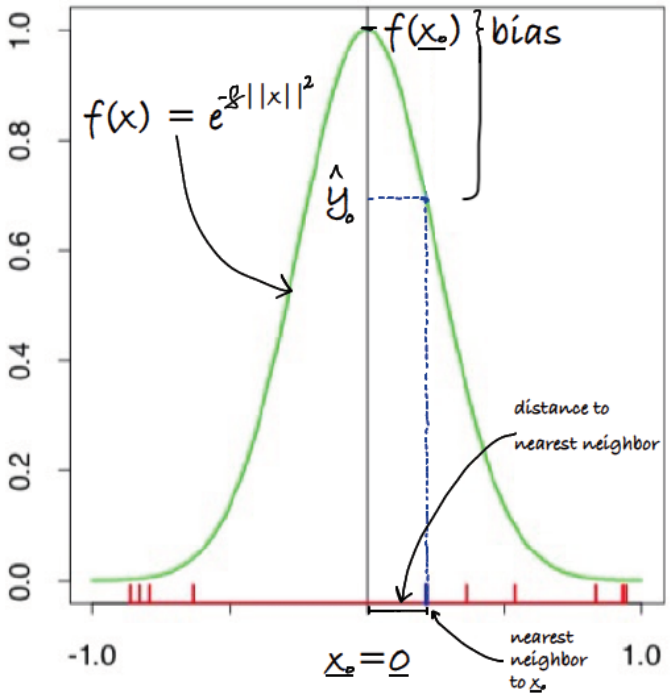
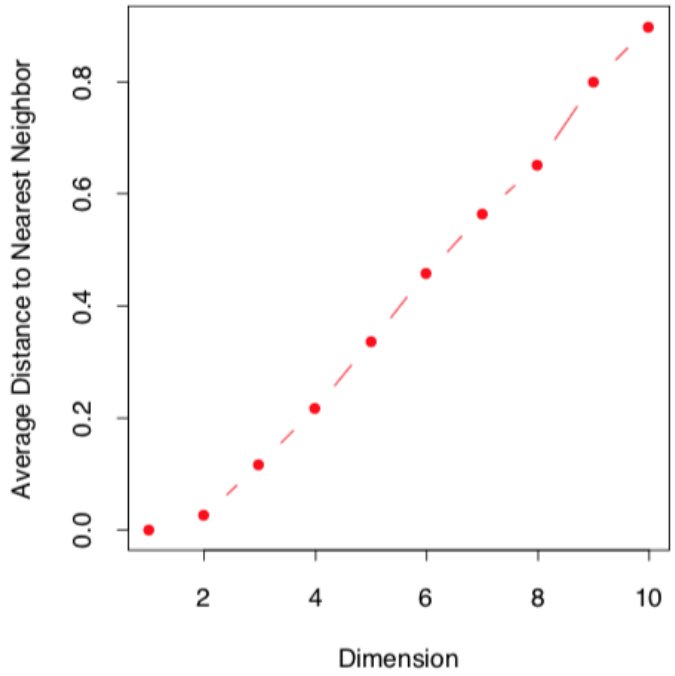
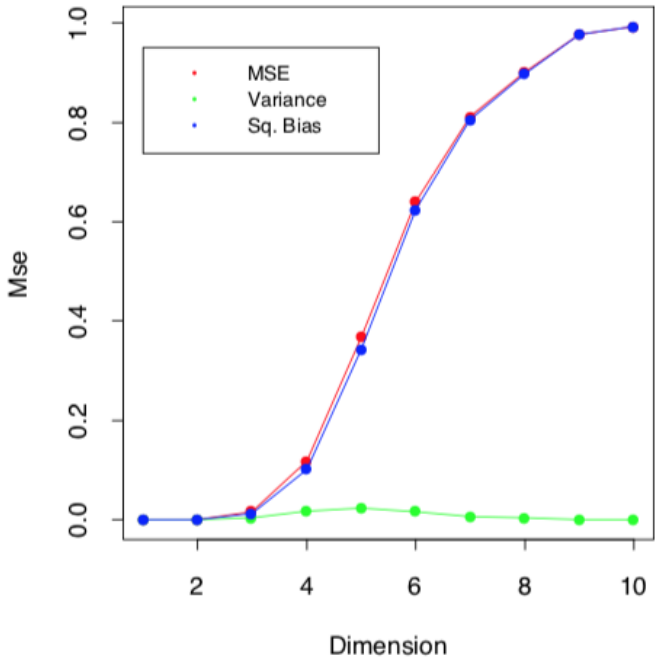


Figure 1: Bias for the 1-nearest neighbors estimator

The bias increases quickly with dimension because the distance to the nearest neighbor increases:



We observe that by $p = 10$ for more than 90% of the samples, the nearest neighbor of the all-zero vector is located further than 0.5 units away from that vector. Therefore, for $p = 10$, the squared bias term is very close to 1:



Also observe that as indicated above, the variance remains small for all p.