1.2 Bias-Variance Analysis

Let's justify this reasoning formally for k-NN applied to regression tasks. Suppose we are given a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where the labels y_i are real valued scalars. We model our hypothesis $h(\mathbf{z})$ as

$$h(\mathbf{z}) = \frac{1}{k}\sum_{i=1}^{n} N(\mathbf{x}_i, \mathbf{z}, k)$$

where the function N is defined as

$$N(\mathbf{x}_i, \mathbf{z}, k) = \begin{cases} y_i & \text{if } \mathbf{x}_i \text{ is one of the } k \text{ closest points to } \mathbf{z} \\ 0 & o.w. \end{cases}$$

Suppose also we assume our labels $y_i = f(\mathbf{x}_i) + \epsilon$, where ϵ is the noise that comes from $\mathcal{N}(0, \sigma^2)$ and f is the true function. Without loss of generality, let $\mathbf{x}_1 \dots \mathbf{x}_k$ be the k closest points. Let's first derive the bias^2 of our model for the given dataset $\mathcal{D}$.

$$\begin{aligned}
\big(\mathbb{E}[h(\mathbf{z})] - f(\mathbf{z})\big)^2 &= \left(\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{n} N(\mathbf{x}_i, \mathbf{z}, k)\right] - f(\mathbf{z})\right)^2 = \left(\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} y_i\right] - f(\mathbf{z})\right)^2 \\
&= \left(\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}[y_i] - f(\mathbf{z})\right)^2 = \left(\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}[f(\mathbf{x}_i) + \epsilon] - f(\mathbf{z})\right)^2 \\
&= \left(\frac{1}{k}\sum_{i=1}^{k} f(\mathbf{x}_i) - f(\mathbf{z})\right)^2
\end{aligned}$$

When $k \longrightarrow \infty$, then $\frac{1}{k}\sum_{i=1}^{k} f(\mathbf{x}_i)$ goes to the average label for $\mathbf{x}$. When $k = 1$, then the bias is simply $f(\mathbf{x}_1) - f(\mathbf{z})$. Assuming $\mathbf{x}_1$ is close enough to $f(\mathbf{z})$, the bias would likely be small when $k = 1$ since it's likely to share a similar label. Meanwhile, when $k \longrightarrow \infty$, the bias doesn't depend on the training points at all which like will restrict it to be higher.

Now, let's derive the variance of our model.

$$\begin{aligned}
\mathrm{Var}[h(\mathbf{z})] &= \mathrm{Var}\left[\frac{1}{k}\sum_{i=1}^{k} y_i\right] = \frac{1}{k^2}\sum_{i=1}^{k} \mathrm{Var}[f(\mathbf{x}_i) + \epsilon] \\
&= \frac{1}{k^2}\sum_{i=1}^{k} \mathrm{Var}[\epsilon] \\
&= \frac{1}{k^2}\sum_{i=1}^{k} \sigma^2 = \frac{1}{k^2} k\sigma^2 = \frac{\sigma^2}{k}
\end{aligned}$$

The variance goes to 0 when $k \longrightarrow \infty$, and is maximized at $k = 1$.