The \textit{k-NN} classifier

\textbf{Assumption:} Similar points share similar labels

\textbf{Classification Rule:} For a test input $x_t$, assign the most common label among its $k$ most similar training inputs.

Formally: $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ training data. Test point $x_t$.

Let $S_k \subseteq D$ such that $|S_k| = k$

and $\forall (x, y) \in D \setminus S_k \\text{ dist}(x, x_t) = \max_{(x', y') \in S_k} \text{ dist}(x', x_t)$

$h(x) = \text{ mode } \{y : (x, y) \in S_k\}$

\textbf{Protip:} In case of a draw decide by reducing $k$ by 1, until you reach a unique mode.

\textbf{Training Error:} Leave-One-Out (LOO) estimate: Take each training point out and estimate its label, pretending it was a test point. (i.e. a point cannot be its own neighbor)

What distance function should we use?

- Common choice: Minkowski's distance: $\text{ dist}(x, x') = \left( \sum_{i=1}^{n} |x_i - x'_i|^p \right)^{\frac{1}{p}}$ for $p \geq 0$
  - special case: $p = 2 \leftarrow \text{ Euclidean distance}$
  - $p = 1 \leftarrow \text{ Manhattan distance}$

\textbf{Quiz:} What if $p \to 0$ or $p \to \infty$?  How does $k$ affect the outcome? How does the classifier behave as $k = 1$, or $k = n$?

\textbf{Bayes Optimal Classifier}

Your data $D$ is drawn from some distribution $(x, y) \sim P(x, y)$. Also: $P(x, y) = P(y|x)P(x)$

Assume you knew $P(y|x)$ (you never do, but just for the sake of the argument).

For some test $x$ what label would you predict?

The most likely label: $h_{\text{opt}}(x) = \arg \max_y P(y|x)$

What is the expected error of the BOC? Let $y^* = h_{\text{opt}}(x)$

$E = P(y^* \neq y)$

You can never do better than the BOC!
Asymptotic error bound for 1-NN (Cover and Hart 1967)

Quiz 1: You have a coin that shows head with probability \( p \).
If you throw it twice, what is the probability \( q \) that both throws lead to different outcomes?

2. Show that \( q \leq 2(1-p) \)

Back to 1-NN. We want to prove that the expected 1-NN test error is less than 2 \( \times \) the BOC error, as \( n \to \infty \). (For binary classification)

Argument: Let \( x \) be the test point and \( \hat{z} \) be its nearest neighbor.

**Claim 1:** As \( n \to \infty \), \( \text{dist}(x, \hat{z}) \to 0 \) \( \Leftarrow \) i.e. The nearest neighbor becomes infinitely close.

**Claim 2:** As \( \text{dist}(x, \hat{z}) \to 0 \), \( \hat{z} \to x \) \( \Leftarrow \) i.e. In fact, the nearest neighbor becomes identical to \( x \). (See Cover & Hart for proof.)

Assume for \( x \), the label \( y^* \) is most likely. Let \( p = P(y^* | x) \).
The BOC would predict \( y^* \) and be wrong with probability \( \epsilon_{\text{BOC}} = 1-p \).
What is the error of 1-NN as \( n \to \infty \)?

1-NN is wrong if the labels of \( x \) and \( \hat{z} \) are different.
By claim 2, we have \( \hat{z} \to x \). And \( p(y^* | x) = P(y^* | \hat{z}) = p \).
Both points \( x \) and \( \hat{z} \) could take on label \( y^* \) with prob. \( p \), and not with \( 1-p \).
Remember Quiz 2. Regard both points as the same coin tossed twice.
They disagree with probability \( 2p(1-p) \leq 2(1-p)^2 = 2\epsilon_{\text{BOC}} \)

\[ \Rightarrow \epsilon_{\text{1-NN}} \leq 2\epsilon_{\text{BOC}} \text{ as } n \to \infty \]
Curse of Dimensionality

Assume $x \in [0,1]^d$ (i.e., the d-dimensional unit hypercube). All data is drawn uniformly at random. Let $k=10$. Let $l$ be the edge length of the smallest hypercube that contains all $k$ nearest neighbors of a test point $x$.

$$l^d = \frac{k}{n} \Rightarrow l = \left(\frac{k}{n}\right)^{\frac{1}{d}}$$

Almost the entire space is needed to fit $10$ nearest neighbors.

If $n=1000$ how big is $l^2$?

<table>
<thead>
<tr>
<th>$d$</th>
<th>$0.1$</th>
<th>$0.2$</th>
<th>$0.3$</th>
<th>$0.4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.00</td>
<td>0.01</td>
<td>0.015</td>
<td>0.02</td>
</tr>
<tr>
<td>100</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>1000</td>
<td>0.02</td>
<td>0.04</td>
<td>0.06</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Almost the entire space is needed to fit $10$ nearest neighbors.

This means nearest neighbors are not similar, violating the k-NN assumption!

How many points would we need for $l$ to be small?

Fix $l = 0.1$.

$$l^d = \frac{k}{n} \Rightarrow n = k \left(\frac{l}{k}\right)^d = k \left(\frac{1}{10}\right)^d \left(\frac{10}{d}\right)^d \left(\frac{1}{10}\right)^d \text{ grows exponentially with } d!$$

Rescue to the curse:

Data can have structure:

- Data can lie on intrinsically low dimensional subspaces or sub-manifolds.
- Data can be clustered (very non-uniform).