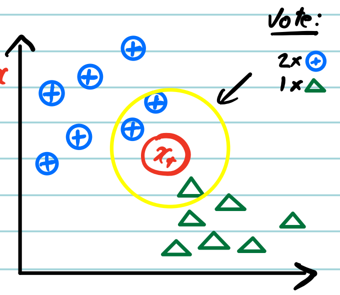
The k-NN classifier

Assumption: Similar points share similar labels

Classification Rule: For a test input x_t assign the most common label among its k most similar training inputs.

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



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

and $\forall (x',y') \in D \setminus S_x \quad dist(x',x) \geq \max_{(x'',y'')\in D\setminus S_x} dist(x'',x)$

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

Protip: In case of a draw decide by reducing k by 1, until you reach a unique mode.

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: Minkowki's distance: $\boxed{dist(x,z) = \left(\sum_\alpha |[x]_\alpha - [z]_\alpha|^p\right)^{1/p}}$ for $p>0$

special case: $p=2 \leftarrow$ Euclidean distance

$p=1 \leftarrow$ Manhattan distance

Quiz: What if $p\to\infty$ or $p \to 0^+$?

How does k affect the outcome? How does the classifier behave as $k=1$, or $k=n$?

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: $h_{opt}(x) = \underset{y}{\operatorname{argmax}}\, P(y|x)$

What is the expected error of the BOC? Let $y^* = h_{opt}(x)$ $\quad \epsilon = 1 - \underbrace{P(y^*|x)}_{}$

The probability that x does not have the most likely label.

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 $q \leq 2(1-p)$.

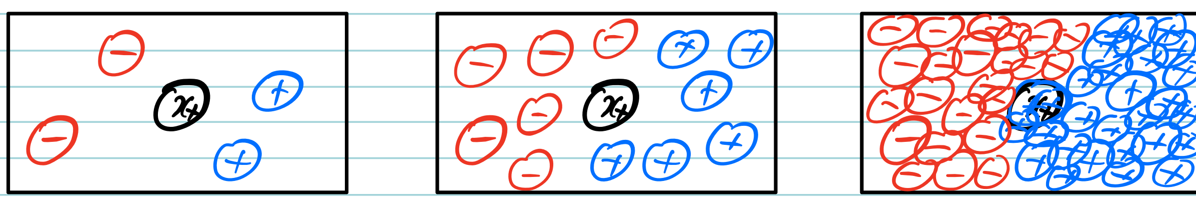
Back to 1-NN. We want to prove that the expected 1-NN test error is less than $2\times$ the BOC error, as $n \to \infty$. (For binary classification.)

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

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

$\boxed{\text{Claim 2: As dist}(x, \hat{x}) \to 0, \ \hat{x} \to x}$ $\leftarrow$ i.e. In fact, the nearest neighbor becomes identical to x.

(See Cover & Hart for proof.)



Assume for x_t the label y^* is most likely. Let $p = P(y^* | x_t)$

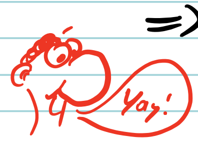
The BOC would predict y^*, and be wrong with probability $\underline{\underline{\varepsilon_{BOC} = 1-p}}$.

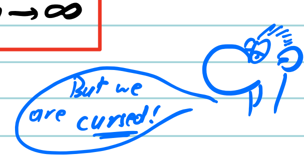
What is the error of 1-NN as $n \to \infty$?

1-NN is wrong if the labels of x and $\hat{x}$ are different.
By claim 2 we have $\hat{x} \to x$. And $p(y^* | \hat{x}) = p(y^* | x) = p$
Both points x and $\hat{x}$ 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 $2p(1-p) \leq 2(1-p) = \underline{\underline{2\varepsilon_{BOC}}}$

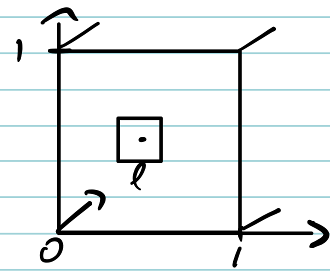
$$\Rightarrow \boxed{\varepsilon_{1\text{-}NN} \leq 2\varepsilon_{BOC} \quad \text{as } n \to \infty}$$





Curse of Dimensionality

Assume $x_i \in [0,1]^d$ (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.

$$\ell^d \approx \frac{k}{n} \Rightarrow \ell \approx \left(\frac{k}{n}\right)^{1/d}$$

volume of mini cube containing the k neighbors

Total volume of hypercube $[0,1]^d$ is $1^d = 1$
$\frac{k}{n}$ is the fraction that k points take up.
(because points are uniformly sampled)

If $n=1000$ how big is ℓ?

d	2	10	100	1000
ℓ	0.1	0.63	0.955	0.9954

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 $\ell = 0.1$

$$\ell^d = \frac{k}{n} \Rightarrow n = k\left(\frac{1}{\ell}\right)^d = k\,10^d$$

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).