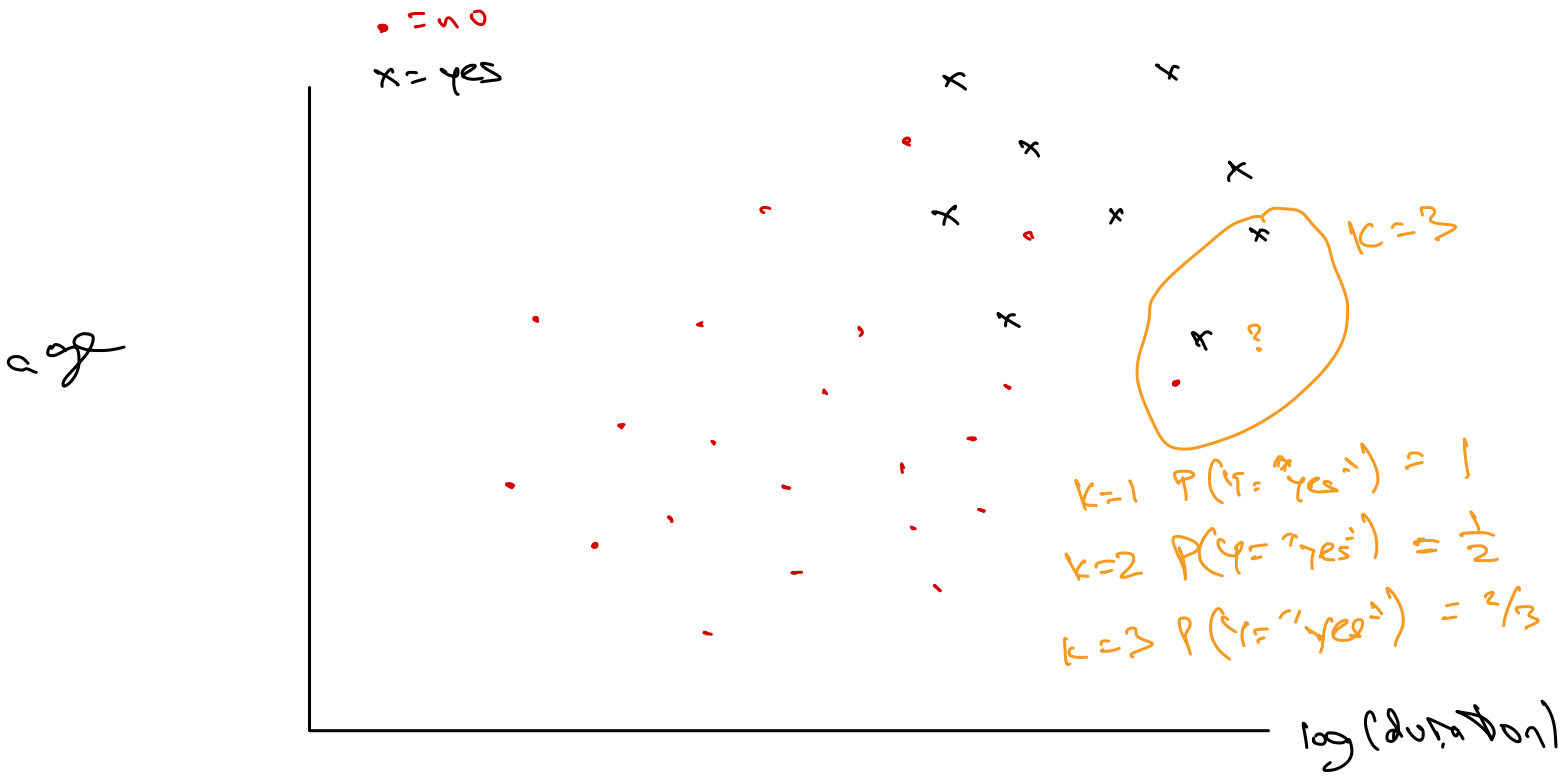
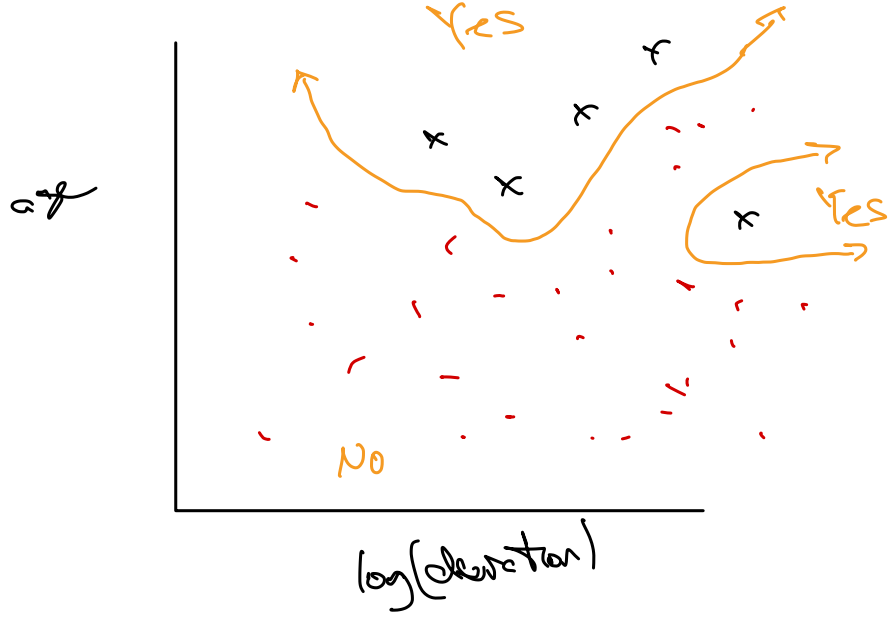
K-nearest neighbors

KNN is a nonparametric classification method that estimates $P(Y=j \mid X=x)$ empirically.

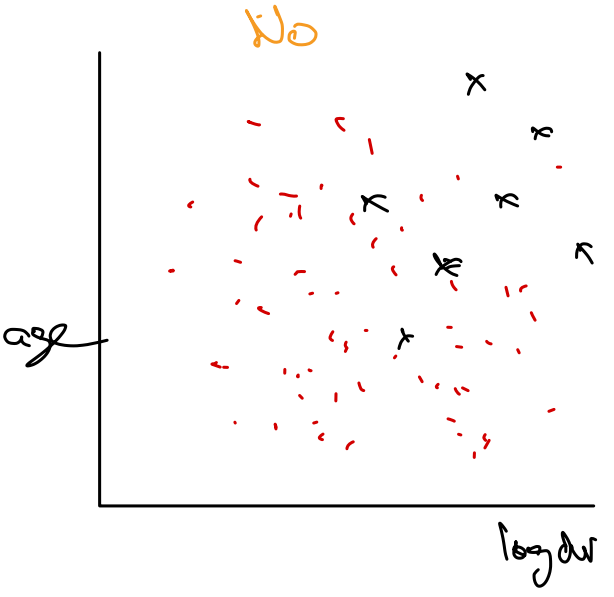
$$\hat{P}(Y=j \mid X=x) = \frac{1}{K} \sum_{i \in \mathcal{N}_x} 1[Y_i = j]$$

for a prespecified integer K & $\mathcal{N}_x$ is the set of K-nearest neighbors to x.


• = no
x = yes
age
log(duration)
K=3
?
K=1 P(y="yes") = 1
K=2 P(y="yes") = 1/2
K=3 P(y="yes") = 2/3



decision boundary for k=1
low bias but high variance



decision boundary for
k really big.
low variability but high bias