K-Nearest Neighbors

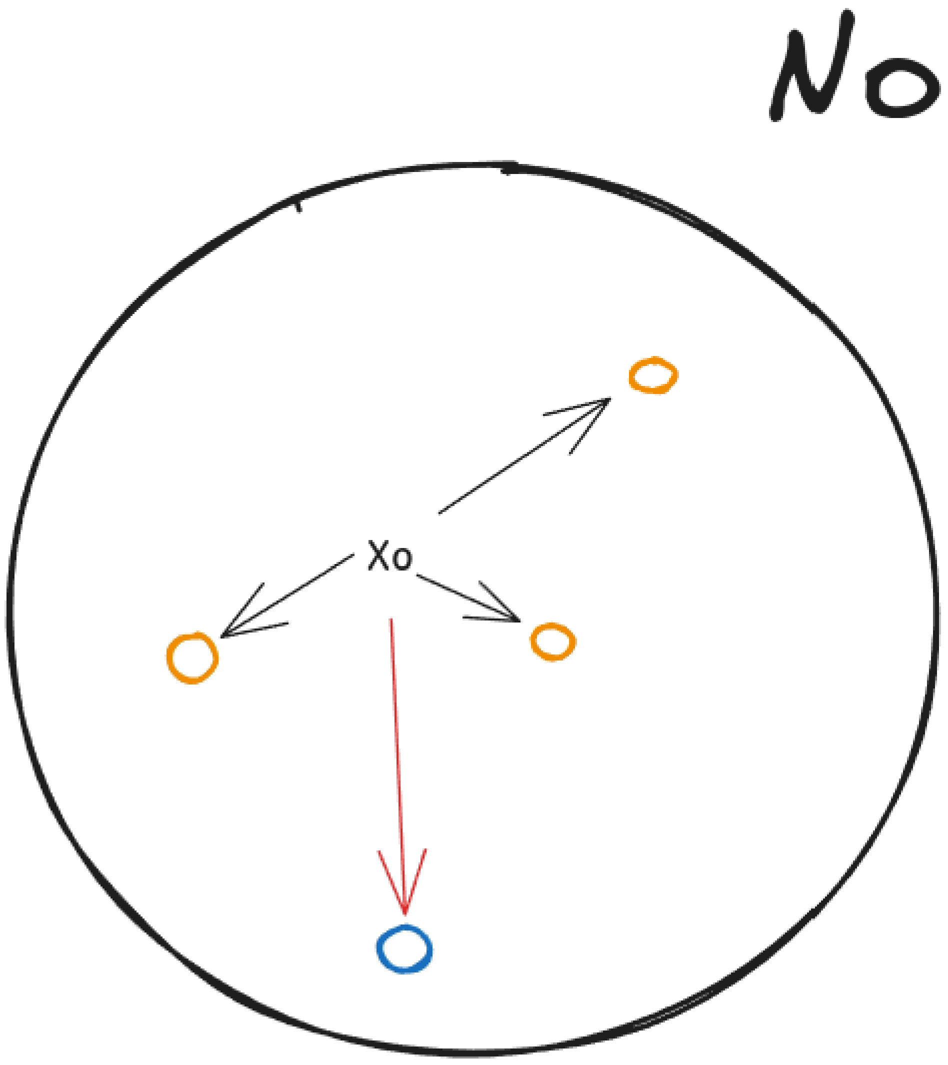
in theory we would like to always use Bayes Classifier mentioned in Assessing-Model-Accuracy.
But its a impossible to compute → $Pr(Y = j|X = x_0)$,we only have finite, noisy dataset
So The **Bayes Classifier** will be the gold standard for our estimations

- K-Nearest Neighbors tries to estimate it

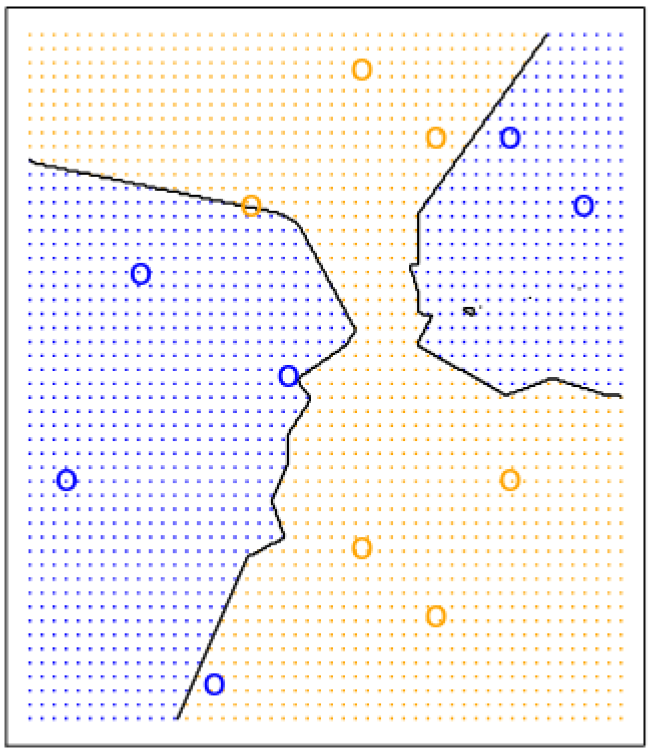
Given :

- K positive integer
- x_0 Observation
  The **KNN** classifier finds the K points in the Training Data closet to x_0
- Then it estimate the Conditional Probability for the class j as the fraction points in $\mathcal{N}_0$

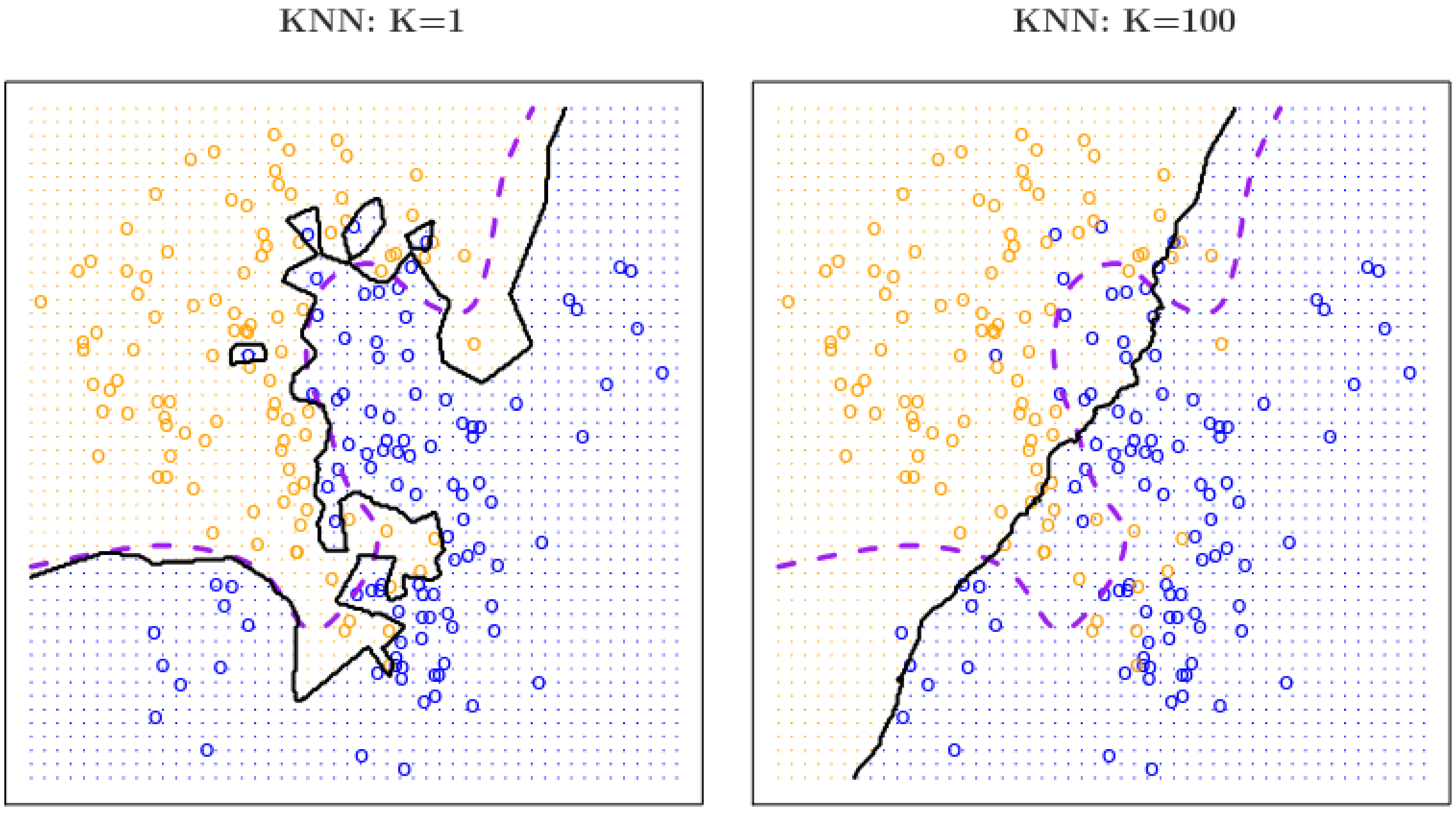
$$\Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i\in\mathcal{N}_0} I(y_i = j)$$



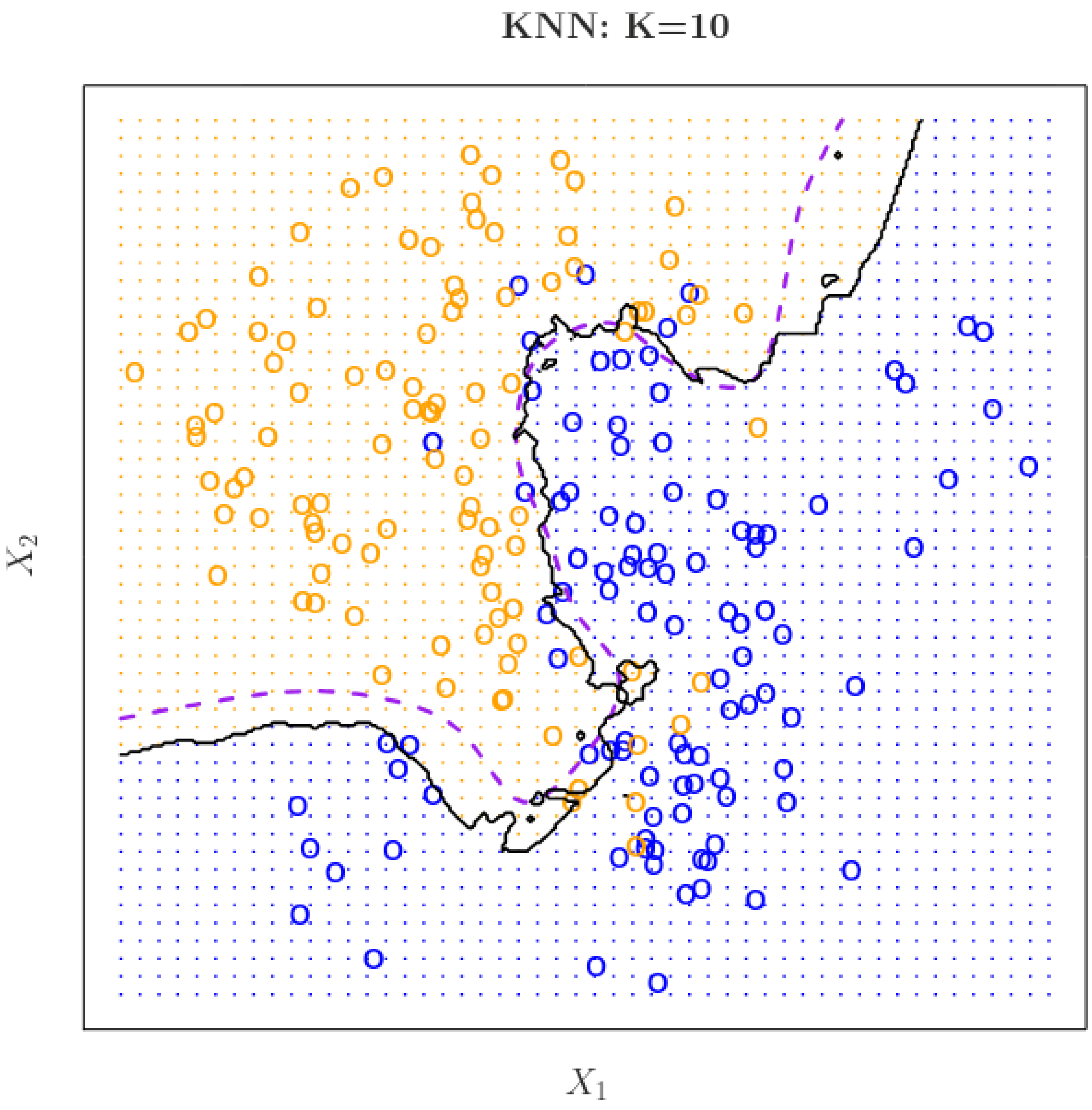
- Here $K = 4$
- So the Classifier find the nearest 3 Training points
  The Probability of x_0 being
- Orange is $\frac{3}{4}$
- Blue is $\frac{1}{4}$
- So **KNN** will predict that x_0 is Class Orange $Y = Orange$



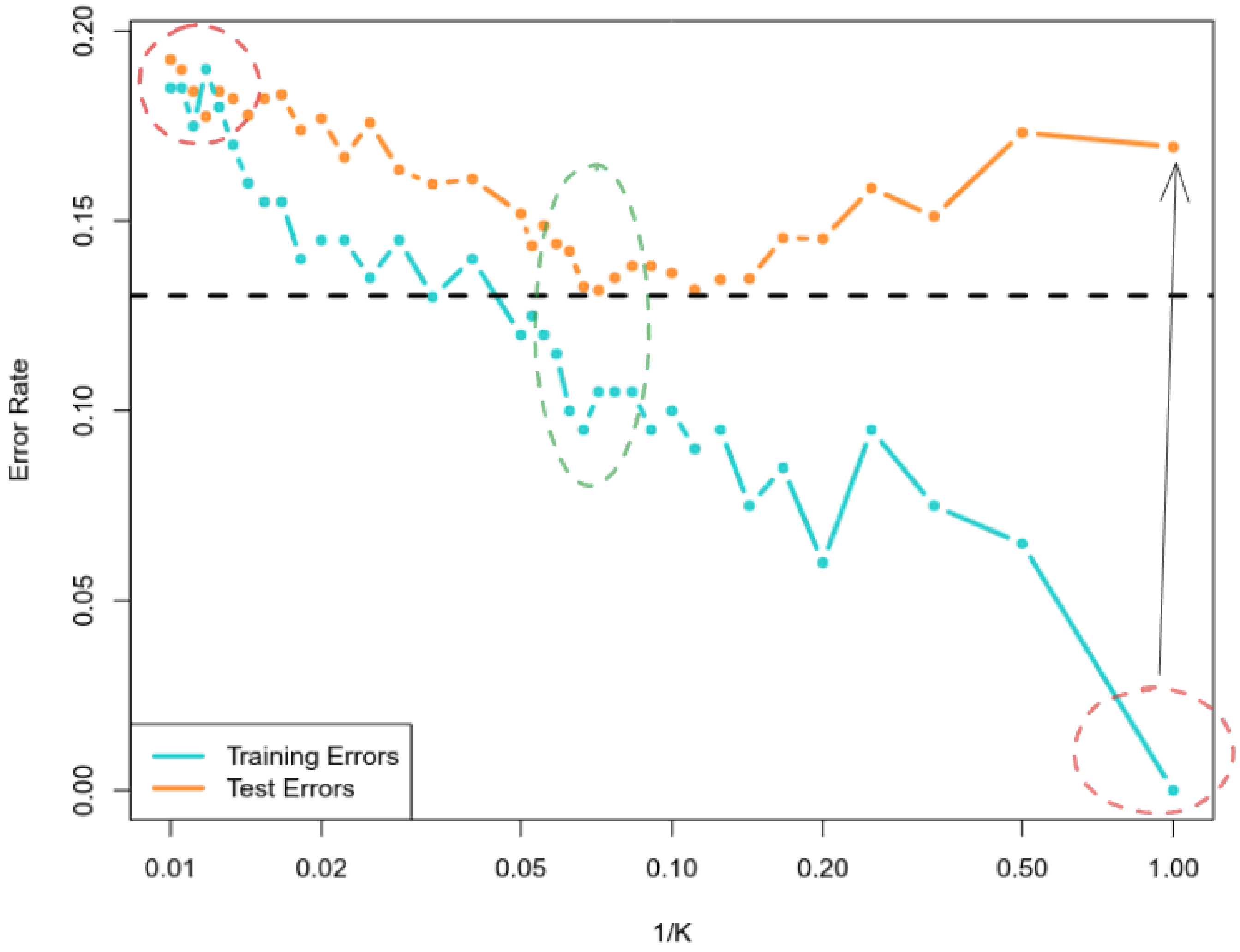
- **KNN** can be very accurate when applied to bigger data
- The KNN error rate is 0.1363 which is very close to Bayes Classifier of 0.1304
  The choice of K effects the predicited results largely, as shown here :



- The Purple dash-line is the Bayes Classifier
- The bigger K value the less flexible and the more Linear it get



- $K = 10$ gets really close to the gold standard



- Same as in Regression $\text{Training error rate} \neq \text{Test error rate}$
- $K = 1$ will result in zero $\text{Training error rate}$ but a very high Test error rate
- and high K Values will also results on very high $\text{Training and Test error rates}$
- Also same as Regression the $Test\ error\ rate$ give a U-shape curve