Today

- Classification case

Regression & classification

- Regression:
  - numerical response variable
  - predict a numerical value given x
  - e.g. number of species given latitude
- Classification:
  - categorical response variable
  - predict the category given x
  - e.g. is it a bird, deer, tree, or mountain lion?
  - e.g. is it dead or alive?; present or absent?

Classification

As before: out-of-sample accuracy

One common, simple measure is the error rate. If we have a *test* dataset of $i = 1 \ldots n$ observations, the out-of-sample error rate is:

$$\frac{1}{n}\sum_{i}^{n} I(y_i \neq \hat{y}_i) = \text{mean}\big(I(y_i \neq \hat{y}_i)\big)$$ = proportion incorrect

$\hat{y}_i$ is the predicted category for test case i.
$I()$ is an indicator function that equals 1 if the prediction is *incorrect* (i.e. if $y_i \neq \hat{y}_i$) and 0 if the prediction is correct.

Theory: best we can do

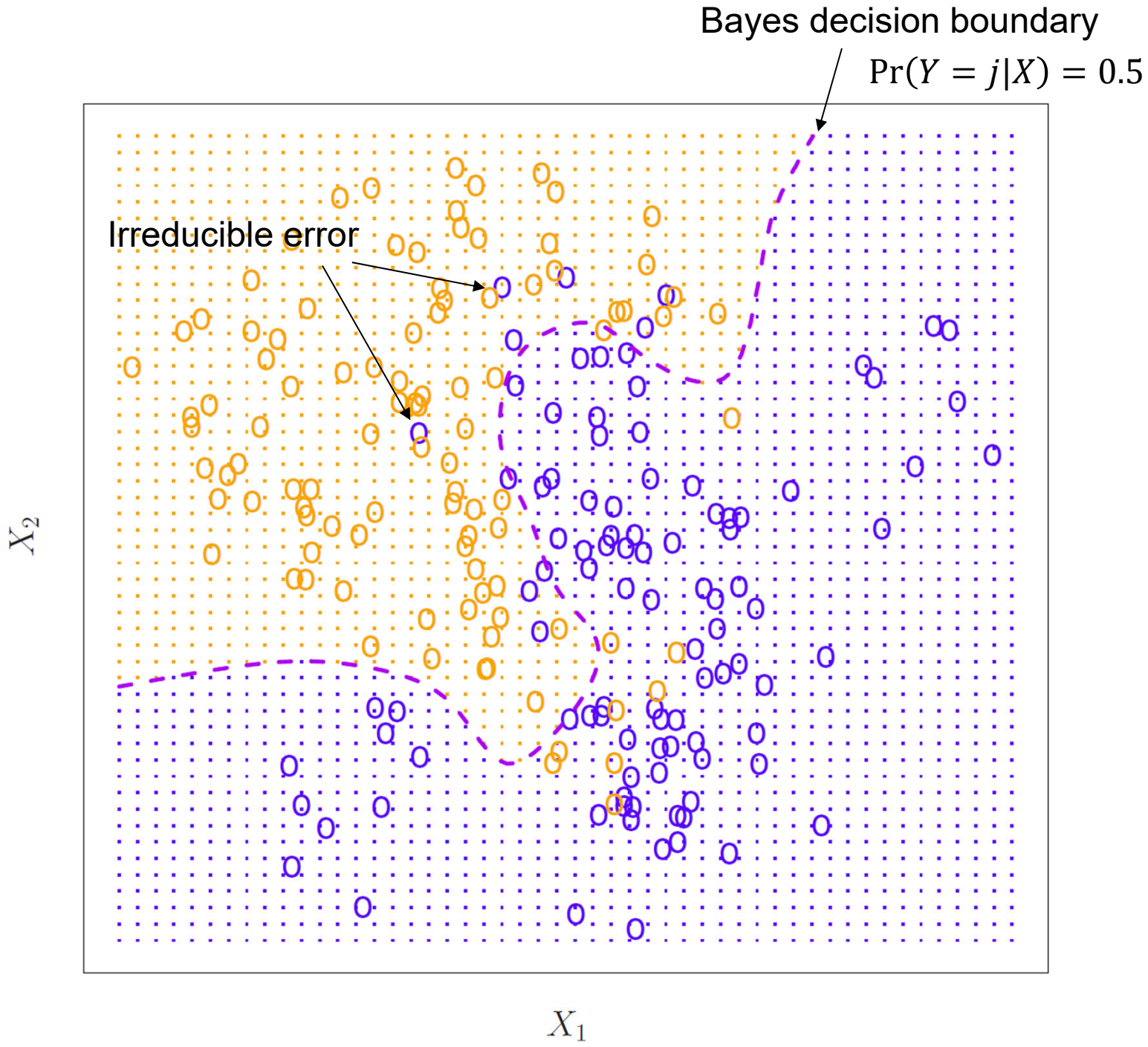
(assuming we know the true data generating process, e.g. in the case of data simulated from a model)

Bayes classifier

We should predict that our new observation belongs to the category with the highest probability. That is, choose category j for which

$$\Pr(Y = j | X = x_{\text{new}})$$

is highest.


Bayes decision boundary
$\Pr(Y = j|X) = 0.5$
Irreducible error
X_2
X_1

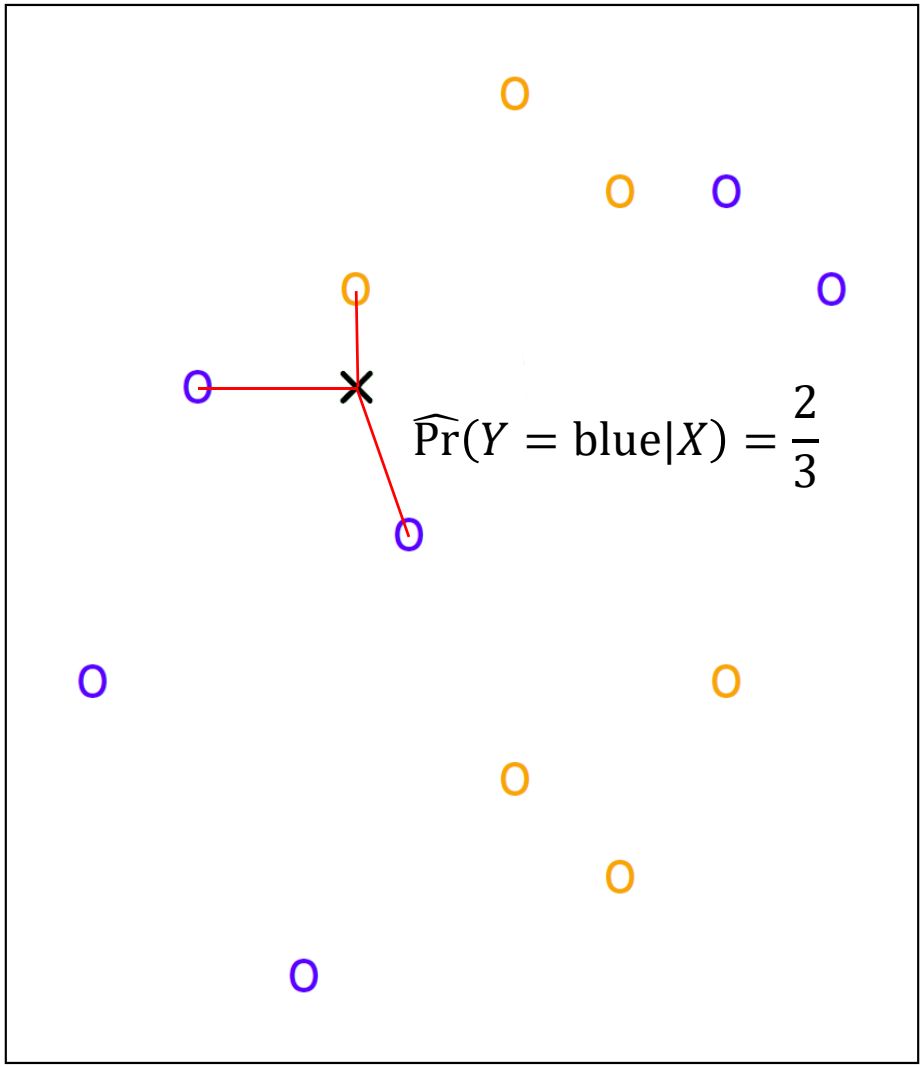
Theory: best we can do

Bayes error rate (irreducible error)

$$1 - E\left(\max_j \Pr(Y = j|X)\right)$$

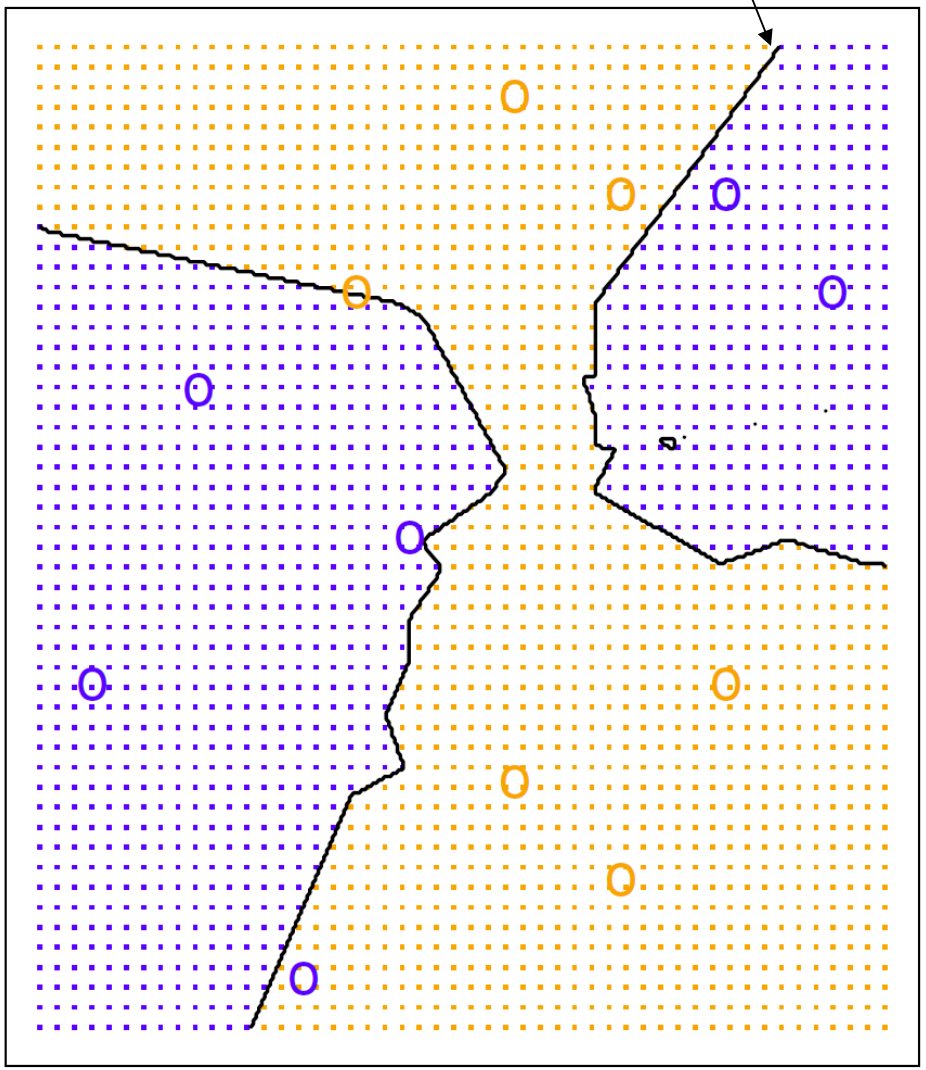
E = expected value = mean over all values of *X*.

KNN
k = 3



KNN decision boundary

$\widehat{\Pr}(Y = j|X) = 0.5$

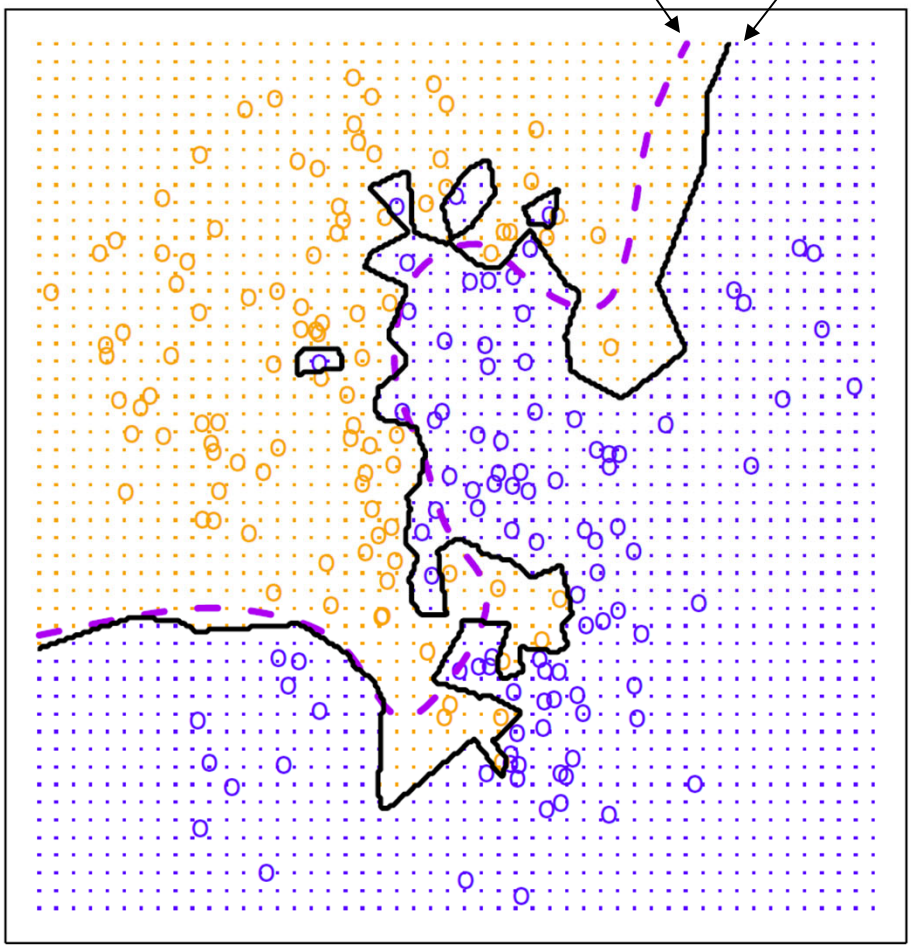


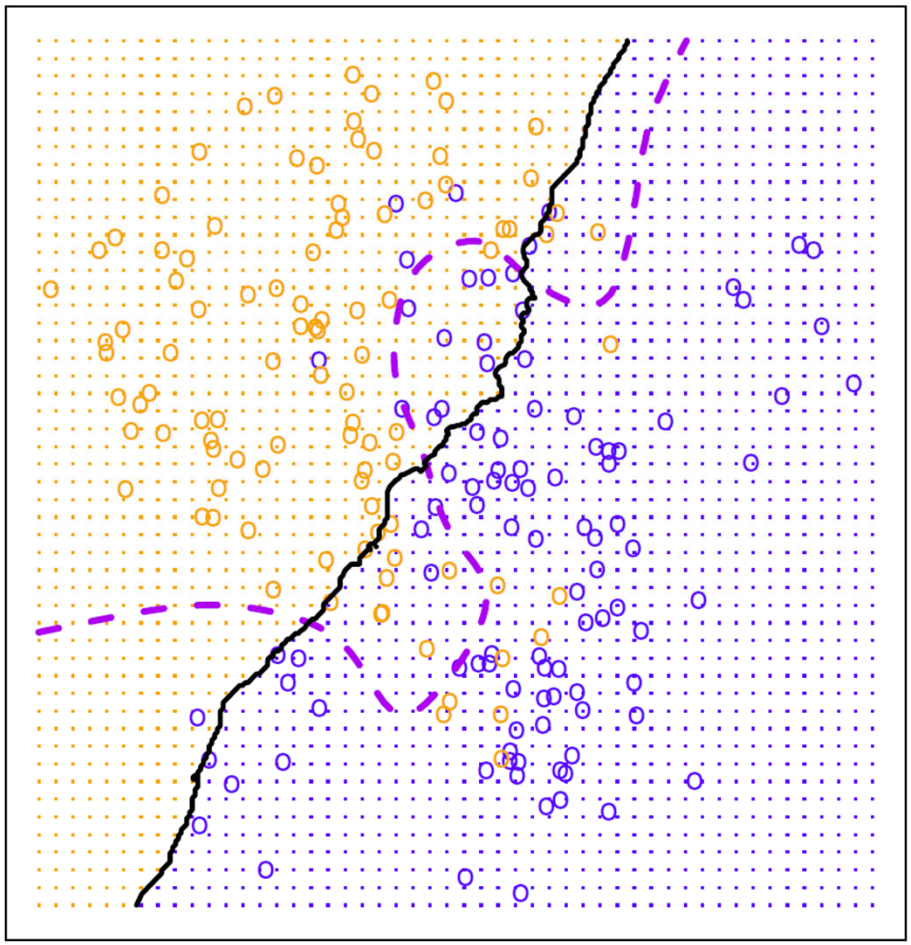
KNN
k = 1

Bayes decision boundary

KNN decision boundary

KNN
k = 100

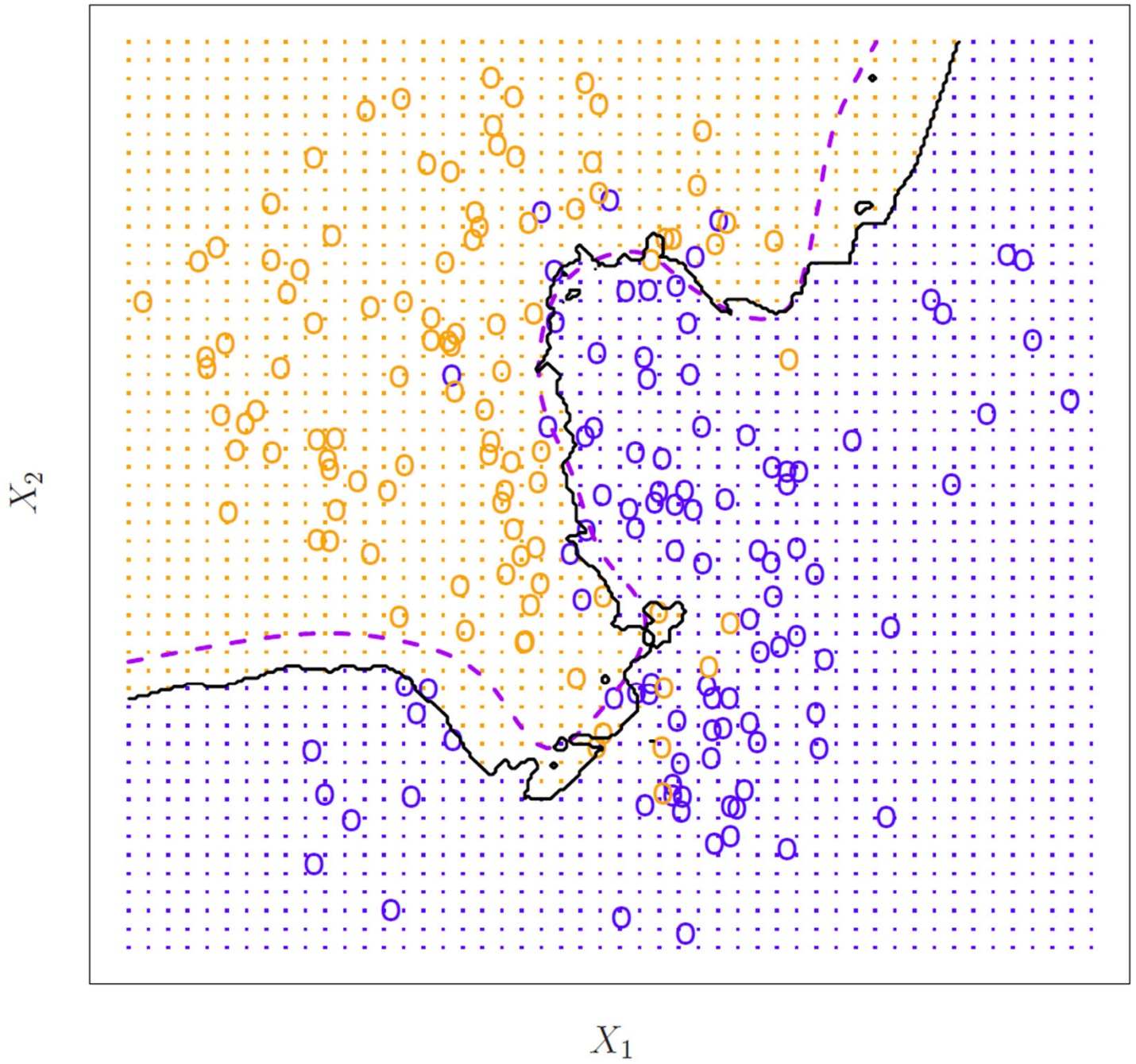


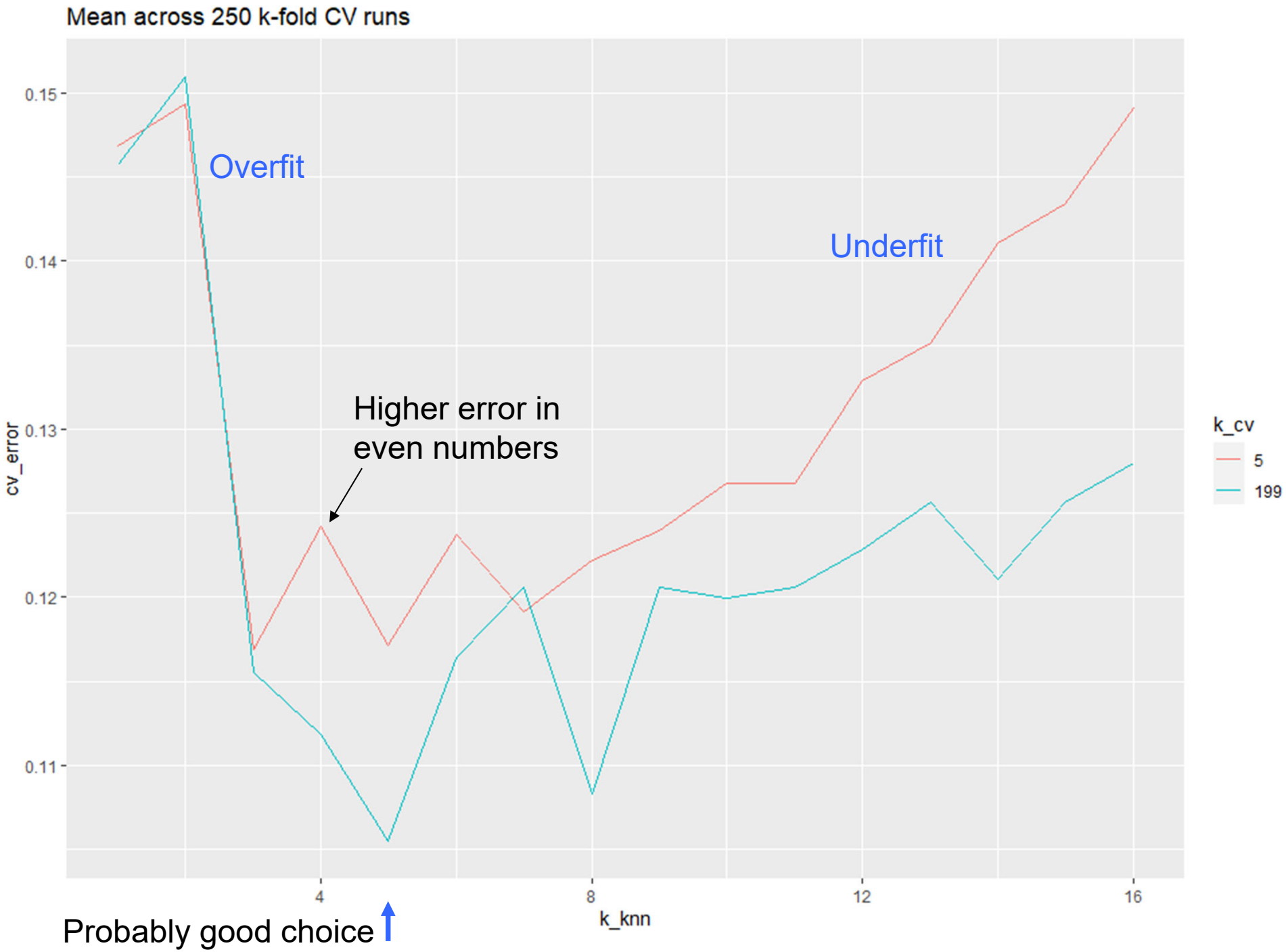


Overfit
high variance

Underfit
high bias

KNN
k = 10




Mean across 250 k-fold CV runs
0.15
0.14
0.13
0.12
0.11
cv_error
Overfit
Underfit
Higher error in
even numbers
4
8
12
16
k_knn
k_cv
5
199
Probably good choice