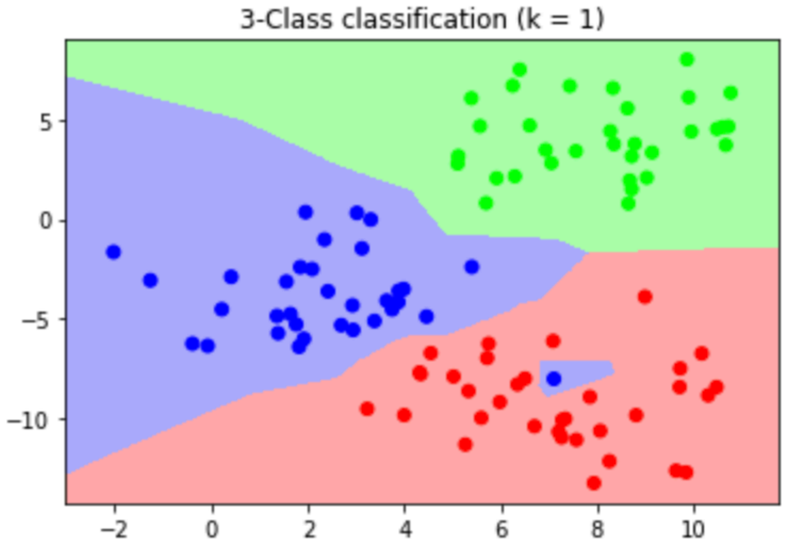
K-Nearest Neighbours

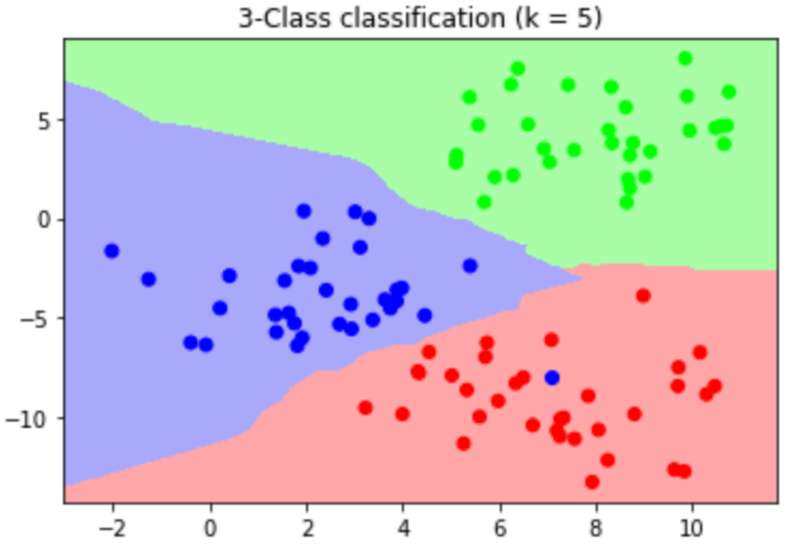
How to choose k?

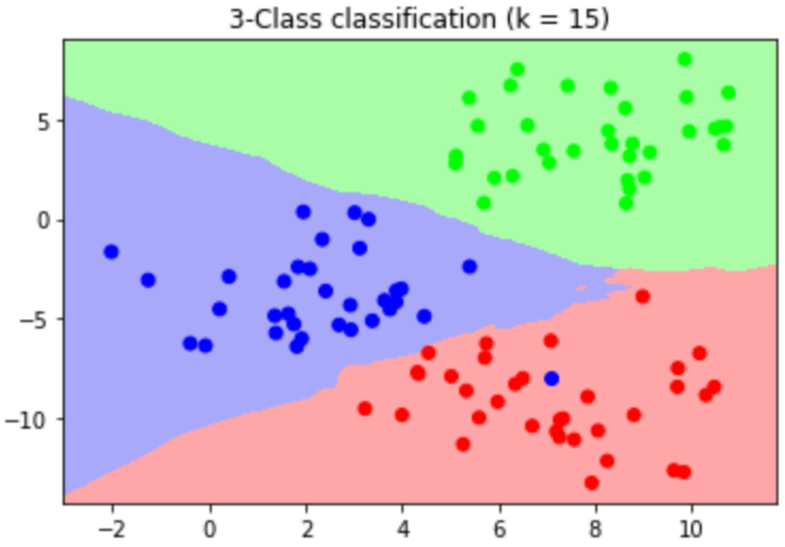
Procheta Sen

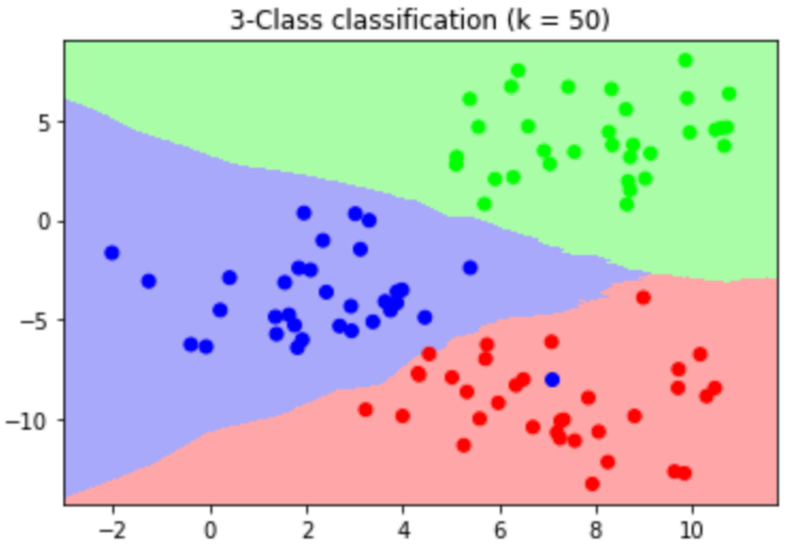
UNIVERSITY OF LIVERPOOL

The parameter k









The parameter k

- k is a “hyperparameter” of the algorithm
- What value to use for k?
  - Depends on the dataset size. Large and diverse datasets need a higher k, whereas a high k for small datasets might cross out of the class boundaries
- How to choose k?
  - **A bad option**: try different values of k when evaluating on **test data**
  - Use **validation dataset** to find a good value of k

Train / Validation / Test approach

Validation data

- Set aside (hold-out) a fraction of train data for validation purposes.
- Using validation data to set hyperparameters reduce overfitting.
- Of course, you CANNOT use test data for any tuning except for testing.

Train / Validation / Test approach

Validation data

- Set aside (hold-out) a fraction of train data for validation purposes.
- Using validation data to set hyperparameters reduce overfitting.
- Of course, you CANNOT use test data for any tuning except for testing.

Cross-validation

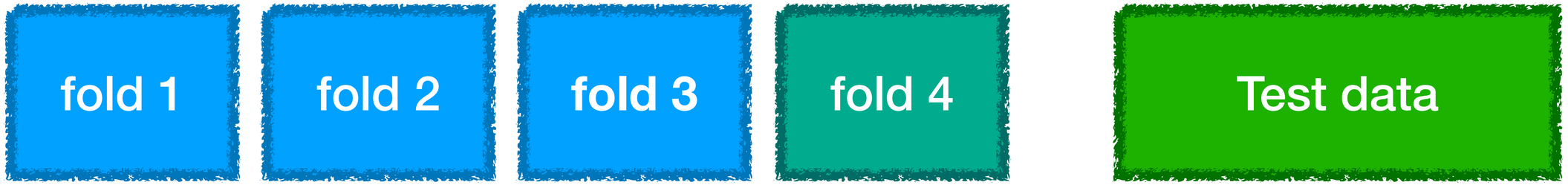
Validation data

- Set aside (hold-out) a fraction of train data for validation purposes.
- Using validation data to set hyperparameters reduce overfitting
- Of course, you CANNOT use test data for any tuning except for testing.

Cross-validation

Validation data

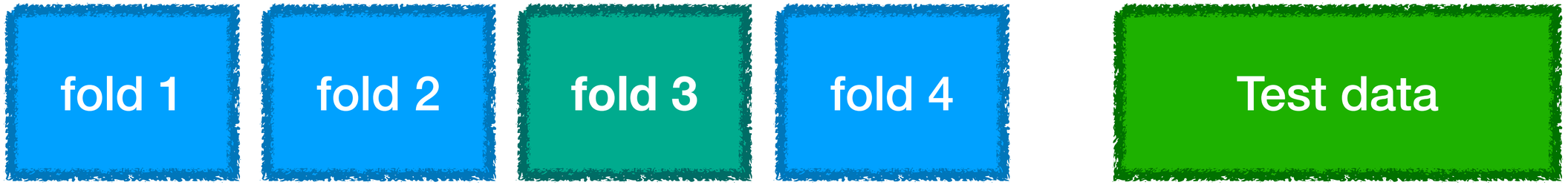
- Set aside (hold-out) a fraction of train data for validation purposes.
- Using validation data to set hyperparameters reduce overfitting
- Of course, you CANNOT use test data for any tuning except for testing.



Cross-validation

Validation data

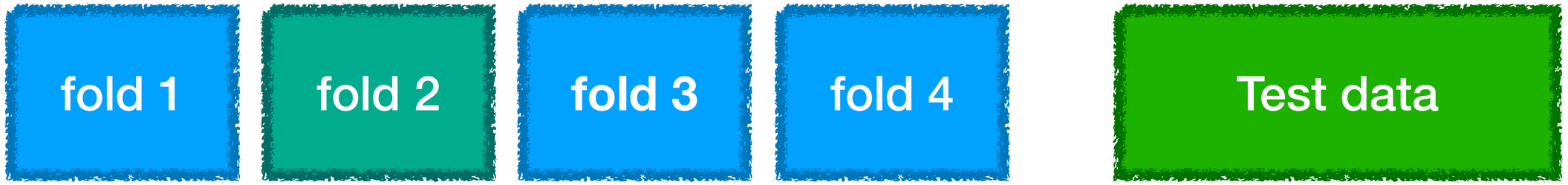
- Set aside (hold-out) a fraction of train data for validation purposes.
- Using validation data to set hyperparameters reduce overfitting
- Of course, you CANNOT use test data for any tuning except for testing.



Cross-validation

Validation data

- Set aside (hold-out) a fraction of train data for validation purposes.
- Using validation data to set hyperparameters reduce overfitting
- Of course, you CANNOT use test data for any tuning except for testing.



Cross-validation

Validation data

- Set aside (hold-out) a fraction of train data for validation purposes.
- Using validation data to set hyperparameters reduce overfitting
- Of course, you CANNOT use test data for any tuning except for testing.