CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

MODULE 2: REGRESSION AND CLASSIFICATION

DAT405, 2019-2020, READING PERIOD 1

Core data science tasks

- Regression
  - Predicting a numerical quantity
- Classification
  - Assigning a label from a discrete set of possibilities
- Clustering
  - Grouping items by similarity

CLASSIFICATION

- Assigning a label from a discrete set of possibilities

Iris data set

*R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics. **7** (2): 179–188.*

- Petal length
- Petal width
- Sepal length
- Sepal width

50 samples from each of three species

Iris setosa

Iris versicolor

Iris virginica

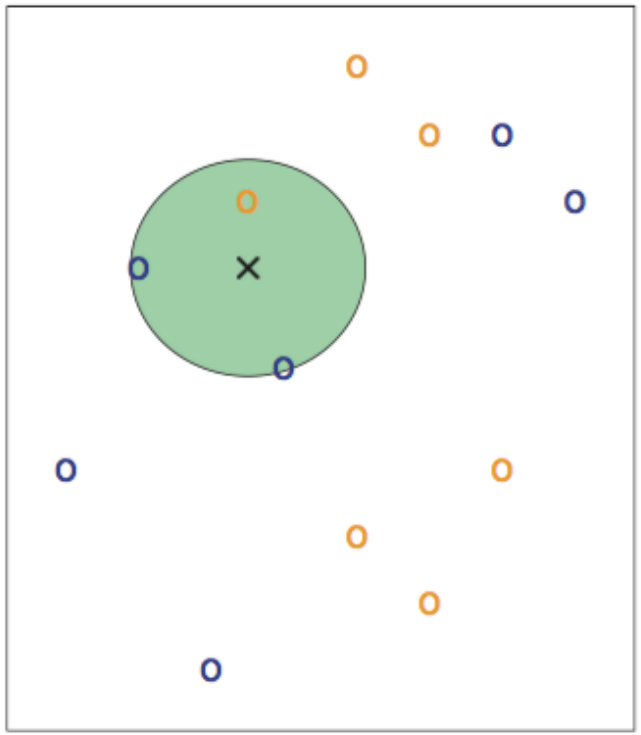
Different types of classifiers

Scikit-Learn provides easy access to numerous different classification algorithms. Among these classifiers are:

- K-Nearest Neighbours
- Logistic Regression
- Support Vector Machines
- Decision Tree Classifiers/Random Forests
- Naive Bayes
- Linear Discriminant Analysis

https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/

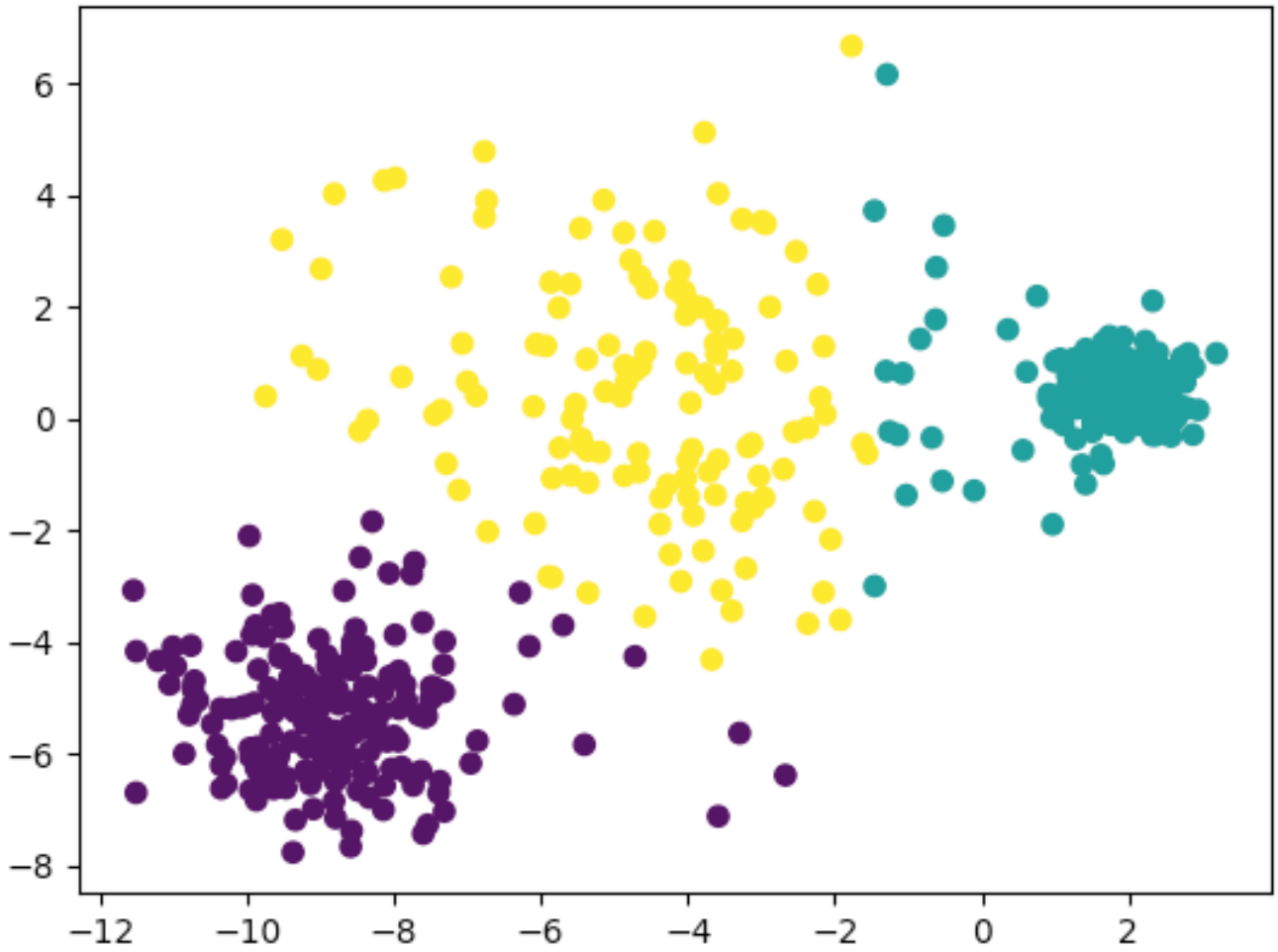
k-nearest neighbour classifier



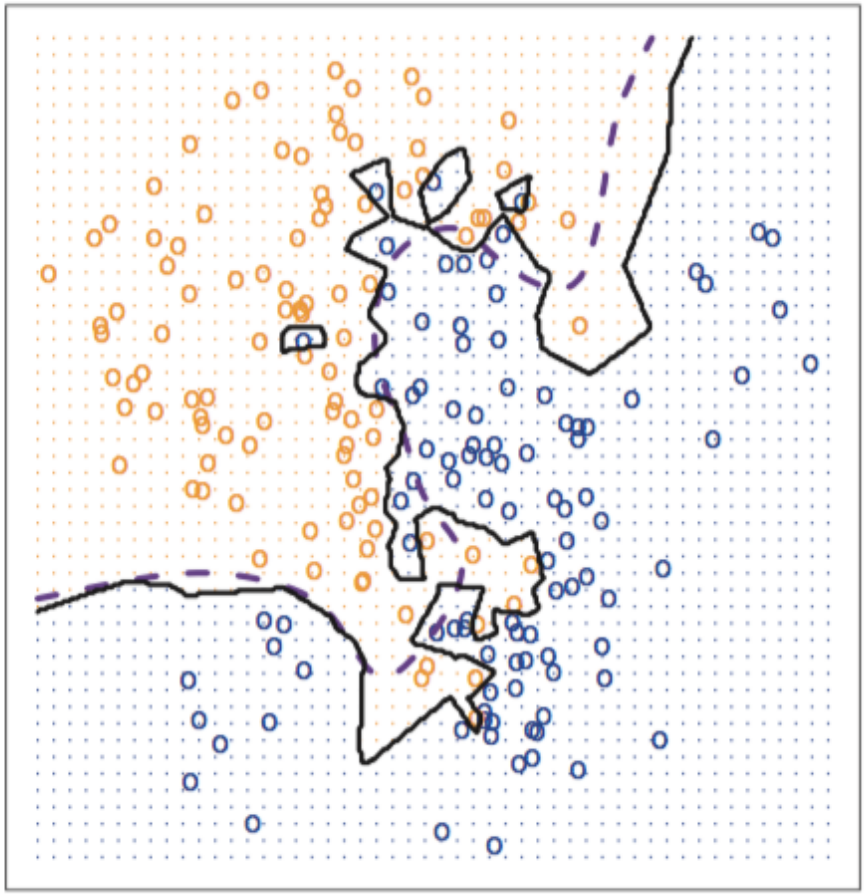
From Intro to Stat. Learning

Majority vote among k closest neighbours

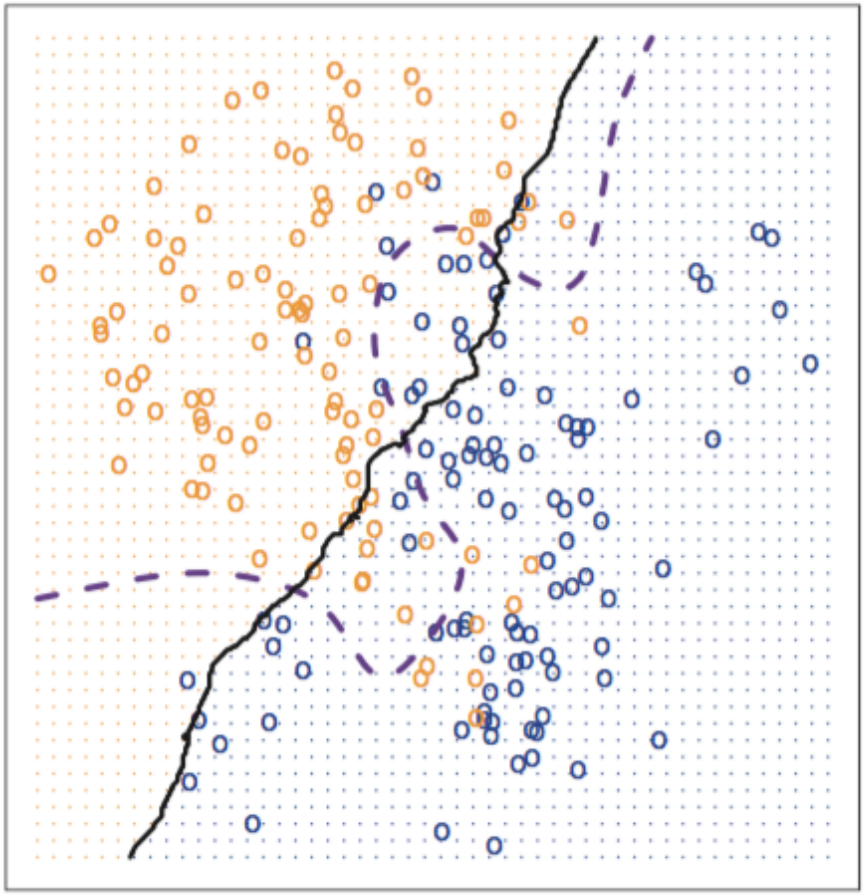
Probability of a class label estimated as proportion of neighbours with this class label.


6
4
2
0
−2
−4
−6
−8
−12
−10
−8
−6
−4
−2
0
2

KNN: K=1



KNN: K=100



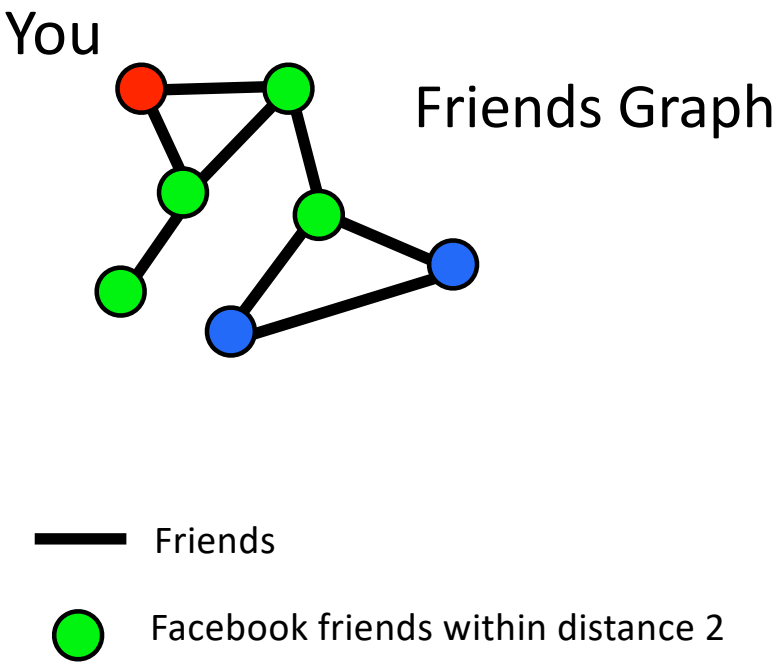
From Intro to Stat. Learning

k-nearest neighbours

- Pros:
  - No training (except choosing k)
  - Works with multiple classes (not only binary)
  - No assumptions about shape

- Cons:
  - Classification can be computationally expensive
  - Classification needs labelled data
  - No insights from classifier

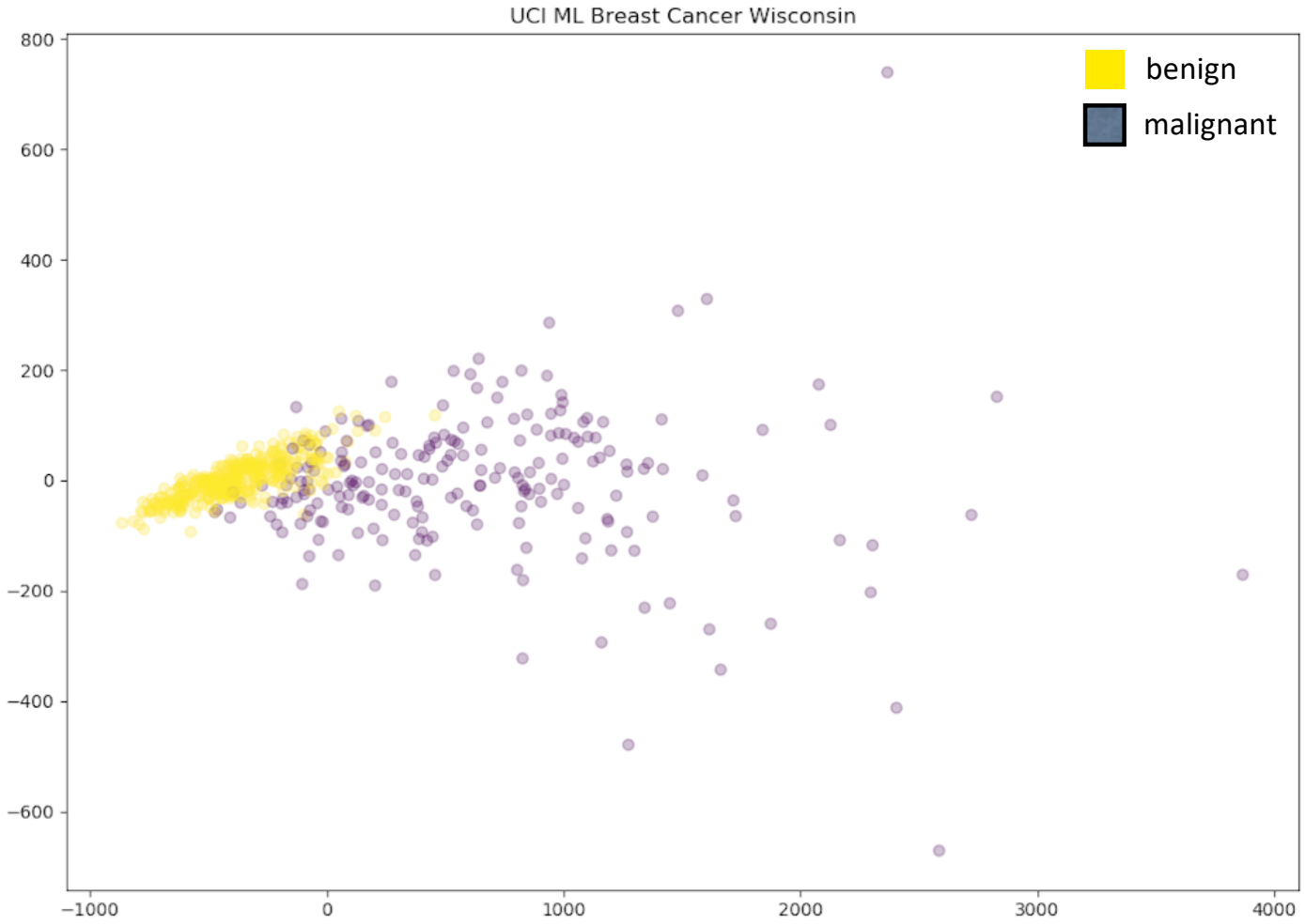
Do you like cat pictures?



Consider:

- k closest friends
- all friends within distance 2

If majority of those like cat pictures ...

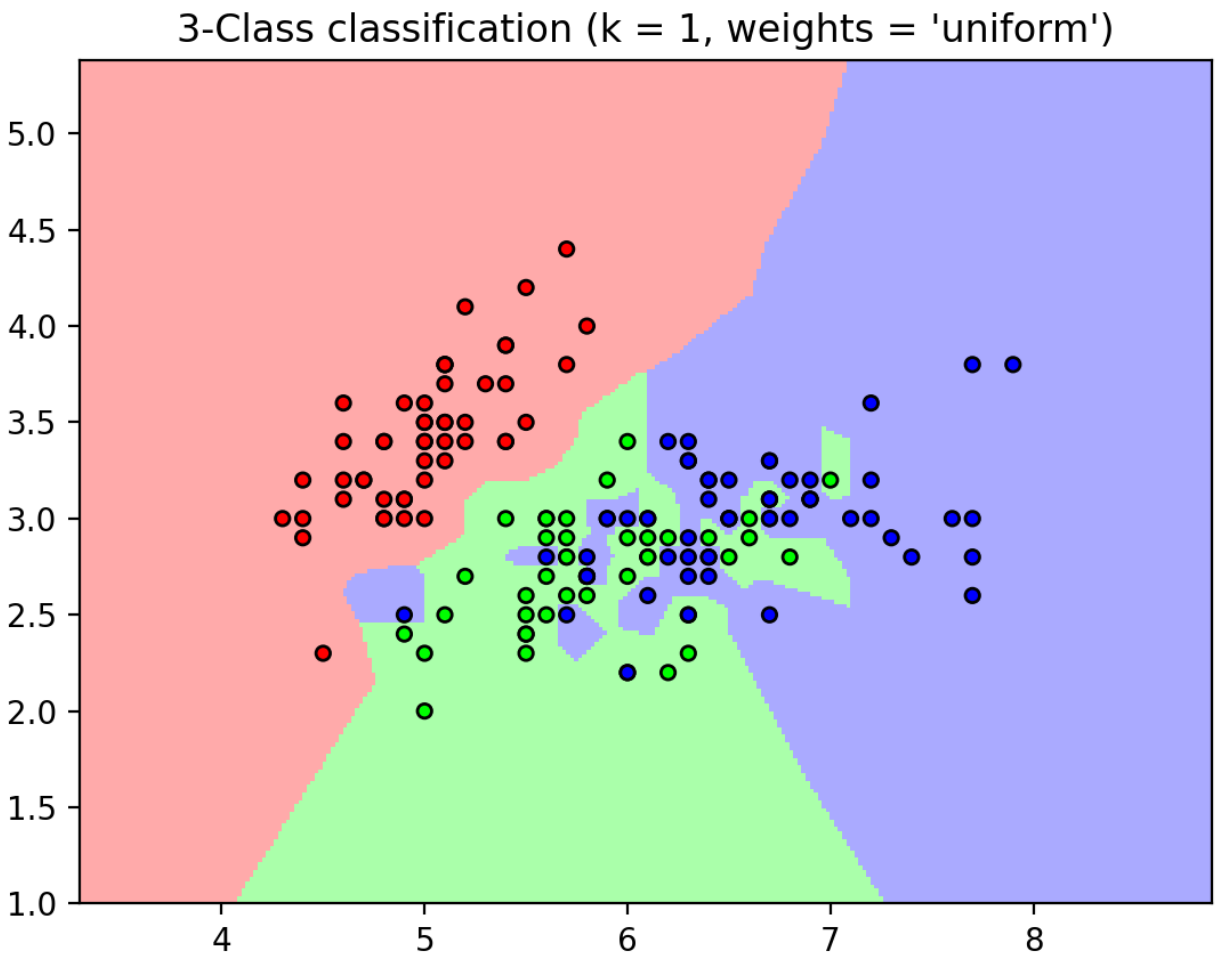

UCI ML Breast Cancer Wisconsin
benign
malignant
800
600
400
200
0
−200
−400
−600
−1000
0
1000
2000
3000
4000

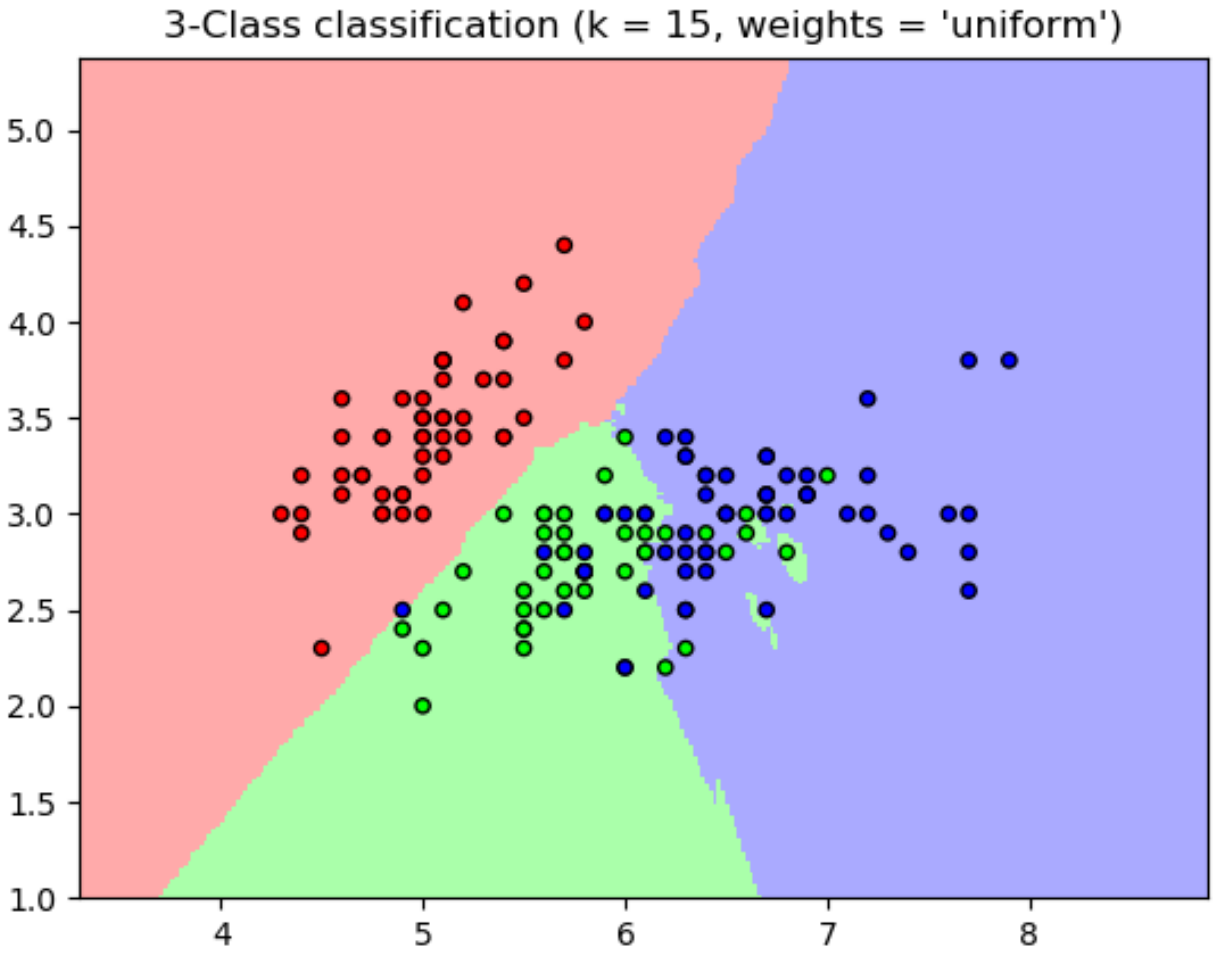
sklearn.neighbors.KNeighborsClassifier

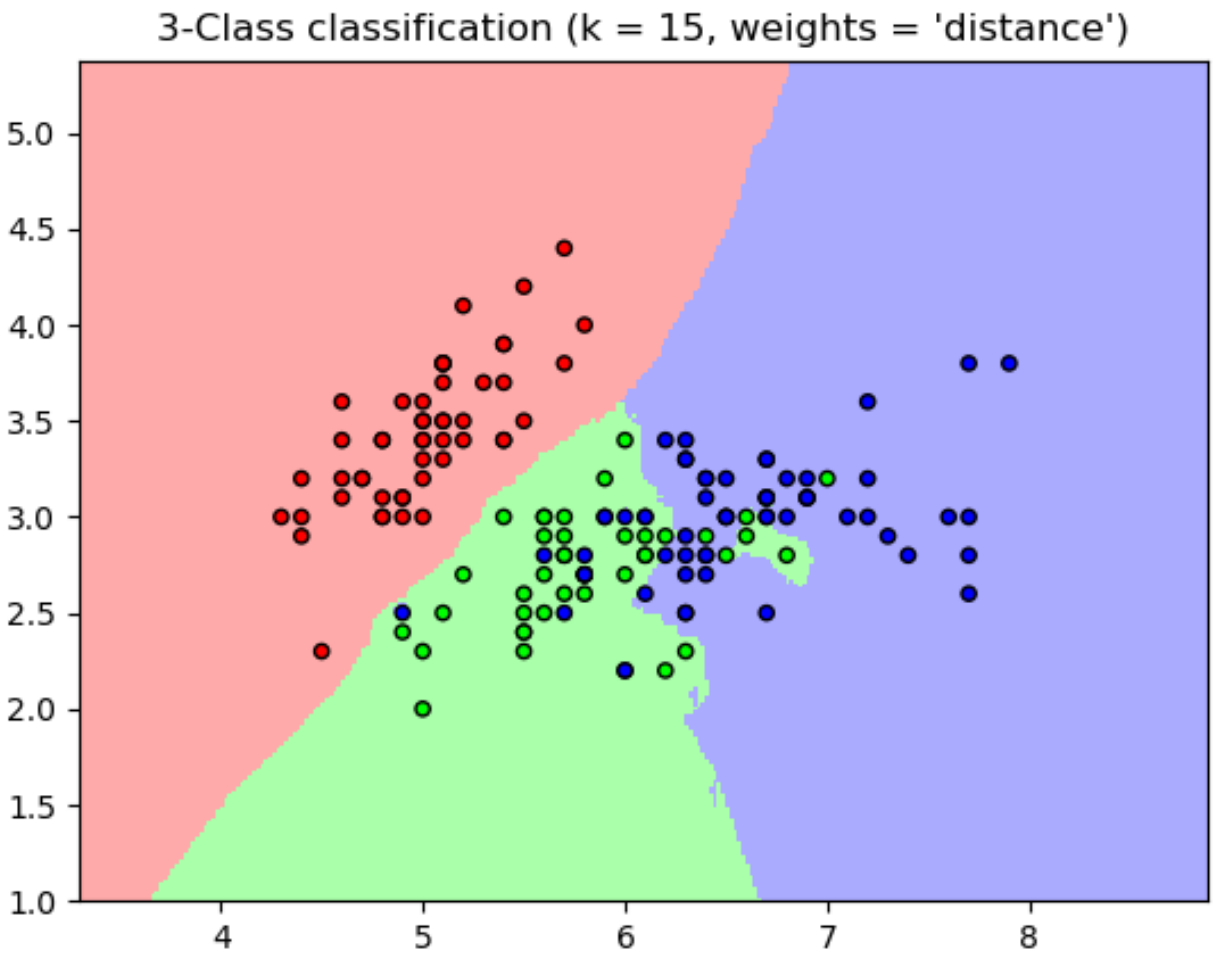
https://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html

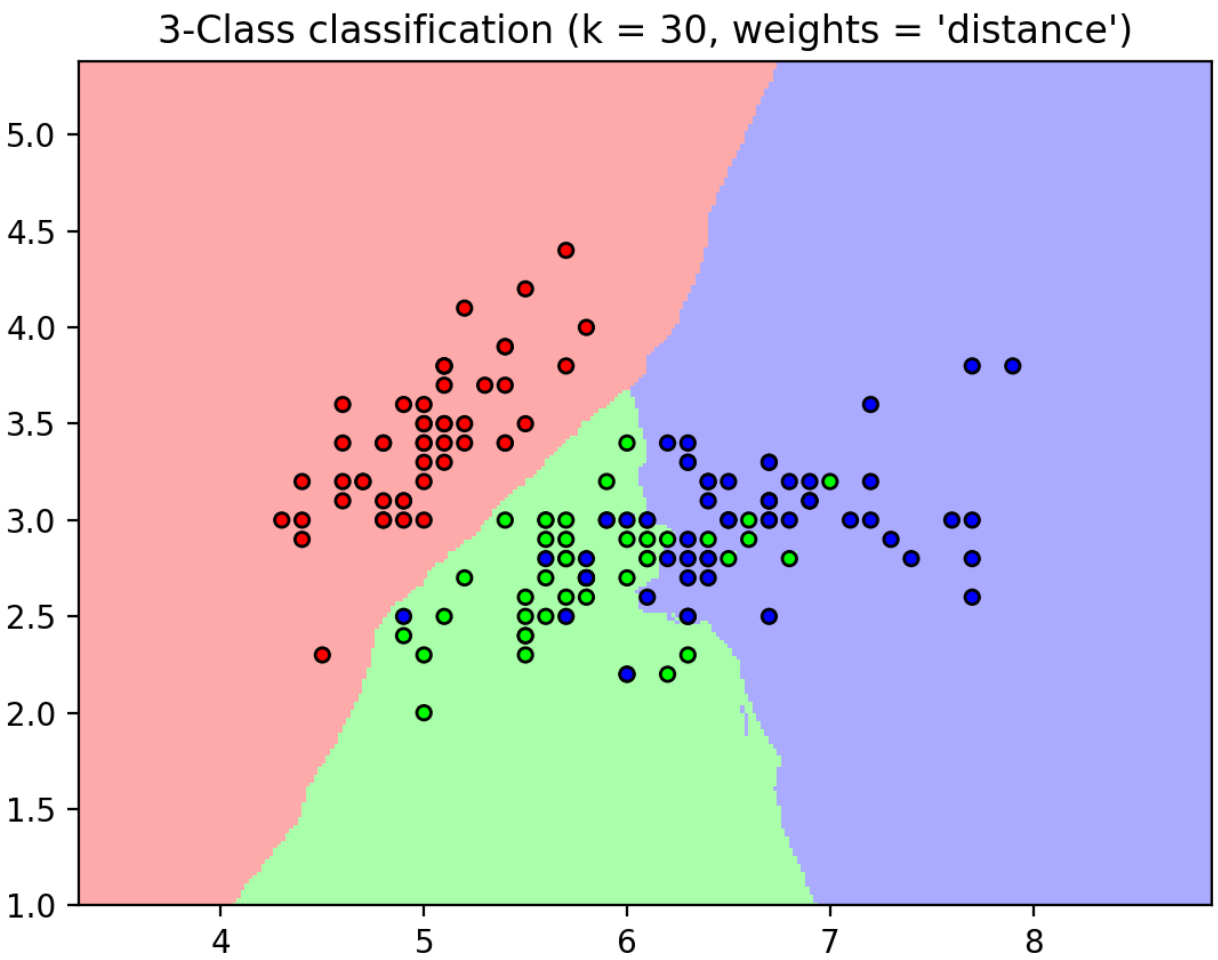
Weights:

- ‘uniform’ : uniform weights.
  - All points in each neighbourhood are weighted equally.
- ‘distance’ : weight points by the inverse of their distance.
  - Closer neighbours of a query point will have a greater influence than neighbours which are further away.


3-Class classification (k = 1, weights = 'uniform')
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
4
5
6
7
8


3-Class classification (k = 15, weights = 'uniform')
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
4
5
6
7
8


3-Class classification (k = 15, weights = 'distance')
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
4
5
6
7
8


3-Class classification (k = 30, weights = 'distance')
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
4
5
6
7
8

Determining k …

- If there are two classes, choosing an odd value for k ensures no ties (although if neighbours are given different weights this is less important – see later)

- If k is small, noise can have a greater influence

- Compute and compare confusion matrices for different values of k

Confusion matrix

		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

- True positive (TP)
  - a hit
- True negative (TN)
  - a correct rejection
- False positive (FP)
  - a false alarm
- False negative (FN)
  - a miss

Accuracy

		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

The number of correct predictions divided by the total number of predictions made.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Precision

		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Also called the positive predictive value (PPV).

$$\frac{TP}{TP + FP}$$

Recall

		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Also called sensitivity, hit rate or true positive rate (TPR)

$$\frac{TP}{TP + FN}$$

Can maximise recall by predicting everything to be positive!

Specificity

		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Also called selectivity or true negative rate (TNR)

$$\frac{TN}{TN + FP}$$

F_1 score

		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

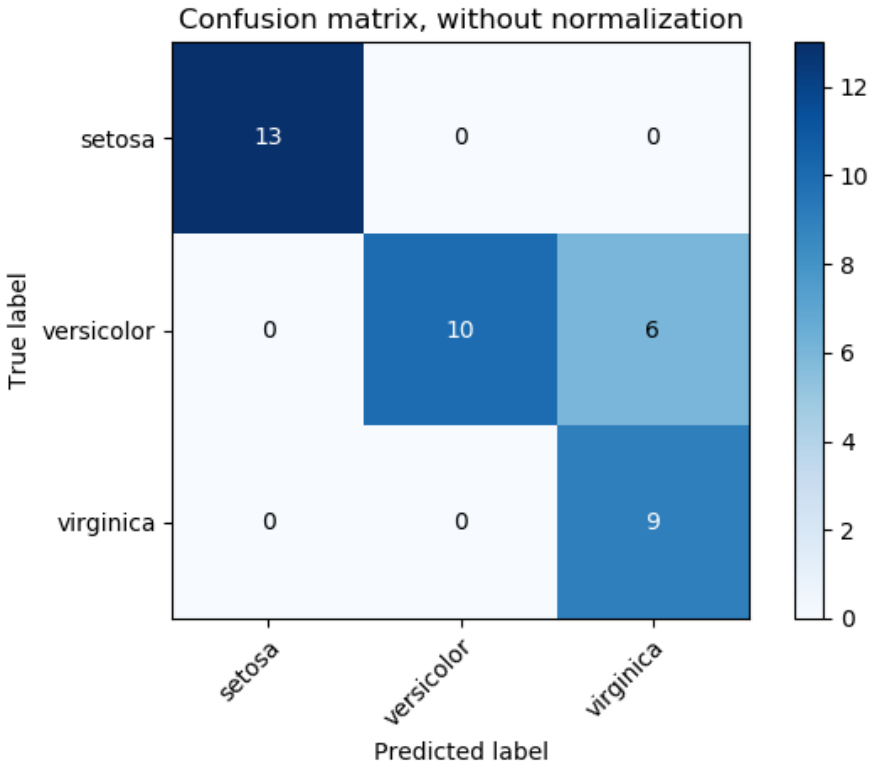
Also called F-score or F-measure.

Can increase precision at the cost of reducing recall, and vice versa.

The F_1 score is the harmonic mean of the precision and the recall.

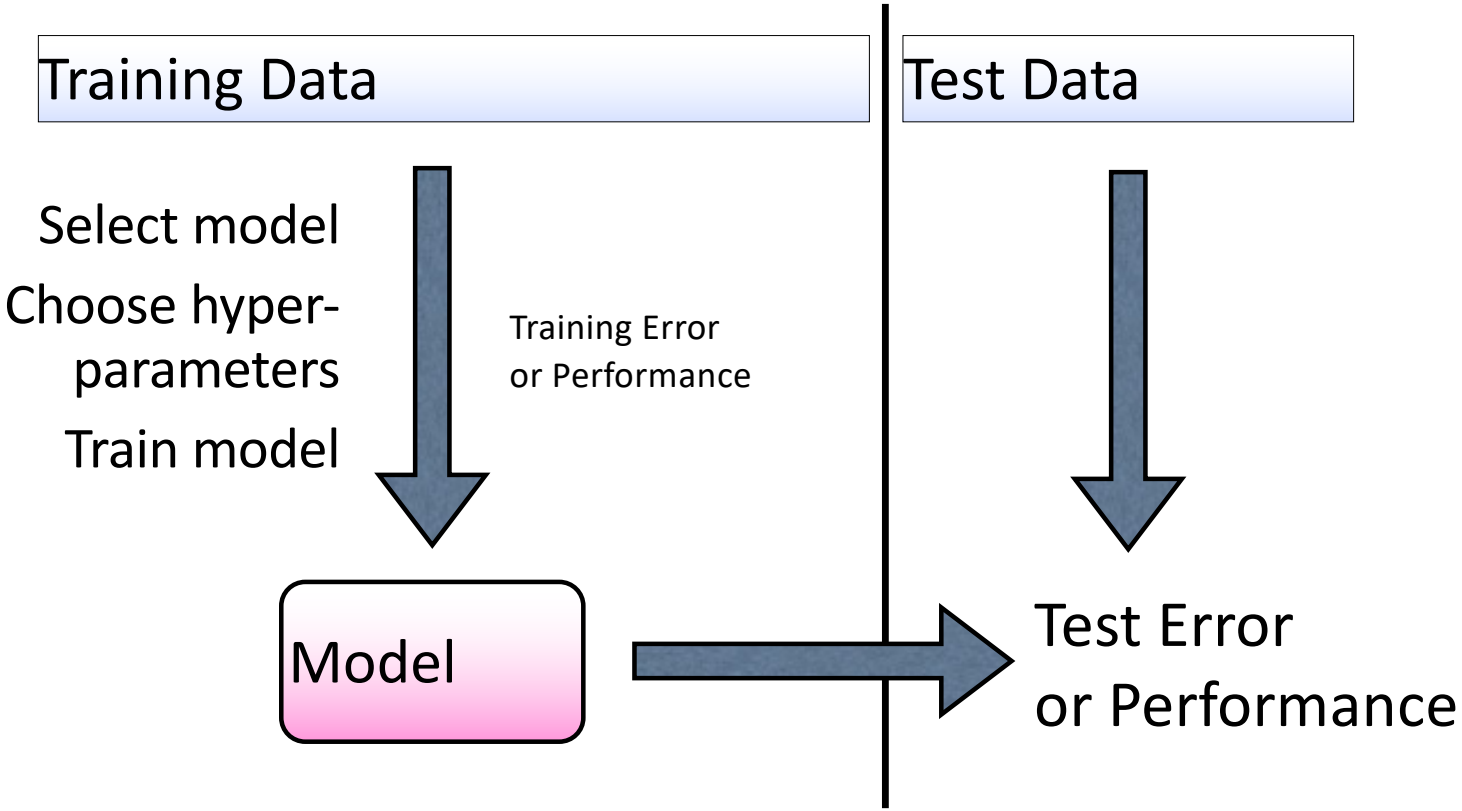
$$\frac{2TP}{2TP + FP + FN}$$

Confusion matrix for multiple classes

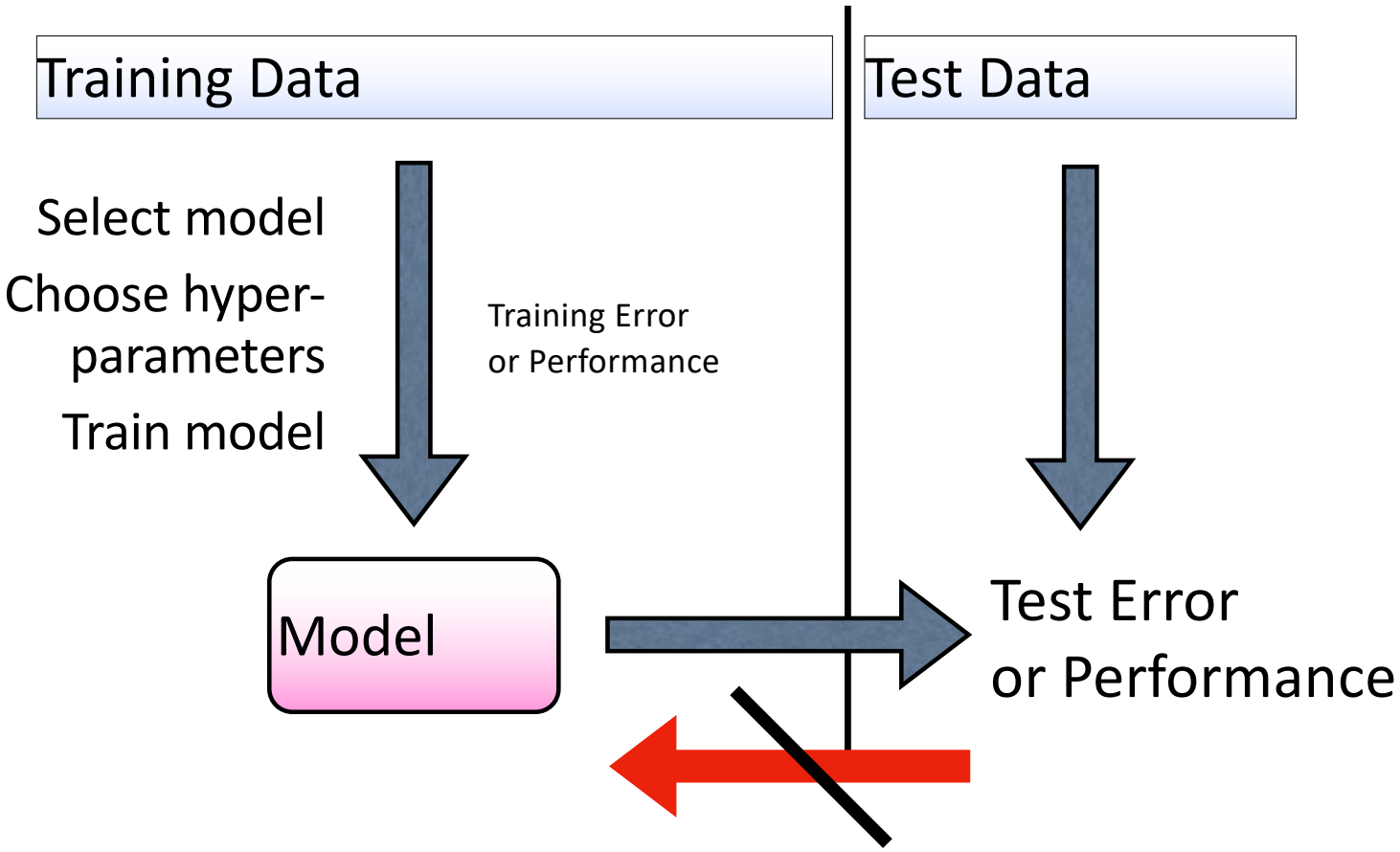


https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

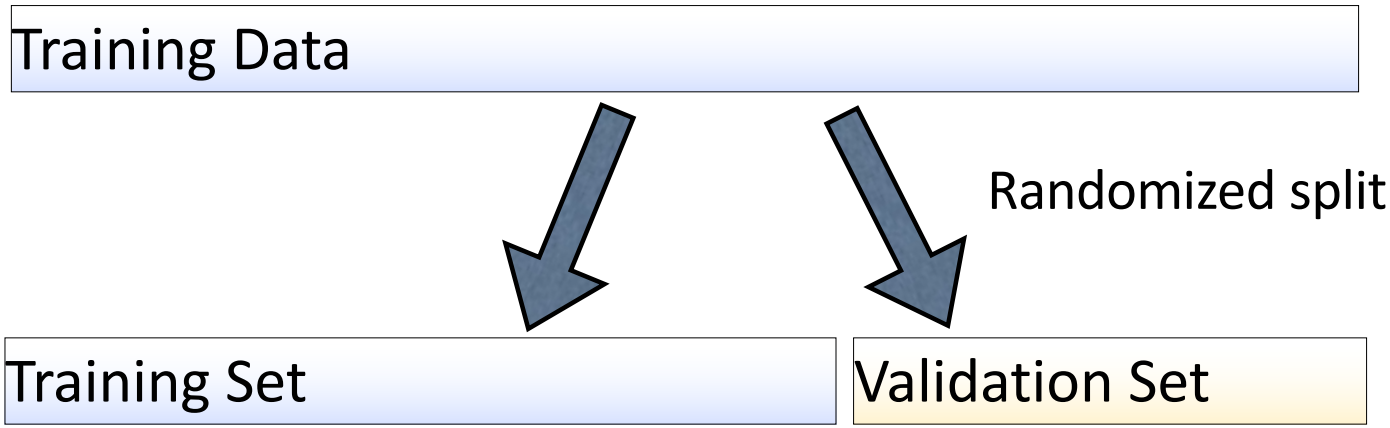
Evaluation strategies



Evaluation strategies



Cross validation: Validation Set



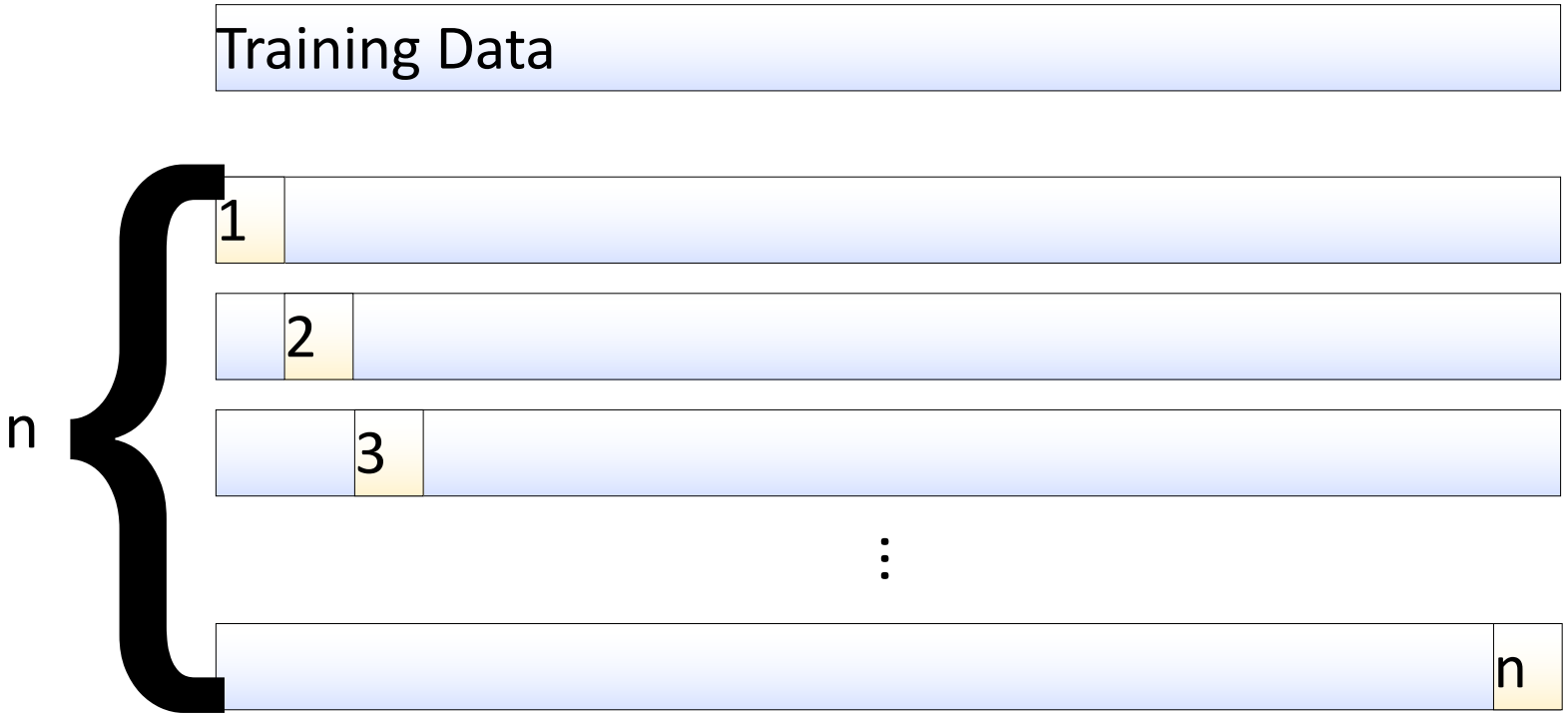
Select model

Choose hyper-parameters

Train model

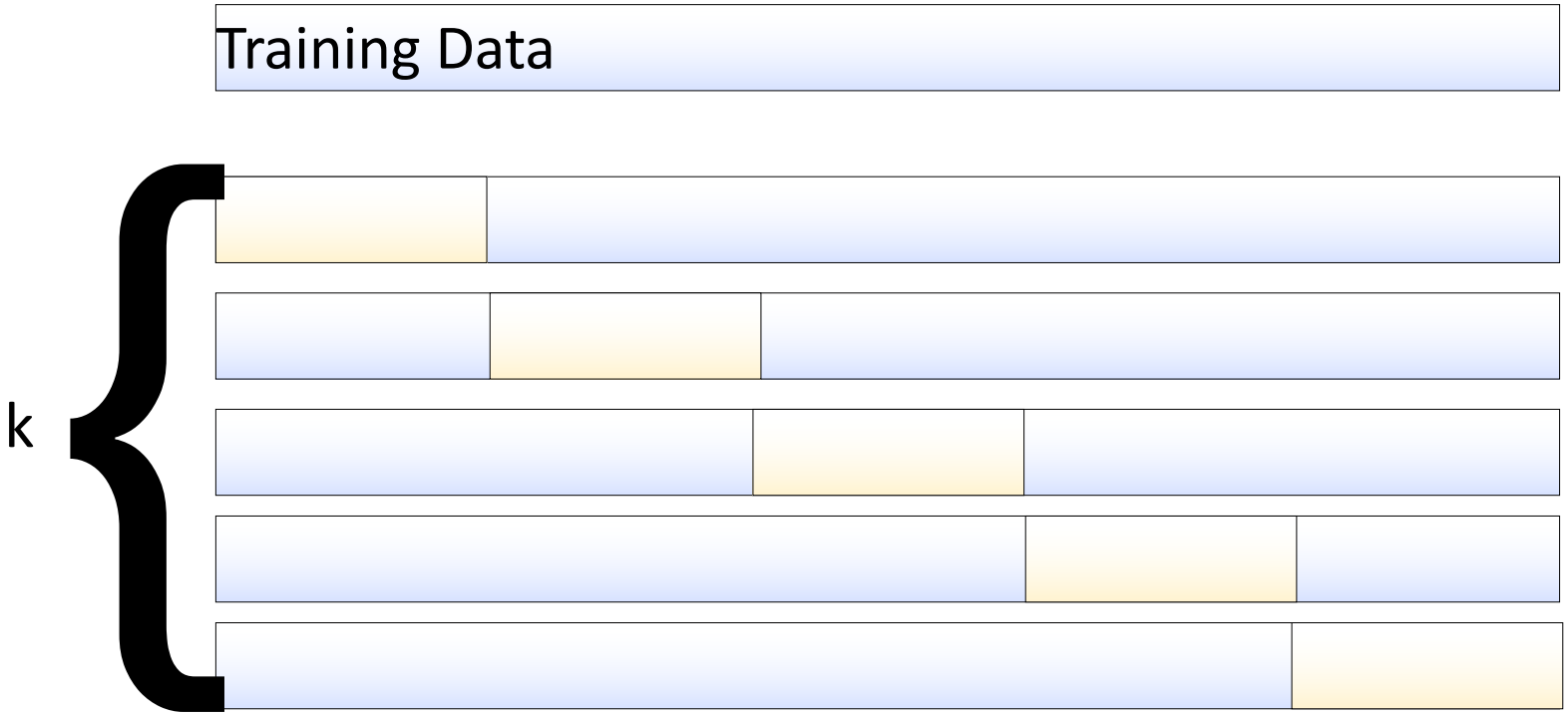
Estimate of Test Error
or Performance

Cross-validation: leave one out



Use D - {i} for training and validate on {i} for all i.
Error/performance averaged over all validations.

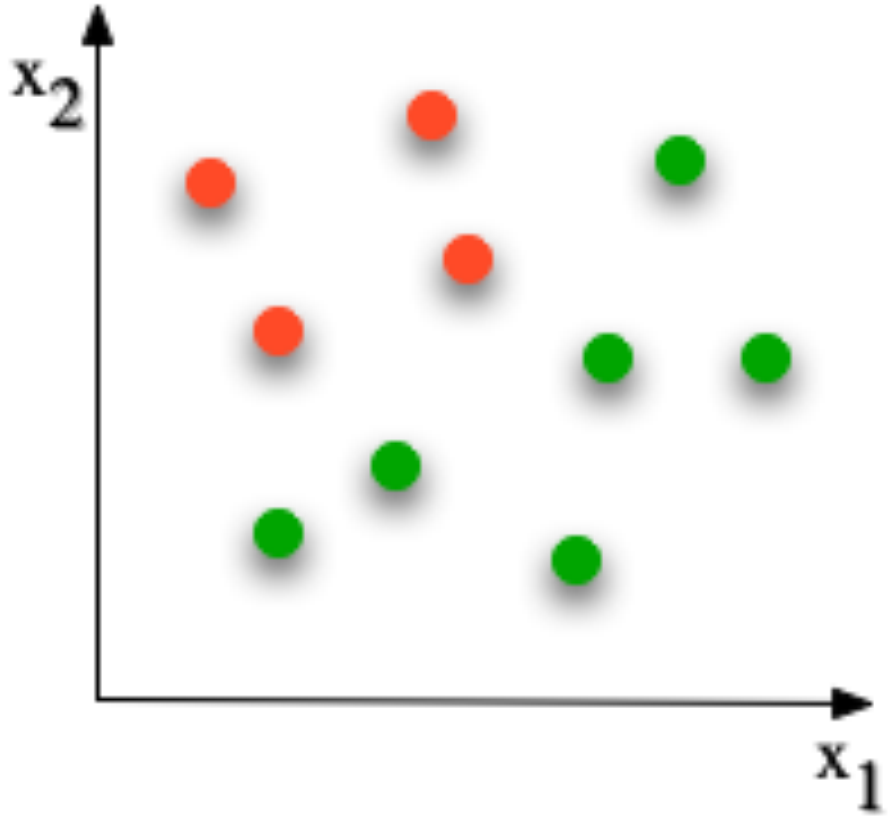
Cross-validation: k-fold cross-validation



Random partition into k non-overlapping groups. Average over k validations.

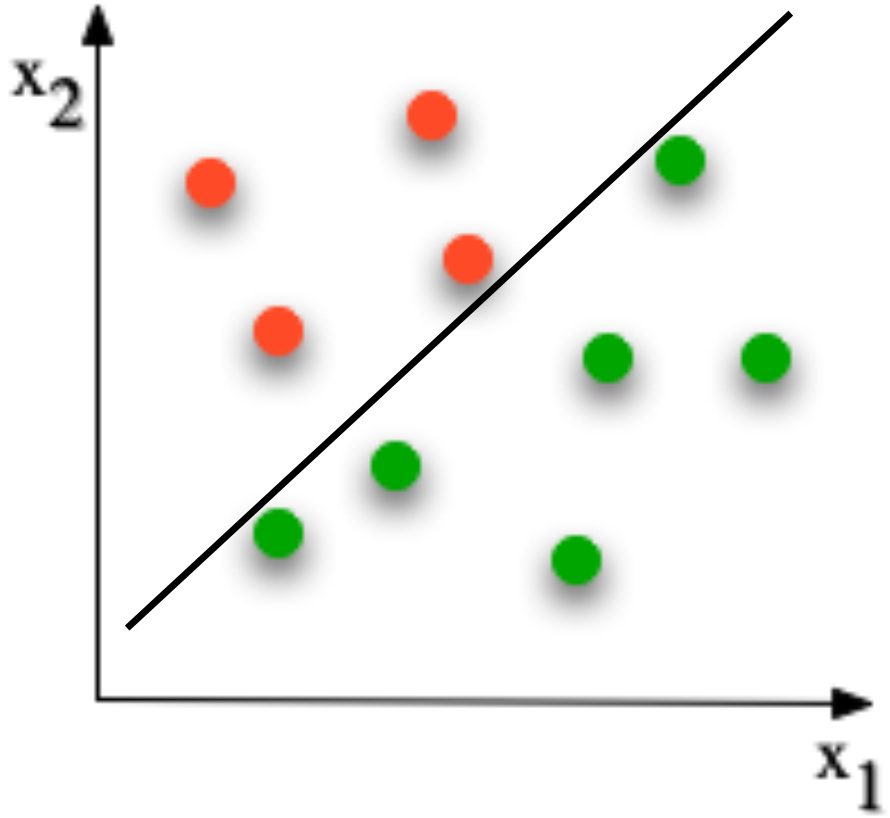
Linear classifier

Idea: find a line separating the classes in the plane

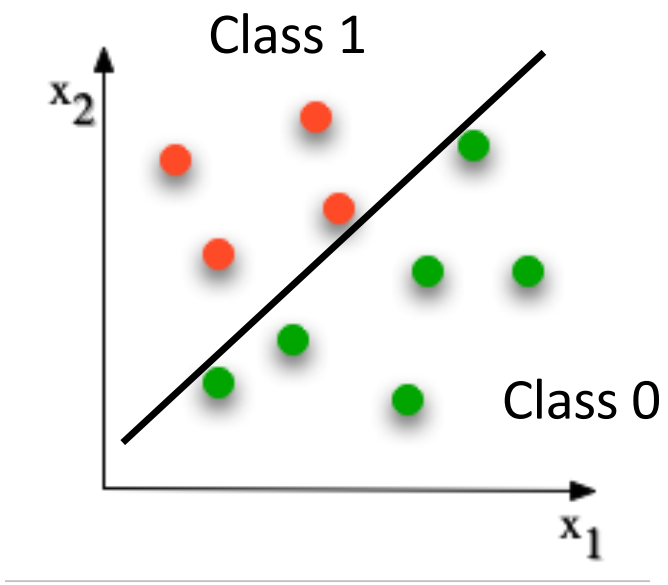


Linear classifier

Idea: find a line separating the classes in the plane



Linear classifier



Line: $x_2 = a\, x_1 + b$

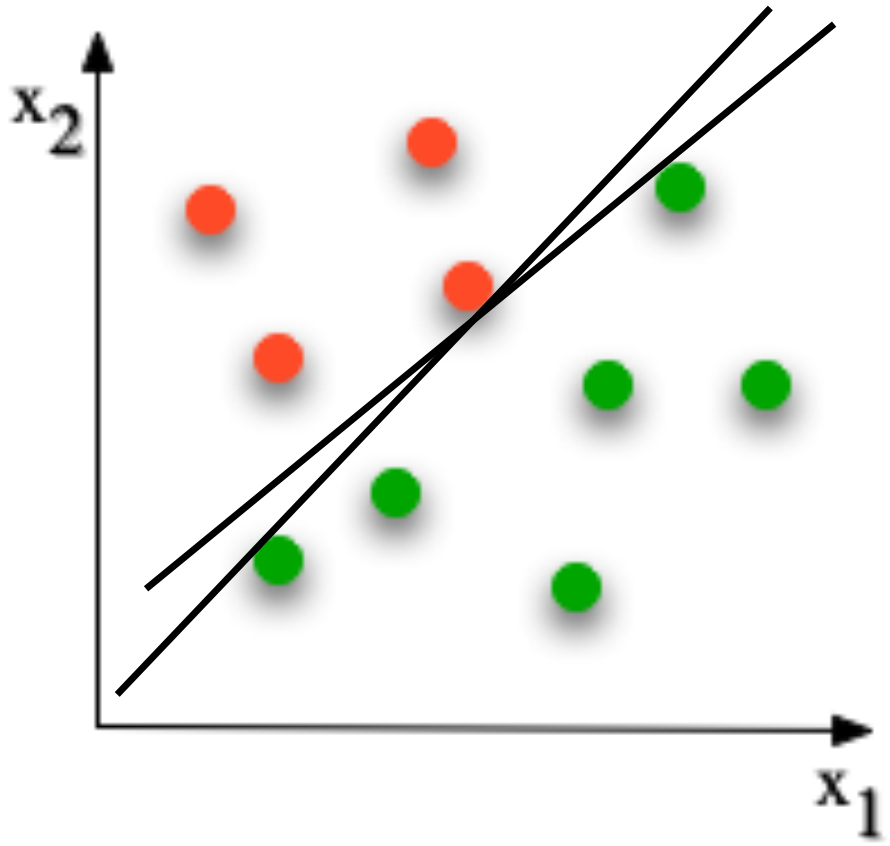
Classifier:

(x_1, y_2) in

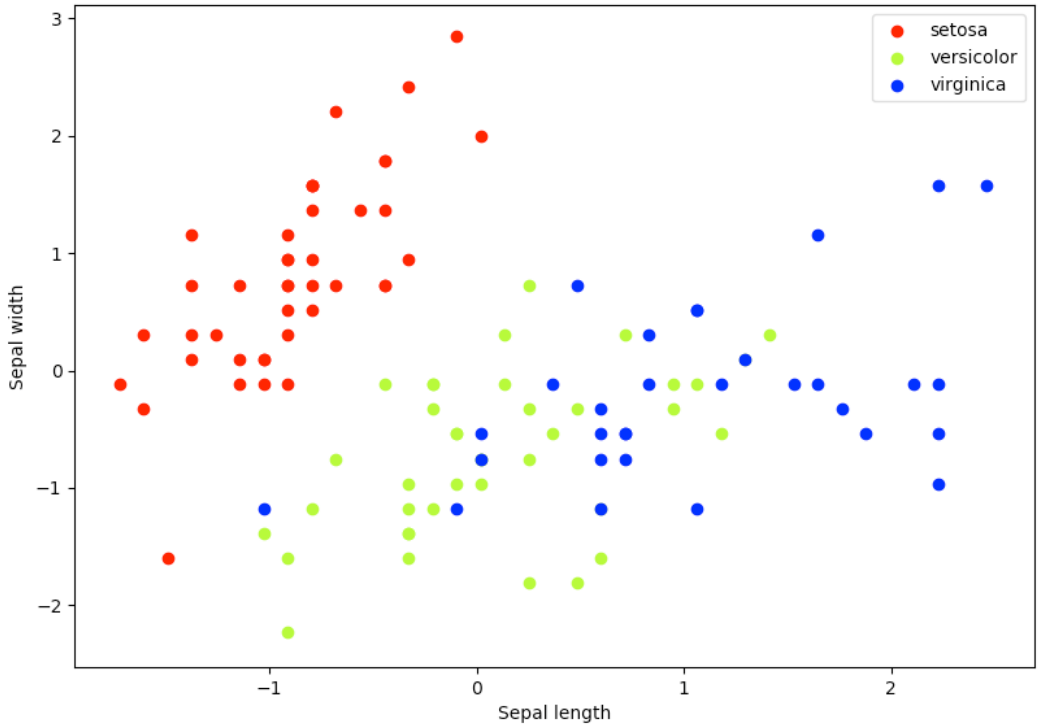
$$\begin{cases} \text{class 0, if } y_2 < a\, x_1 + b \\ \text{class 1 else} \end{cases}$$

Linear classifier

Note: not unique



Iris dataset



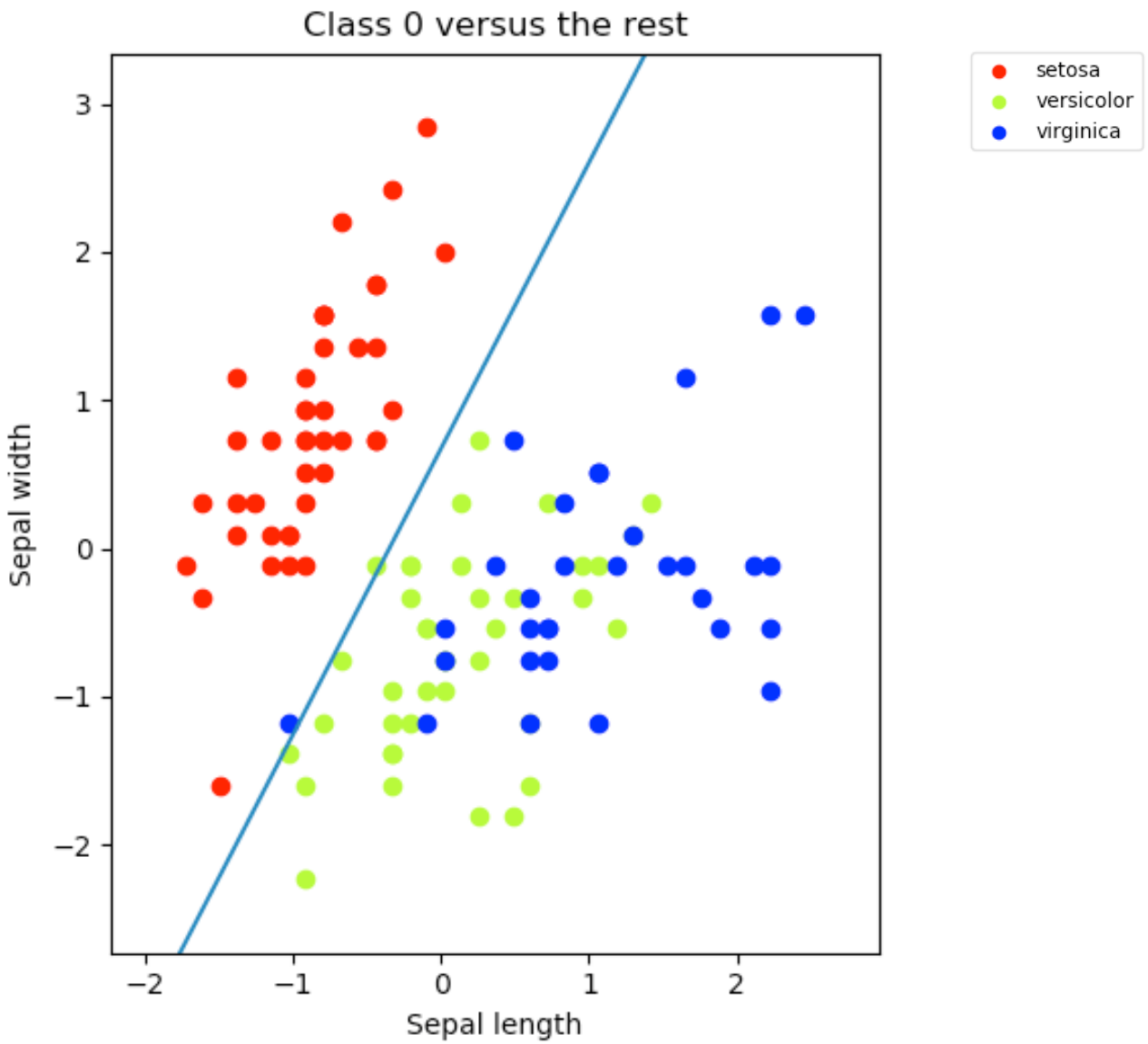
Iris Setosa

Sepal

Iris Versicolor

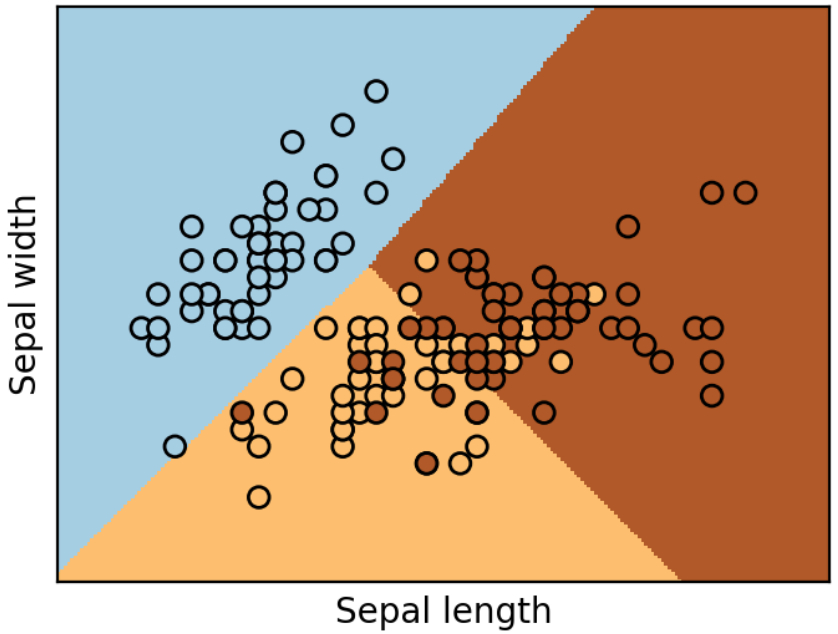
Iris Virginica

Data from Ronald Fisher (1936),
Images Wikipedia


Class 0 versus the rest
setosa
versicolor
virginica
Sepal width
Sepal length
3
2
1
0
−1
−2
−2
−1
0
1
2

Logistic regression 3-class classifier

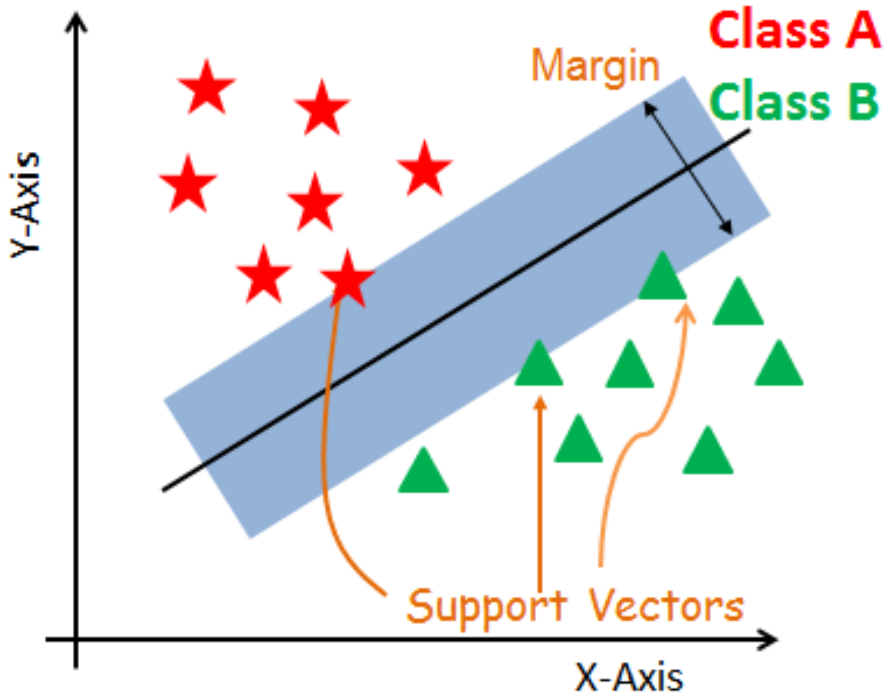
A linear model for classification.



https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/auto_examples/tutorial/plot_iris_logistic.html

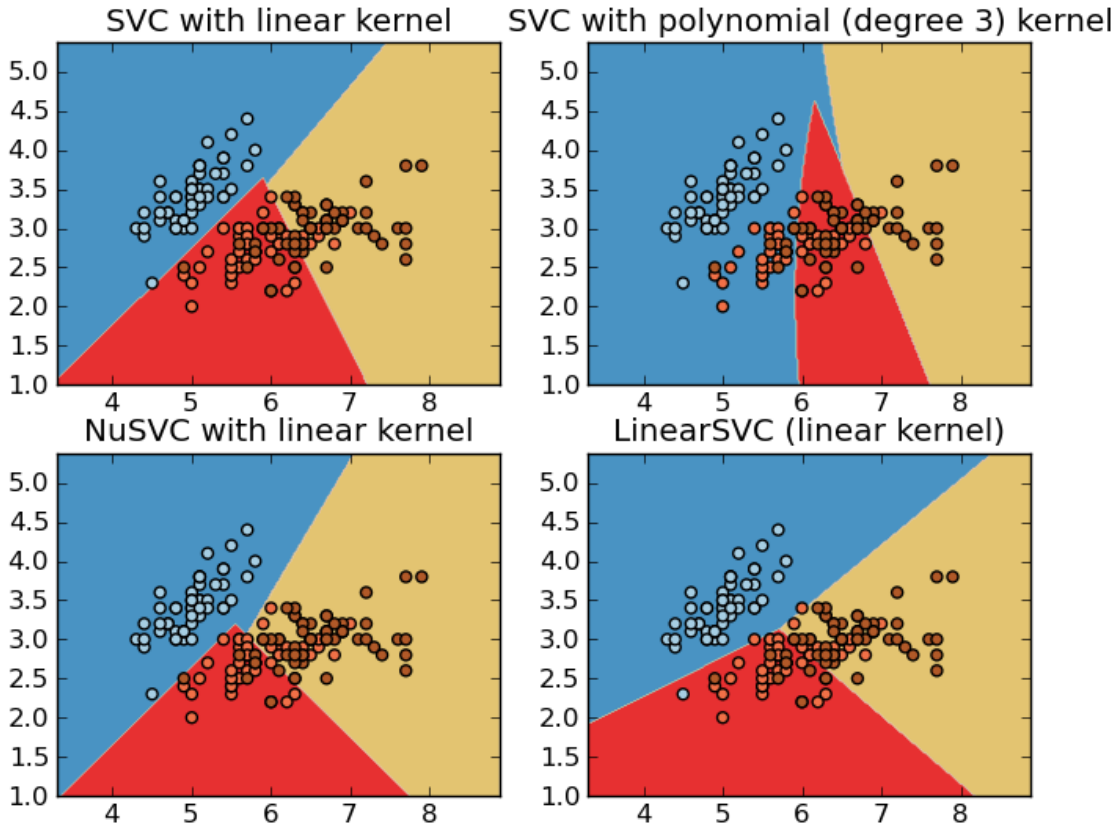
Support vector machines

- Find the line (or plane or hyperplane in higher dimensions) that separate the two classes with the largest margin.



https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

SVMs and the iris dataset



http://scikit-learn.sourceforge.net/0.5/auto_examples/svm/plot_iris.html