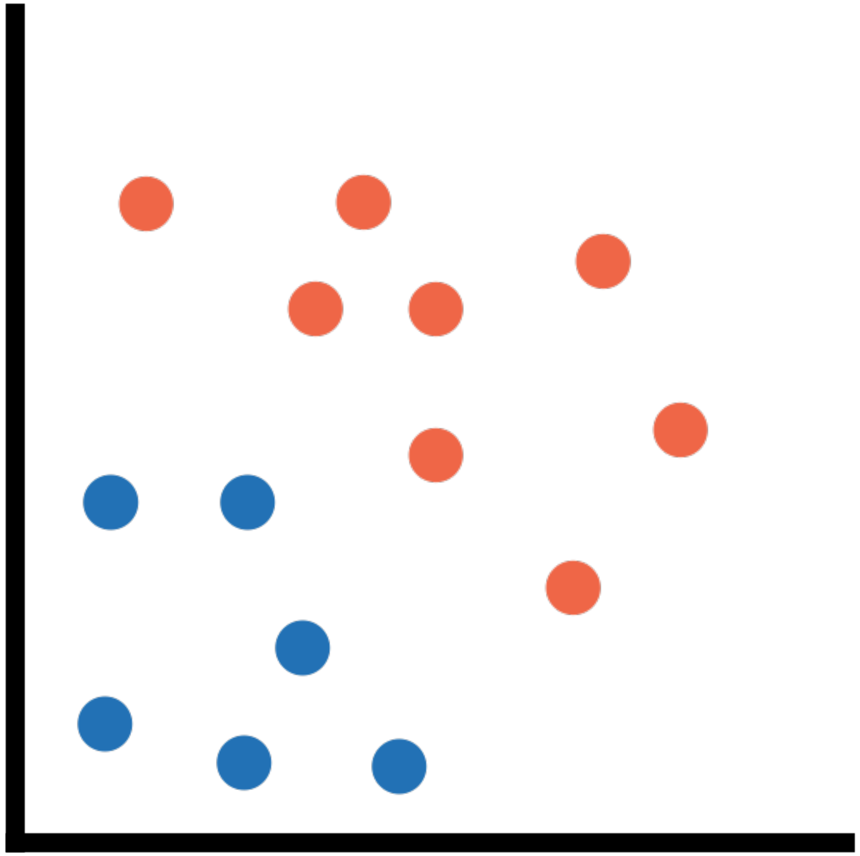
Advanced Topics in Social and Geographic Data Science

GPS data classification

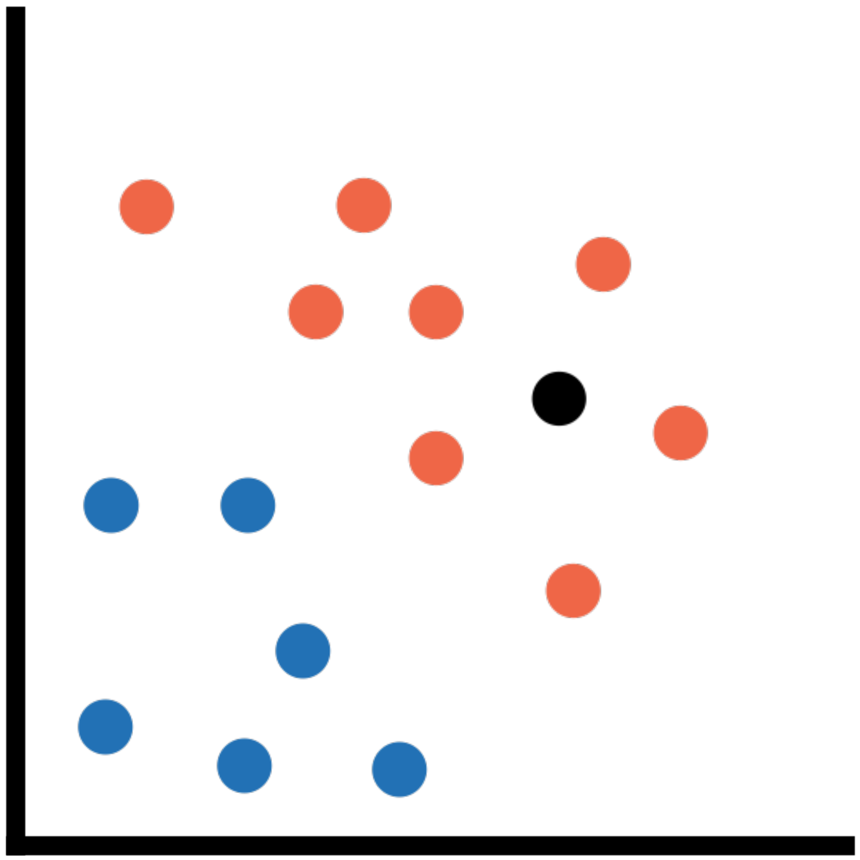
GPS data classification

- classification (supervised learning problem involving predicting a class label)
- generative model (using a decision boundary)
- tree-based methods (boosted decision tree, random forest)
- support vector machines

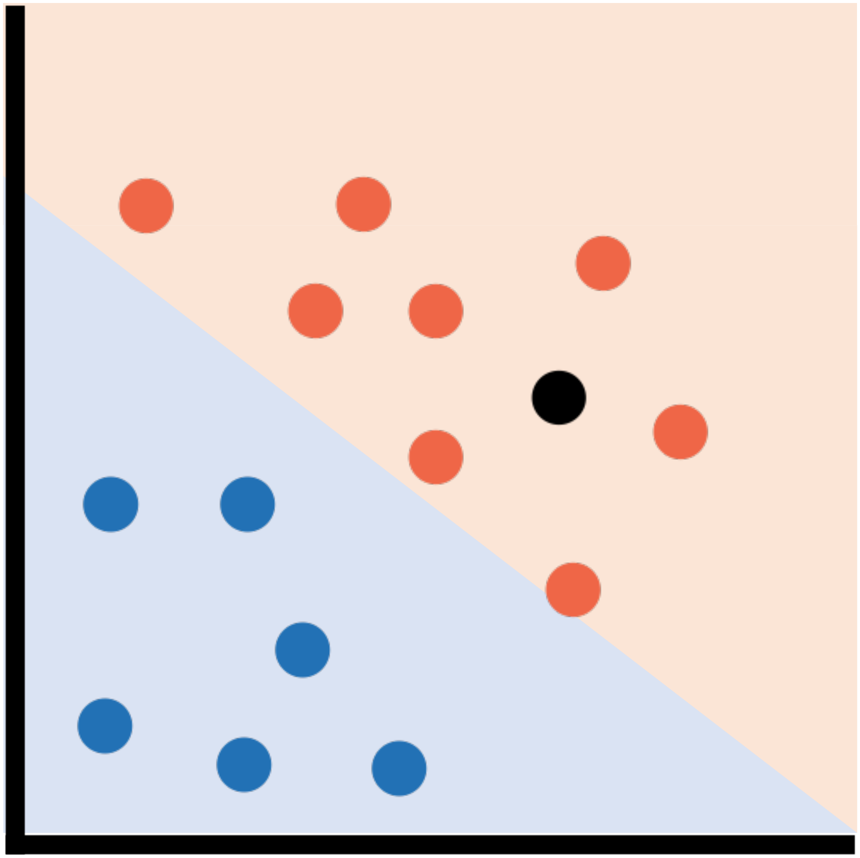
Decision surface



Decision surface



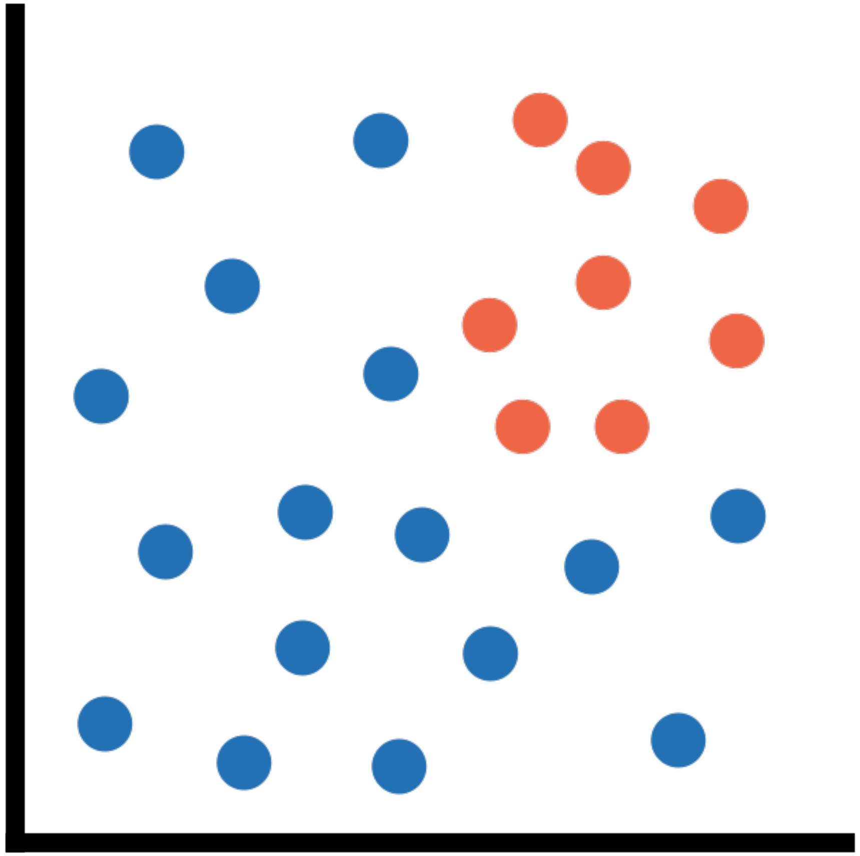
Decision surface



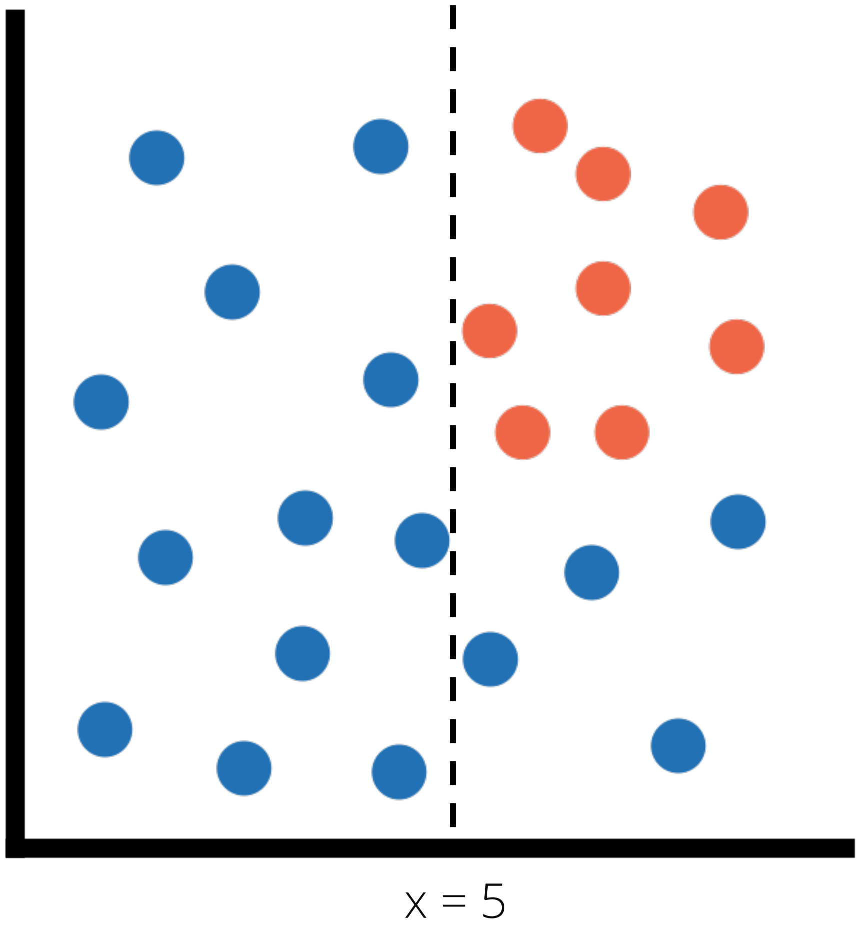
Processing pipeline

- split into train and test
- develop model on the train data set (decision surface)
- predict classes on the test data set
- measure of accuracy: percentage correctly predicted or Kappa value for unbalanced classes (comparison of prediction to class probabilities)

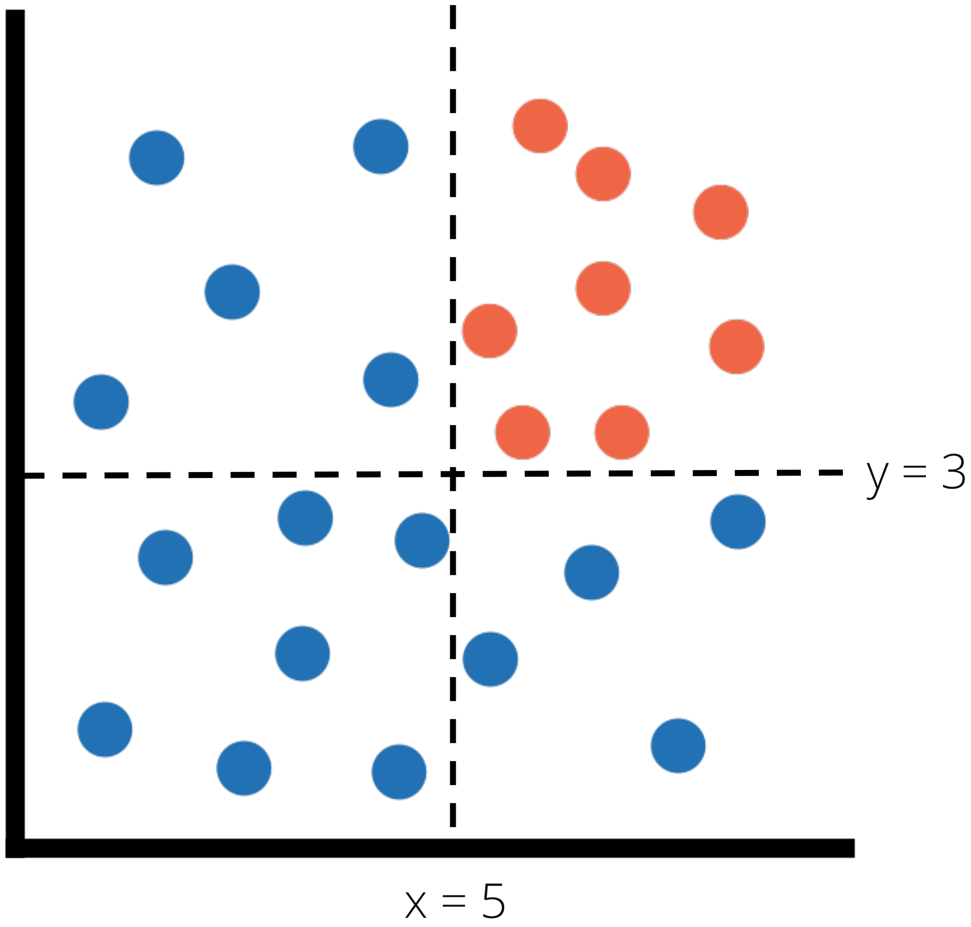
Decision trees



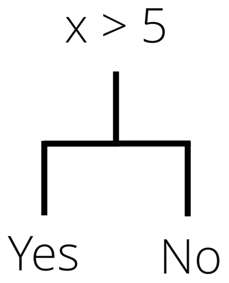
Decision trees

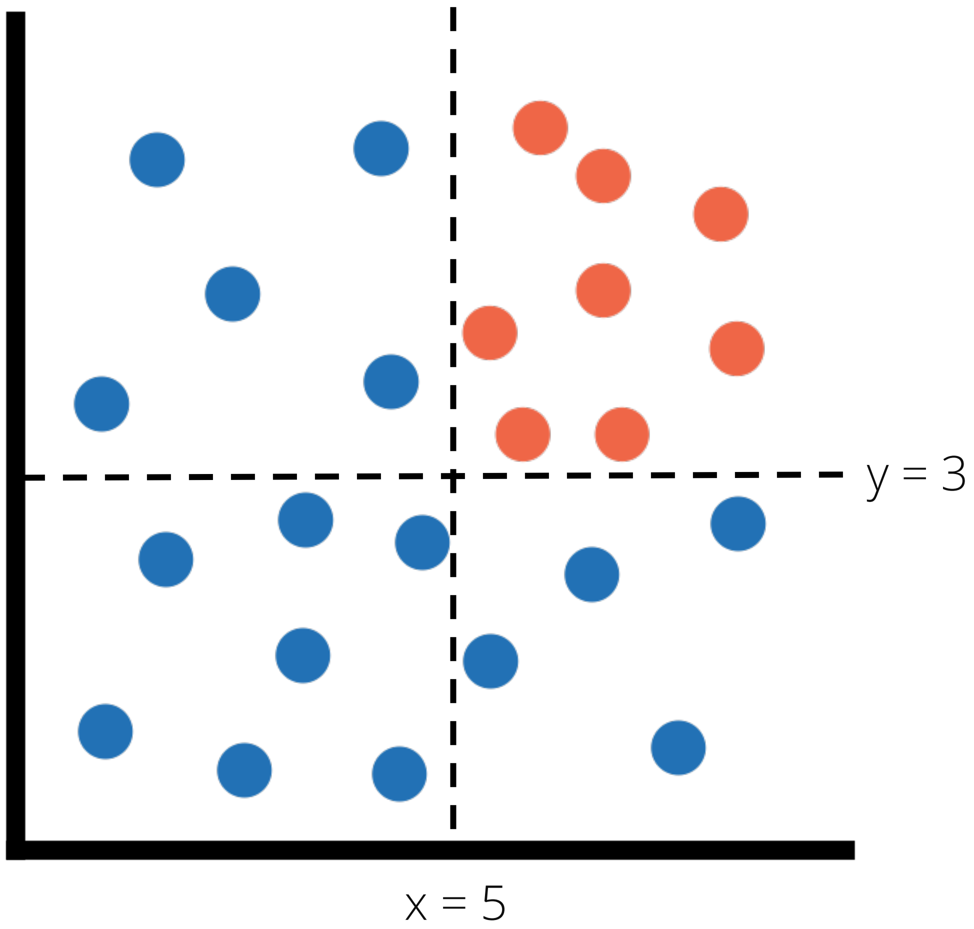


Decision trees

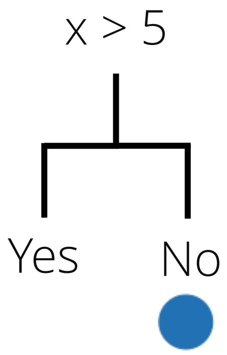


Decision trees





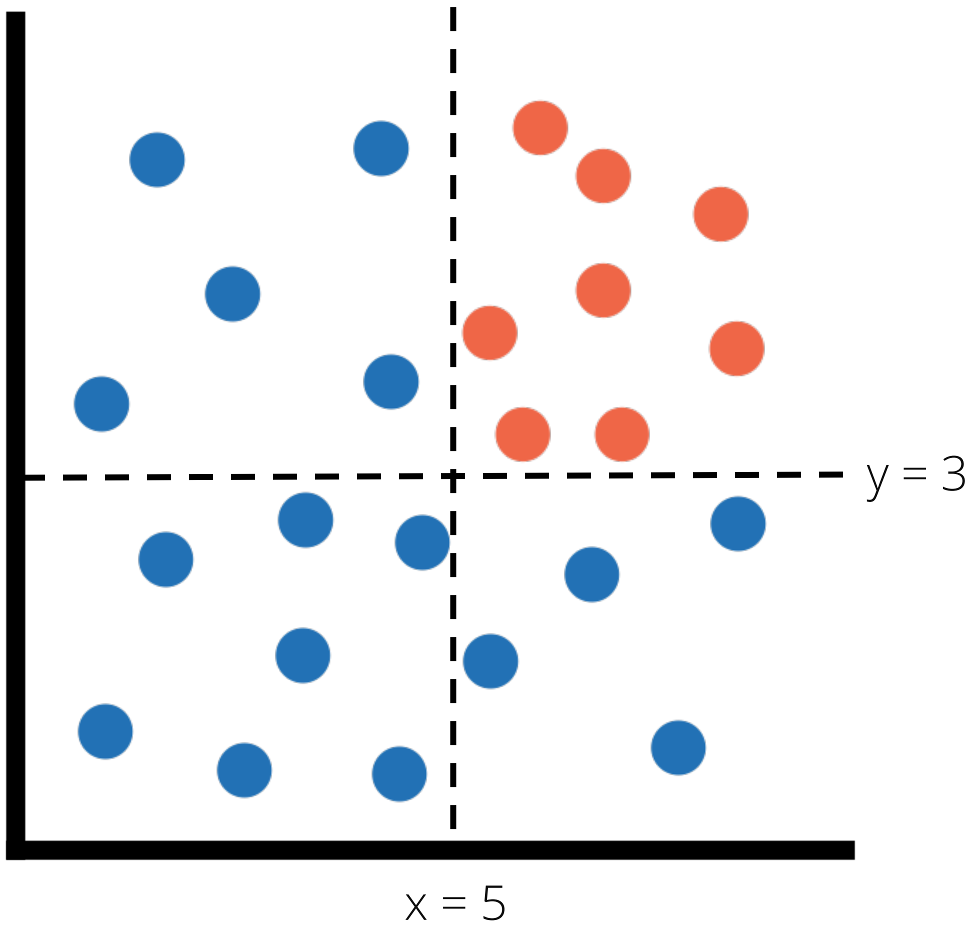
Decision trees



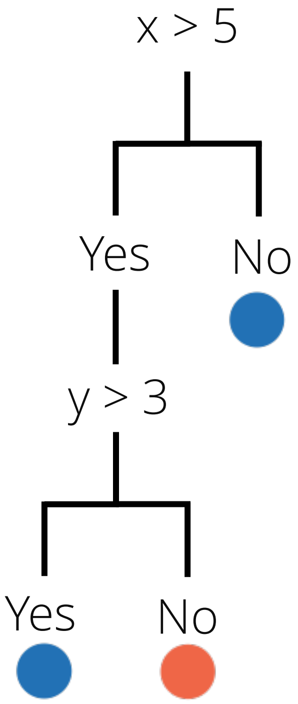
x > 5

Yes

No



Decision trees



x > 5

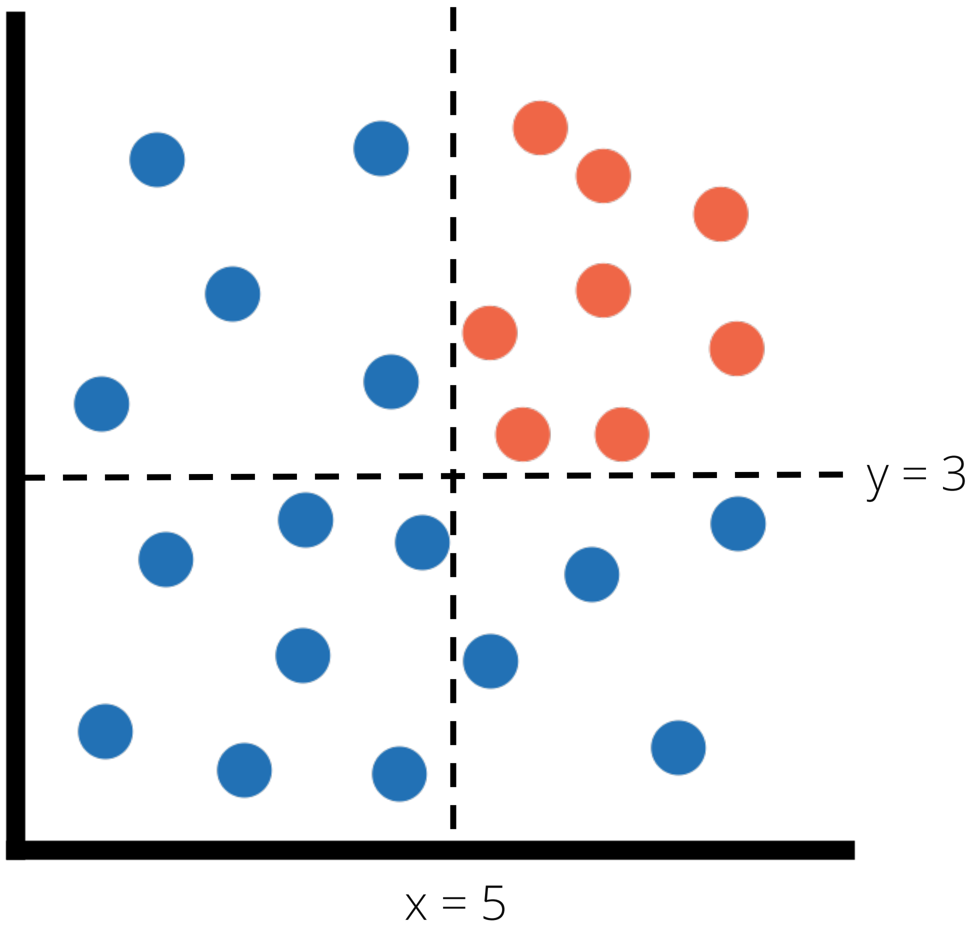
Yes

No ●

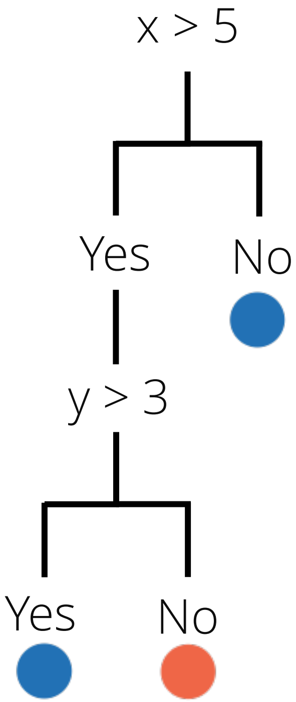
y > 3

Yes ●

No ●



Decision trees



x > 5

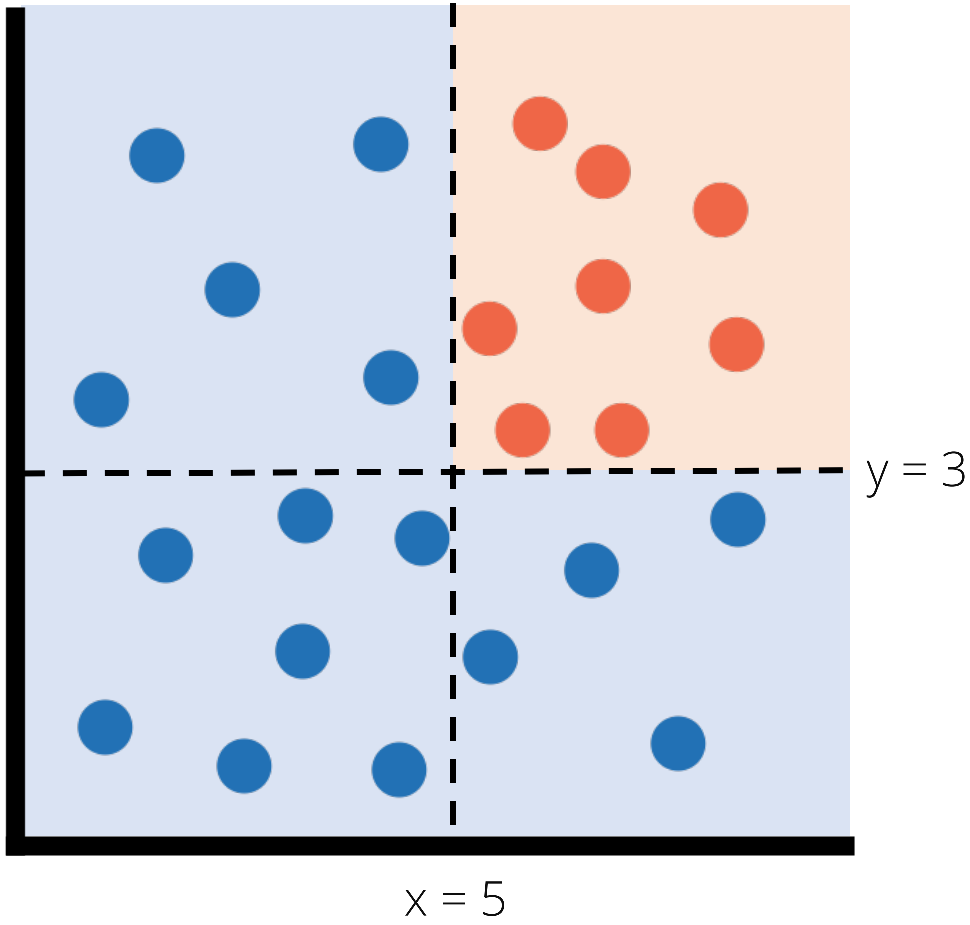
Yes

No ●

y > 3

Yes ●

No ●



Splitting the data

- entropy used to split the data (how to decide where to split the data): measure of impurity (which split will give the best results), i.e. is there a split possible where one of branches gets all class label of one group
    - all examples in same class: entropy is 0
    - all examples equally split: entropy is 1
- entropy used to calculate information gain – which split will lead to the largest decrease in entropy

Typical parameters

- minimum sample split (when to stop splitting the data to avoid overfitting): minimum number of samples in your test data at the leaves of your tree
- pruning the tree (removing sections with little predictive power) to avoid overfitting

Boosted decision trees

- grow a full decision tree
- test the resulting tree against the training data
- try to improve the misclassified class labels with a new decision tree
- test the resulting tree against the training data
- try to improve the misclassified class labels with a new decision tree
- …and repeat till maximum number of boosting iterations has been reached

- for prediction: majority vote for final class label

Random forest

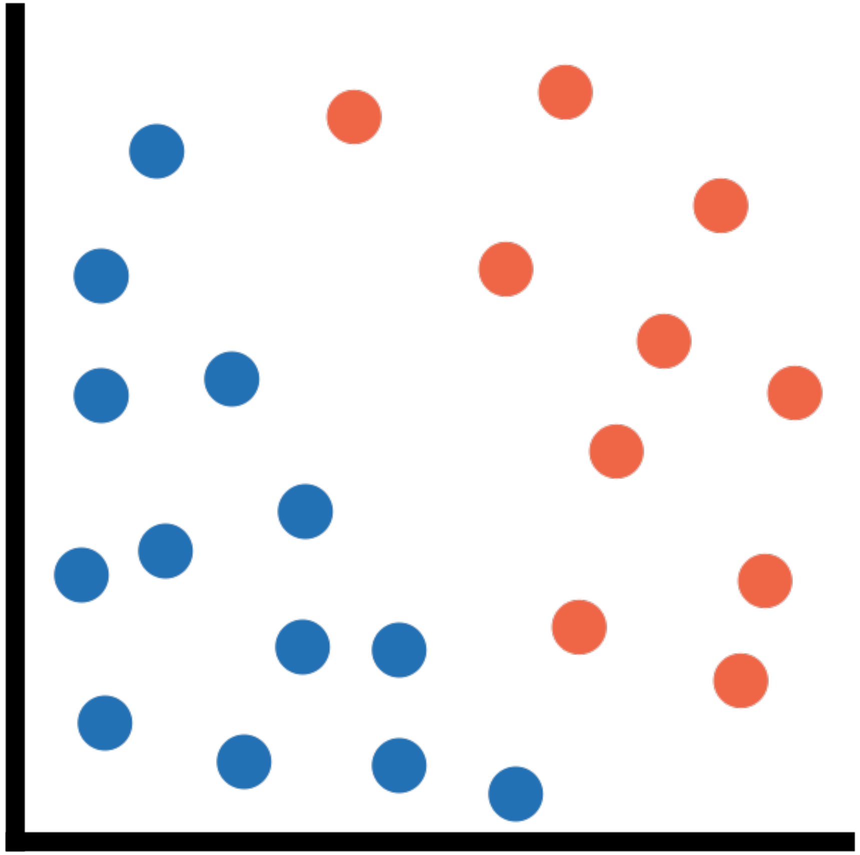
- grow many decisions trees (an ensemble of trees) but on a subset of the training data using bootstrap sampling (random sampling with replacement)
- at every split an *n* number of randomly selected variables is used to calculate the next split
- result is a forest of trees that have been grown using varying features on different subsets of the training data

- for prediction: majority vote for final class labels

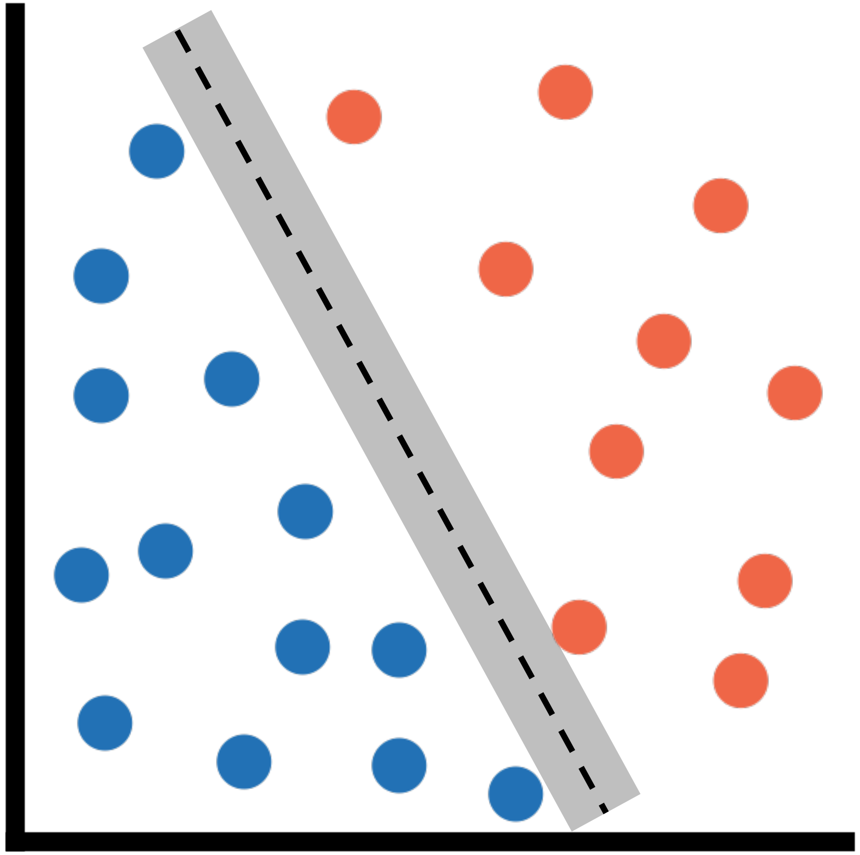
Tree-based methods

- many different implementations, different parameters
- different ways of calculating information gain
- typically involve a form of boosting, bagging

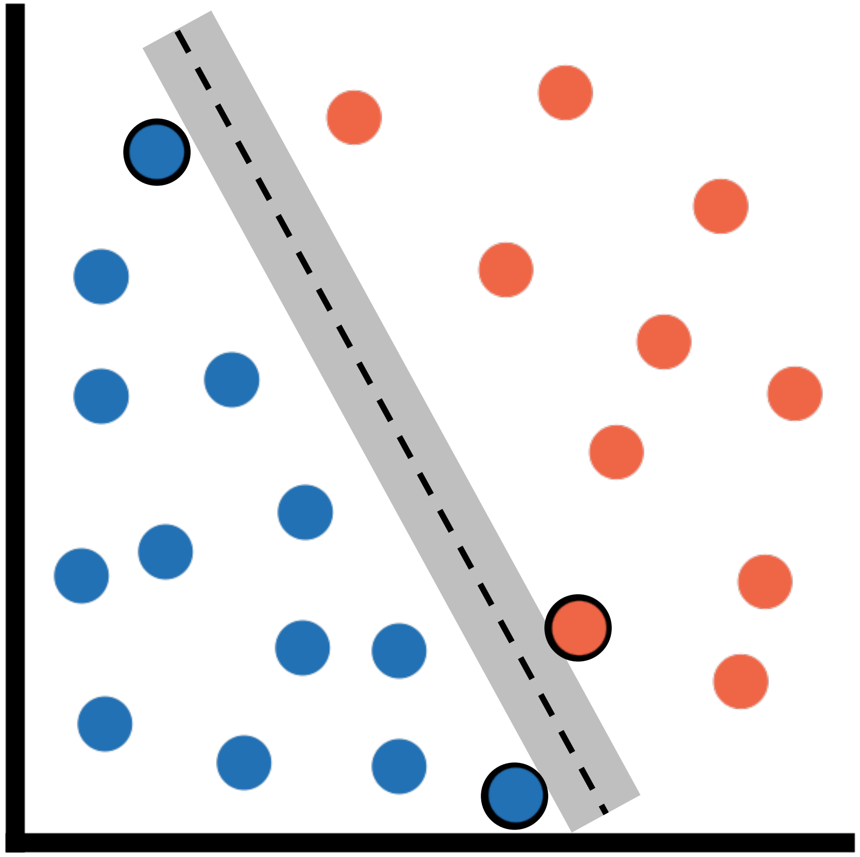
Support Vector Machines



Support Vector Machines

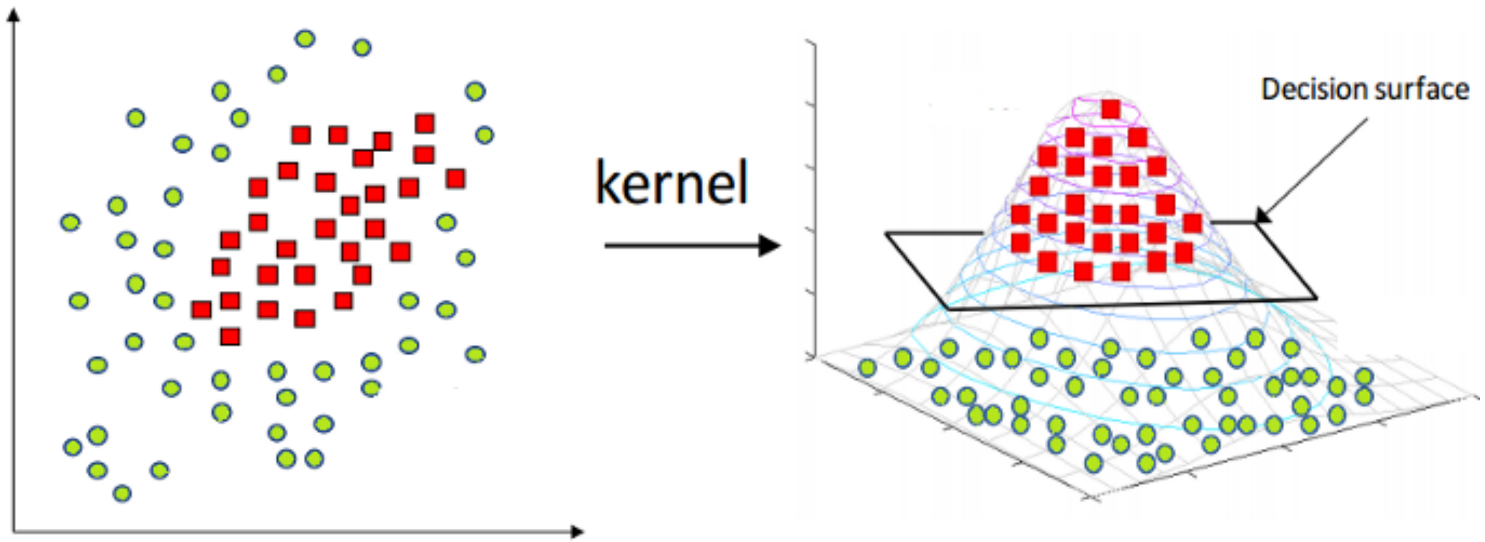


Support Vector Machines



Support Vector Machines

- tries to find the largest possible hyperplane to linearly separate the data by maximising the distance to the nearest points (margin)
- different way to deal with outliers (ignoring)
- SVM is a classifier built on linear-separation but it can generate non-linear decision boundaries by transforming the input data to make them linear separatable through something called the kernel trick
- kernel trick transforms low-dimensional data into high-dimensional data to make them linearly separatable


kernel
Decision surface

Support Vector Machines

- different kernels
- C parameter (controls the tradeoff between smooth decision boundary and one that classifies all the training points correctly)
- Gamma parameter (defines the range of influence of a single training example has on the decision boundaries)

GPS data classification

- brief introduction to two sets of supervised machine learning classifiers that could be used to impute trips (moves) and activities (stays) from raw GPS trajectories
- like all supervised machine learning classifiers: labelled data are required to train the classifiers