Imperial College London

Supervised Learning: classification

Machine Learning

QSRI summer school – July 2022

Sarah Filippi

Department of Mathematics
Imperial College London

Classification task

We have already looked at the regression task, and how it can be tackled using linear regression models.

Classification is another type of supervised learning task in which the labels are from a *discrete* set of values.

We assume we are given a dataset $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$ such that, for all $1 \leq i \leq N$, $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \{1 \dots K\}$.

Binary classification: exactly 2 label types.

Multiclass classification: more than 2 label types.

What is Classification?

Automatic methods for deciding in which group a new object should be categorised.

Very important class of problems in the fields of Machine Learning and Statistics.

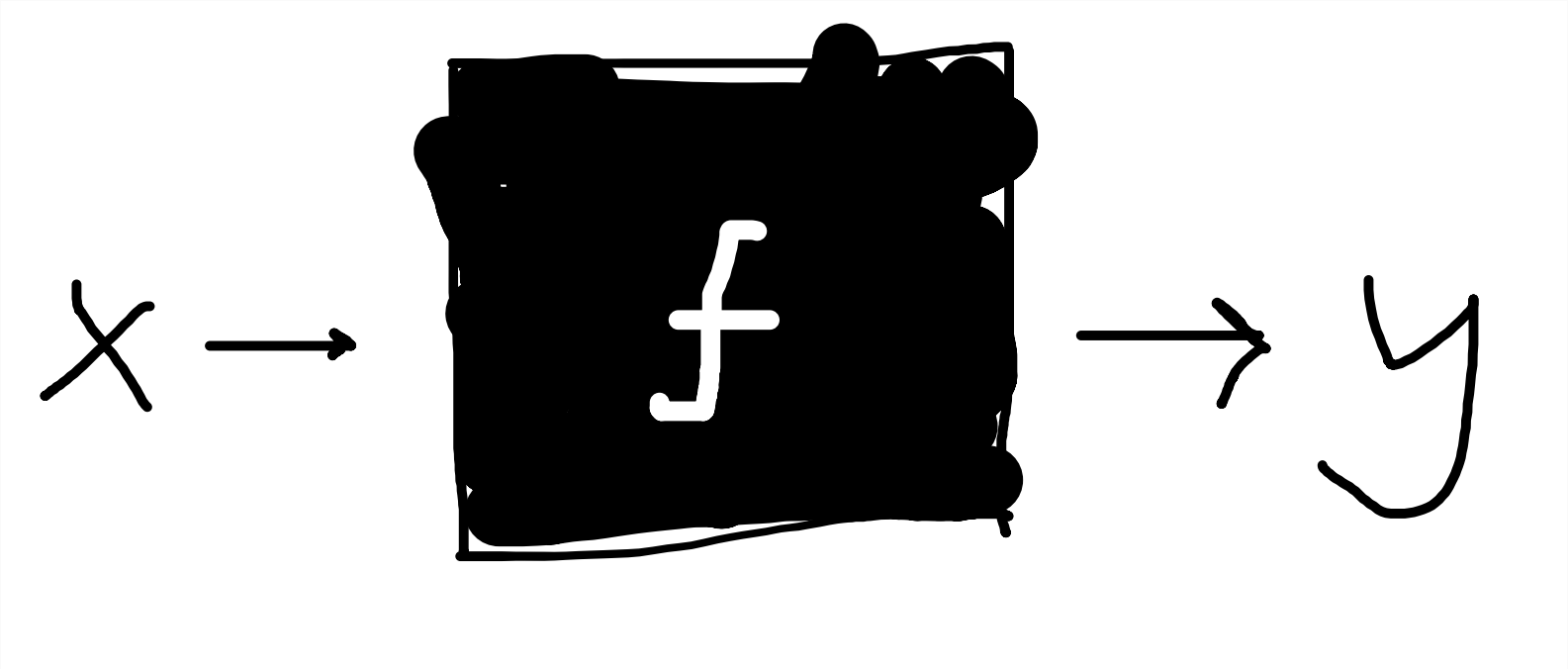
Examples include:

- Spam filtering
- Fraud detection (e.g. credit card transaction fraud)
- Character/face recognition
- Medical diagnosis

In the first lecture, we considered a very simple classifier: the k-nearest neighbours classifier.

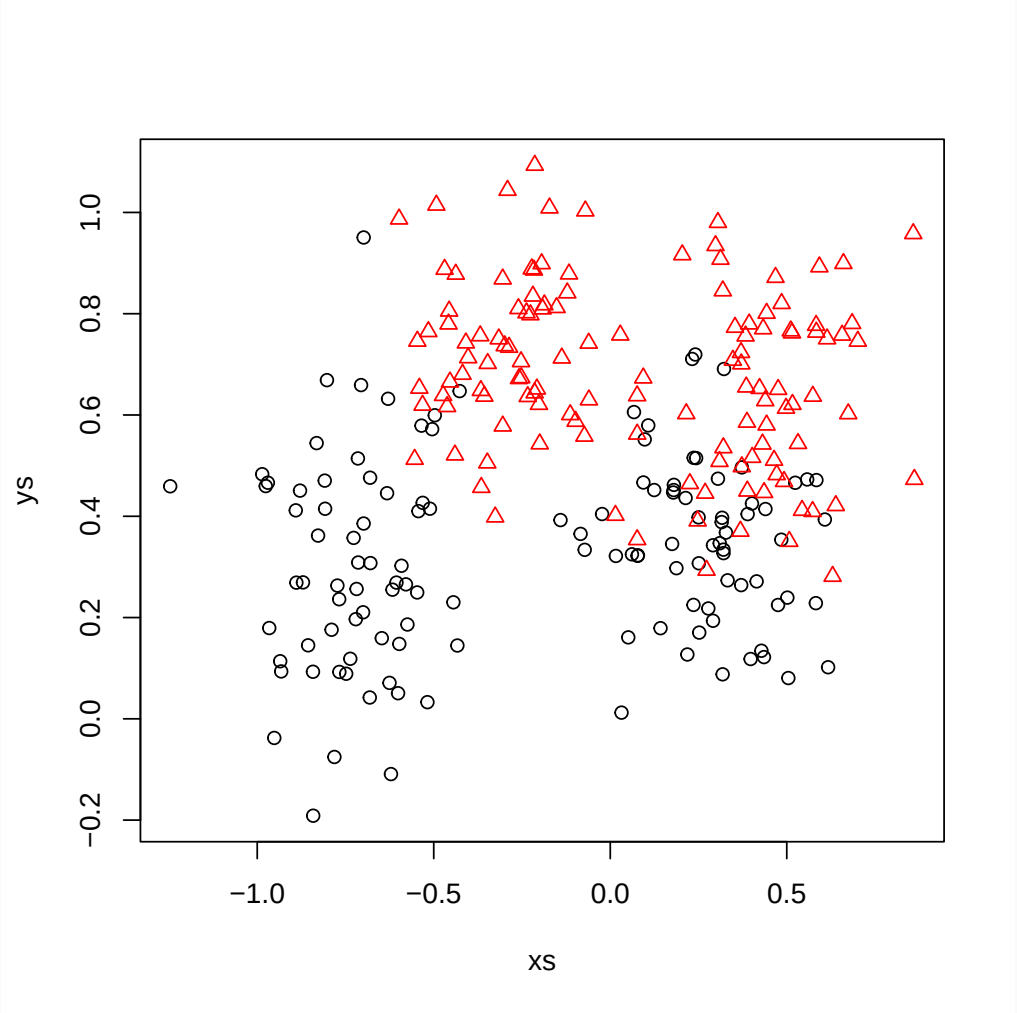
We will consider now how to train parametric models for classification.

Supervised learning



The function (algorithm, black box, decision rule, classifier, probability distribution) $f : \mathcal{X} \mapsto \mathcal{Y}$ encapsulates our assumption regarding the underlying mechanism that generates the observed data.

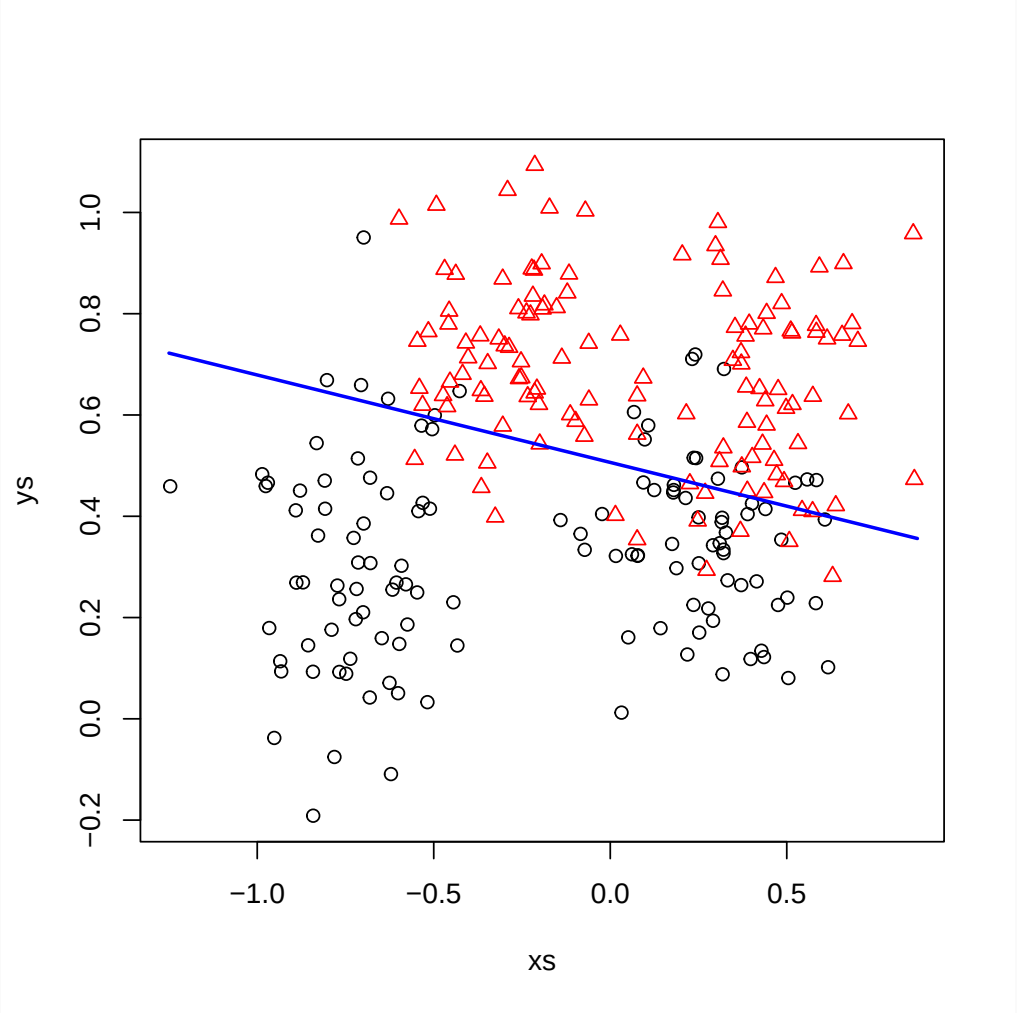
Classification using a decision boundary



Can we identify a decision boundary that would delimit the two classes?

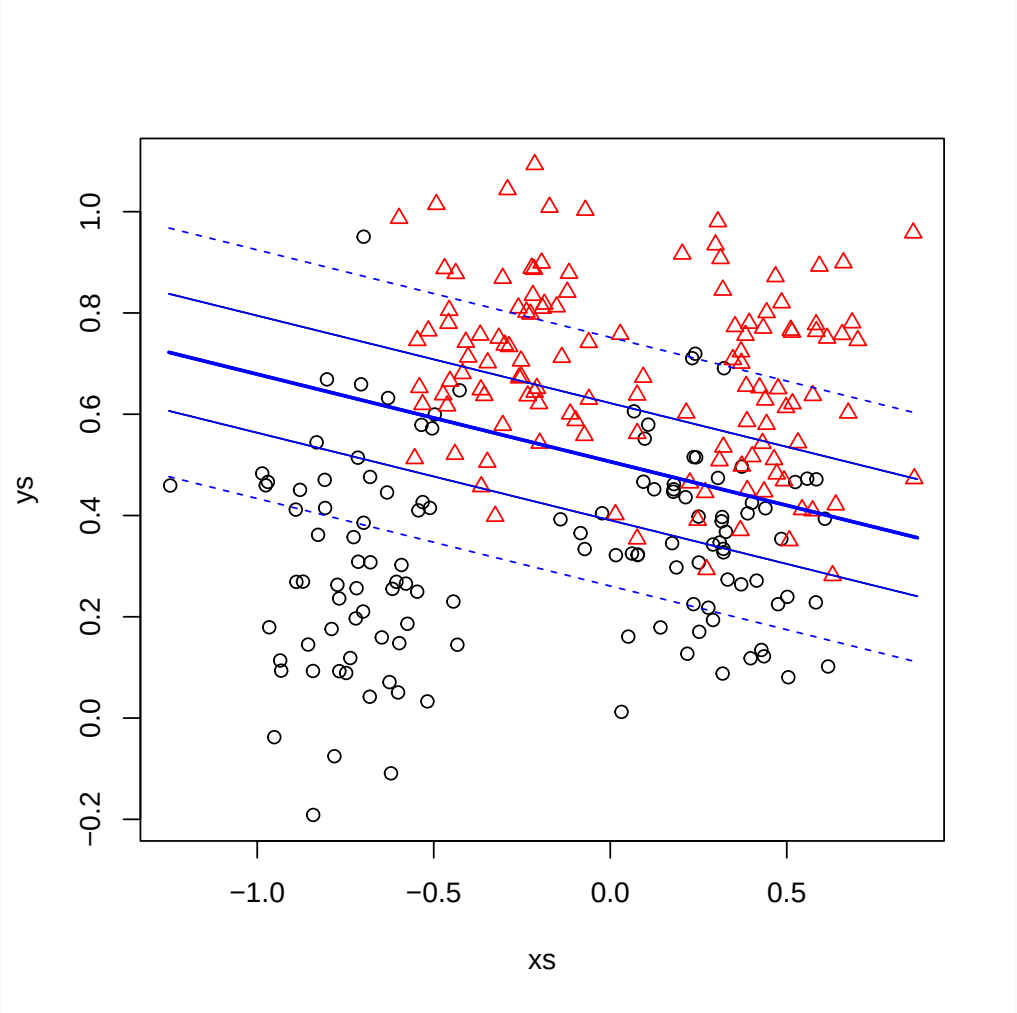
Logistic regression

Logistic regression finds a decision boundary using a linear model.



Logistic regression

Logistic regression models the probability of a new data point to be classified in one of the two classes.



Logistic regression

Consider a dataset $\{x^{(i)}, y^{(i)}\}$ where $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \{0, 1\}$.

- Linear model on the log-odds :

$$\log\left(\frac{\mathbb{P}(y=1|x)}{\mathbb{P}(y=0|x)}\right) = x^T\beta$$

  with $\mathbb{P}(y=1|x) = 1 - \mathbb{P}(y=0|x)$

- Rewriting the terms we have:

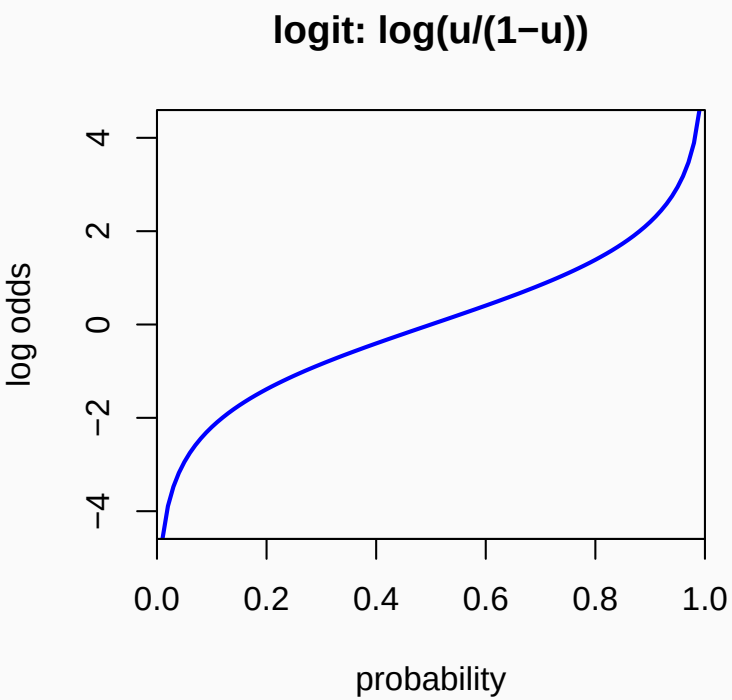
$$\mathbb{P}(y=1) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$$

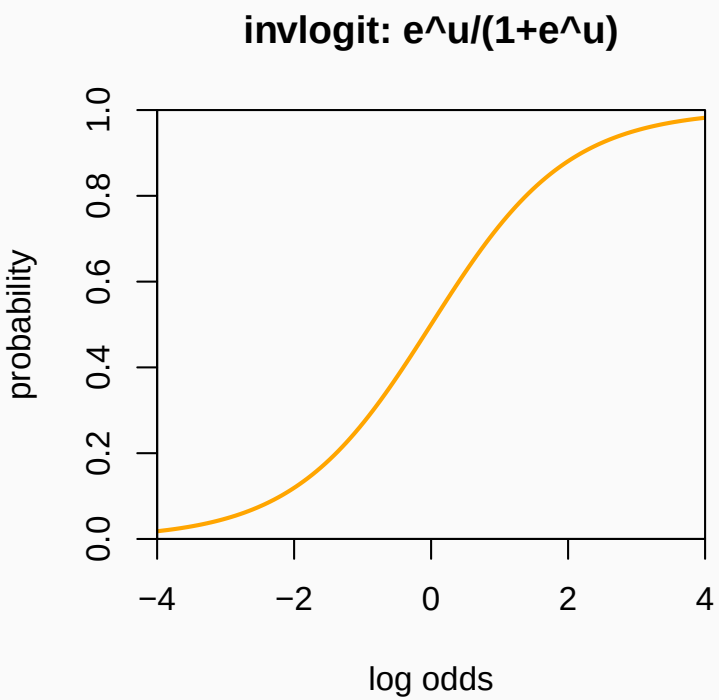
- Define logit and inverse logit:

$$\sigma(u) = \log\left(\frac{u}{1-u}\right) \quad \text{and} \quad \sigma^{-1}(u) = \frac{e^u}{1+e^u}$$

  - $\sigma(u)$ goes from probability $[0, 1]$ to $\mathbb{R}$.
  - $\sigma^{-1}(u)$ goes from $\mathbb{R}$ to probability $[0, 1]$.

Logistic regression





Inferring model parameters

Recall that in Machine Learning, parameters are inferred by minimising a loss function on the training set $\{x^{(i)}, y^{(i)}\}$.

Let denote $\hat{y}^{(i)}$ the output of the classifier for input $x^{(i)}$.

Examples of loss function for classification task:

- Misclassification loss: assuming $\hat{y}^{(i)}$ is binary

$$I(y^{(i)} \neq \hat{y}^{(i)})$$

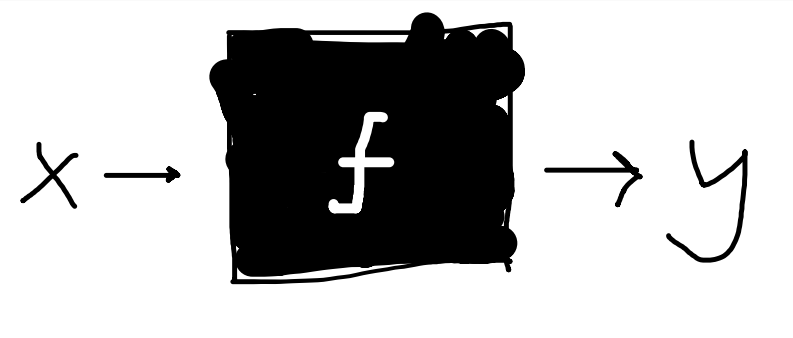
- Log loss: assuming $\hat{y}^{(i)}$ is a probability

$$-y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

  (Negative log likelihood for a Bernoulli.)
- Possible to add a regularisation such as ridge, lasso, elasticnet...

Parameters of the logistic regression model are learnt minimising the log loss on a training dataset.

Optimization: where does it fit?



Many popular ML methods:

- Specify a model for f: parameters β and hyperparameters θ
- Specify a loss function $\mathcal{L}$
- Use optimization to find the best choice β^* to minimize $\mathcal{L}$ on the training data
- Use crossvalidation to find the best choice θ to minimize $\mathcal{L}$ on the testing data

Optimization methods

In some specific cases, we can compute the minimiser of the loss function analytically (see linear model and mean square error) but for most models, one needs to use optimisation methods such as:

- One-dimensional methods: `optimize`
- Multidimensional methods: `optim`
- Using a gradient (first derivative): conjugate gradient, gradient descent, stochastic gradient descent
- Using a Hessian (second derivative): Newton, quasi-Newton, BFGS (also in `optim`)
- Global optimization methods: many methods to choose from; Bayesian optimization

Be careful of possible local optima.

Gradient descent

Iteratively follow the gradient (direction of greatest change) of the loss surface, taking steps of size *epsilon* (learning rate):

$$\beta^t = \beta^{t-1} - \epsilon \nabla_\beta \mathcal{L}(\beta^{t-1})$$

Some optimisation demos:

- Gradient descent
  https://github.com/lilipads/gradient_descent_viz
  https://jermwatt.github.io/machine_learning_refined/notes/4_Second_order_methods/4_4_Newtons.html
- Bayesian optimization
  https://github.com/fmfn/BayesianOptimization/blob/master/examples/visualization.ipynb

Logistic regression in R

Parameters of the logistic regression model are learnt minimising the log loss on a training dataset – convex function so no local optima.

The `glm` function in `R` can be used to fit generalised linear models – which include logistic regression.

For logistic regression the function is the following:

```
fit <- glm(y~.,data=dataset,family="binomial")
```

The `predict` function can then be used to obtain the probability that a new data point belongs to each class, i.e. $\mathbb{P}(y = 1|x) = 1 - \mathbb{P}(y = 0|x)$.

Based on this probability, one can classify the new point using a threshold. For example: if $\mathbb{P}(y = 1|x) \geq 0.5$, classify the new observation in class 1, otherwise in class 0.

Performance measures

Confusion matrices are often used to evaluate performances of a binary classifier on a test set:

	True state 0	True state 1
Predicted 0	True negative	False negative
Predicted 1	False positive	True positive

The following performance measures are derived from it:

- Accuracy: $(TP + TN)/(TP + TN + FP + FN)$.
- Error rate: $(FP + FN)/(TP + TN + FP + FN)$.
- Sensitivity/Recall (true positive rate): $TP/(TP + FN)$.
- Specificity (true negative rate): $TN/(TN + FP)$.
- False positive rate (1-Specificity): $FP/(TN + FP)$.
- Precision: $TP/(TP + FP)$.

ROC curve

As we vary the prediction threshold $p(y = 0|x) = c$ from 0 to 1:

- Specificity varies from 0 to 1
- Sensitivity goes from 1 to 0

Received Operating Characteristic (ROC) curves plot the true positive rate (or sensitivity) vs the false positive rate (1-specificity)

It is a common practice to compare the area under the curve (AUC) of the ROC curves produced by different classifiers.

See demo `logisticRegression.R`

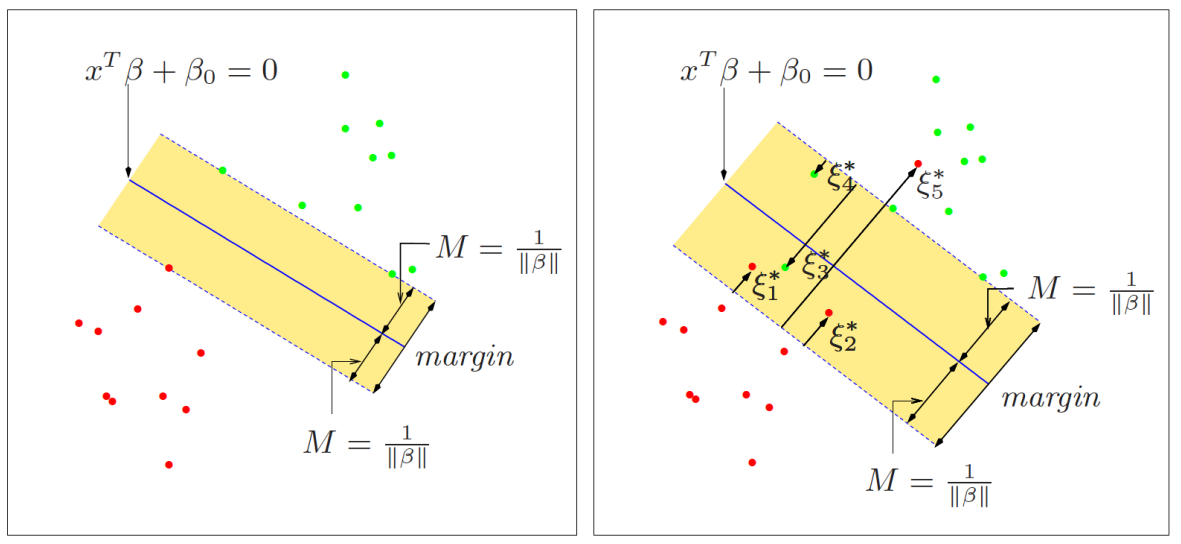
Summary

- Parametric models such as logistic regression can be used for classification.
- They typically consist in identifying a decision boundary between classes.
- As for any supervised learning approach, the parameters of the classifier are inferred by minimising a loss function.
- Optimisation methods can be used if the loss function can not be minimised analytically.
- Classifiers can be evaluated based on their performance on a test set comparing accuracy or based on the ROC curves.
- Care needs to be taken when dealing with unbalanced dataset.

Other classification algorithms

There exists many other algorithms for classification.

- Linear Discriminant Analysis
- Naive Bayes classifier
- A support vector classifier aims at identifying an hyperplane that creates the biggest margin between training points of different classes.



[Figure from Hastie, Tibshirani and Friedman's book]

Other classification algorithms

There exists many other algorithms for classification.

- Linear Discriminant Analysis
- Naive Bayes classifier
- A support vector classifier aims at identifying an hyperplane that creates the biggest margin between training points of different classes.
- As with other linear methods, it is possible to make the classification procedure more flexible by enlarging the feature space using basis expansion (such as polynomials or splines). The support vector machine classifier allows to consider features in a very large – or even infinite – space using the so-called kernel trick.
- Neural networks and methods based on decision trees can be used for both regression and classification. We will discuss these later.