Machine Learning for Exploration Geophysics

Th4: Classification Algorithms

10. - 12. March 2020

Hamburg

Outline

- Logistic Regression
- K-Nearest Neighbors (K-NN)
- Support Vector Machine (SVM)
- Naive Bayes
- Decision Tree Classification
- Random Forest Classification
- XGBoost Classification

Classification

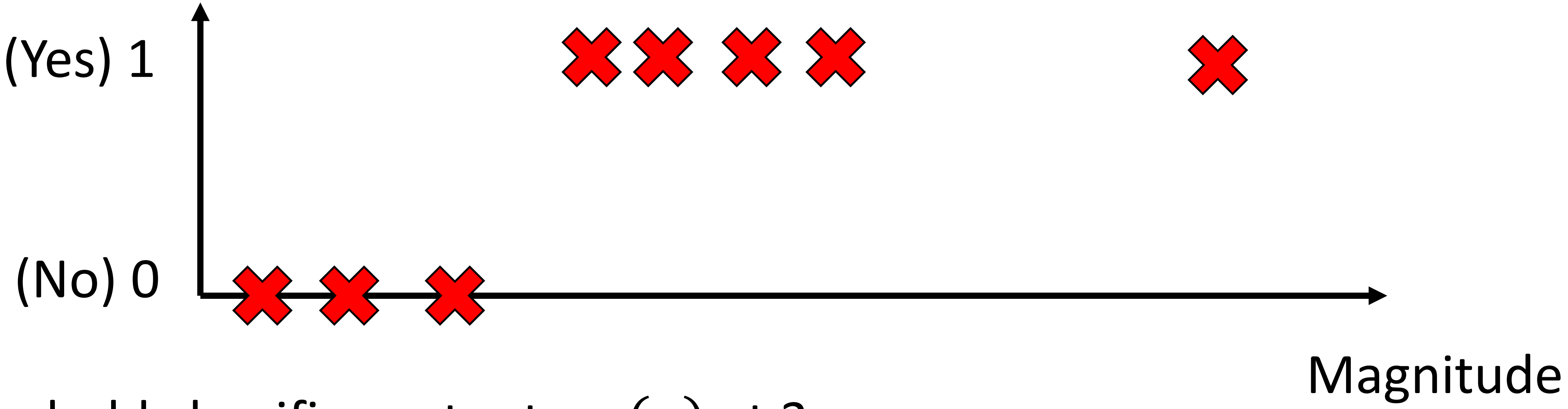
- Microseismic & Seismology: event/noise?
- Seismic Imaging: Diffraction body (yes/no; if yes – what type?)
- Seismic Facies Classification: sand/shale/cemented sand?

$y = \{0, 1\}$

0: “Negative Class” (e.g., noise)
1: “Positive Class” (e.g., event)

Classification

natural tectonic
earthquake



Threshold classifier output $y_m(x)$ at 3:

- If $y_w(x) \geq 3$ predict "y=1"
- If $y_w(x) < 3$ predict "y=0"

Classification

Regression:

$$y \in R$$

$$y_w(\boldsymbol{x}) \in R$$

Classification:

$$y \in [0, 1]$$

$$0 \leq y_w(\boldsymbol{x}) \leq 1$$

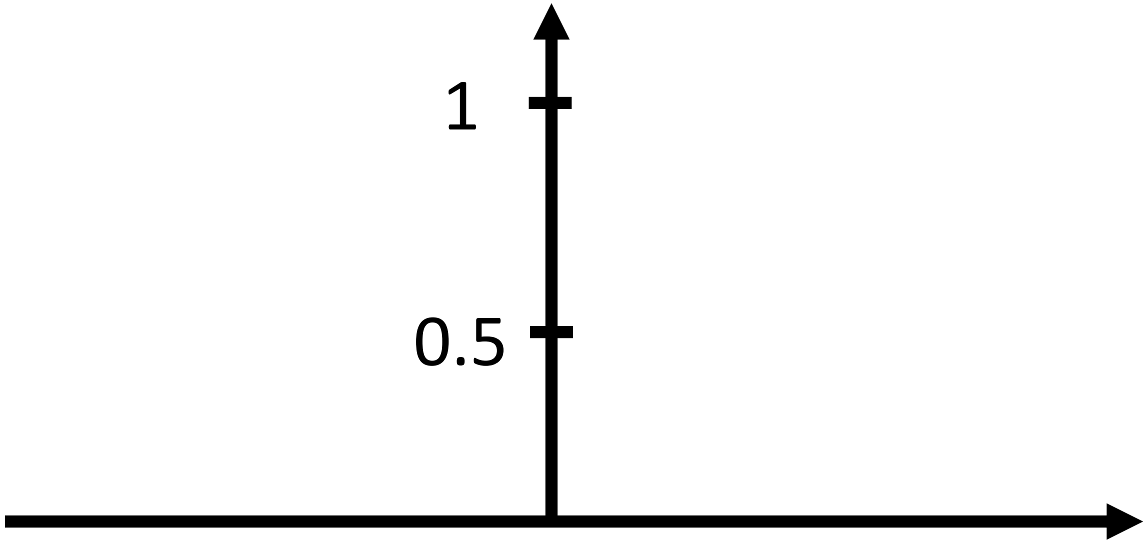
Logistic Regression

We want:

$$0 \le y_w(\boldsymbol{x}) \le 1$$

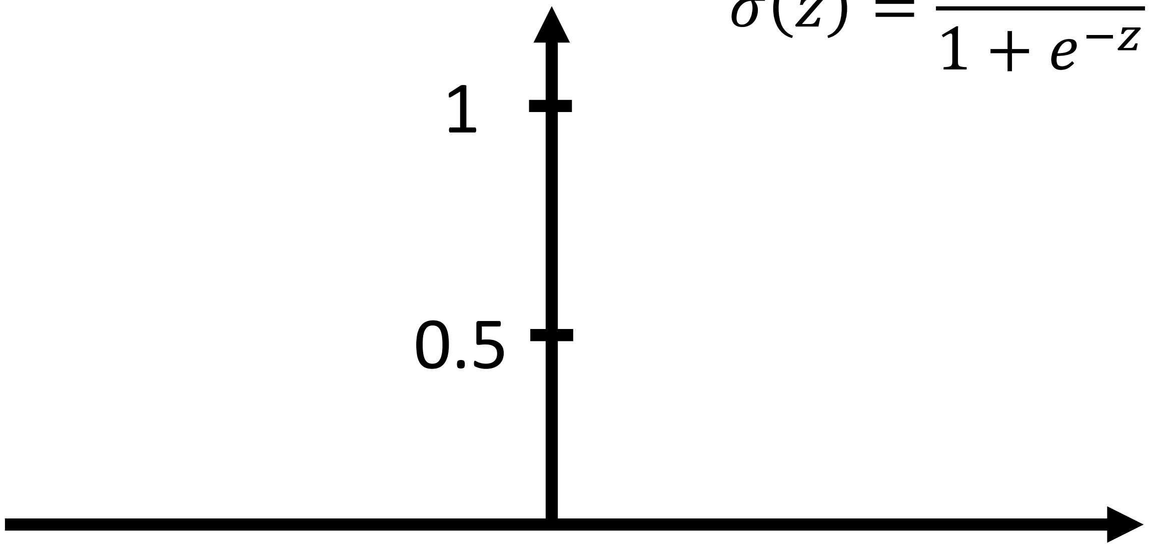
$$y_w(x) = \quad b + w_1 x + w_2 x_2 + \cdots$$

Sigmoid function
Logistic function



Logistic Regression

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



We want:

$$0 \leq y_w(\boldsymbol{x}) \leq 1$$

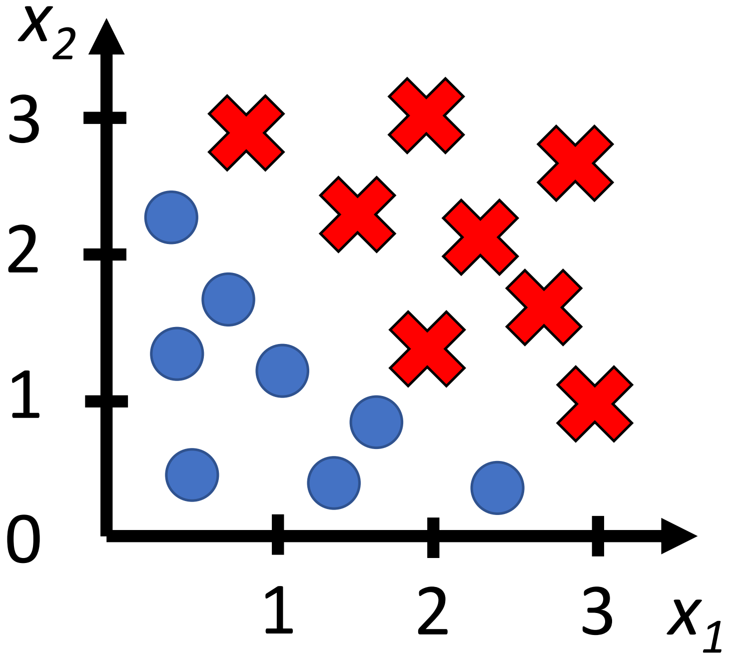
$$y_w(x) = \sigma[b + w_1 x + w_2 x_2 + \cdots]$$

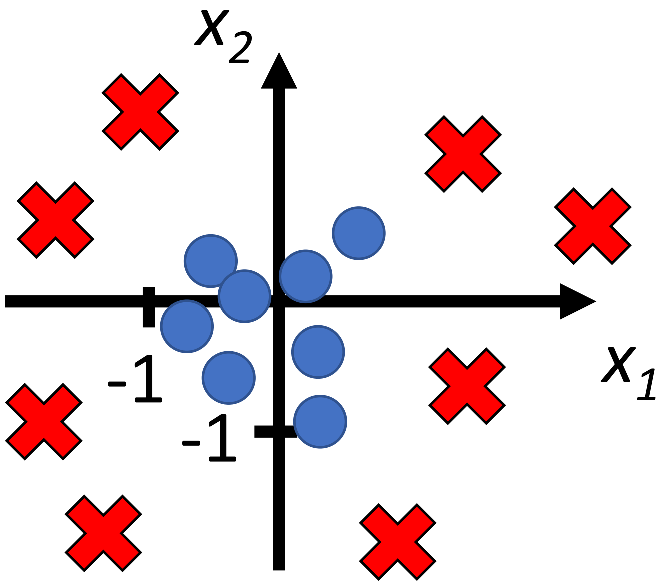
Predict "y=1" if $y_w(x) \geq 0.5$

Predict "y=0" if $y_w(x) < 0.5$

Sigmoid function
Logistic function

Linear & Non-linear Decision Boundaries





$$y_w(x) = \sigma(b + w_1 x_1 + w_2 x_2)$$

$$y_w(x) = \sigma\begin{pmatrix} b + w_1 x_1 + w_2 x_2 + \\ w_3 x_1^2 + w_4 x_1 x_2 + w_5 x_2^2 \end{pmatrix}$$

Training data: $\left\{ \left(x^{(i)}, y^{(i)} \in \{0,1\}\right),\ i = 1, \dots N \right\}$

$$\boldsymbol{x}^{(i)} = \begin{bmatrix} 1 \\ x_1 \\ \dots \\ x_m \end{bmatrix},\ \boldsymbol{w} = \begin{bmatrix} b \\ w_1 \\ \dots \\ w_m \end{bmatrix}$$

$$y_w(x) = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}}}$$

How to find parameters $\boldsymbol{w}$?

Linear regression loss function

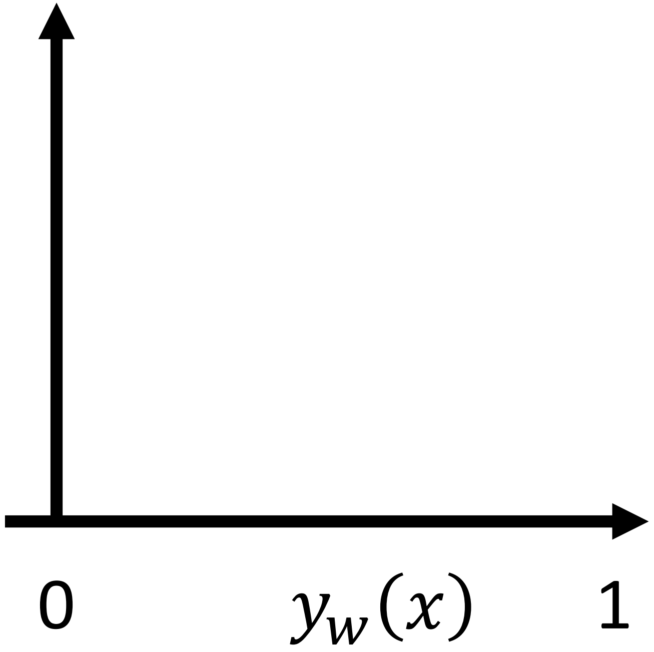
$$J(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(y_w\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$\boldsymbol{w} = \operatorname{argmin} J(\boldsymbol{w})$$

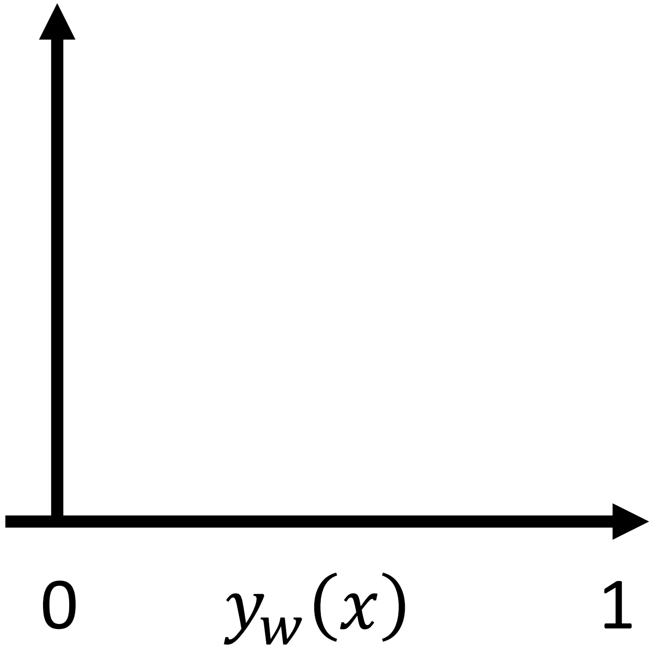
Logistic regression cost function

$$J(\boldsymbol{w}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}\ln y_w\left(x^{(i)}\right) + \left(1 - y^{(i)}\right)\ln\left(1 - y_w\left(x^{(i)}\right)\right)\right]$$

If $y = 1$



If $y = 0$



Logistic regression cost function

$$J(\boldsymbol{w}) = -\frac{1}{N}\sum_{i=1}^{N} \left[y^{(i)} \ln y_w\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \ln\left(1 - y_w\left(x^{(i)}\right)\right)\right]$$

$$\boldsymbol{w} = \operatorname{argmin} J(\boldsymbol{w})$$

Logistic regression with Regularization

$$J(\boldsymbol{w}) = -\frac{1}{N}\sum_{i=1}^{N} \left[y^{(i)} \ln y_w\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \ln\left(1 - y_w\left(x^{(i)}\right)\right)\right] + \frac{\lambda}{2}\sum_{i=1}^{M} w_i^2$$

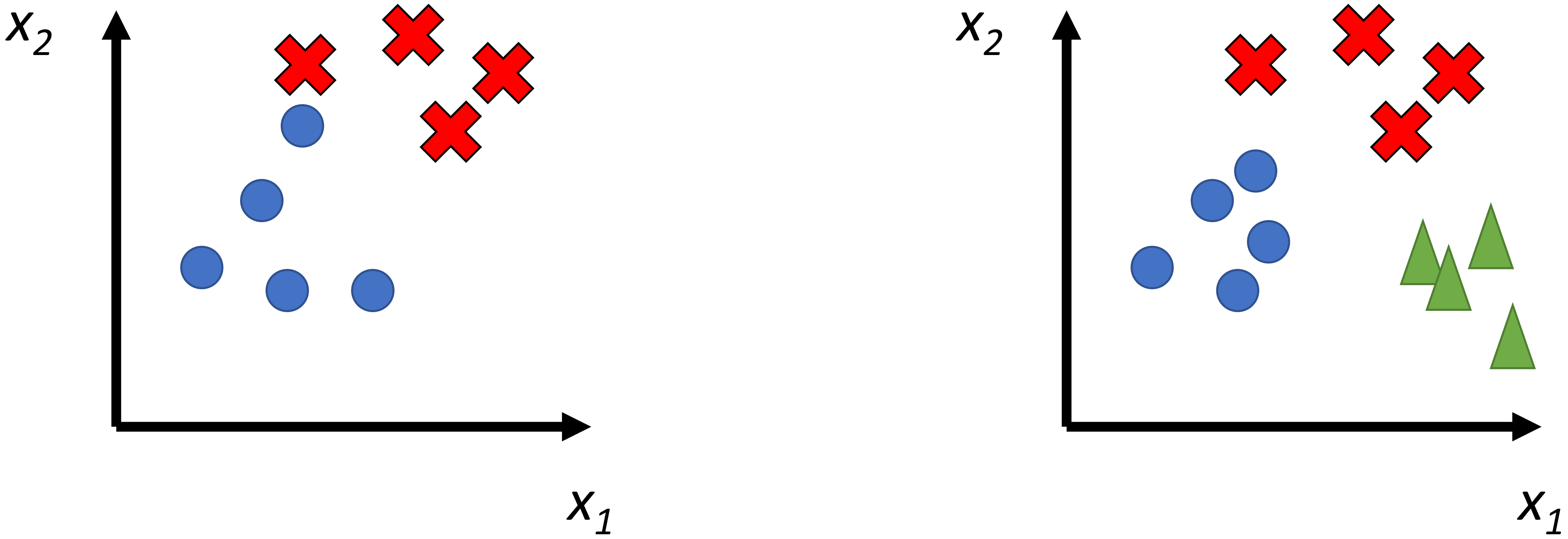
Multiclass classification

Seismic Facies Classification: sand/shale/cemented sand

Event detection: Reflection, Edge diffraction, Point diffraction, Noise

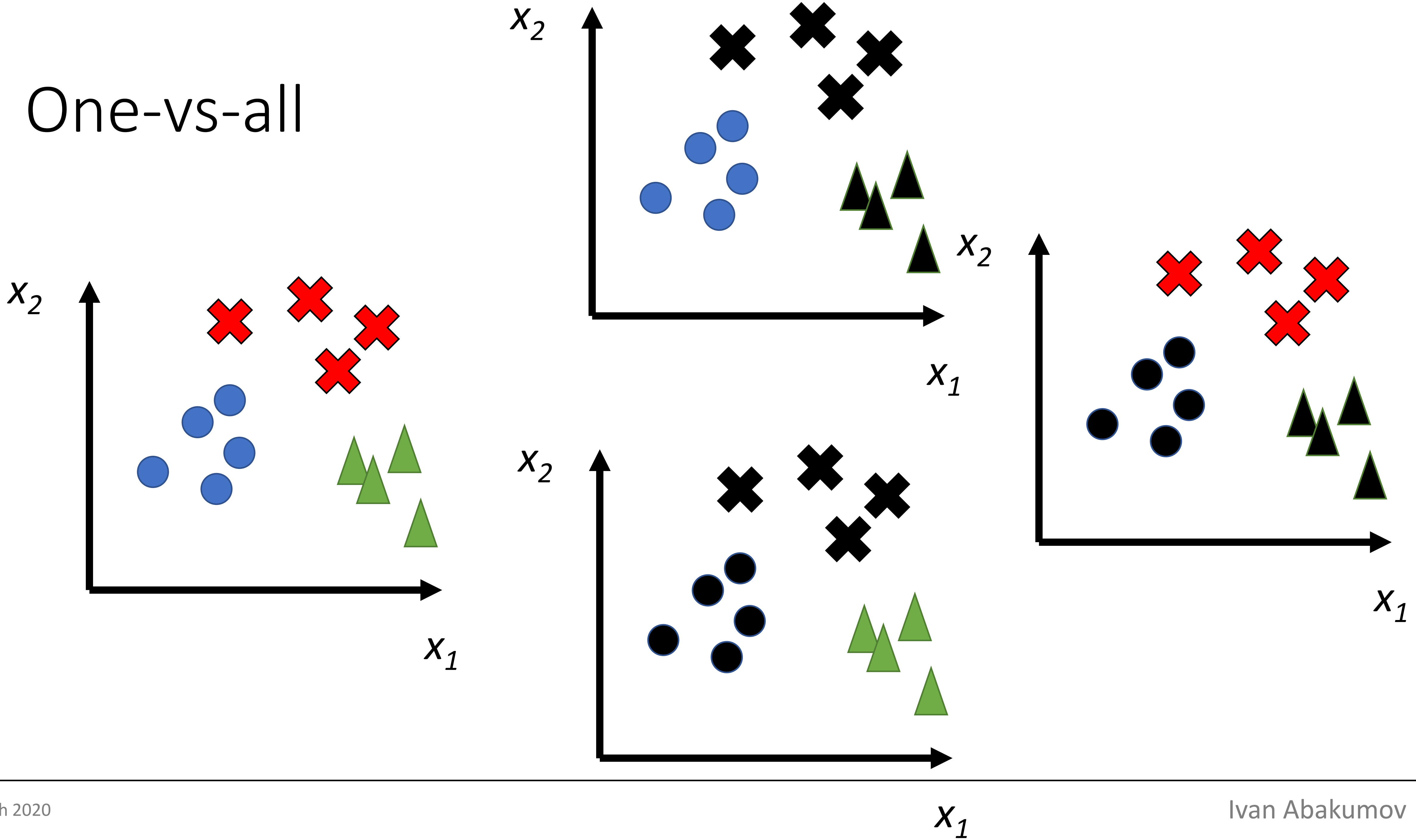
Number of bedrooms: 1, 2, 3…

Multiclass classification



Ivan Abakumov

One-vs-all

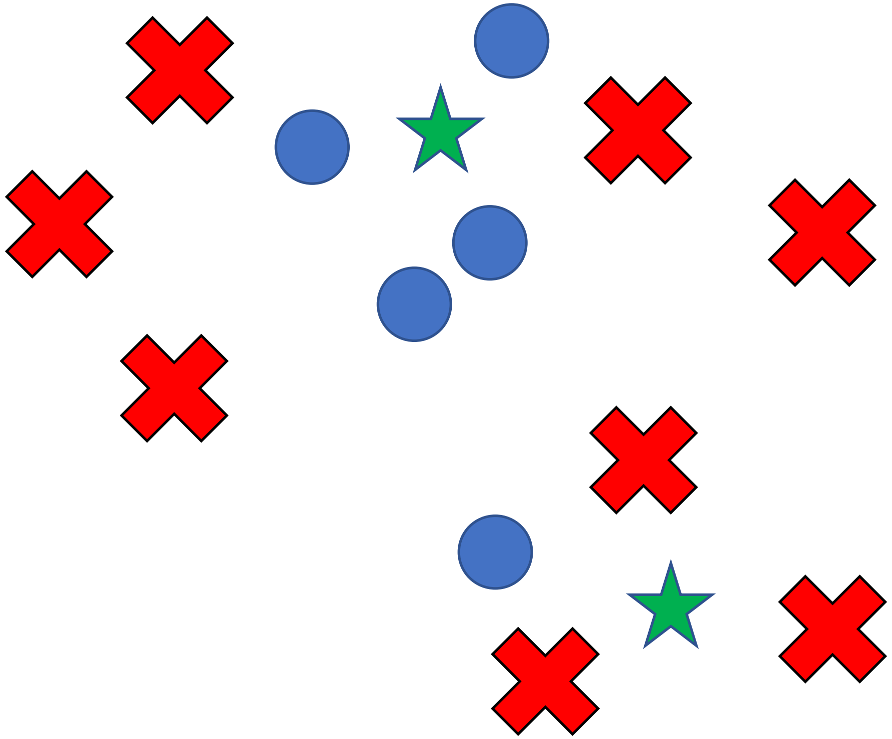


Outline

- Logistic Regression
- K-Nearest Neighbors (K-NN)
- Support Vector Machine (SVM)
- Naive Bayes
- Decision Tree Classification
- Random Forest Classification
- XGBoost Classification

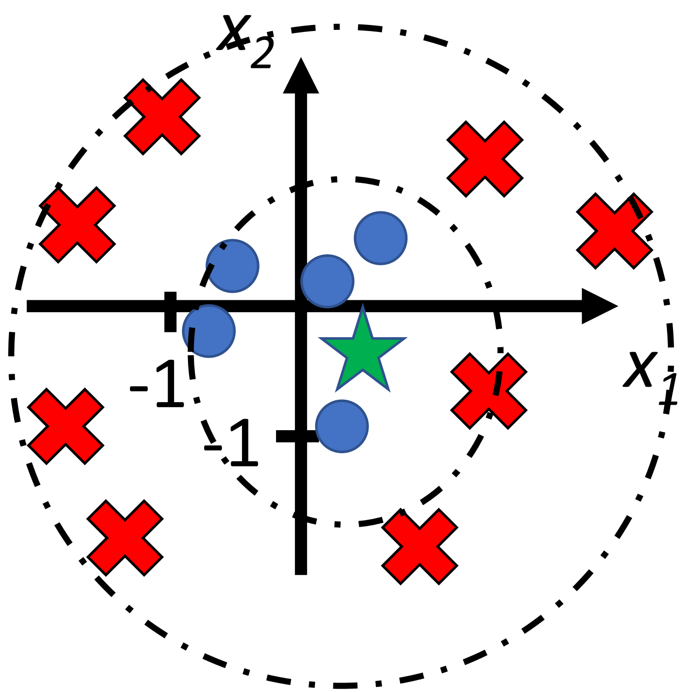
KNN Classification approach

- An object is classified by a majority votes for its neighbor classes
- The object is assigned to the most common class among its K nearest neighbors



How to choose K?

- If K is too small the algorithm is sensitive to noise points
- Larger K works well. But too large K may include majority points from other classes
- Rule of thumb is $K < \sqrt{n}$, n is number of examples



Strengths and weakness of KNN

- Strengths of KNN
  - Very simple and intuitive
  - Can be applied to the data from any distribution
  - Good classification if the number of samples is large enough

- Weakness of KNN
  - Takes more time to classify a new example
  - need to calculate and compare the distance from new example to all other examples
  - Choosing k may be tricky
  - Need a large number of samples for accuracy

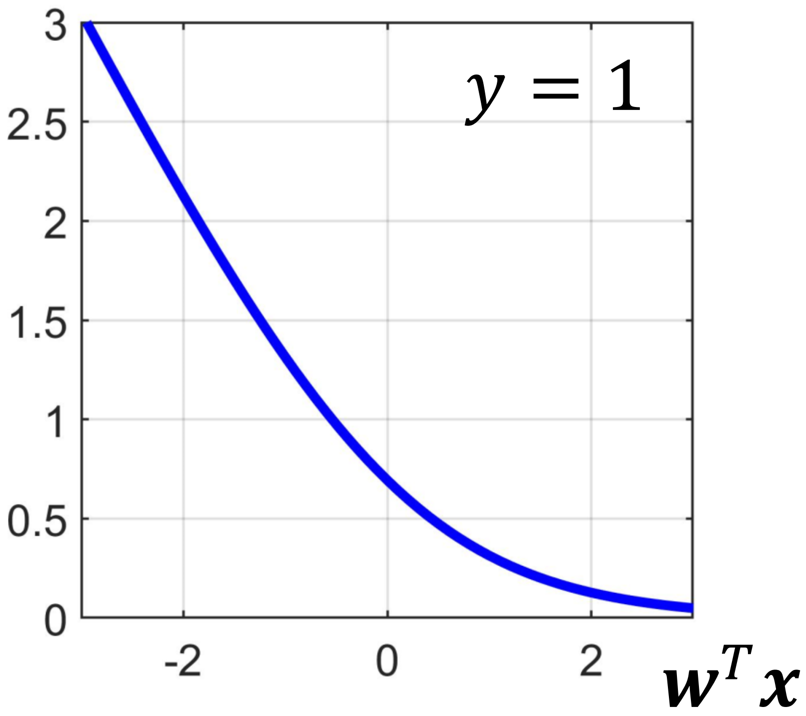
Outline

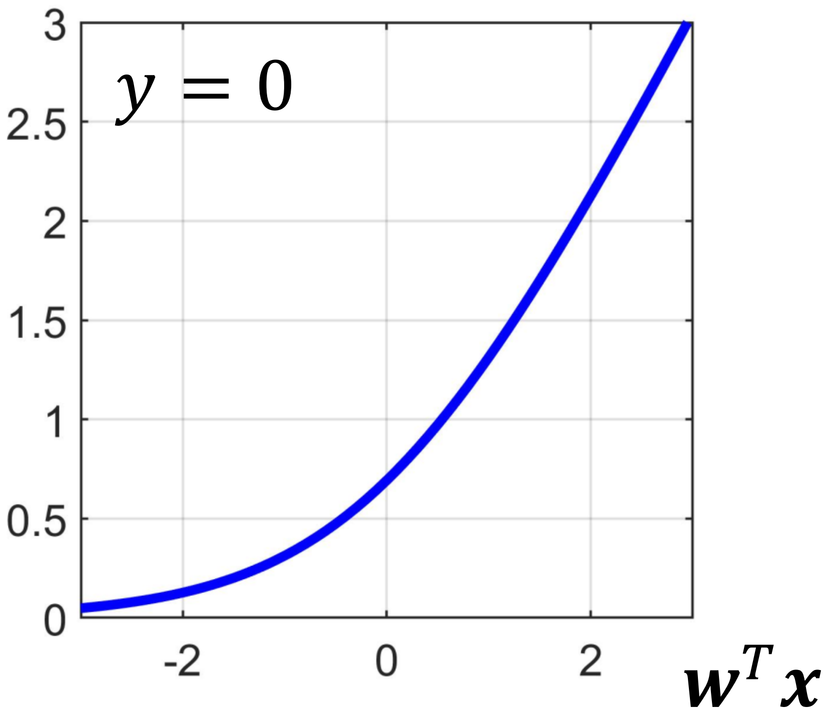
- Logistic Regression
- K-Nearest Neighbors (K-NN)
- Support Vector Machine (SVM)
- Naive Bayes
- Decision Tree Classification
- Random Forest Classification
- XGBoost Classification

Logistic regression cost function

$$J(\boldsymbol{w}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)} \ln y_w\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \ln\left(1 - y_w\left(x^{(i)}\right)\right)\right]$$

$$y_w(x) = \frac{1}{1 + e^{-\boldsymbol{w}^T\boldsymbol{x}}}$$

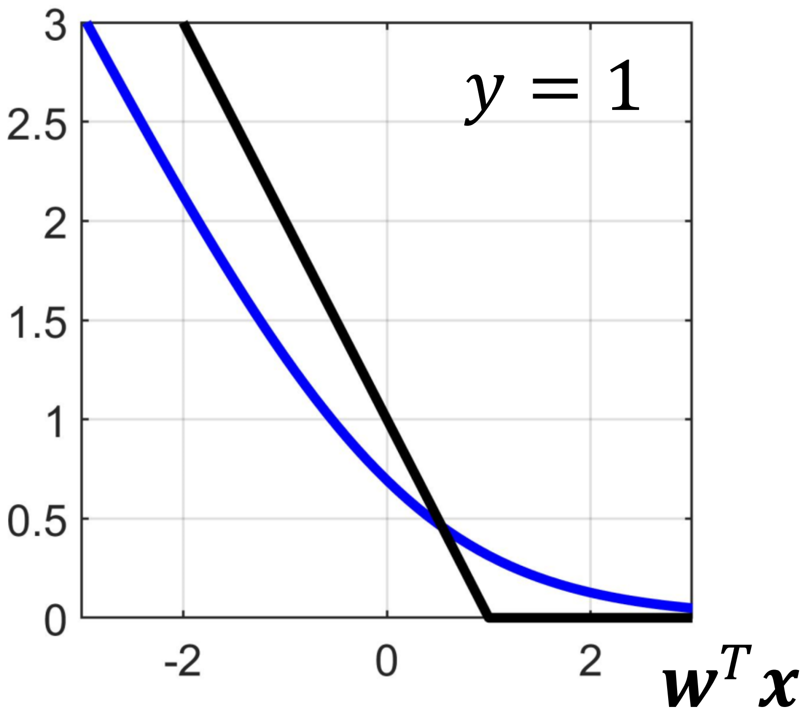


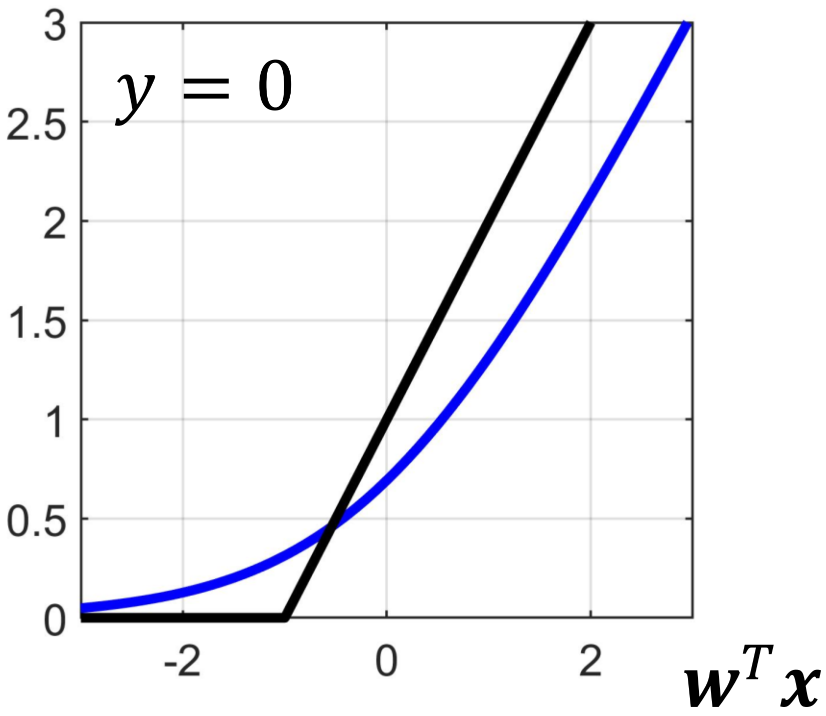


Logistic regression cost function

$$J(\boldsymbol{w}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)} \ln y_w\left(x^{(i)}\right) + \left(1 - y^{(i)}\right)\ln\left(1 - y_w\left(x^{(i)}\right)\right)\right]$$

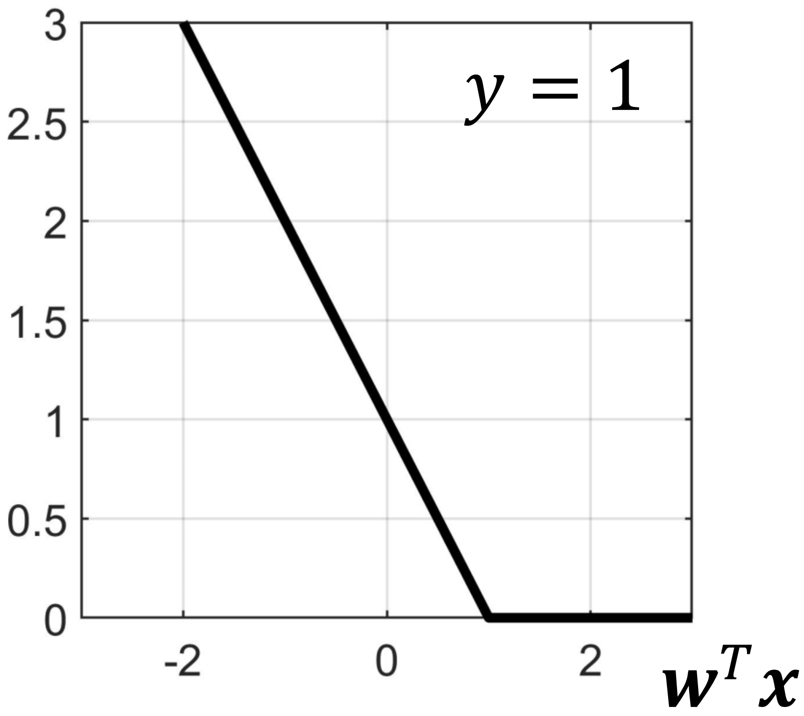
$$y_w(x) = \frac{1}{1 + e^{-\boldsymbol{w}^T\boldsymbol{x}}}$$

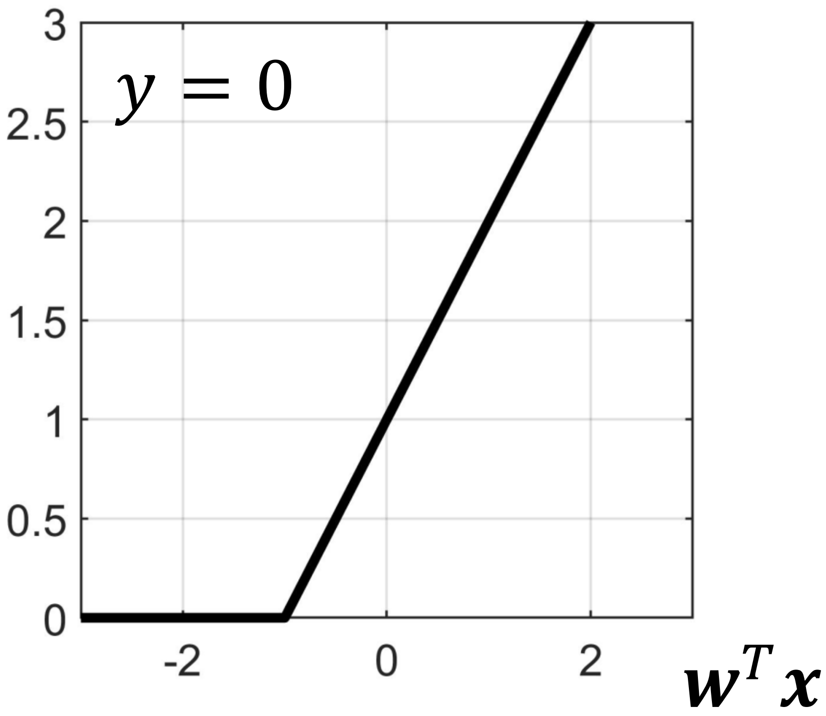




SVM cost function

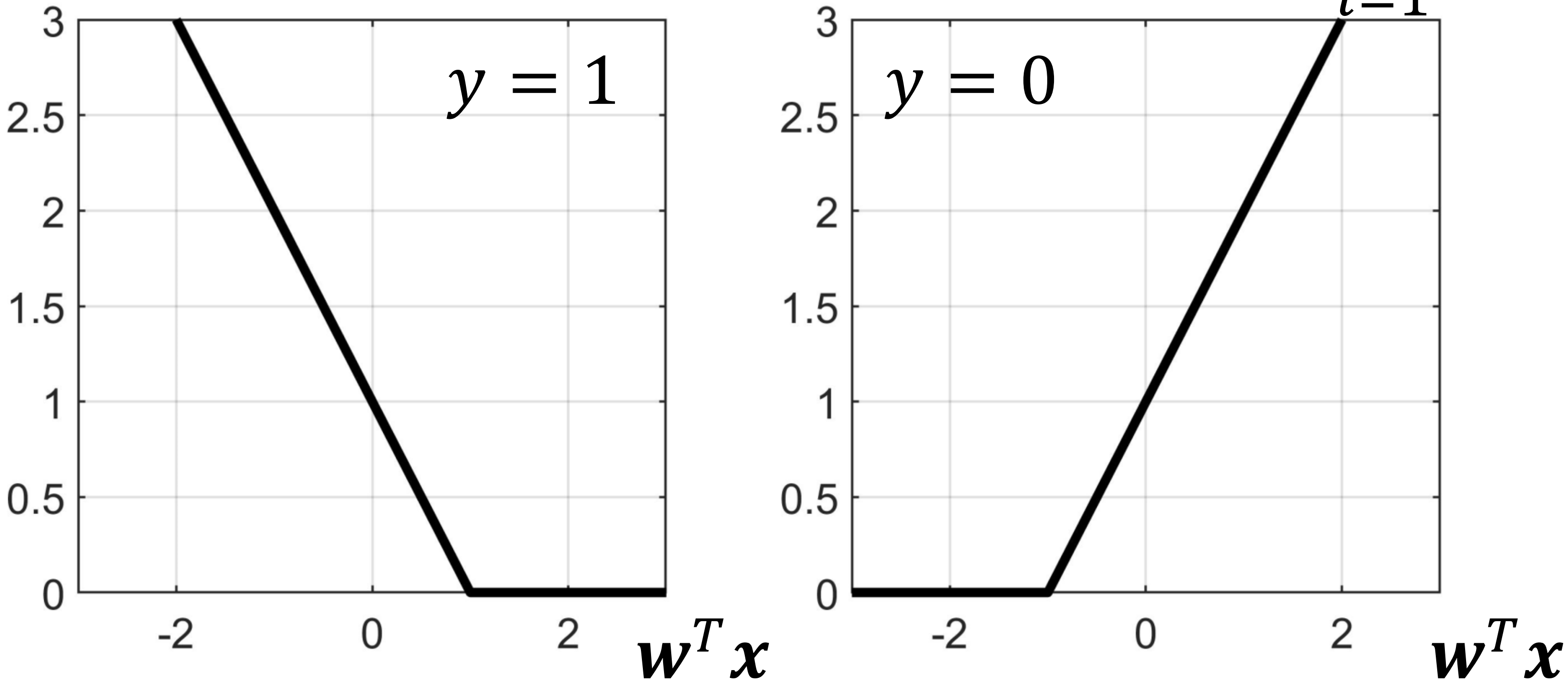
$$J(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}\, cost_1(w^T x) + \left(1 - y^{(i)}\right) cost_0(w^T x)\right]$$





SVM cost function

$$J(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}\, cost_1(w^T x) + \left(1 - y^{(i)}\right) cost_0(w^T x)\right] + \frac{\lambda}{2}\sum_{i=1}^{M} w_i^2$$



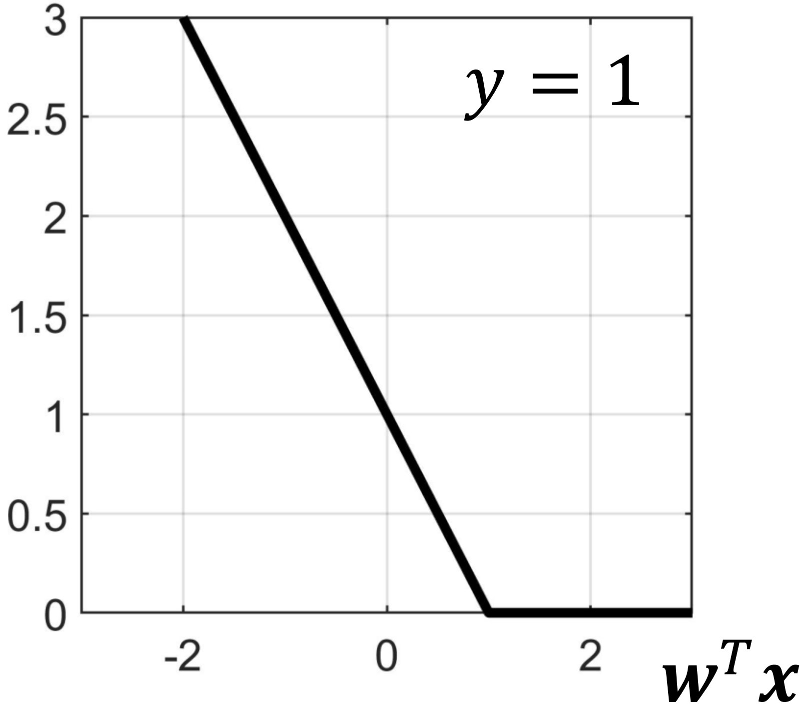
Ivan Abakumov

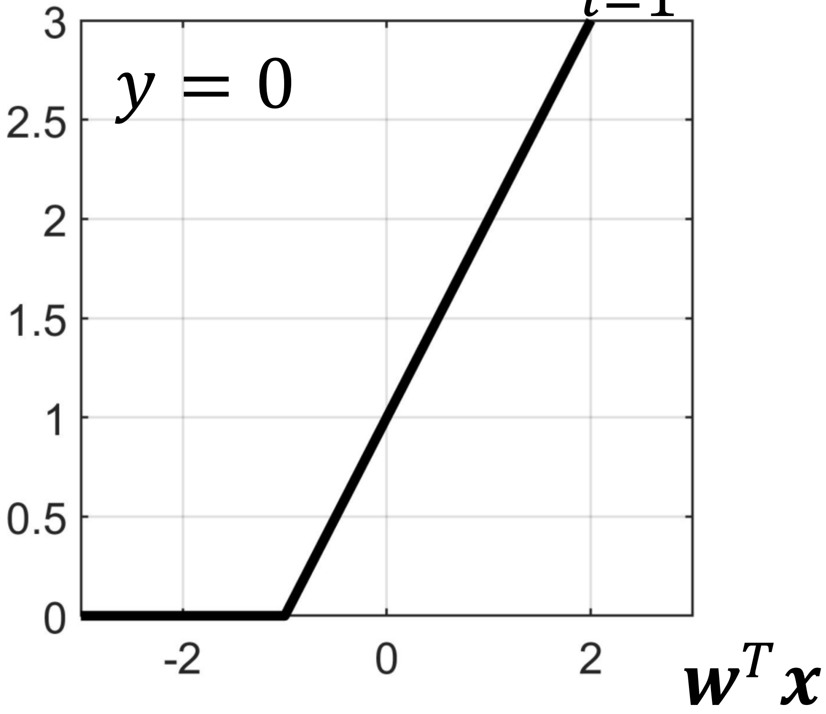
SVM cost function

$$J(\boldsymbol{w}) = C\sum_{i=1}^{N}\left[y^{(i)}\, cost_1(w^T x) + \left(1-y^{(i)}\right) cost_0(w^T x)\right] + \frac{1}{2}\sum_{i=1}^{M} w_i^2$$

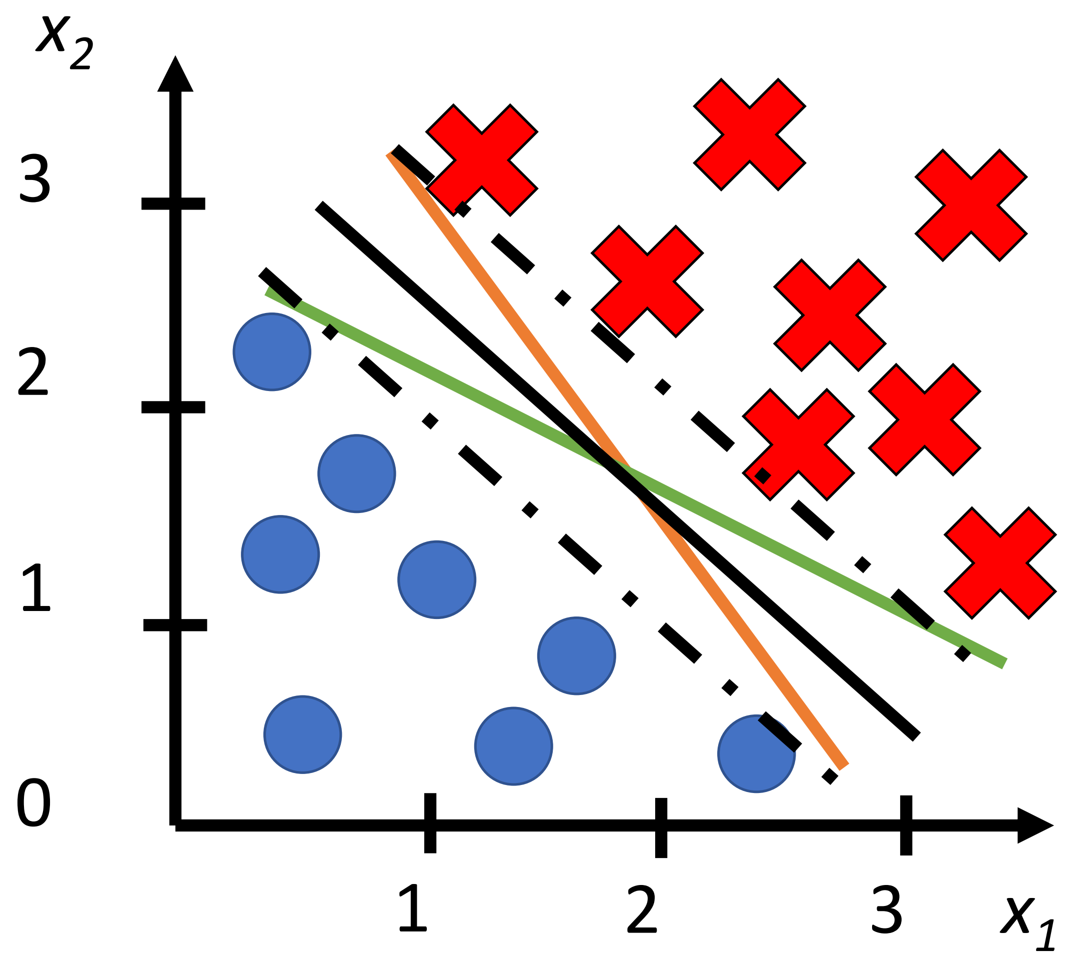
If $y = 1$, we want $\boldsymbol{w}^T\boldsymbol{x} > 1$ (not just>0)

If $y = 0$, we want $\boldsymbol{w}^T\boldsymbol{x} \leq -1$ (not just<0)

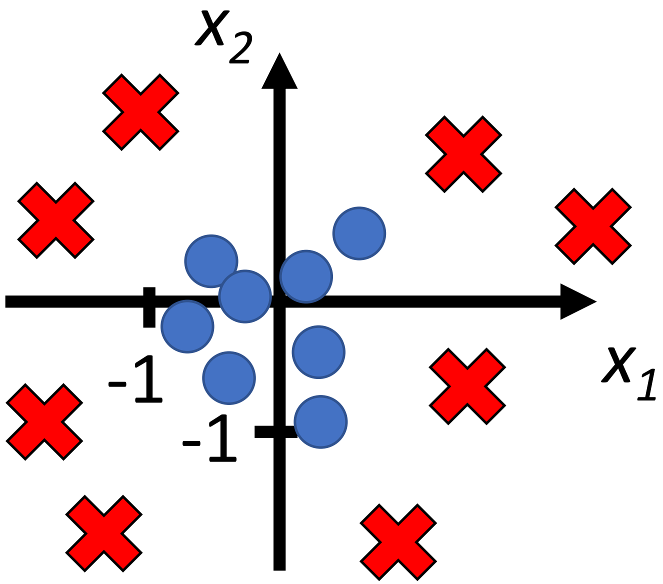




SVM Decision Boundary



Non-linear Decision Boundaries



$$y_w(x) = \sigma\left(\begin{array}{c} b + w_1 x_1 + w_2 x_2 + \\ w_3 x_1^2 + w_4 x_1 x_2 + w_5 x_2^2 + \\ \dots \end{array}\right)$$

Kernel

$$f_i(x) = \exp\left(-\frac{\left|x - x^{(i)}\right|^2}{2\sigma^2}\right)$$

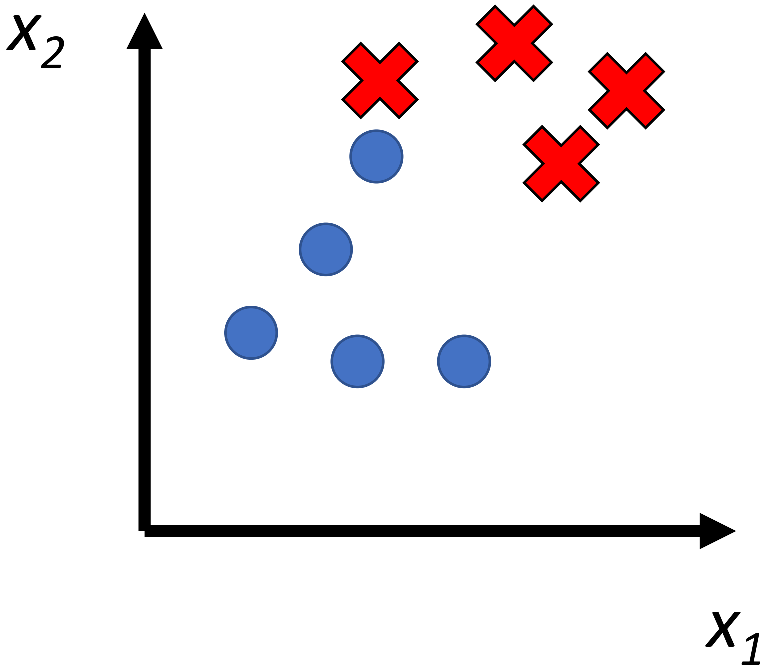
Is there a better choice of the features?

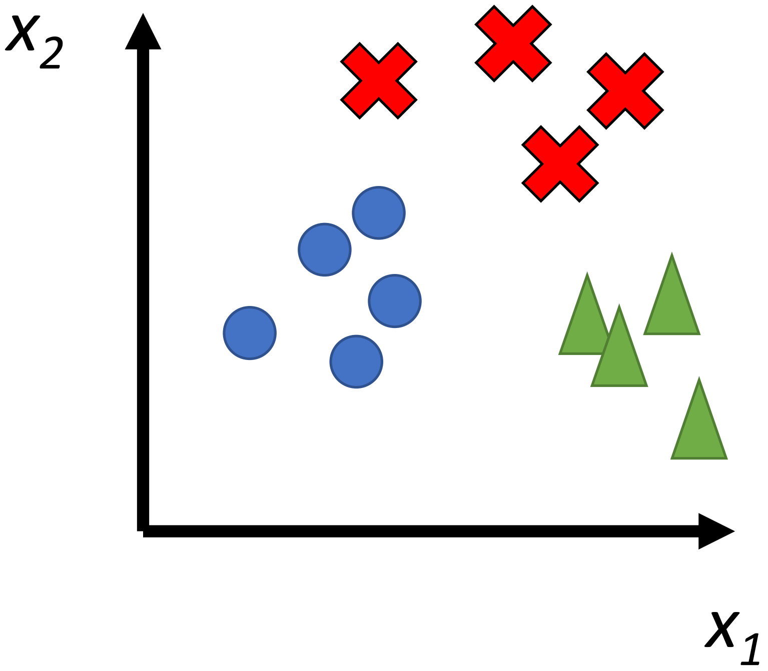
SVM hyperparameters

- Parameter C:
- Kernel
  - No kernel (“linear kernel”)
  - Gaussian kernel
    - Need to choose sigma

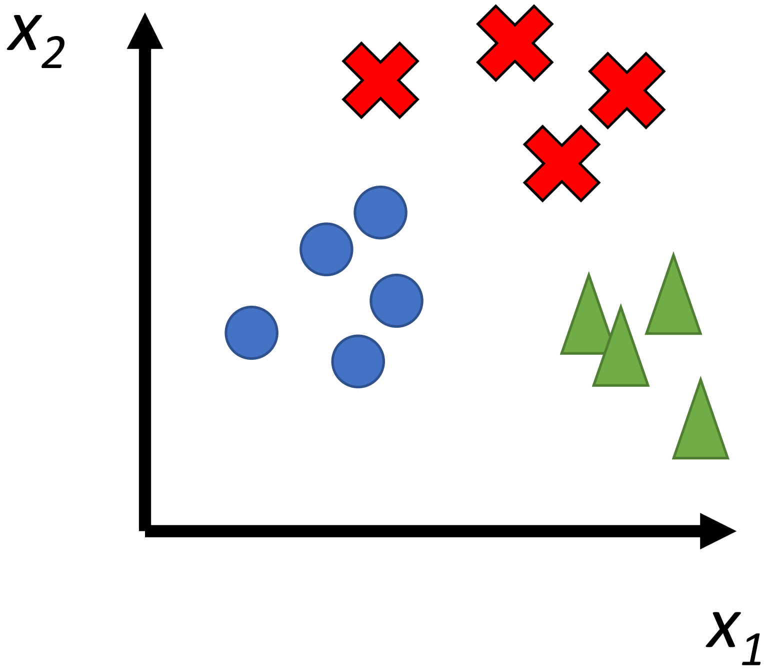
- Important: feature scaling in case of Gaussian kernel!

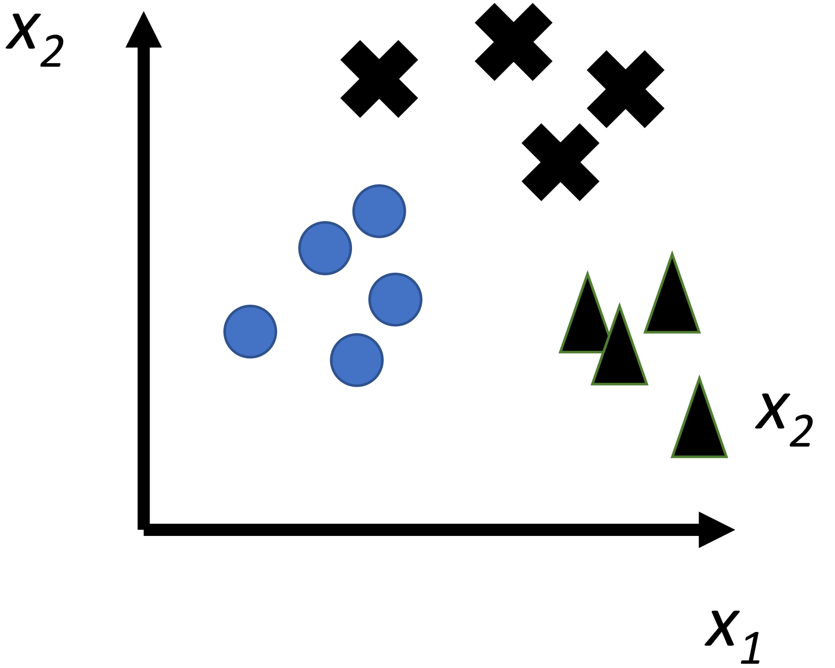
Multiclass classification

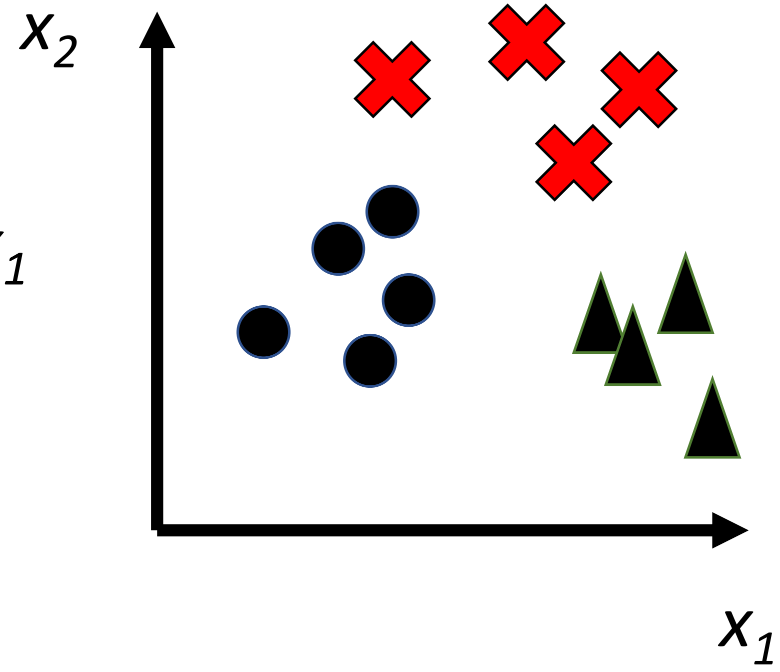


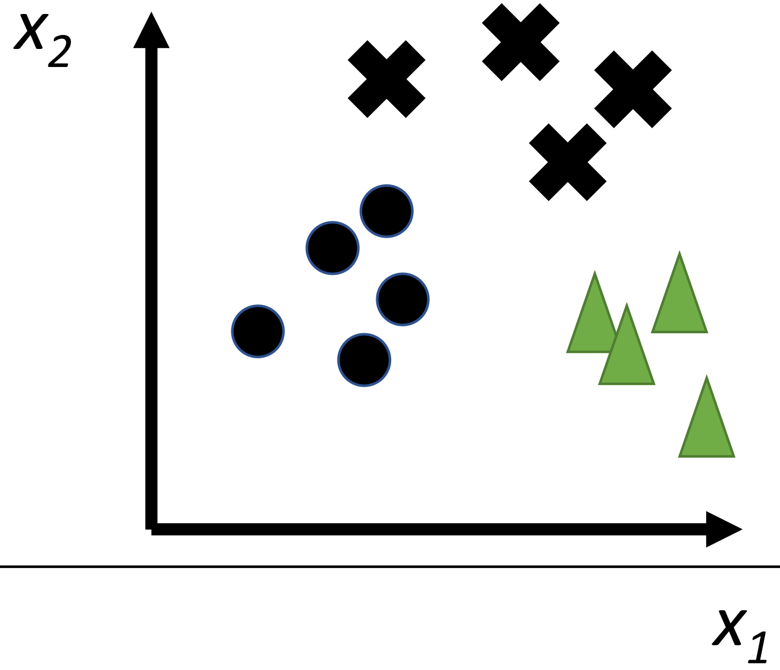


One-vs-all

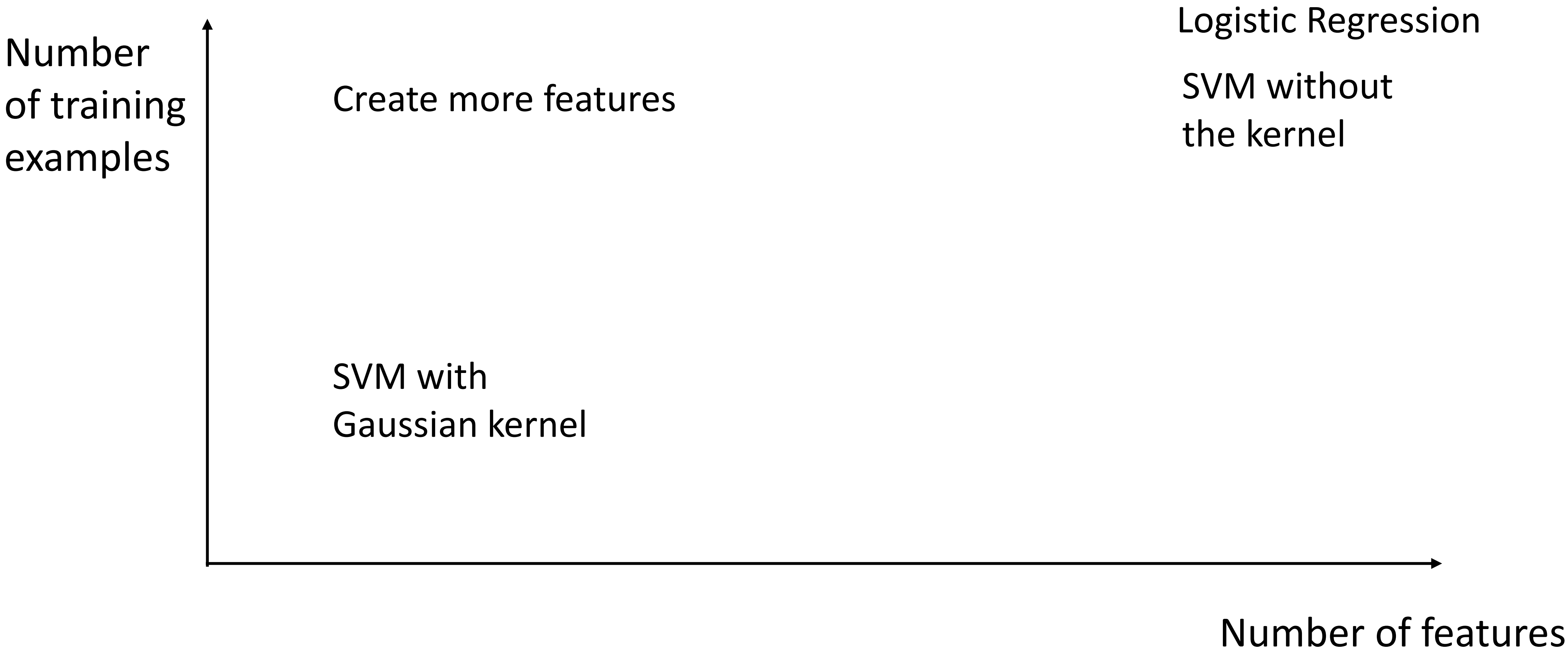








Logistic regression vs SVM



Outline

- Logistic Regression
- K-Nearest Neighbors (K-NN)
- Support Vector Machine (SVM)
- Naive Bayes
- Decision Tree Classification
- Random Forest Classification
- XGBoost Classification

Probability of event

BOX 1

What is the probability of picking a red/blue truffle?

P(red) =

P(blue) =

Probability of event

BOX 2

What is the probability of picking a red/blue truffle?

P(red) =

P(blue) =

Probability of event

BOX 3

What is the probability of picking a red/blue truffle?

P(red) =

P(blue) =

BOX 1

BOX 2

BOX 3

What is the probability of picking a truffle from box 1, 2 and 3?

$P(H_1) =$
$P(H_2) =$
$P(H_3) =$

BOX 1

BOX 2

BOX 3

I have picked a red truffle. What is the probability that I picked the truffle from box 1, 2 and 3?

$P(H_1|\text{red}) =$

$P(H_2|\text{red}) =$

$P(H_3|\text{red}) =$

Bayes Theorem

Likelihood

How probable is the evidence given that our hypothesis is true

Prior

How probable was our hypothesis before observing the evidence?

$$P(H_i|e) = \frac{P(e|H_i)P(H_i)}{P(e)}$$

Posterior

How probable is our hypothesis given the observed evidence?
(Not directly computable)

Marginal

How probable is the new evidence under all possible hypotheses:
$P(e) = \sum P(e|H_i)P(H_i)$

Bayes Theorem

Likelihood

probability for the data to be actually observed if model m is the true model

Prior

Prior probability of model

$$p(m|d) \propto p(d|m)P(m)$$

Posterior

Pobability of model

Marginal

How probable is the new evidence under all possible hypotheses:

$$P(e) = \sum P(e|H_i)P(H_i)$$