Aalto University
School of Electrical Engineering

CS-C3240 – Machine Learning D

Classification

Stephan Sigg

Department of Communications and Networking
Aalto University, School of Electrical Engineering
stephan.sigg@aalto.fi

Version 1.0, January 11, 2022

Learning goals

- Logistic Regresion
  - Logistic Loss
- Support Vector Machines
  - Hinge loss
  - Maximum margin principle
- The perceptron algorithm
- Multiclass and multilabel problems

Outline

Recap: linear regression

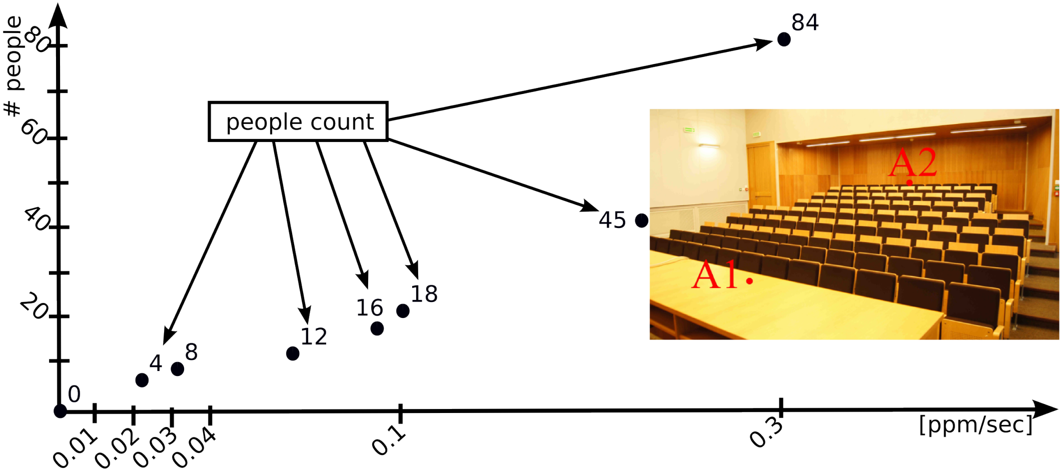
Logistic regression

Support Vector Machines

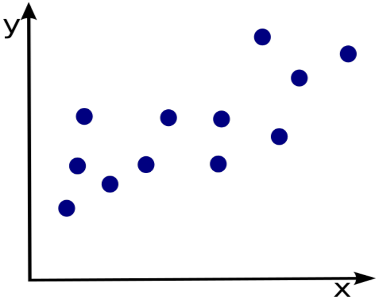
The Perceptron algorithm

Multiclass classification

Recap: linear regression

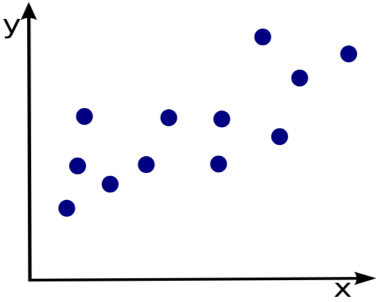


Recap: linear regression



Hypothesis: $h(x) = w_0 + w_1 x$

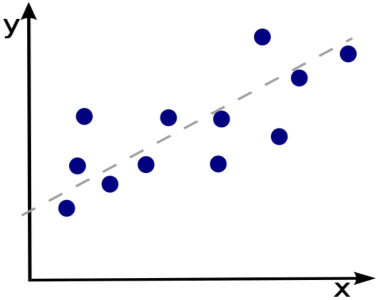
Recap: linear regression



What do we try to find with linear regression?

Hypothesis: $h(x) = w_0 + w_1 x$

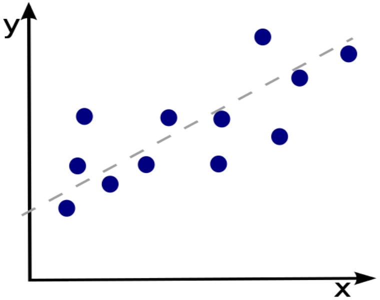
Recap: linear regression



Hypothesis: $h(x) = w_0 + w_1 x$

What do we try to find with linear regression?

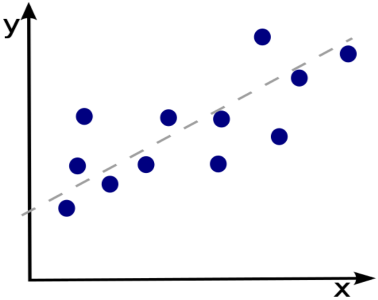
Recap: linear regression



Hypothesis: $h(x) = w_0 + w_1 x$

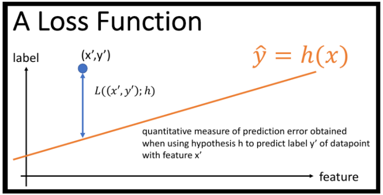
- What do we try to find with linear regression?
- How do we find proper parameters w_0 and w_1 ?

Recap: linear regression

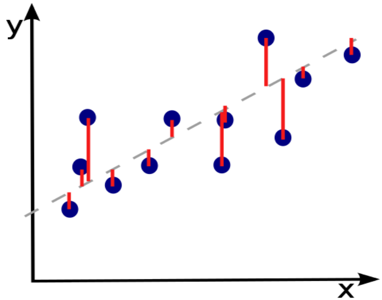


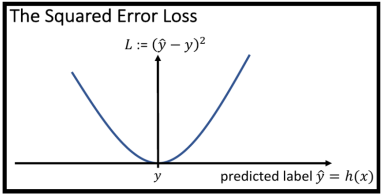
Hypothesis: $h(x) = w_0 + w_1 x$

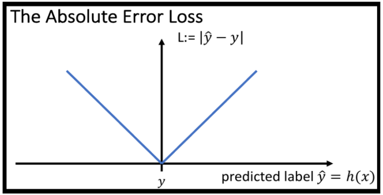
- What do we try to find with linear regression?
- How do we find proper parameters w_0 and w_1 ?



Recap: linear regression

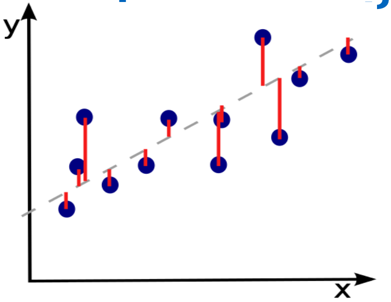






Hypothesis: $h(x) = w_0 + w_1 x$

Recap: linear regression



Loss function: estimates quality of current solution;

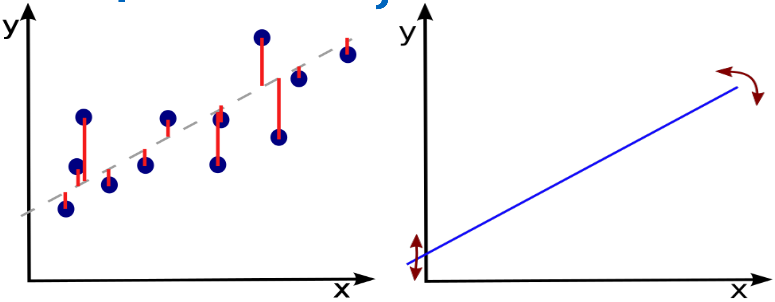
sometimes called *error function* or *cost function*.

Hypothesis: $h(x) = w_0 + w_1 x$

$$\text{minimize } E[w_0, w_1] = L[(X, Y), h(x)] = \frac{1}{2n}\sum_{i=1}^{n}(h(x_i) - y_i)^2$$

$$w_1 = w_1 - \delta \cdot \frac{\partial}{\partial w_1} E[w_0, w_1]$$

Recap: linear regression



Loss function: estimates quality of current solution;

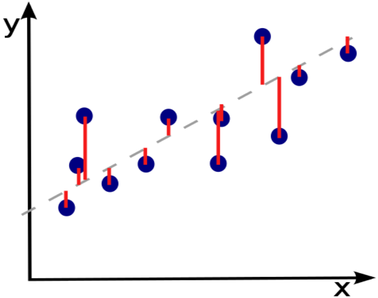
sometimes called *error function* or *cost function*.

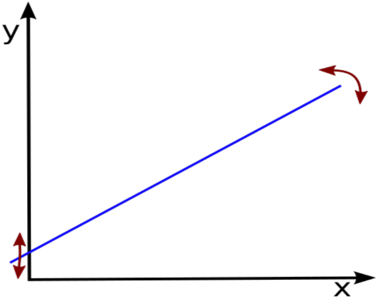
Hypothesis: $h(x) = w_0 + w_1 x$

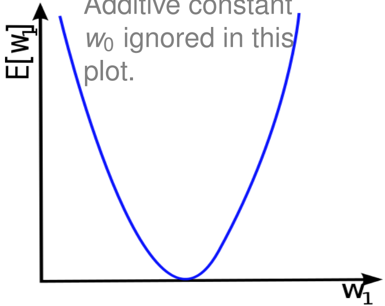
$$\text{minimize } E[w_0, w_1] = L[(X,Y), h(x)] = \frac{1}{2n}\sum_{i=1}^{n}(h(x_i) - y_i)^2$$

$$w_1 = w_1 - \delta \cdot \frac{\partial}{\partial w_1} E[w_0, w_1]$$

Recap: linear regression







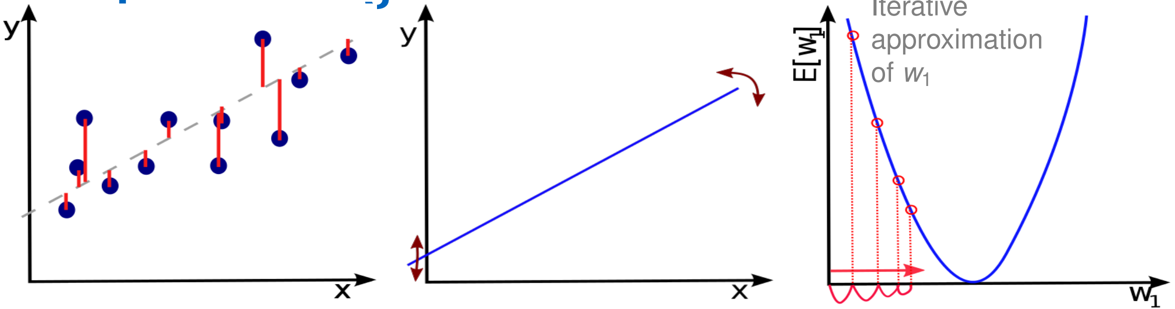
Additive constant w_0 ignored in this plot.

Hypothesis: $h(x) = w_0 + w_1 x$

$$\text{minimize } E[w_0, w_1] = L[(X,Y), h(x)] \quad = \quad \frac{1}{2n}\sum_{i=1}^{n}(h(x_i) - y_i)^2$$

$$w_1 \quad = \quad w_1 - \delta \cdot \frac{\partial}{\partial w_1} E[w_0, w_1]$$

Recap: linear regression



Hypothesis: $h(x) = w_0 + w_1 x$

$$\text{minimize } E[w_0, w_1] = L[(X, Y), h(x)] = \frac{1}{2n}\sum_{i=1}^{n}(h(x_i) - y_i)^2$$

$$w_1 = w_1 - \delta \cdot \frac{\partial}{\partial w_1} E[w_0, w_1]$$

Outline

Recap: linear regression

Logistic regression

Support Vector Machines

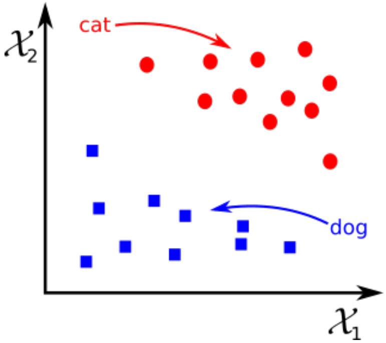
The Perceptron algorithm

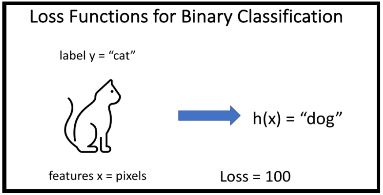
Multiclass classification

Logistic regression

Nominal classes

Classes might be nominal in real-world problems

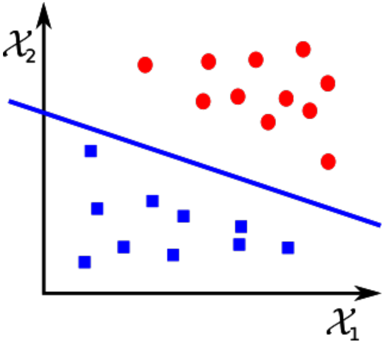


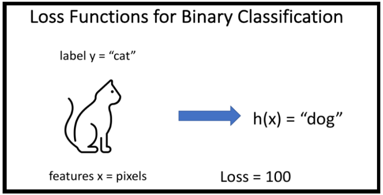


Logistic regression

Nominal classes

Classes might be nominal in real-world problems



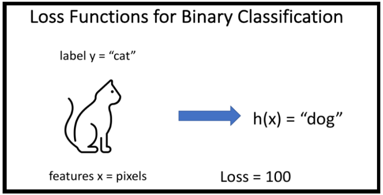


Logistic regression

Nominal classes

Classes might be nominal in real-world problems

Weather	Sunny, rainy
Medical	positive diagnosis, negative diagnosis
Localisation	indoor, outdoor



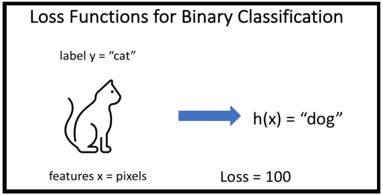
Logistic regression

Nominal classes

Classes might be nominal in real-world problems

Weather	Sunny, rainy
Medical	positive diagnosis, negative diagnosis
Localisation	indoor, outdoor

In such case, classification is binary: $y \in \{0, 1\}$



Logistic regression

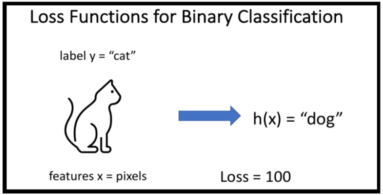
Nominal classes

Classes might be nominal in real-world problems

Weather	Sunny, rainy
Medical	positive diagnosis, negative diagnosis
Localisation	indoor, outdoor

In such case, classification is binary: $y \in \{0, 1\}$

Linear regression: $h(x)$ can be smaller than 0 or greater than 1



Logistic regression

Nominal classes

Classes might be nominal in real-world problems

Weather Sunny, rainy

Medical positive diagnosis, negative diagnosis

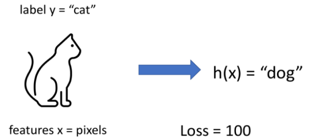
Localisation indoor, outdoor

In such case, classification is binary: $y \in \{0, 1\}$

Linear regression: $h(x)$ can be smaller than 0 or greater than 1

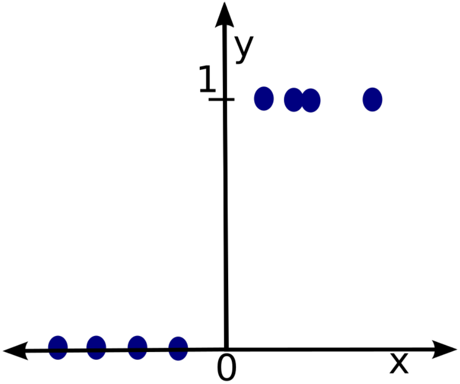
Logistic regression: $0 \leq h(x) \leq 1$

Loss Functions for Binary Classification



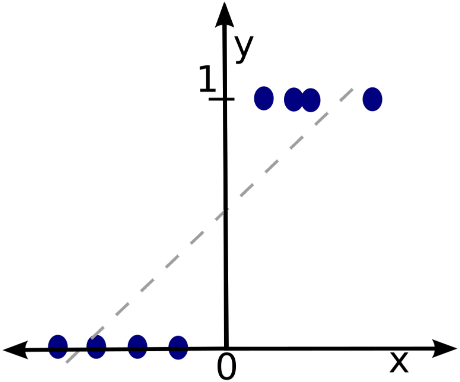
Logistic regression

Nominal classes



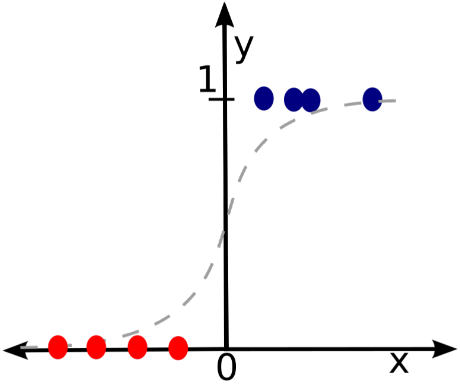
Logistic regression

Nominal classes



Logistic regression

Loss function



Logistic regression

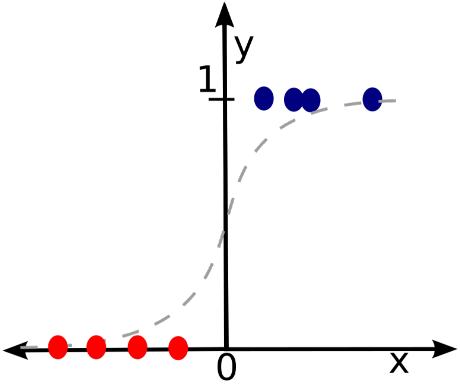
Loss function

Linear regression

$$h(x) = W^T x$$

Logistic regression

$$h(x) = \frac{1}{1+e^{-W^T x}}$$



Logistic regression

Loss function

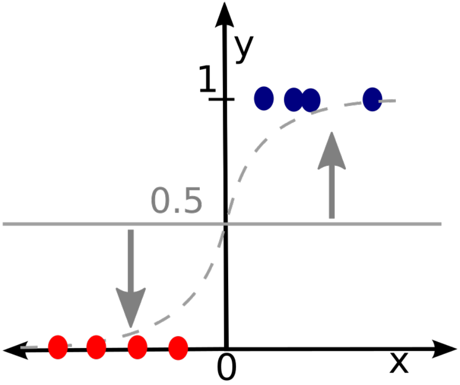
Linear regression

$$h(x) = W^T x$$

Logistic regression

$$h(x) = \frac{1}{1+e^{-W^T x}}$$

$$y = \begin{cases} 1 & \text{if } h(x) \geq 0.5 \\ 0 & \text{else} \end{cases}$$

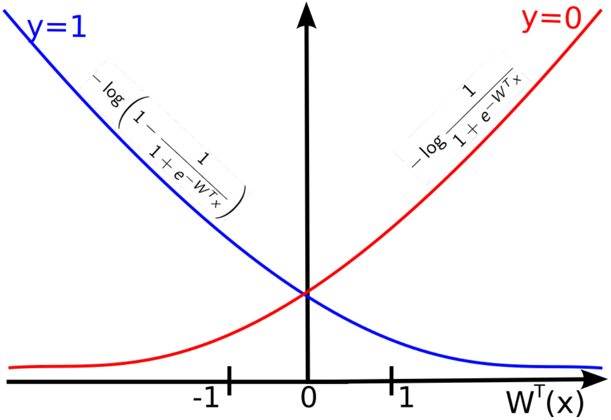


Logistic regression

Loss function

$$y = \begin{cases} 1 & \text{if } h(x) \geq 0.5 \\ 0 & \text{else} \end{cases}$$

$$E[h(x), y] = \begin{cases} -\log(h(x)) & \text{if } y = 1 \\ -\log(1 - h(x)) & \text{else} \end{cases}$$



Outline

Recap: linear regression

Logistic regression

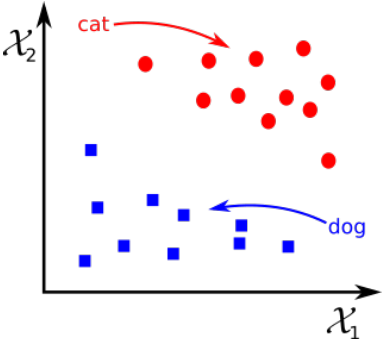
Support Vector Machines

The Perceptron algorithm

Multiclass classification

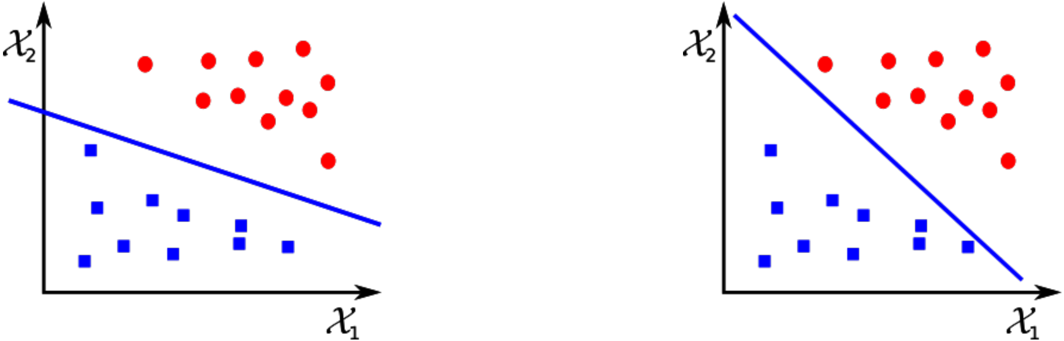
Support vector machines (SVM)

Large margin classifier



Support vector machines (SVM)

Large margin classifier

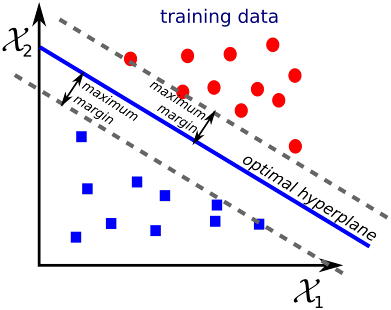


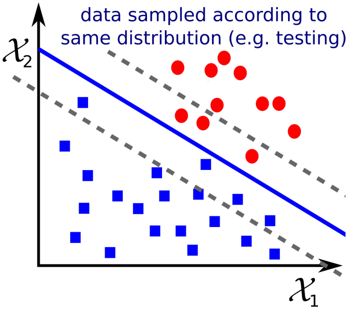
Support vector machines (SVM)

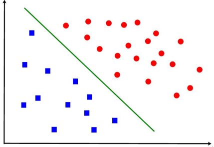
Large margin classifier

The goal for support vector machines is to find a linear and separating hyperplane with the largest margin to the outer points in all sets

If needed, map all points into a higher dimensional space until such a plane exists





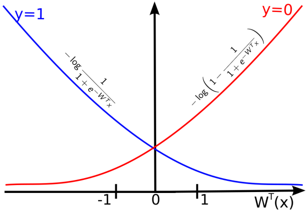


Support vector machines (SVM)

Contribution of a single sample to the overall loss:

Logistic regression

$$-y \cdot \log\left(1 - \frac{1}{1 + e^{-W^T x}}\right) - (1 - y) \cdot \log \frac{1}{1 + e^{-W^T x}}$$

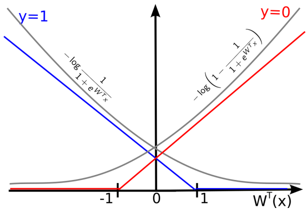


Support vector machines (SVM)

Contribution of a single sample to the overall loss:

SVM

$$-y \cdot \mathrm{cost}_{y=1}(W^T x) + -(1-y) \cdot \mathrm{cost}_{y=0}(W^T x)$$

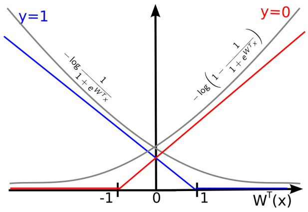


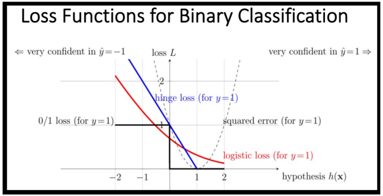
Support vector machines (SVM)

Contribution of a single sample to the overall loss:

SVM

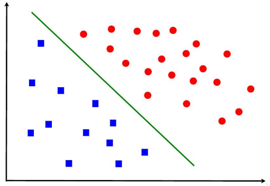
$$-y \cdot \text{cost}_{y=1}(W^T x) + -(1-y) \cdot \text{cost}_{y=0}(W^T x)$$





Support vector machines (SVM)

Cost function



Logistic regression

$$\min_W \quad \frac{1}{m}\left[\sum_{i=1}^{m} y_i\left(-\log\left(1-\frac{1}{1+e^{-W^T x_i}}\right)\right)+(1-y_i)\left(-\log\frac{1}{1+e^{-W^T x_i}}\right)\right]+\frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$$

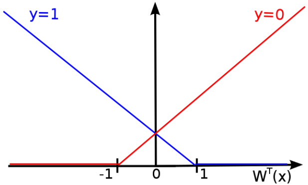
SVM

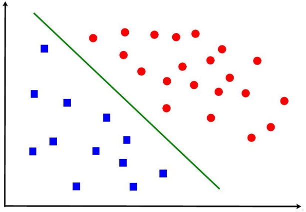
$$\min_W \quad C\sum_{i=1}^{m}\left[y_i \mathrm{cost}_{y=1}(W^T x_i)+(1-y_i)\mathrm{cost}_{y=0}(W^T x_i)\right]+\frac{1}{2}\sum_{j=1}^{n} w_j^2$$

C here plays a similar role as $\frac{1}{\lambda}$

Support vector machines (SVM)

SVM hypothesis

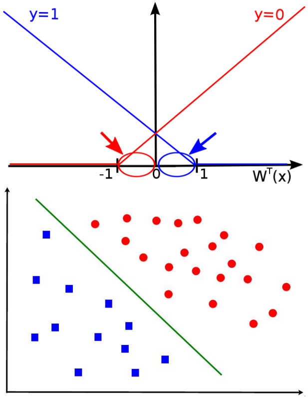




$$\min_{W} C \sum_{i=1}^{m} \left[y_i \text{cost}_{y=1}(W^T x_i) + (1 - y_i)\text{cost}_{y=0}(W^T x_i) \right] + \frac{1}{2} \sum_{j=1}^{n} w_j^2$$

Support vector machines (SVM)

SVM hypothesis

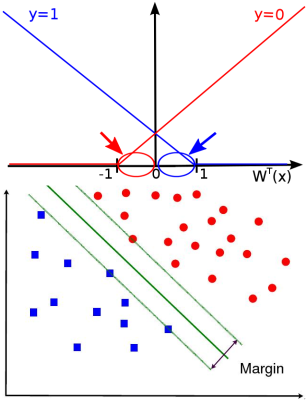


$$W^T x \begin{cases} \geq 0 \\ < 0 \end{cases} \text{ sufficient}$$

$$\min_W C \sum_{i=1}^{m} \left[y_i \text{cost}_{y=1}(W^T x_i) + (1 - y_i)\text{cost}_{y=0}(W^T x_i) \right] + \frac{1}{2} \sum_{j=1}^{n} w_j^2$$

Support vector machines (SVM)

SVM hypothesis



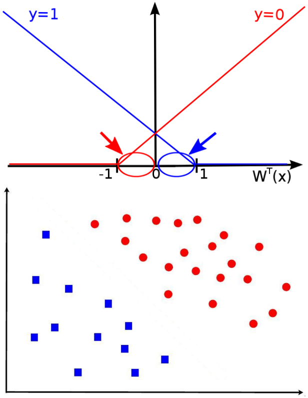
$$W^T x \begin{cases} \geq 0 \\ < 0 \end{cases} \text{sufficient}$$

$$W^T x \begin{cases} \geq 1 \\ \leq -1 \end{cases} \Rightarrow \text{confidence}$$

$$\min_W C \sum_{i=1}^{m} \left[y_i \text{cost}_{y=1}(W^T x_i) + (1 - y_i)\text{cost}_{y=0}(W^T x_i) \right] + \frac{1}{2} \sum_{j=1}^{n} w_j^2$$

Support vector machines (SVM)

SVM hypothesis



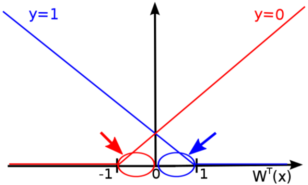
$$W^T x \begin{cases} \geq & 1 \\ \leq & -1 \end{cases} \Rightarrow \text{confidence}$$

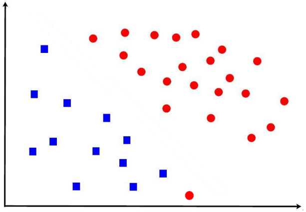
Outliers: Elastic decision boundary

$$\min_W C \sum_{i=1}^{m} \left[y_i \text{cost}_{y=1}(W^T x_i) + (1 - y_i)\text{cost}_{y=0}(W^T x_i) \right] + \frac{1}{2} \sum_{j=1}^{n} w_j^2$$

Support vector machines (SVM)

SVM hypothesis





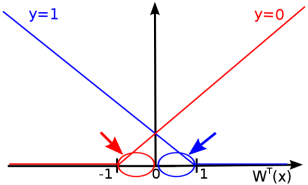
$$W^T x \begin{cases} \geq 1 \\ \leq -1 \end{cases} \Rightarrow \text{confidence}$$

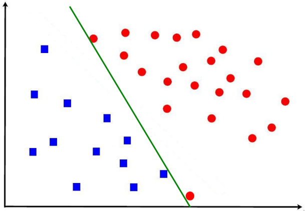
Outliers: Elastic decision boundary

$$\min_W C \sum_{i=1}^{m} \left[y_i \text{cost}_{y=1}(W^T x_i) + (1 - y_i)\text{cost}_{y=0}(W^T x_i) \right] + \frac{1}{2} \sum_{j=1}^{n} w_j^2$$

Support vector machines (SVM)

SVM hypothesis





$$W^T x \begin{cases} \geq & 1 \\ \leq & -1 \end{cases} \Rightarrow \text{confidence}$$

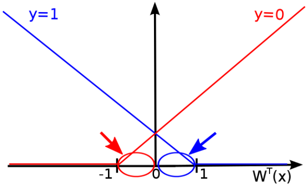
Outliers: Elastic decision boundary

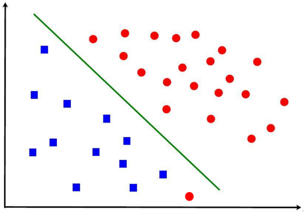
large C stricter boundary at the cost of smaller margin

$$\min_W C \sum_{i=1}^{m} \left[y_i \text{cost}_{y=1}(W^T x_i) + (1 - y_i)\text{cost}_{y=0}(W^T x_i) \right] + \frac{1}{2} \sum_{j=1}^{n} w_j^2$$

Support vector machines (SVM)

SVM hypothesis





$$W^T x \begin{cases} \geq & 1 \\ \leq & -1 \end{cases} \Rightarrow \text{confidence}$$

Outliers: Elastic decision boundary

small C tolerates outliers

$$\min_W C \sum_{i=1}^{m} \left[y_i \text{cost}_{y=1}(W^T x_i) + (1 - y_i)\text{cost}_{y=0}(W^T x_i) \right] + \frac{1}{2} \sum_{j=1}^{n} w_j^2$$

Support vector machines (SVM)

Large margin classifier

$$\min_W C \sum_{i=1}^{m} \left[y_i \text{cost}_{y=1}(W^T x_i) + (1 - y_i)\text{cost}_{y=0}(W^T x_i) \right] + \frac{1}{2} \sum_{j=1}^{n} w_j^2$$

Support vector machines (SVM)

Large margin classifier

$$\min_W C \sum_{i=1}^{m} \left[y_i \text{cost}_{y=1}(W^T x_i) + (1 - y_i)\text{cost}_{y=0}(W^T x_i) \right] + \frac{1}{2} \sum_{j=1}^{n} w_j^2$$

Rewrite the SVM optimisation problem as

$$\begin{aligned} \min_W \quad & \tfrac{1}{2} \textstyle\sum_{j=1}^{n} w_j^2 \\ s.t. \quad & W^T x_i \geq 1 \quad \text{if } y_i = 1 \\ & W^T x_i \leq -1 \quad \text{if } y_i = 0 \end{aligned}$$

Support vector machines (SVM)

Large margin classifier

$$
\begin{aligned}
\min_{W} \quad & \frac{1}{2}\sum_{j=1}^{n} w_j^2 \\
s.t. \quad & W^T x_i \geq 1 \ \text{ if } y_i = 1 \\
& W^T x_i \leq -1 \text{ if } y_i = 0
\end{aligned}
$$

Support vector machines (SVM)

Large margin classifier

$$\begin{aligned} \min_{W} \quad & \tfrac{1}{2}\textstyle\sum_{j=1}^{n} w_j^2 = \tfrac{1}{2}\left(\sqrt{w_1^2 + \cdots + w_n^2}\right)^2 \\ s.t. \quad & W^T x_i \geq 1 \;\text{ if } y_i = 1 \\ & W^T x_i \leq -1 \text{ if } y_i = 0 \end{aligned}$$

Support vector machines (SVM)

Large margin classifier

$$\min_{W} \quad \tfrac{1}{2}\sum_{j=1}^{n} w_j^2 = \tfrac{1}{2}\left(\sqrt{w_1^2+\cdots+w_n^2}\right)^2 \quad = \frac{1}{2}||W||^2$$

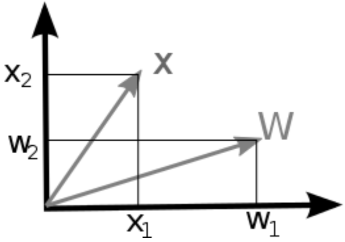
$$\begin{aligned} s.t. \quad & W^T x_i \geq 1 \;\text{ if } y_i = 1 \\ & W^T x_i \leq -1 \text{ if } y_i = 0 \end{aligned}$$

Support vector machines (SVM)

Large margin classifier

$$\min_{W} \quad \frac{1}{2}\sum_{j=1}^{n} w_j^2 = \frac{1}{2}\left(\sqrt{w_1^2 + \cdots + w_n^2}\right)^2 \quad = \frac{1}{2}||W||^2$$

$$\begin{aligned} s.t. \quad & W^T x_i \geq 1 \text{ if } y_i = 1 \\ & W^T x_i \leq -1 \text{ if } y_i = 0 \end{aligned}$$



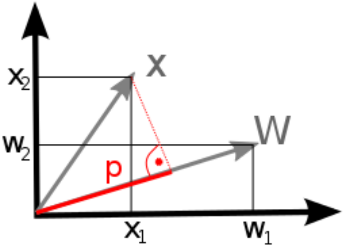
$$W^T x = w_1 x_1 + w_2 x_2$$

Support vector machines (SVM)

Large margin classifier

$$\min_{W} \quad \frac{1}{2}\sum_{j=1}^{n} w_j^2 = \frac{1}{2}\left(\sqrt{w_1^2 + \cdots + w_n^2}\right)^2 \quad = \frac{1}{2}||W||^2$$

$$\begin{aligned} s.t. \qquad & W^T x_i \geq 1 \text{ if } y_i = 1 \\ & W^T x_i \leq -1 \text{ if } y_i = 0 \end{aligned}$$



$$W^T x = w_1 x_1 + w_2 x_2 = ||W|| \cdot p$$

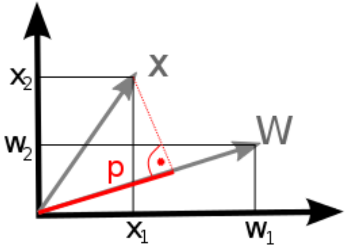
Support vector machines (SVM)

Large margin classifier

$$\min_{W} \quad \frac{1}{2}\sum_{j=1}^{n} w_j^2 = \frac{1}{2}\left(\sqrt{w_1^2 + \cdots + w_n^2}\right)^2 \quad = \frac{1}{2}||W||^2$$

$$s.t. \quad W^T x_i \geq 1 \text{ if } y_i = 1 \quad \rightarrow ||W|| \cdot p_i \geq 1$$

$$W^T x_i \leq -1 \text{ if } y_i = 0 \quad \rightarrow ||W|| \cdot p_i \leq -1$$



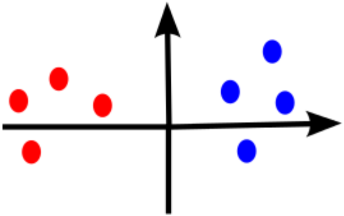
$$W^T x = w_1 x_1 + w_2 x_2 = ||W|| \cdot p$$

Support vector machines (SVM)

Large margin classifier

$$\min_{W} \quad \frac{1}{2}\sum_{j=1}^{n} w_j^2 = \frac{1}{2}\left(\sqrt{w_1^2 + \cdots + w_n^2}\right)^2 \quad = \frac{1}{2}||W||^2$$

$$\begin{aligned} s.t. \quad & W^T x_i \geq 1 \text{ if } y_i = 1 && \rightarrow ||W|| \cdot p_i \geq 1 \\ & W^T x_i \leq -1 \text{ if } y_i = 0 && \rightarrow ||W|| \cdot p_i \leq -1 \end{aligned}$$



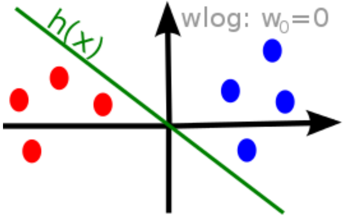
Which decision boundaray is found?

Support vector machines (SVM)

Large margin classifier

$$\min_{W} \quad \frac{1}{2}\sum_{j=1}^{n} w_j^2 = \frac{1}{2}\left(\sqrt{w_1^2+\cdots+w_n^2}\right)^2 \quad = \frac{1}{2}||W||^2$$

$$\begin{aligned} s.t. \quad & W^T x_i \geq 1 \ \text{ if } y_i = 1 && \rightarrow ||W|| \cdot p_i \geq 1 \\ & W^T x_i \leq -1 \text{ if } y_i = 0 && \rightarrow ||W|| \cdot p_i \leq -1 \end{aligned}$$



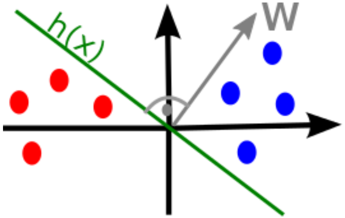
Which decision boundaray is found?

Support vector machines (SVM)

Large margin classifier

$$\min_{W} \quad \frac{1}{2}\sum_{j=1}^{n} w_j^2 = \frac{1}{2}\left(\sqrt{w_1^2 + \cdots + w_n^2}\right)^2 \qquad = \frac{1}{2}||W||^2$$

$$\begin{aligned} s.t. \qquad & W^T x_i \geq 1 \text{ if } y_i = 1 && \rightarrow ||W|| \cdot p_i \geq 1 \\ & W^T x_i \leq -1 \text{ if } y_i = 0 && \rightarrow ||W|| \cdot p_i \leq -1 \end{aligned}$$



Which decision boundaray is found?

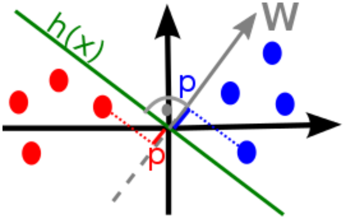
$$h(x) = w_1 x_1 + w_2 x_2$$

$\rightarrow$ W orthogonal to all x with $h(x) = 0$

Support vector machines (SVM)

Large margin classifier

$$
\begin{aligned}
\min_{W} \quad & \tfrac{1}{2}\textstyle\sum_{j=1}^{n} w_j^2 = \tfrac{1}{2}\left(\sqrt{w_1^2+\cdots+w_n^2}\right)^2 && = \frac{1}{2}||W||^2 \\
s.t. \quad & W^T x_i \geq 1 \ \text{ if } y_i = 1 && \rightarrow ||W|| \cdot p_i \geq 1 \\
& W^T x_i \leq -1 \text{ if } y_i = 0 && \rightarrow ||W|| \cdot p_i \leq -1
\end{aligned}
$$



Which decision boundaray is found?

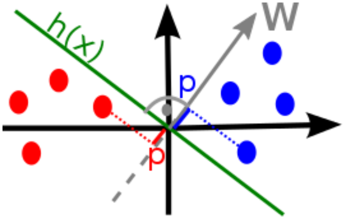
$$h(x) = w_1 x_1 + w_2 x_2$$

$\rightarrow$ W orthogonal to all x with $h(x) = 0$

Support vector machines (SVM)

Large margin classifier

$$
\begin{aligned}
\min_{W} \quad & \tfrac{1}{2}\sum_{j=1}^{n} w_j^2 = \tfrac{1}{2}\left(\sqrt{w_1^2+\cdots+w_n^2}\right)^2 && = \frac{1}{2}||W||^2 \\
s.t. \quad & W^T x_i \geq 1 \ \text{ if } y_i = 1 && \rightarrow ||W|| \cdot p_i \geq 1 \\
& W^T x_i \leq -1 \text{ if } y_i = 0 && \rightarrow ||W|| \cdot p_i \leq -1
\end{aligned}
$$



Which decision boundaray is found?

$$h(x) = w_1 x_1 + w_2 x_2$$

$\rightarrow$ W orthogonal to all x with $h(x) = 0$

$\Rightarrow$ $\min \frac{1}{2}||W||^2$ and $||W|| \cdot p_i \geq 1$ necessitate larger p_i

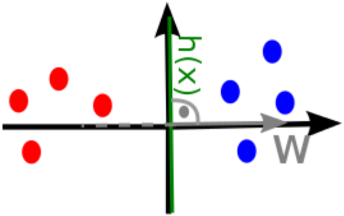
Support vector machines (SVM)

Large margin classifier

$$\min_{W} \quad \frac{1}{2}\sum_{j=1}^{n} w_j^2 = \frac{1}{2}\left(\sqrt{w_1^2+\cdots+w_n^2}\right)^2 \quad = \frac{1}{2}||W||^2$$

$$s.t. \quad W^T x_i \geq 1 \;\text{ if } y_i = 1 \quad \rightarrow ||W|| \cdot p_i \geq 1$$

$$W^T x_i \leq -1 \text{ if } y_i = 0 \quad \rightarrow ||W|| \cdot p_i \leq -1$$



Which decision boundaray is found?

$h(x) = w_1 x_1 + w_2 x_2$

$\rightarrow$ W orthogonal to all x with $h(x) = 0$

$\Rightarrow$ $\min \frac{1}{2}||W||^2$ and $||W|| \cdot p_i \geq 1$ necessitate larger p_i

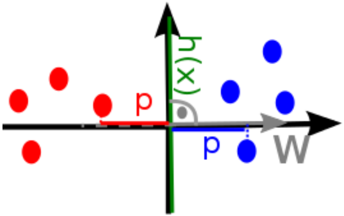
Support vector machines (SVM)

Large margin classifier

$$\min_{W} \quad \frac{1}{2}\sum_{j=1}^{n} w_j^2 = \frac{1}{2}\left(\sqrt{w_1^2+\cdots+w_n^2}\right)^2 \quad = \frac{1}{2}||W||^2$$

$$s.t. \quad W^T x_i \geq 1 \;\text{ if } y_i = 1 \quad \rightarrow ||W|| \cdot p_i \geq 1$$

$$W^T x_i \leq -1 \text{ if } y_i = 0 \quad \rightarrow ||W|| \cdot p_i \leq -1$$



Which decision boundaray is found?

$h(x) = w_1 x_1 + w_2 x_2$

$\rightarrow$ W orthogonal to all x with $h(x) = 0$

$\Rightarrow$ $\min \frac{1}{2}||W||^2$ and $||W|| \cdot p_i \geq 1$ necessitate larger p_i

Outline

Recap: linear regression

Logistic regression

Support Vector Machines

The Perceptron algorithm

Multiclass classification

The perceptron algorithm

Two-class model ($\mathcal{C} \in \{-1, 1\}$) in which $\overrightarrow{x}$ is the feature vector:

$$y(x) = f(w^T \overrightarrow{x})$$

The perceptron algorithm

Two-class model ($\mathcal{C} \in \{-1, 1\}$) in which $\overrightarrow{x}$ is the feature vector:

$$y(x) = f(w^T \overrightarrow{x})$$

nonlinear activation function defined as a step-function:

$$f(a) = \begin{cases} +1, & a \geq 0 \\ -1, & a < 0. \end{cases}$$

The perceptron algorithm

Training: Find error function for $\overrightarrow{w}$ to enforce

$$\begin{aligned} x_i \in \mathcal{C}_1 : &\quad \overrightarrow{w}^T \overrightarrow{x_i} > 0 \\ x_i \in \mathcal{C}_{-1} : &\quad \overrightarrow{w}^T \overrightarrow{x_i} < 0 \end{aligned}$$

For the set $\mathcal{D}$ of all misclassified patterns, the error function is

$$E(\overrightarrow{w}) = \begin{cases} -\sum_{i \in \mathcal{D}} \overrightarrow{w}^T \overrightarrow{x_i} \mathcal{C}(x_i) & x_i \in \mathcal{D} \\ 0 & \text{else} \end{cases}$$

The perceptron algorithm

For the set $\mathcal{D}$ of all misclassified patterns, the error function is

$$E(\vec{w}) = \begin{cases} -\sum_{i\in\mathcal{D}} \vec{w}^T\vec{x_i}\mathcal{C}(x_i) & x_i \in \mathcal{D} \\ 0 & \text{else} \end{cases}$$

$E(\vec{w})$ is piecewise linear:

linear in regions of $\vec{w}$-space where pattern is misclassfied

0 in regions where it is classified correctly

The perceptron algorithm

For the set $\mathcal{D}$ of all misclassified patterns, the error function is

$$E(\vec{w}) = \begin{cases} -\sum_{i\in\mathcal{D}} \vec{w}^T \vec{x_i}\mathcal{C}(x_i) & x_i \in \mathcal{D} \\ 0 & \text{else} \end{cases}$$

$E(\vec{w})$ is piecewise linear:

linear in regions of $\vec{w}$-space where pattern is misclassfied

0 in regions where it is classified correctly

Apply stochastic gradient descent to this error function:

$$\vec{w}^{t+1} = \vec{w}^{t} - \begin{cases} \delta \frac{\partial}{\partial \vec{w}} E(\vec{w}) & x_i \in \mathcal{D} \\ 0 & \text{else} \end{cases}$$

$$= \vec{w}^{t} + \begin{cases} \delta \vec{x_i}\mathcal{C}(x_i) & x_i \in \mathcal{D} \\ 0 & \text{else} \end{cases}$$

The perceptron algorithm

Interpretation of the learning function

$$\overrightarrow{w}^{t+1} = \overrightarrow{w}^{t} + \begin{cases} \delta \overrightarrow{x_i} \mathcal{C}(x_i) & x_i \in \mathcal{D} \\ 0 & \text{else} \end{cases}$$

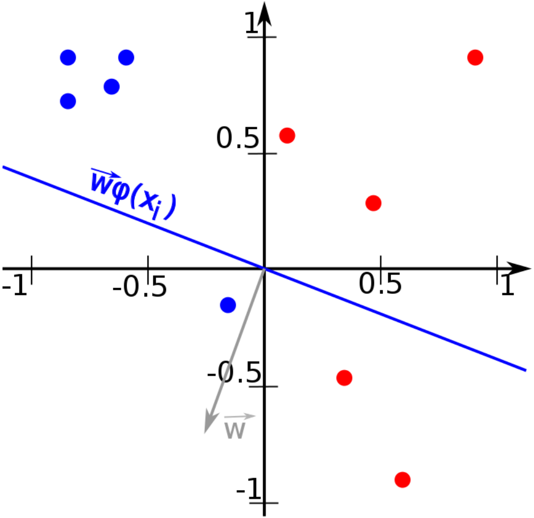
for each x_i:

correct classification: weight vector remains unchanged

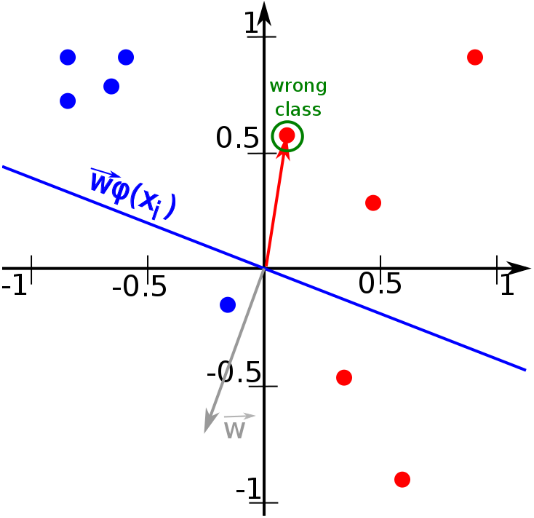
incorrect classification:

$$\begin{aligned} \text{Class } \mathcal{C}_1 &: \quad \text{add vector } \overrightarrow{x_i} \\ \text{Class } \mathcal{C}_{-1} &: \quad \text{subtract vector } \overrightarrow{x_i} \end{aligned}$$

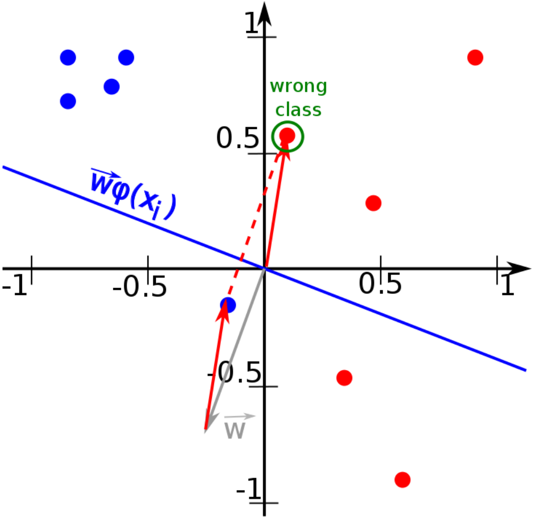
The perceptron algorithm



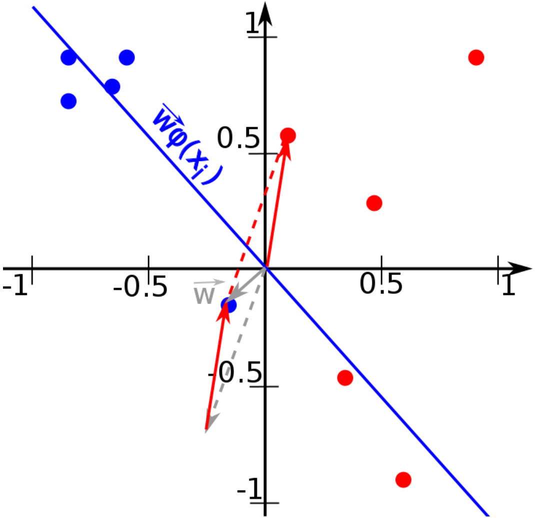
The perceptron algorithm



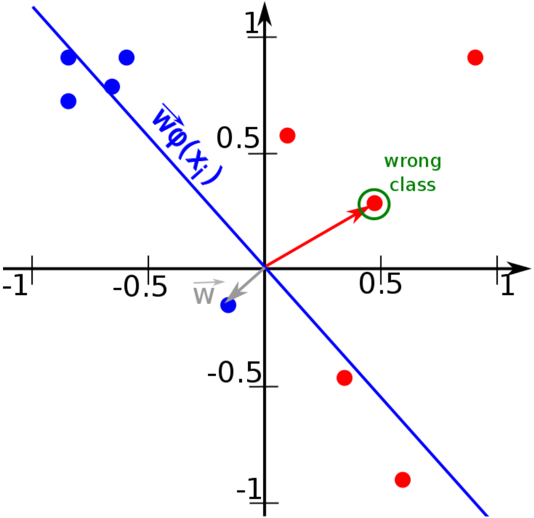
The perceptron algorithm



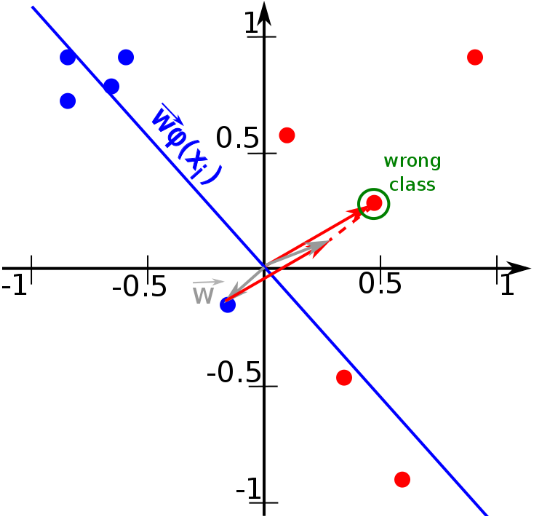
The perceptron algorithm



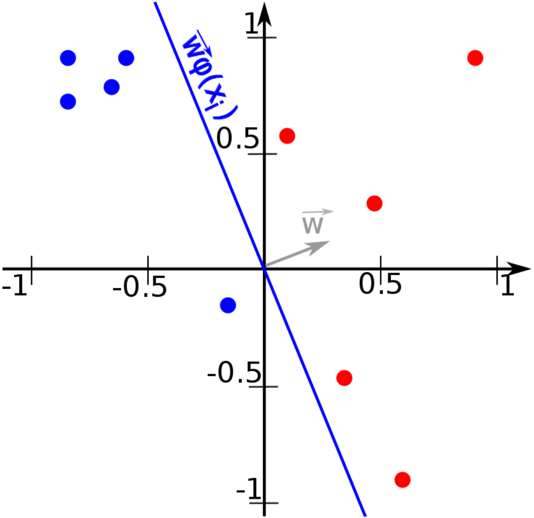
The perceptron algorithm



The perceptron algorithm



The perceptron algorithm



The perceptron algorithm

Perceptron convergence theorem

IFF the training data is linearly separable, then the perceptron learning algorithm will always find an exact solution in finite number of steps.

$\rightarrow$ Number of steps required might be large

$\rightarrow$ Until convergence, not possible to distinguish separable problem from non-separable

$\rightarrow$ For non-separable data sets the algorithm will never converge

Outline

Recap: linear regression

Logistic regression

Support Vector Machines

The Perceptron algorithm

Multiclass classification

Multiclass classification

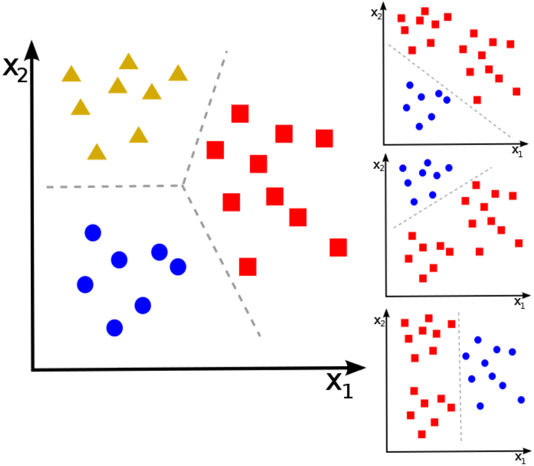
Multi-class: One-versus all:

Train classifiers for each class to obtain probability that x belongs to class i:

$$h_i(x) = P(y = i | \vec{x}, \vec{W})$$

then, choose

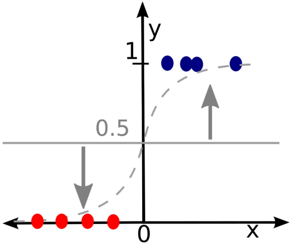
$$max_i\,(h_i(x))$$

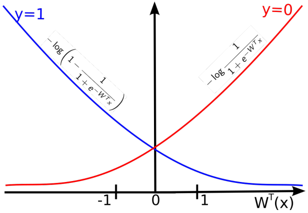


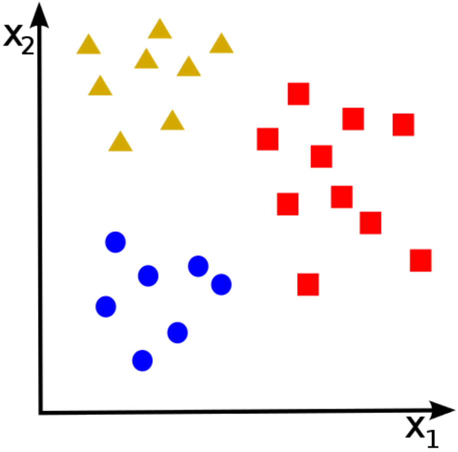
Multiclass classification

Multiple classes

Can we use logistic regression for problems with more than two classes?





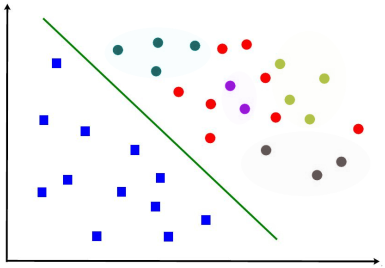


Application to several classes iteratively: One-versus-all

belongs to class 1 or not?

belongs to class 2 or not?

...



Questions?

Stephan Sigg
stephan.sigg@aalto.fi

Si Zuo
si.zuo@aalto.fi

Literature

- C.M. Bishop: Pattern recognition and machine learning, Springer, 2007.
- R.O. Duda, P.E. Hart, D.G. Stork: Pattern Classification, Wiley, 2001.