

# Methods 3: Multilevel Statistical Modeling and Machine Learning

Week 7: *Logistic regression (machine learning)*  
November 16, 2021

*by:* Lau Møller Andersen

These slides are distributed according to the CC  
BY 4.0 licence:

<https://creativecommons.org/licenses/by/4.0/>



# Remember to vote today!

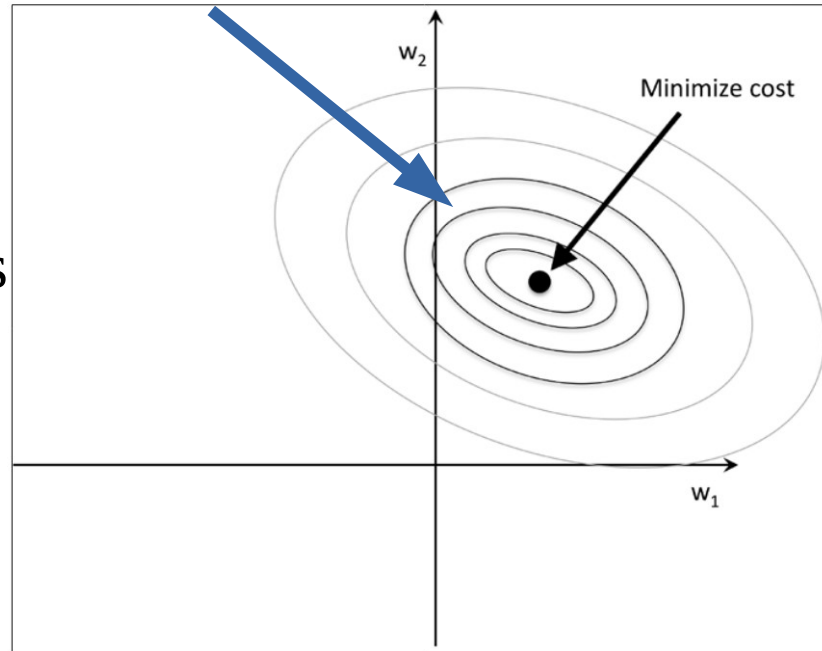
# Regularization - recap

the cost function (what we want to minimize)

$$J(w) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Solution space is infinite;  
 $w$  can be any set of values

*Solution space*



(p. 113: Raschka, 2015)

# L2 regularization

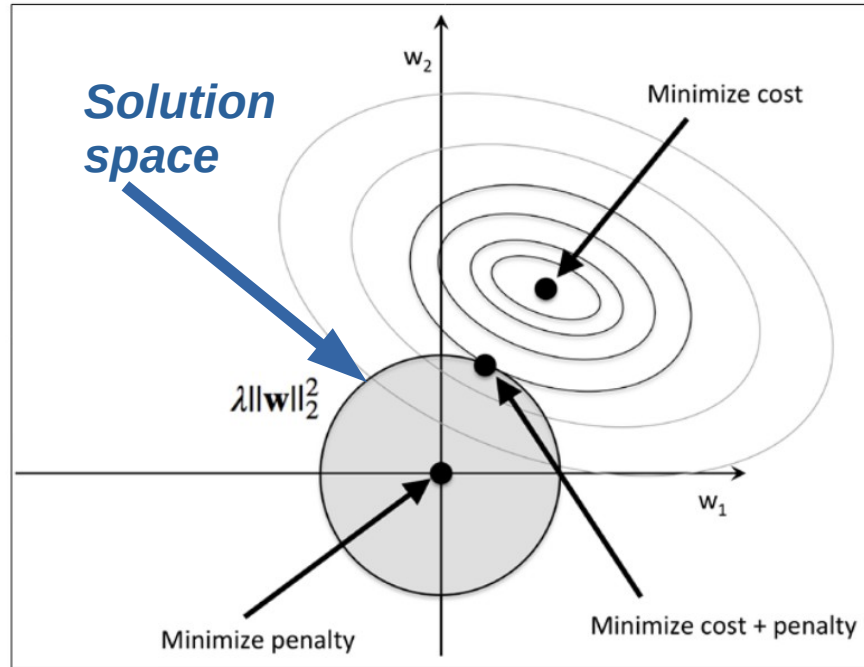
Why is the *solution space* round?

Compare with a circle centred at (0,0)

$$w_1^2 + w_2^2 = r^2$$

$$L_2 \text{ norm: } \|w\|_2 = \sqrt{w_1^2 + w_2^2}$$

(p. 114: Raschka, 2015)



$$J(w)_{\text{Ridge}} = \sum_{i=1}^n \left( y^{(i)} - \hat{y}^{(i)} \right)^2 + \lambda \|w\|_2^2$$

# L1 regularization

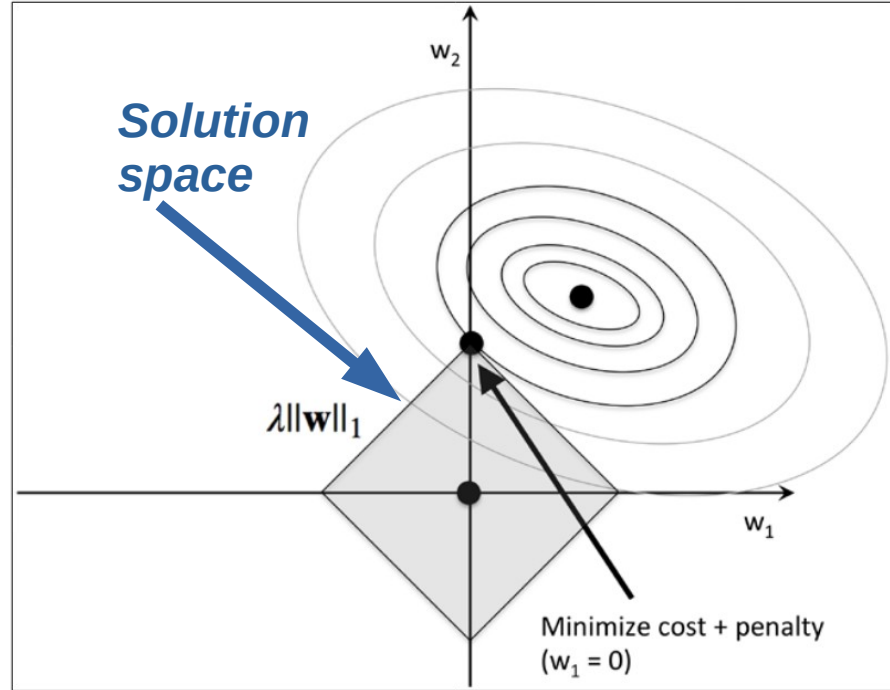
Why is the *solution space* square?

$L_1$  norm:  $\|w\|_1 = |w_1| + |w_2|$

if  $w_1 = \max(w_1)$  then  $w_2 = 0$

if  $w_2 = \max(w_2)$  then  $w_1 = 0$

(p. 115: Raschka, 2015)



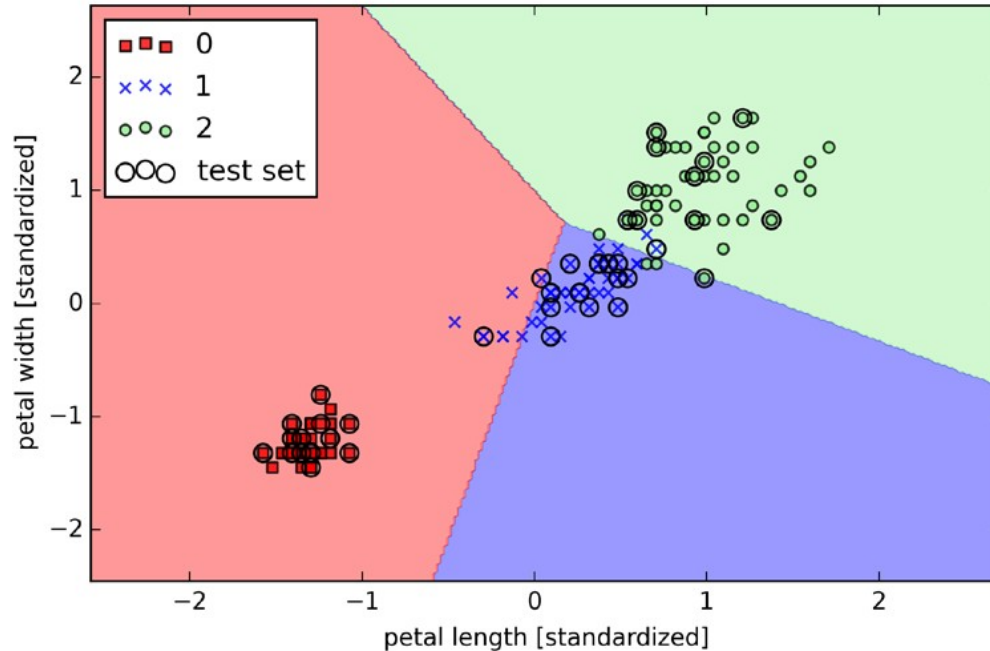
$$J(w)_{LASSO} = \sum_{i=1}^n \left( y^{(i)} - \hat{y}^{(i)} \right)^2 + \lambda \|w\|_1$$

# Learning goals

*Logistic regression (machine learning)*

- 1) Understanding of how logistic regression can be adapted to a classification framework
- 2) Understanding the idea of a Support Vector Machine ??????????????
- 3) Getting acquainted with how Support Vector Machines can solve non-linear problems

# The problem (Perceptron)



(p. 55: Raschka, 2015)

the lines that separate the data categories are linear (lines),  
and in this case a squiggly line would probably separate it better (more bias)  
→ we should instead use logistic regression (next next slide)

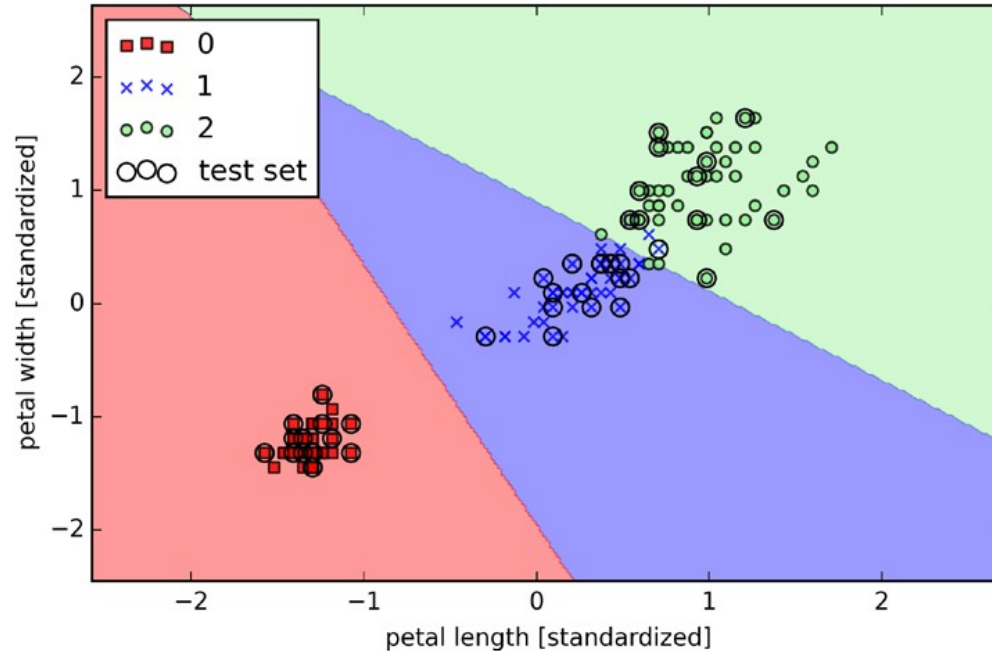
*Not linearly separable → never converges*



**WANTED:** an algorithm that converges  
with linearly separable regions that  
minimise the number of errors made

# Something like this

## LOGISTIC REGRESSION



(p. 63: Raschka, 2015)

*Separates flowers, while keeping errors at a minimum*

# Odds ratio

the ratio of probability of something happening and NOT happening

$$\text{odds ratio} = \frac{p}{1-p}$$

the odds ratio on logscale (with this we can then model it linearly)

$$\log \text{ odds} = \log \left( \frac{p}{1-p} \right) = \text{logit}(p)$$

$$\text{logit}(p(y=1 | \mathbf{x})) = w_0 x_0 + w_1 x_1 + \dots + w_m x_m = \sum_{i=0}^m w_i x_i = \mathbf{w}^T \mathbf{x}$$

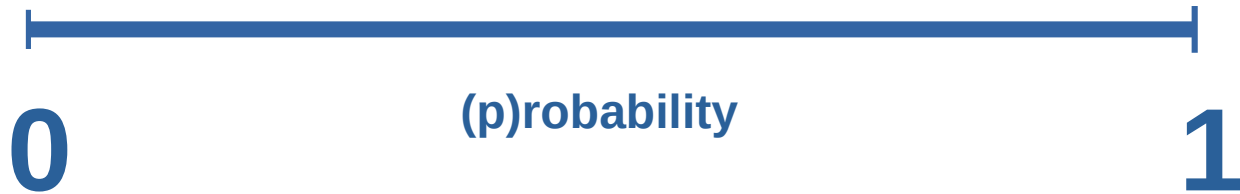
og så får vi en weight vektor

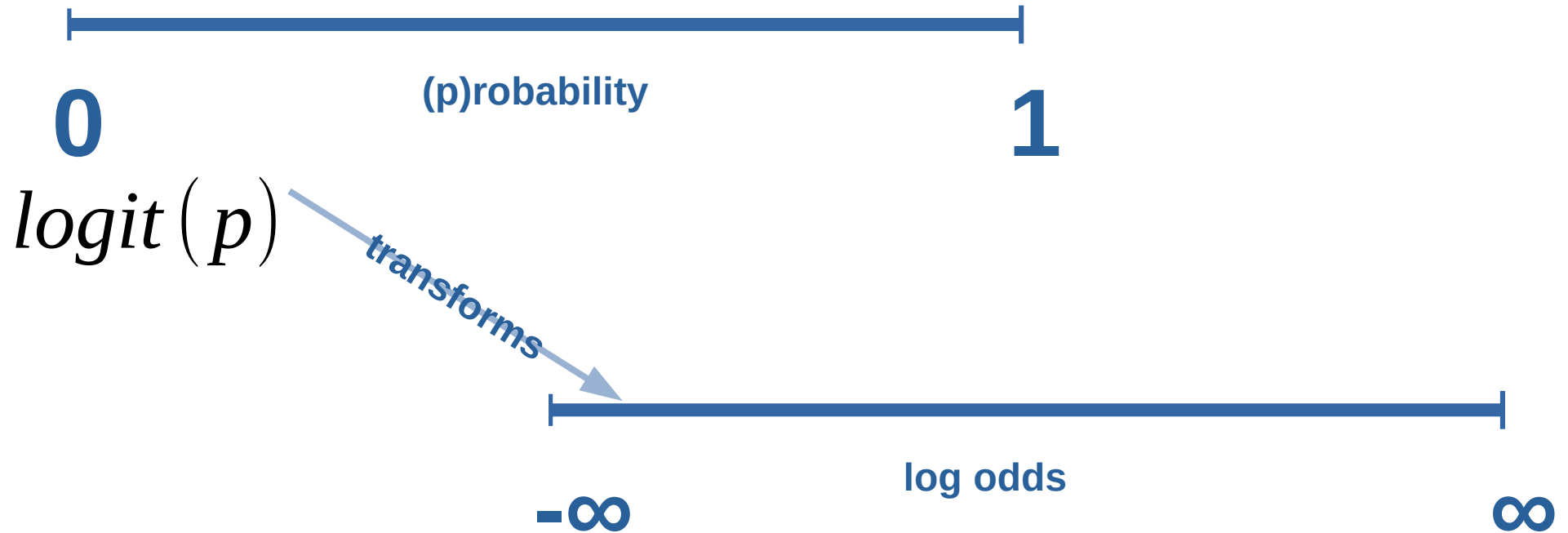
the probability of y being 1 given data punkt x1, osv osv osv

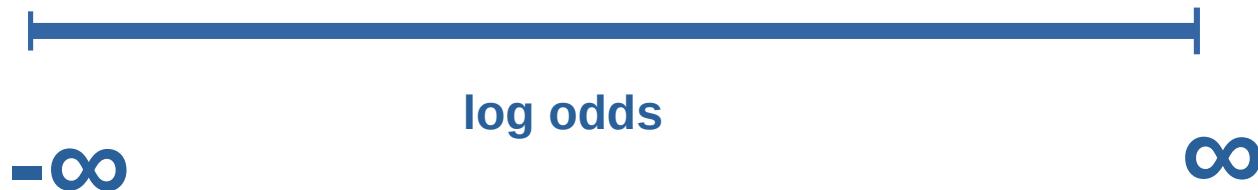
**Example:** an odds ratio of 7 to 1 (7:1) means that it is 7 times more likely that something is going to happen than that it is not going to happen

$$\text{logit}(p(y=1|\mathbf{x})) = w_0 x_0 + w_1 x_1 + \dots + w_m x_m = \sum_{i=0}^m w_i x_i = \mathbf{w}^T \mathbf{x}$$

$p(y=1|\mathbf{x})$ : is the probability that  $y=1$ , given  $\mathbf{x}$







$$\phi(z) = p$$

IN R this would be the link-logit function

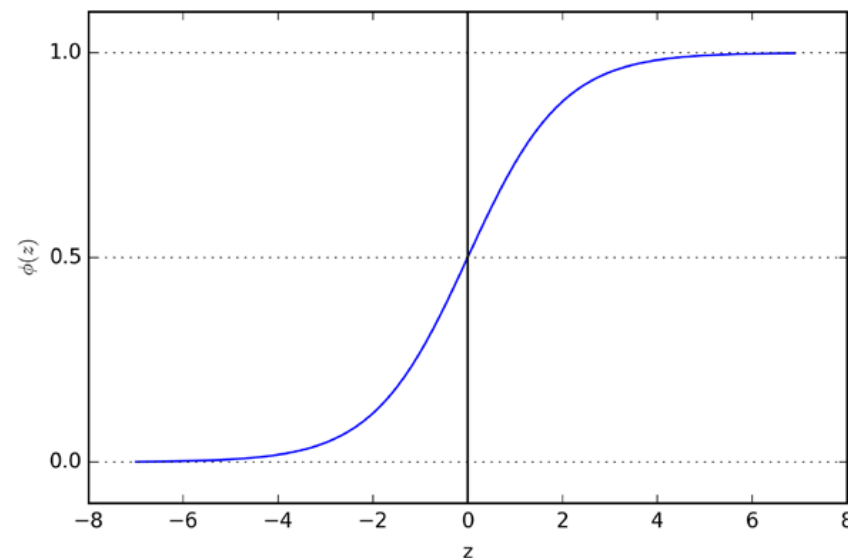
$$\text{logit}^{-1}(z) = \phi(z) = \frac{1}{1 + e^{-z}}$$

(sigmoid-shaped)

$$\text{net input: } z = \mathbf{w}^T \mathbf{x} = w_0 x_0 + w_1 x_1 + \dots + w_m x_m$$

z will be a real number from -inf to inf ON THE X AXIS

ON THE Y AXIS IT IS PROBABILITY



(p. 58: Raschka, 2015)

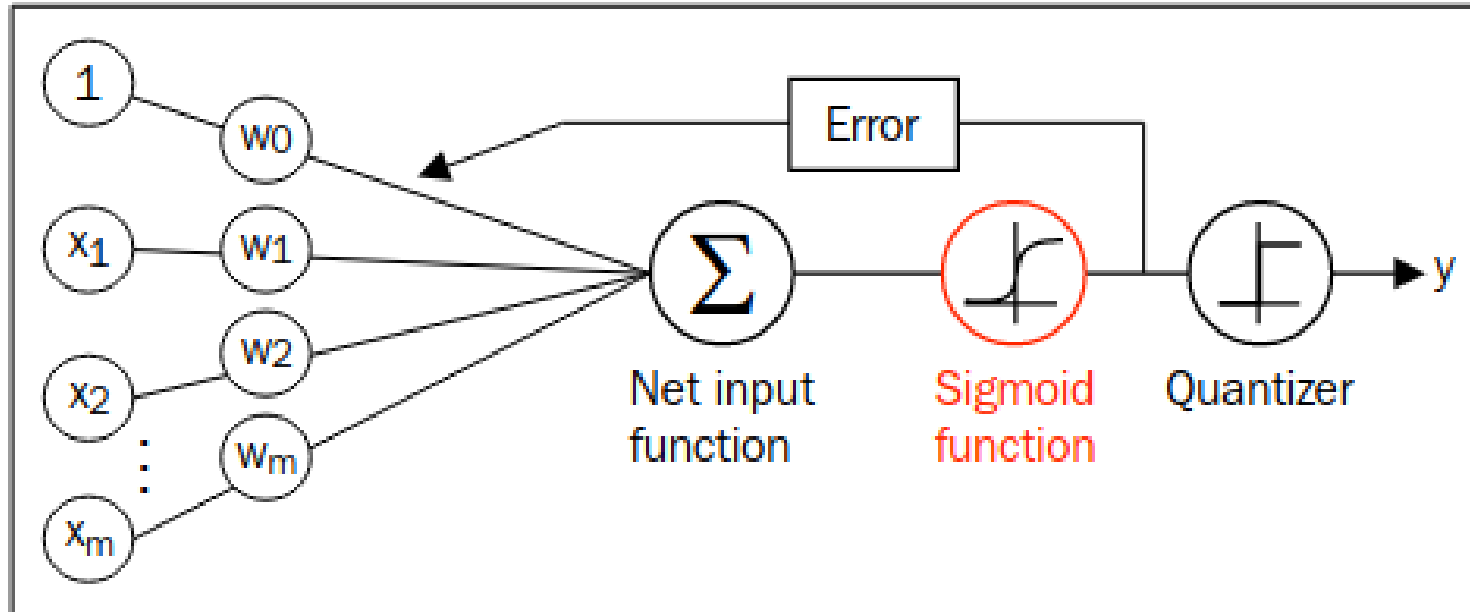
# Quantizer

$$\hat{y} = \begin{cases} 1 & \text{if } \phi(z) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{y} = \begin{cases} 1 & \text{if } z \geq 0.0 \\ 0 & \text{otherwise} \end{cases}$$

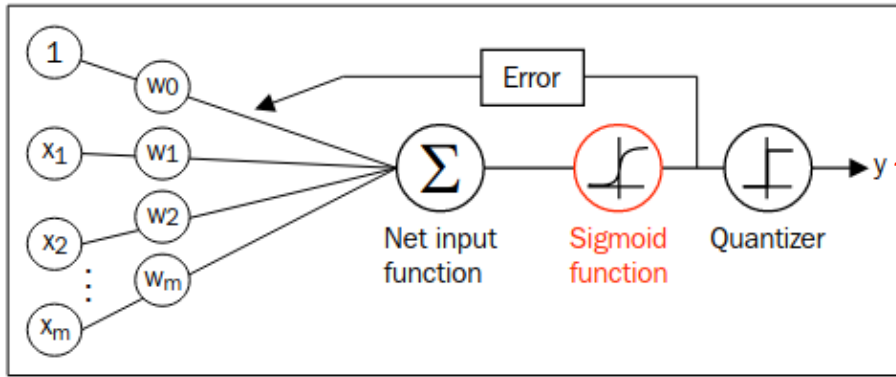


# Logistic regression



(p. 58: Raschka, 2015)

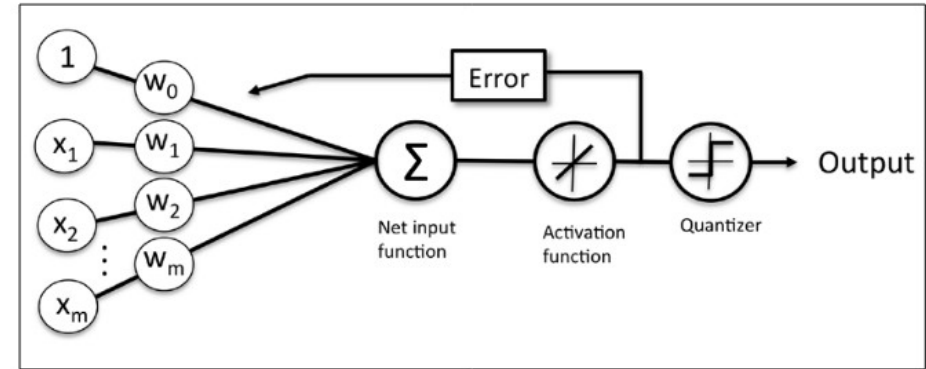
# Comparison with ADALINE



(p. 58: Raschka, 2015)

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

logit function  
logistic regression



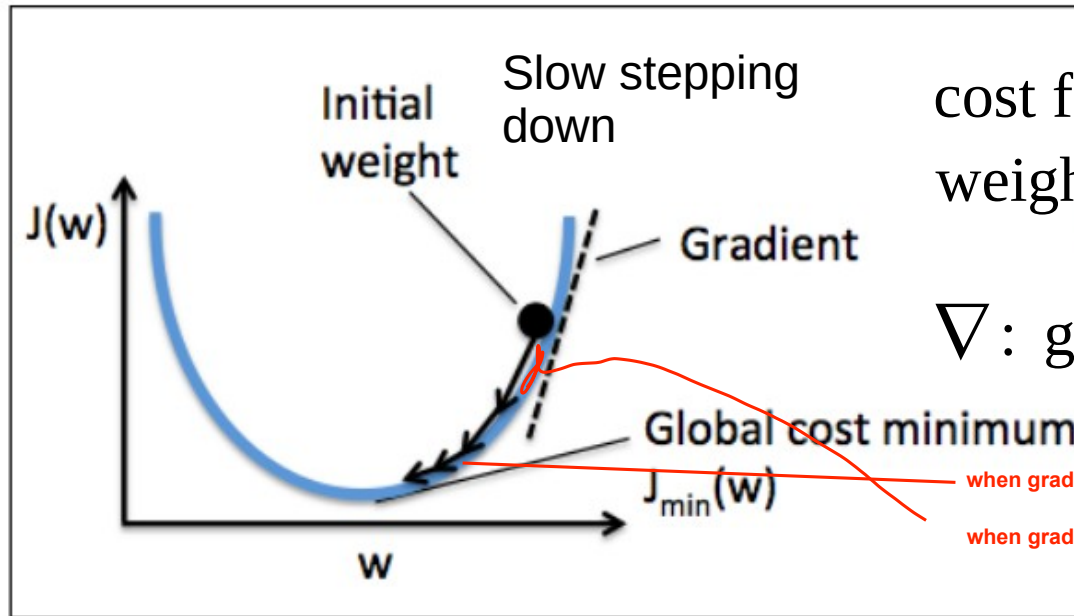
(p. 33: Raschka, 2015)

$$\phi(z) = z$$

identity function (what comes in, comes out)  
linear regression

# Still updating with **Gradient Descent**

# Gradient descent $\phi(z) = z$



cost function:  $J(\mathbf{w}) = (\sum (y - \hat{y})^2) / 2$   
 weight change:  $\Delta \mathbf{w} = -\eta \nabla J(\mathbf{w})$

$\nabla$ : gradient (rate of change)

when gradient is low, low learning rate (small steps in gradient movement)

when gradient is high, high learning rate (greater steps in gradient movement)

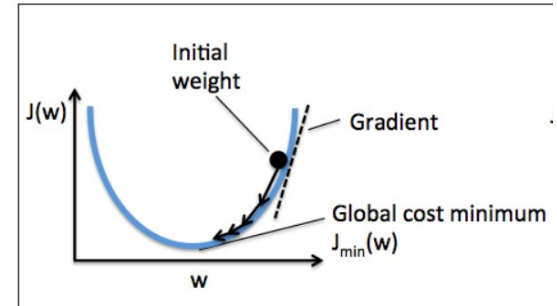
(p. 40: Raschka, 2015)

# Gradient descent $\phi(z) = \frac{1}{1 + e^{-z}}$

cost function:  $J(\mathbf{w}) = -\sum_i^n y^{(i)} \log(\phi(z^i)) + (1 - y^{(i)}) \log(1 - \phi(z^i))$

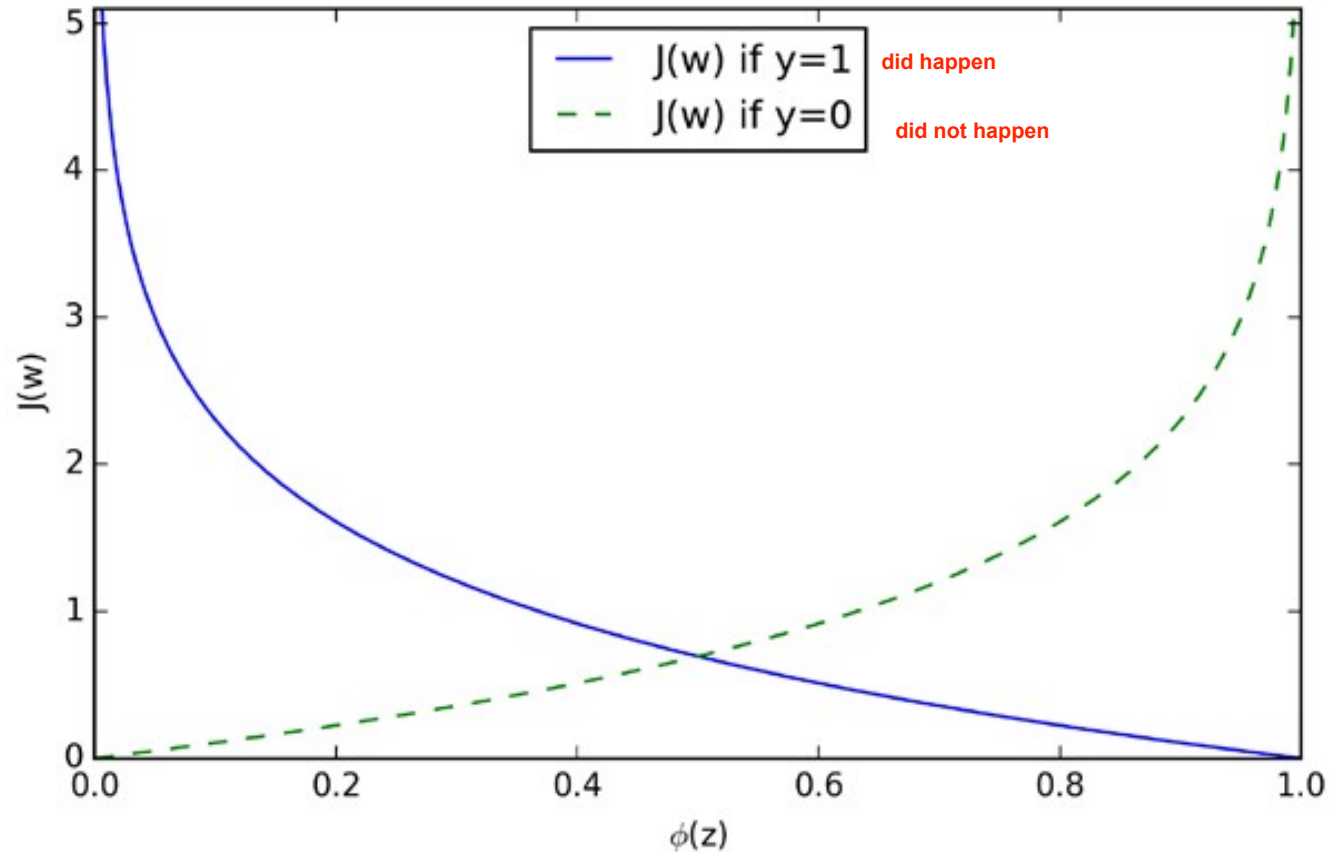
weight change:  $\Delta \mathbf{w} = -\eta \nabla J(\mathbf{w})$

What's this part equal to?



(p. 40: Raschka, 2015)

# A closer look at the cost function



# A closer look at the updating of weights

$$\Delta \mathbf{w} = -\eta \nabla J(\mathbf{w})$$

$\eta$  = learning rate

This is common between ADALINE and Linear and Logistic Regression

but  $J(\mathbf{w})$  is not the same across these different methods?!

A general formulation: 
$$\Delta w_j = \eta \sum_i^n (y^{(i)} - \phi(z^{(i)})) x_j^{(i)}$$

$\phi$  = derivative / the change in the likelihood function


It can be shown that: 
$$\frac{\partial}{\partial w_j} l(\mathbf{w}) = (y - \phi(z)) x_j$$

$\frac{\partial}{\partial w_j} l(\mathbf{w})$ : (partial) derivative of log-likelihood function (proof on p. 64)

# A clarification on ADALINE

$$\Delta w_j = \eta \sum_i (y^{(i)} - \phi(z^{(i)})) x_j^{(i)}$$

$\phi(z) = z = \hat{y}$



```
def fit(self, X, y):  
    """ Fit training data.  
  
    Parameters  
    -----  
    X : {array-like}, shape = [n_samples, n_features]  
        Traing vectors, where n_samples  
        is the number of samples and  
        n_features is the number of features.  
    y : array-like, shape = [n_samples]  
        Target values.  
  
    Returns  
    -----  
    self : object  
  
    """  
    self.w_ = np.zeros(1 + X.shape[1])  
    self.cost_ = []  
  
    for i in range(self.n_iter):  
        output = self.net_input(X)  
        errors = (y - output)  
        self.w [1:] += self.eta * X.T.dot(errors)  
        self.w_[0] += self.eta * errors.sum()  
        cost = (errors**2).sum() / 2.0  
        self.cost_.append(cost)  
    return self
```



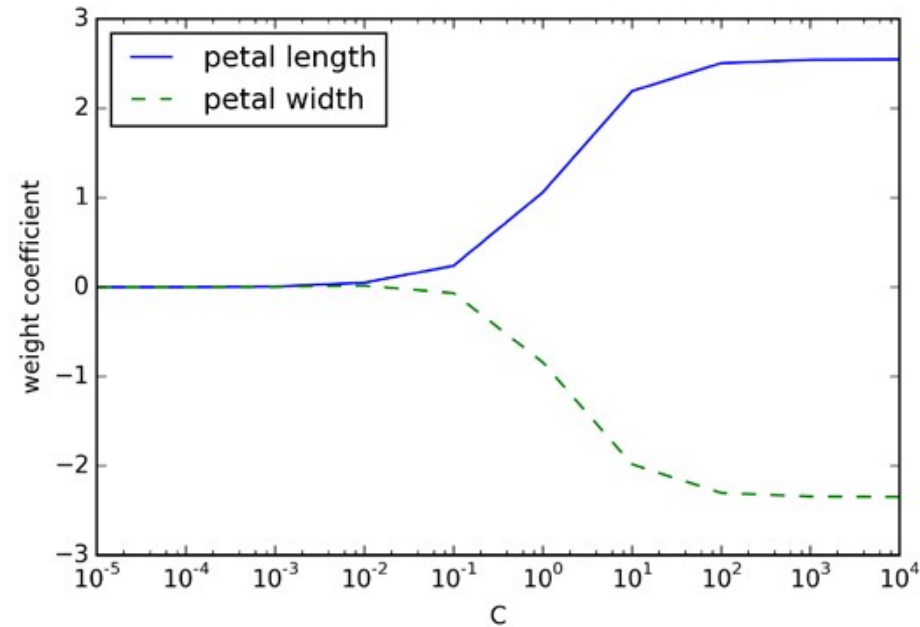
# With regularisation

$$J(\mathbf{w}) = -\sum_i^n y^{(i)} \log(\phi(z^{(i)})) + (1 - y^{(i)}) \log(1 - \phi(z^{(i)})) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

NB! Regularisation in *sklearn.linear\_model.LogisticRegression* is controlled by *C*

$$C = \frac{1}{\lambda}$$

# Effect of C parameter



# Now for the **classification**

```
score(X, y, sample_weight=None)
```

[\[source\]](#)

Return the mean accuracy on the given test data and labels.

In multi-label classification, this is the subset accuracy which is a harsh metric since you require for each sample that each label set be correctly predicted.

Parameters:	<b>X : array-like of shape (n_samples, n_features)</b> Test samples.
	<b>y : array-like of shape (n_samples,) or (n_samples, n_outputs)</b> True labels for X.
	<b>sample_weight : array-like of shape (n_samples,), default=None</b> Sample weights.
Returns:	<b>score : float</b> Mean accuracy of <code>self.predict(X)</code> wrt. y.

```
## LOGISTIC REGRESSION

from sklearn.linear_model import LogisticRegression
logR = LogisticRegression(penalty='none') # no regularisation
logR.fit(X_train_std, y_train)
print(logR.score(X_test_std, y_test))  how many correct predictions from trainig to test set
```

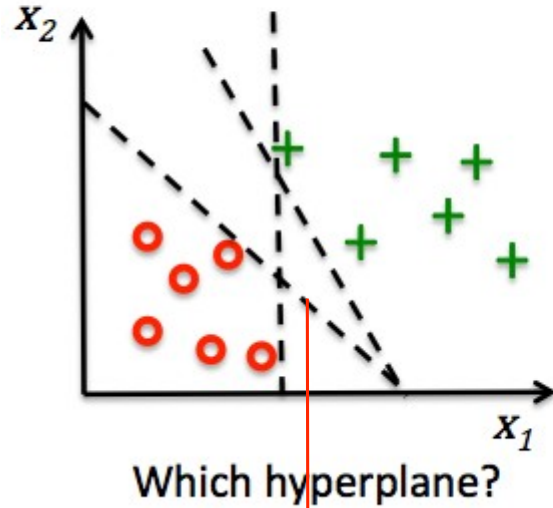
# Live coding

## EXAMPLE\_00

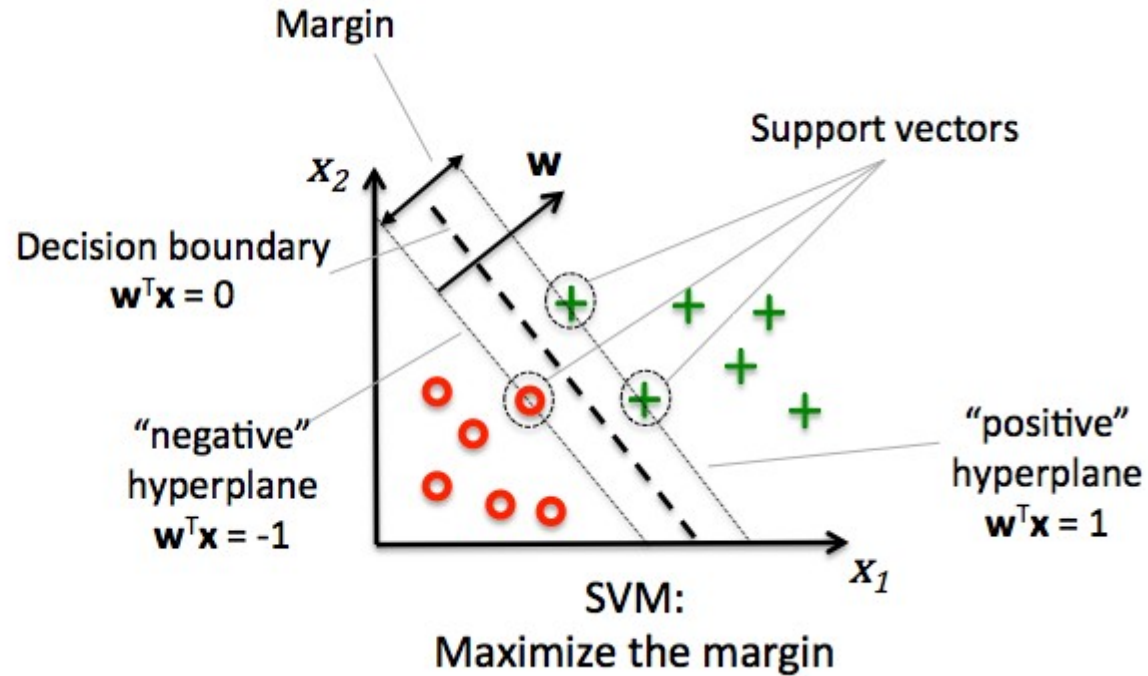
# SUPPORT VECTOR MACHINES

we want to maximize the margin(the larger the more space is between the two data)

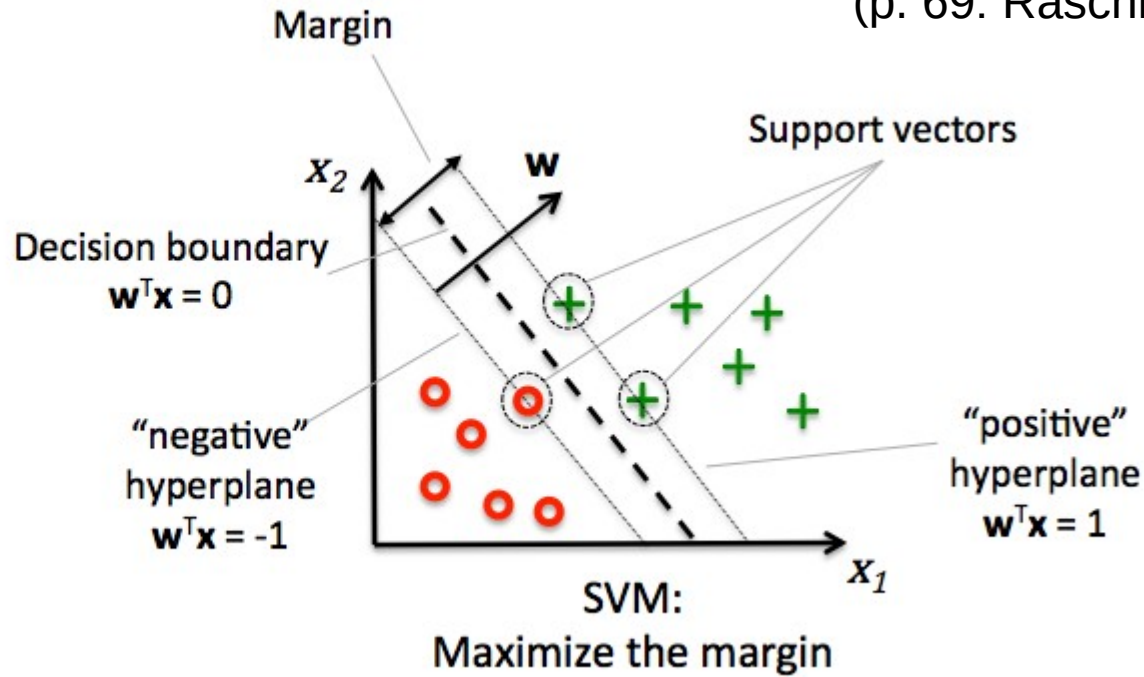
different lines that can separate the two data



fx er den her margin  
dårlig, fordi den ligger  
tæt på den ene data  
og langt væk fra den  
anden (vi  
skal have den til at  
ligge i midten)



(p. 69: Raschka, 2015)



$$w_0 + w^T x_{\text{pos}} = 1 \quad (1)$$

$$w_0 + w^T x_{\text{neg}} = -1 \quad (2)$$

**QUESTION:** What is the subtraction of equation (2) from equation (1) equal to on the figure?



# Subtraction

$$\mathbf{w}^T (\mathbf{x}_{\text{pos}} - \mathbf{x}_{\text{neg}}) = 2$$

normalizing by  $\mathbf{w}$ 's length:  $\|\mathbf{w}\| = \sqrt{\left(\sum_{j=1}^m w_j^2\right)}$

$$\dots \text{ we get: } \mathbf{w}^T \frac{(\mathbf{x}_{\text{pos}} - \mathbf{x}_{\text{neg}})}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$$

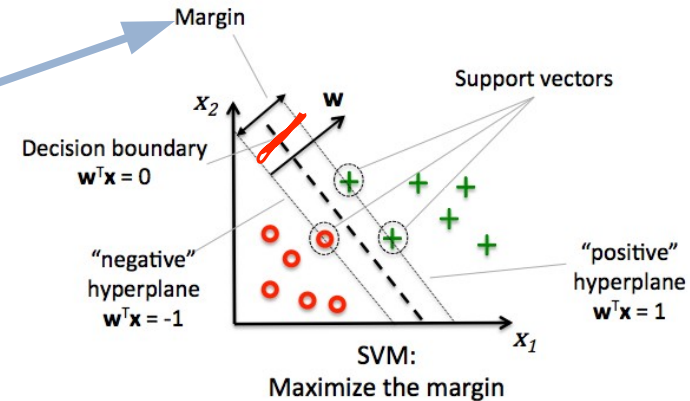
the larger width margin between the data and decision boundary = the better classification This will optimise the width of the margin

The greater width the better the training set is for predicting the test set... but this comes with problems = fails to converge

So we want to maximize:  $\frac{2}{\|\mathbf{w}\|}$

under the constraints:

$$w_0 + \mathbf{w}^T \mathbf{x}^{(i)} \geq 1 \text{ if } y^{(i)} = 1$$
$$w_0 + \mathbf{w}^T \mathbf{x}^{(i)} < -1 \text{ if } y^{(i)} = -1$$



(p. 69: Raschka, 2015)

In practice, the following is minimized:  $\frac{1}{2} \|\mathbf{w}\|^2$

This faces a problem that the  
**Perceptron** also faces

Which?

SOLUTION: introducing a *slack*  
variable  $\xi$  ( $x_i$ )

# Introducing $\xi$

$$w_0 + \mathbf{w}^T \mathbf{x}_{\text{pos}} = 1 - \xi^{(i)}$$

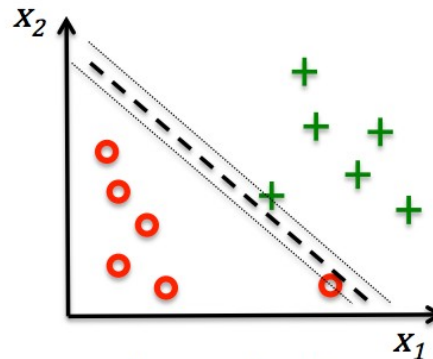
$$w_0 + \mathbf{w}^T \mathbf{x}_{\text{neg}} = -1 + \xi^{(i)}$$

remember C is lemada for logistic regression

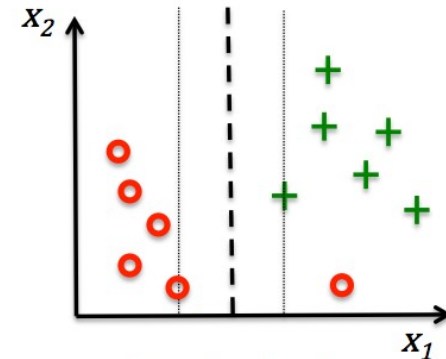
now we need to minimize:  $\frac{1}{2} \|\mathbf{w}\|^2 + C \left( \sum_i^n \xi^{(i)} \right)$  (p. 72: Raschka, 2015)

to maximise the margin

Large C: narrow margin –  
penalises errors harshly  
Small C: wide margin –  
penalises errors mildly



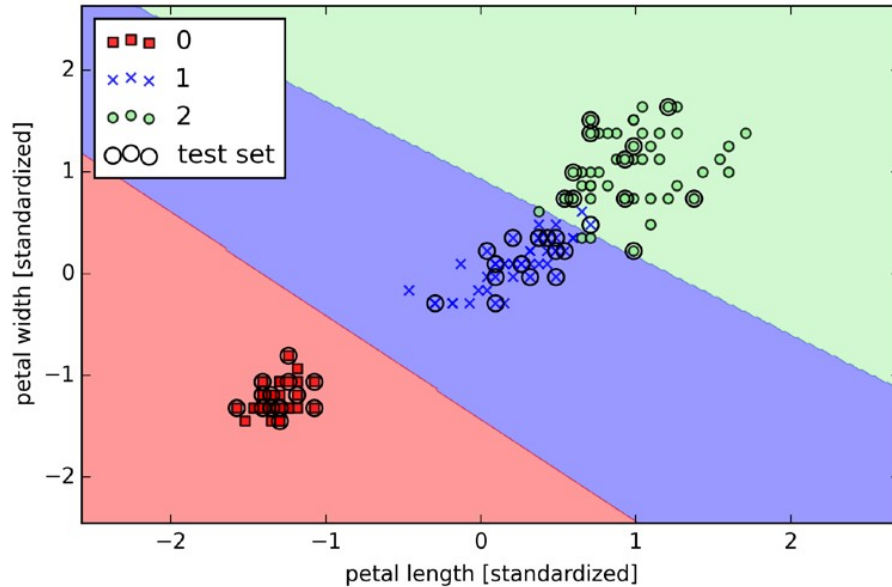
Large value for  
parameter C



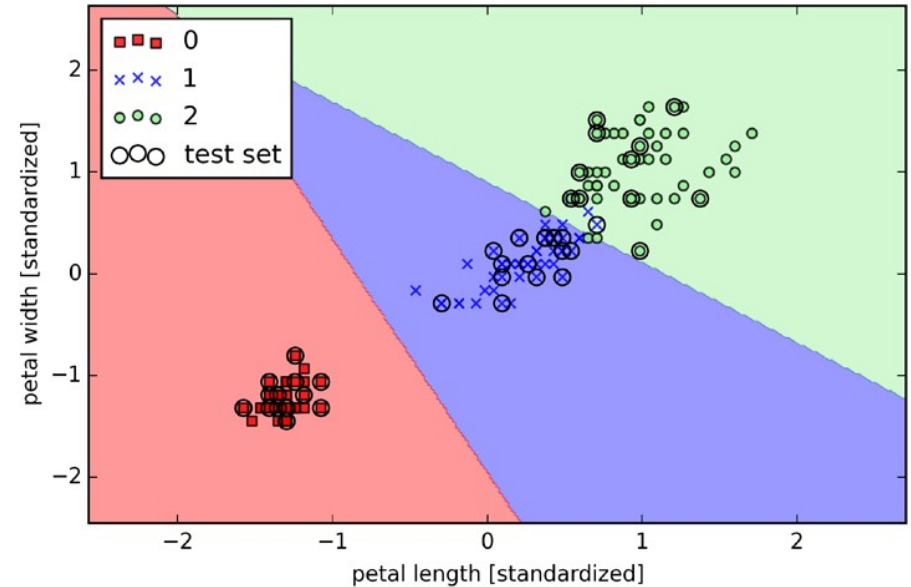
Small value for  
parameter C

better :))

# Comparison to logistic regression



Support Vector Machine (p.763,  
Raschka 2015)



Logistic regression (p. 63,  
Raschka 2015)



## Logistic regression versus SVM

In practical classification tasks, linear logistic regression and linear SVMs often yield very similar results. Logistic regression tries to maximize the conditional likelihoods of the training data, which makes it more prone to outliers than SVMs. The SVMs mostly care about the points that are closest to the decision boundary (support vectors). On the other hand, logistic regression has the advantage that it is a simpler model that can be implemented more easily. Furthermore, logistic regression models can be easily updated, which is attractive when working with streaming data.

logistic regression are more prone to outliers..

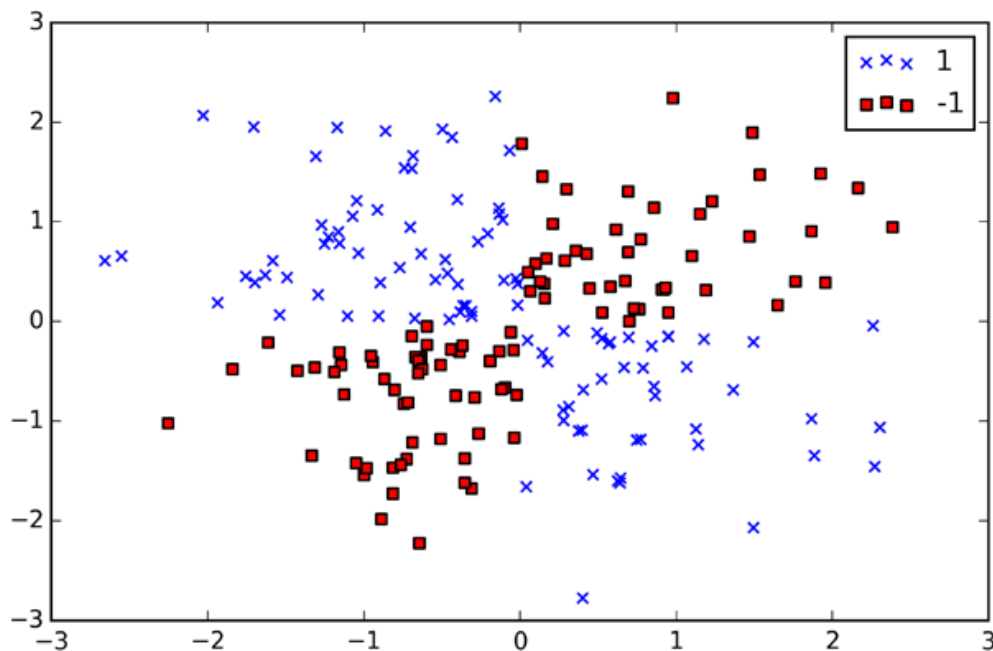
(p. 74: Raschka, 2015)

# Live coding

## EXAMPLE\_00



# A brief look at a non-linear problem

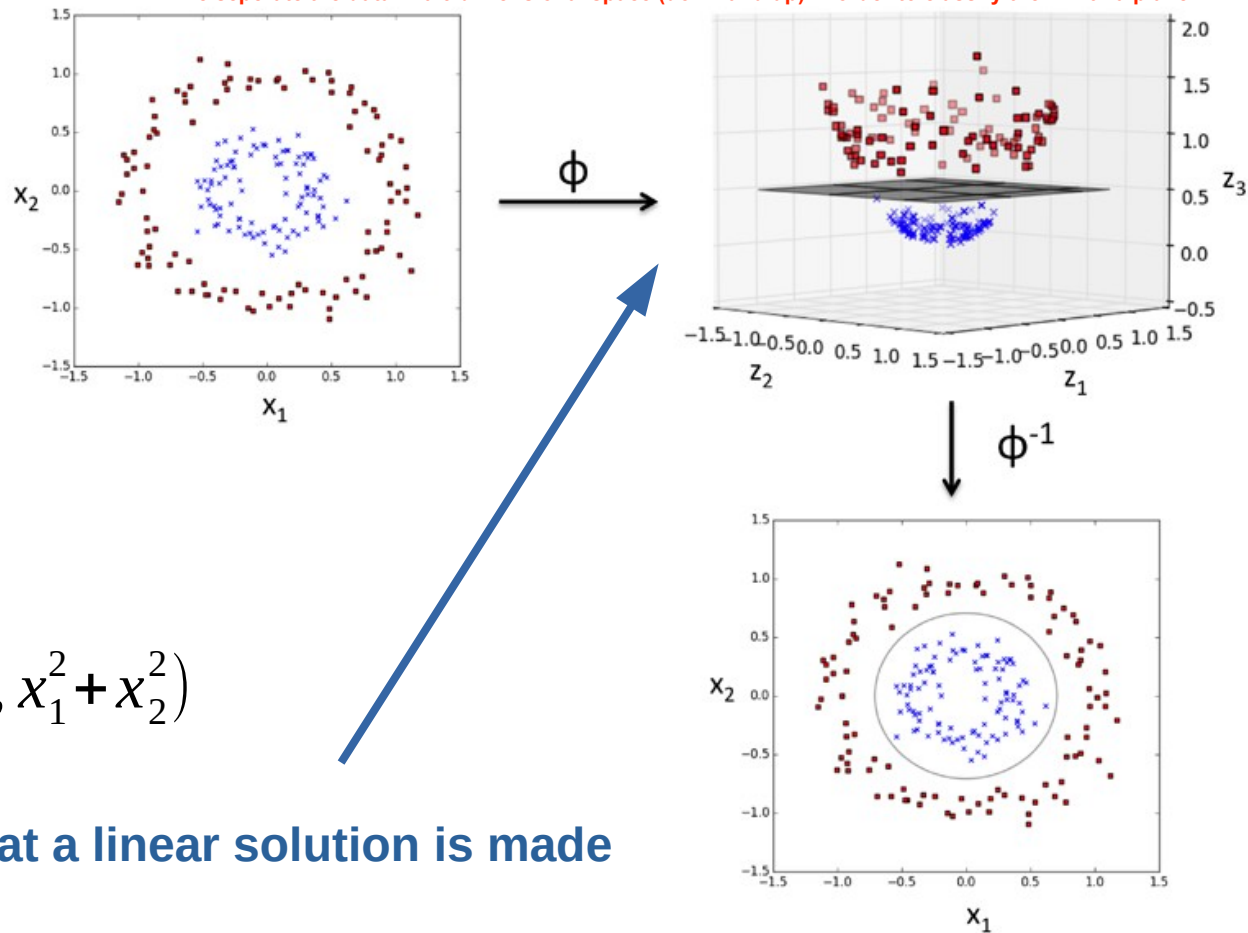


(p. 75: Raschka, 2015)

# Map onto higher-dimensional space

$$\phi(\cdot)$$

we separate the data in a 3 dimensional space (down and up) in order to classify them with a plane



$$\phi(x_1, x_2) = (z_1, z_2, z_3) = (x_1, x_2, x_1^2 + x_2^2)$$

**Note that a linear solution is made**

$$\phi(x_1, x_2) = (z_1, z_2, z_3) = (x_1, x_2, x_1^2 + x_2^2)$$

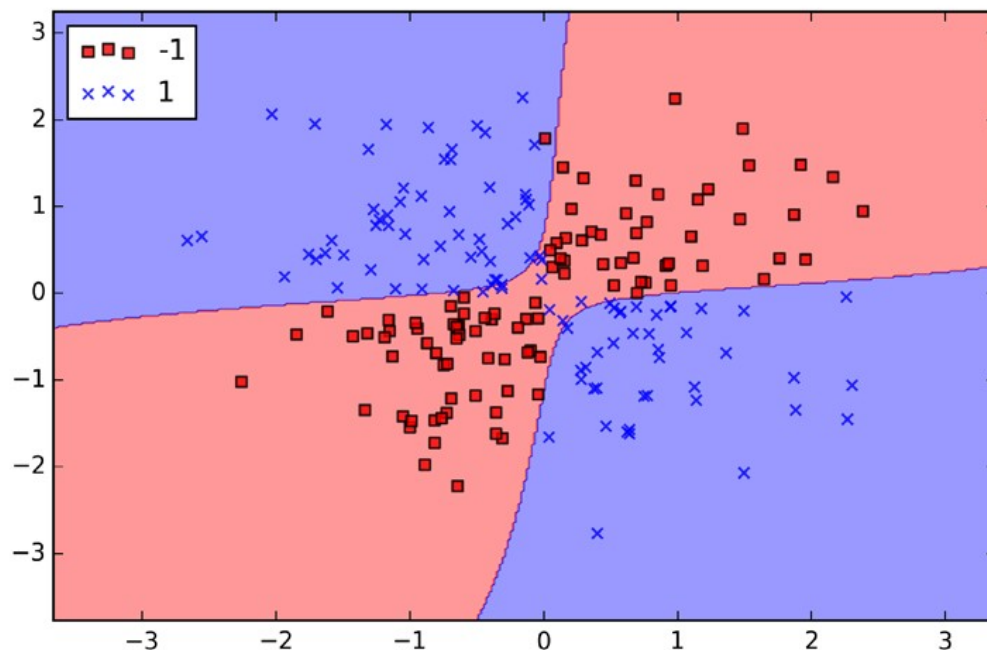
**Creating the higher dimensions  
can be computationally expensive**

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \phi(\mathbf{x}^{(i)})^T \phi(\mathbf{x}^{(j)})$$

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = e^{(-\gamma \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2)}$$

$$(\gamma = \frac{1}{2\sigma^2}, \text{ also called the precision})$$

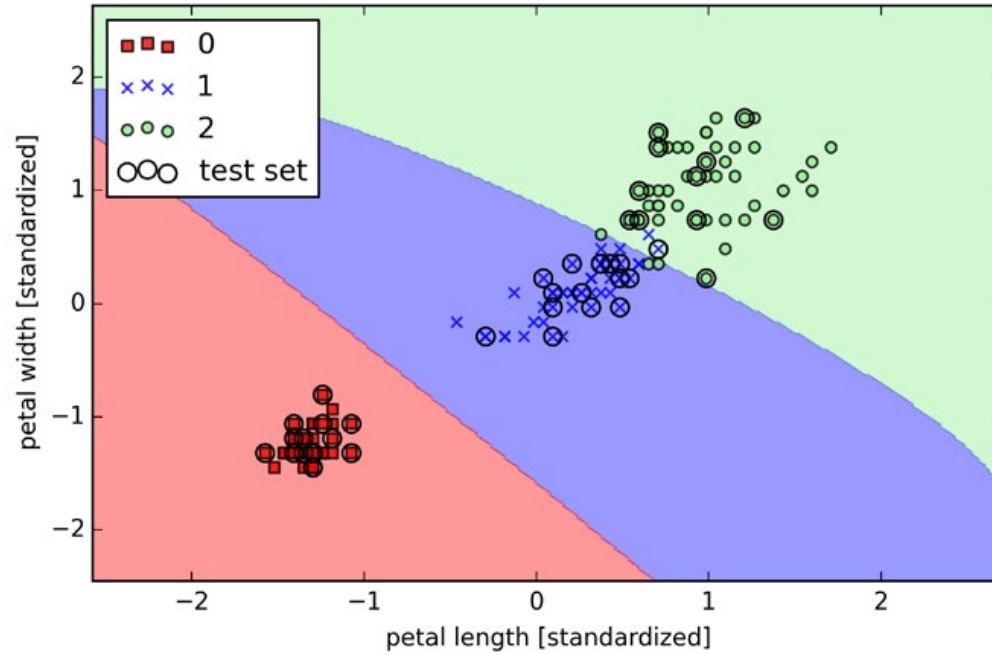
# Non-linear decision boundaries



(p. 78: Raschka, 2015)

# Low $\gamma$ - soft boundary

low gamma = high precision

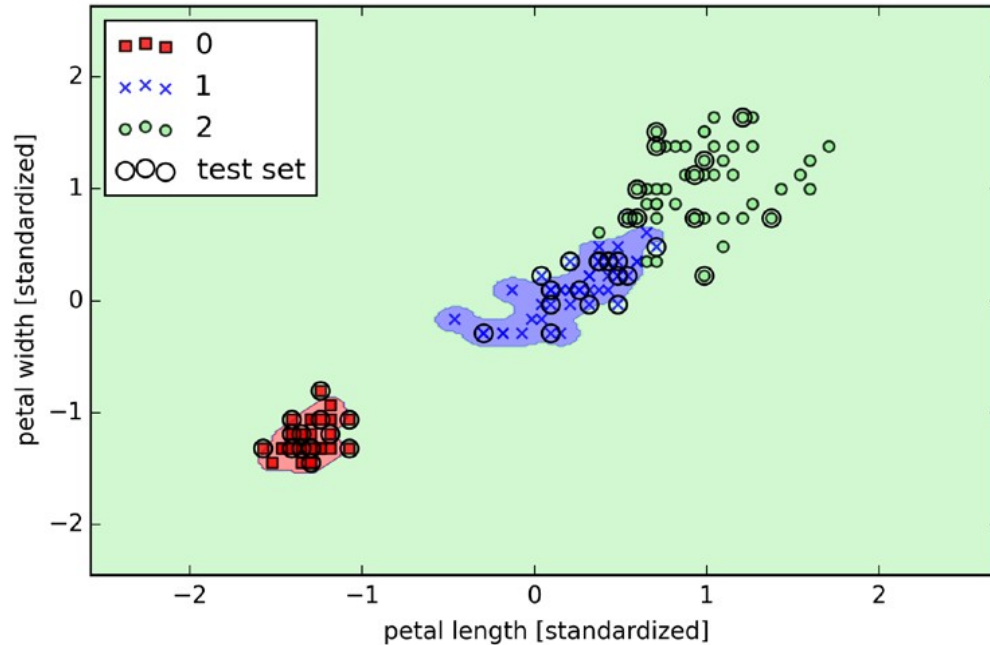


(p. 79: Raschka, 2015)

# High $\gamma$ - tight boundary

high gamma = low precision

its a really good fit to your training data, but not good for test data



(p. 80: Raschka, 2015)

# Live coding

## EXAMPLE\_00

# Did you learn?

*Logistic regression (machine learning)*

- 1) Understanding of how logistic regression can be adapted to a classification framework
- 2) Understanding the idea of a Support Vector Machine
- 3) Getting acquainted with how Support Vector Machines can solve non-linear problems



# References

- Raschka, S., 2015. Python Machine Learning. Packt Publishing Ltd.