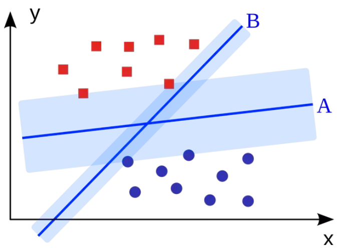
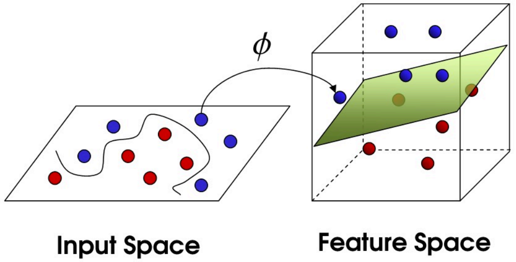
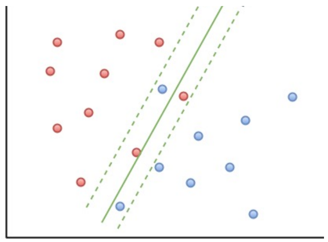
Section 9. Support vector machines and kernel methods

- Support Vector Machines
- Lagrange multiplier
- Kernel Methods
- Regularization







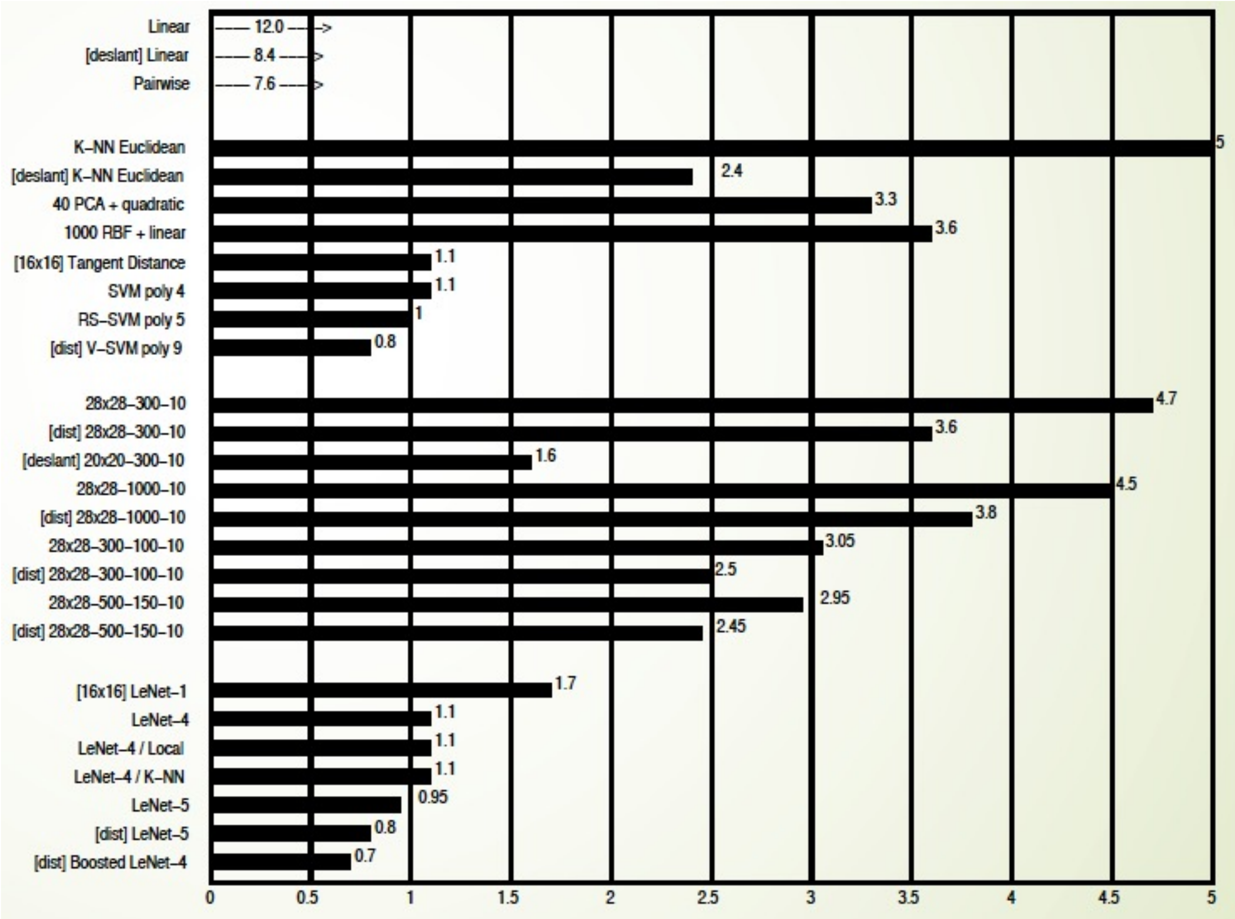
➢ Support Vector Machines (SVM)

SVM was Developed at AT&T Bell Laboratories by Vladimir Vapnik with colleagues in 1994.

- Support vector machine is one of the most popular machine learning methodologies.
- Empirically successful, with well developed theory.
- One of the best off-the-shelf methods.
- We mainly address classification.

Simple SVM performs as well as Multilayer Convolutional Neural Networks which need careful tuning (LeNets) Second dark era for NN: 2000s

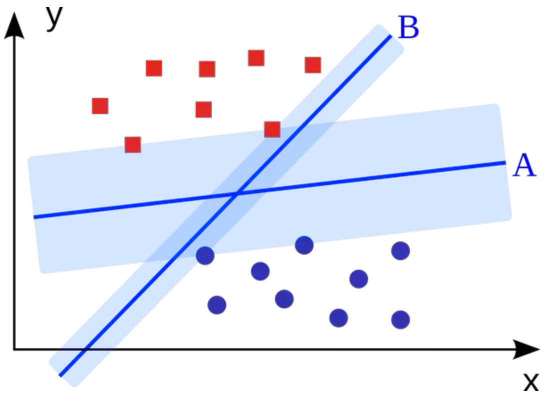
MNIST Dataset Test Error: SVM vs. CNN



LeCun et al. 1998

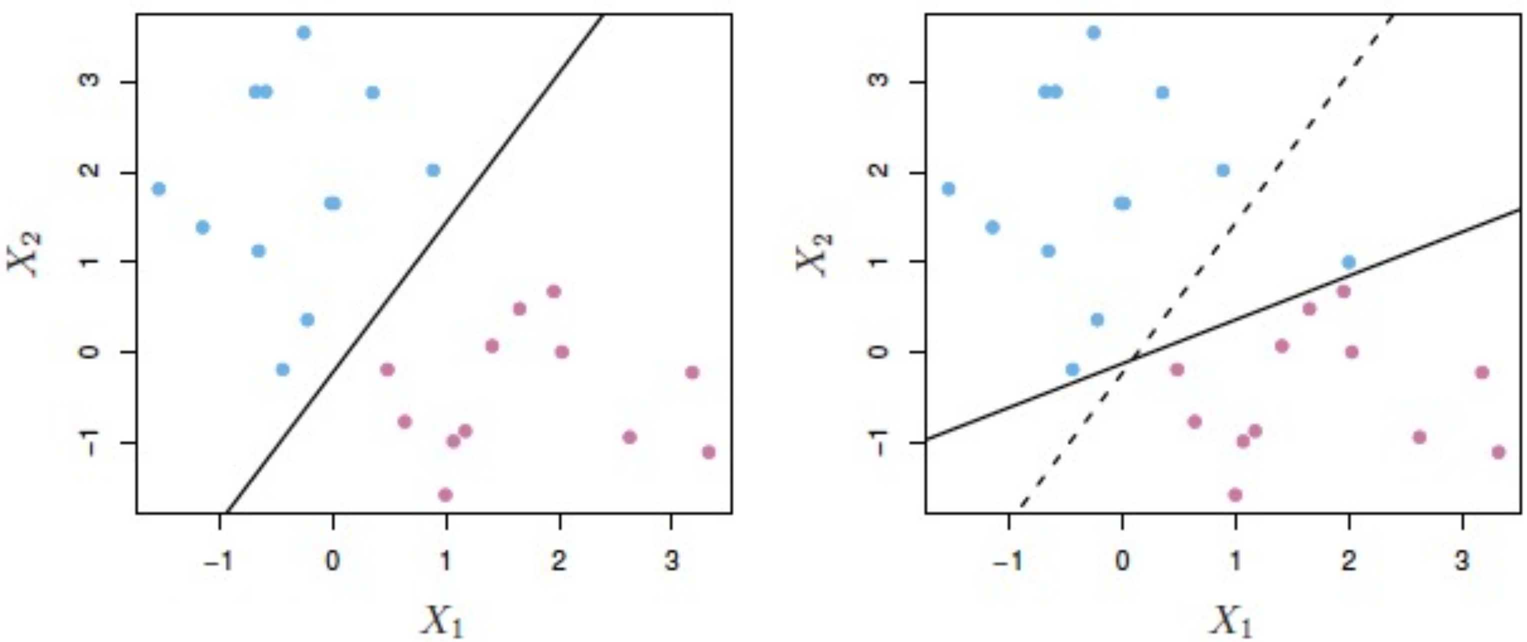
➢ **Support Vector Machines (SVM)** for binary classification. (Max-Margin Classifier)

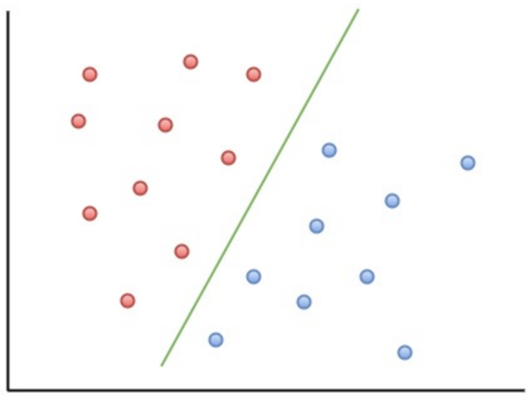
Assume the datasets are *linearly separable.*

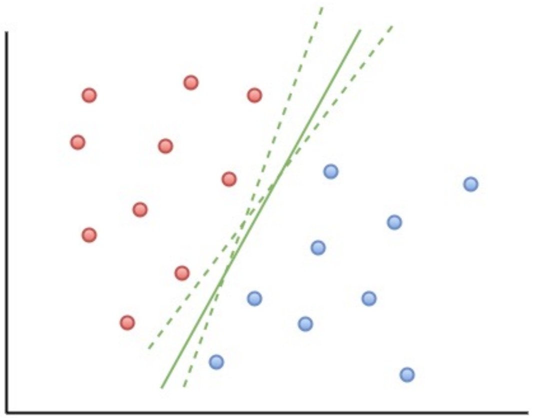


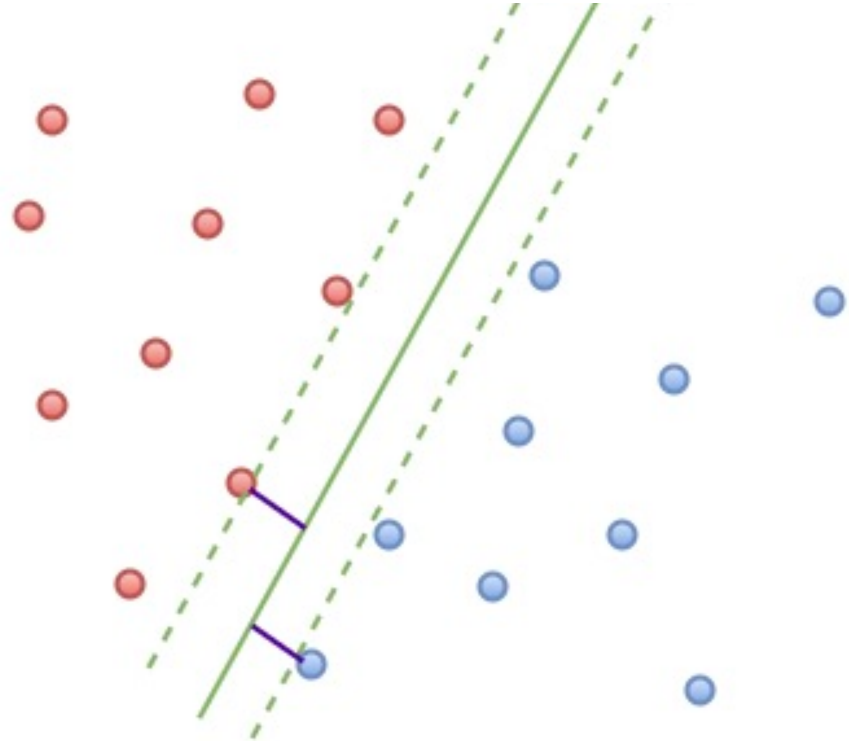
- **Maximal margin hyperplane:** the separating hyperplane that optimal separating hyperplane is **farthest** from the training observations.
- The separating hyperplane such that the **minimum** distance of any training point to the hyperplane is the largest.
- Creates a **widest gap** between the two classes.
- Points on the boundary hyperplane, those with smallest distance to the max margin hyperplane, are called **support vectors**. They support the maximal margin hyperplane in the sense vector that if these points were moved slightly then the maximal margin hyperplane would move as well.

- Note that margin M > 0 is the half of the width of the strip separating the two classes.
- The eventual solution, the max margin hyperplane is determined by the support vectors.
- If x_i on the correct side of the trip varies, the solution would remain same.
- The max margin hyperplane may vary a lot when the support vectors vary. (high variance)



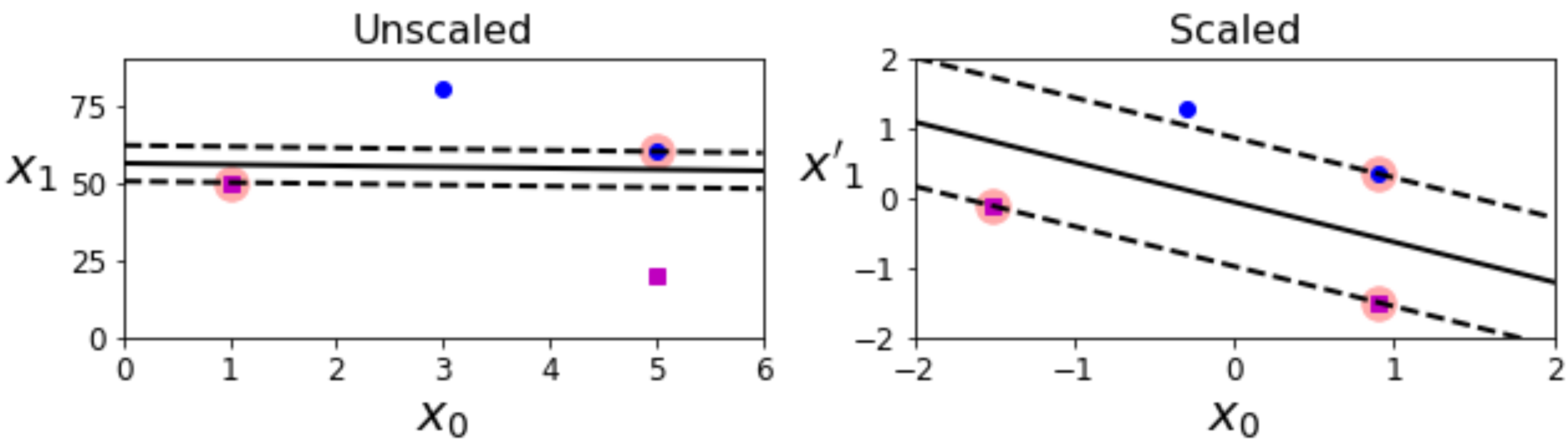




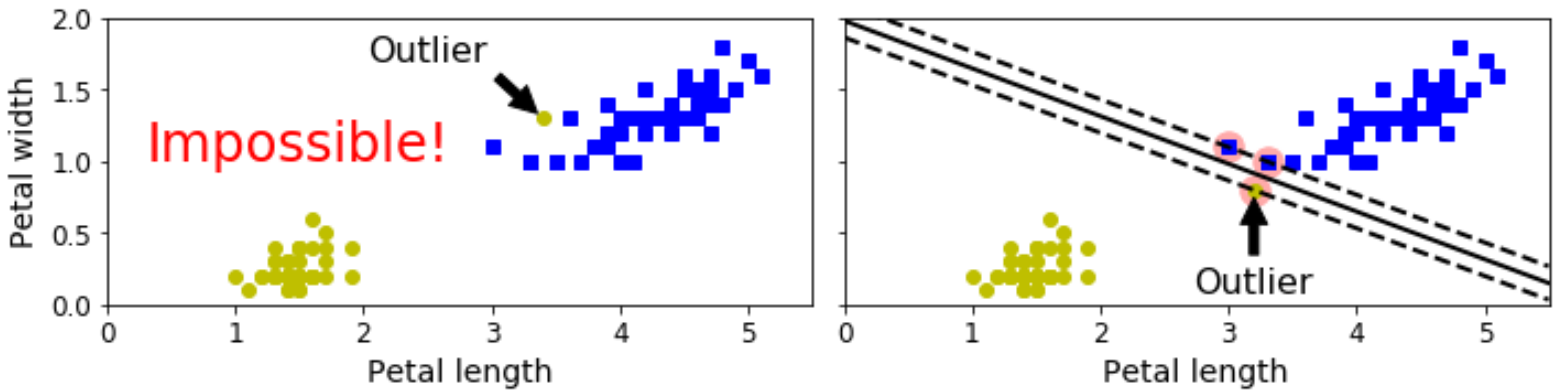


- Lagrange Multipliers (optimization)

Sensitivity to feature scales



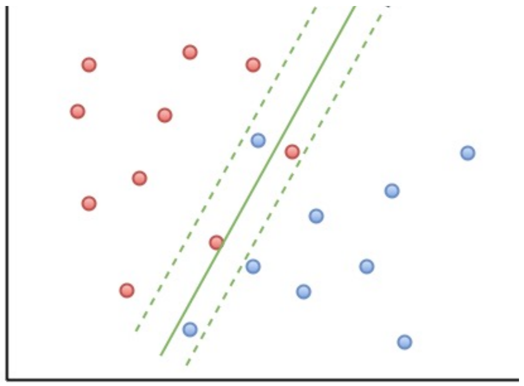
Outlier:



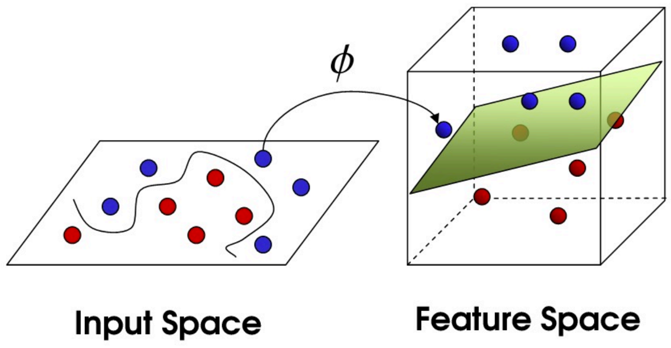
➢ Non-separable cases:

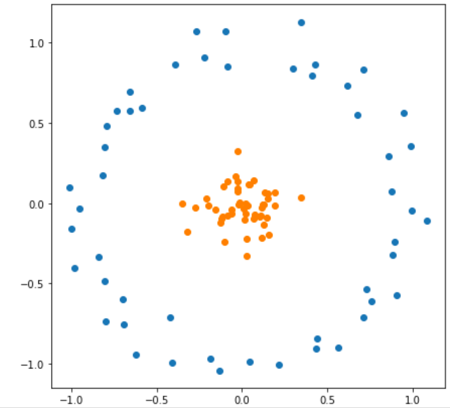
- In general, the two classes are usually not separable by any hyperplane.
- Even if they are, the max margin may not be desirable because of its high variance, and thus possible over-fit.
- The generalization of the maximal margin classier to the non-separable case is known as the support vector classier.
- Use a soft-margin in place of the max margin.
- Soft-margin classier (support vector classier) allow some violation of the margin: some can be on the wrong side of the margin (in the river) or even wrong side of the hyperplane.

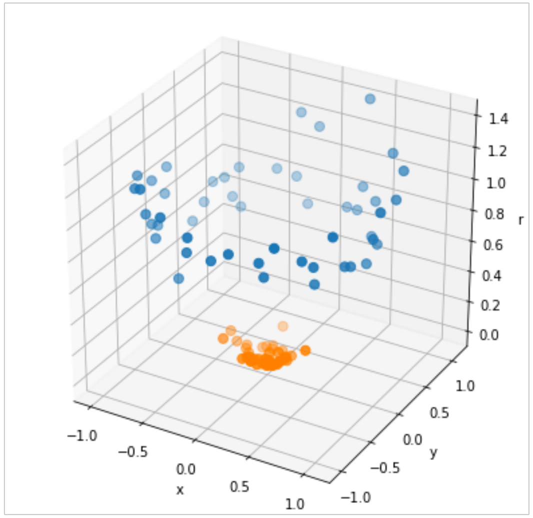
If the datasets are not *linearly separable, or less sensitive to outliers.*



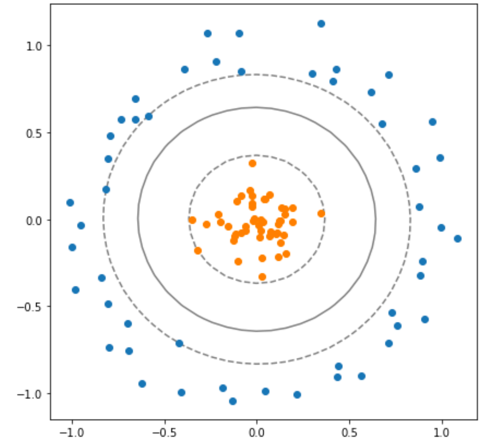
- The kernel method







x
y
r



scikit-learn

https://scikit-learn.org/stable/modules/svm.html#svm

Support Vector Machine - Regression (SVR)

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin).

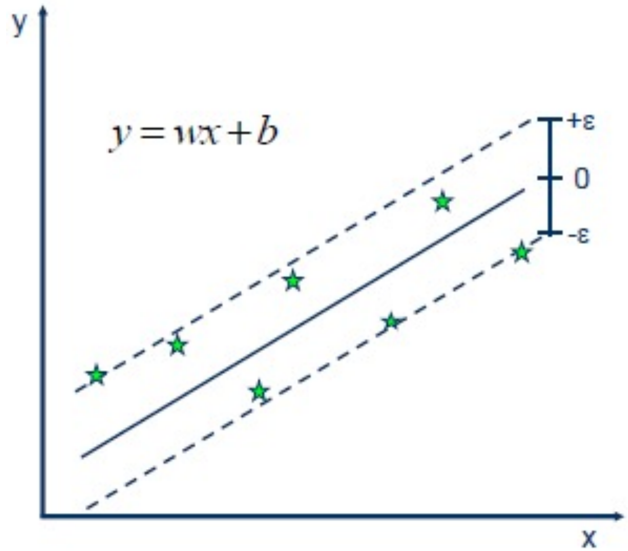
First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities.

In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem.

But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration.

However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

Support Vector Machine - Regression (SVR)

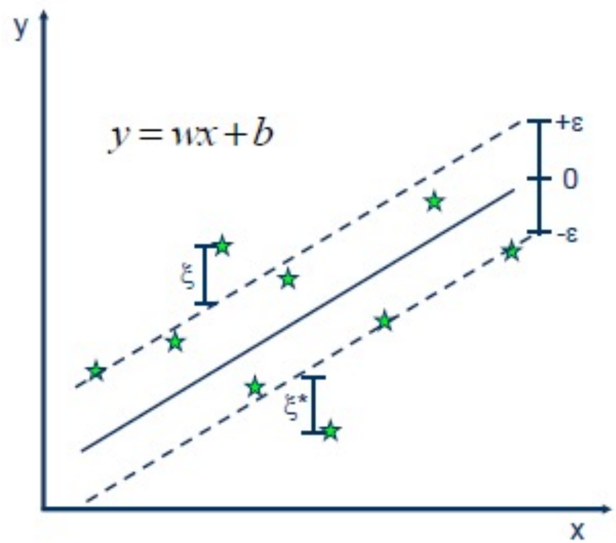


- Solution:

$$\min \frac{1}{2}\|w\|^2$$

- Constraints:

$$y_i - wx_i - b \le \varepsilon$$
$$wx_i + b - y_i \le \varepsilon$$



- Minimize:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)$$

- Constraints:

$$y_i - wx_i - b \le \varepsilon + \xi_i$$
$$wx_i + b - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0$$

Linear SVR

$$y = \sum_{i=1}^{N} \left(\alpha_i - \alpha_i^*\right) \cdot \langle x_i, x \rangle + b$$

Non-linear SVR

The kernel functions transform the data into a higher dimensional feature space to make it possible to perform the linear separation.

$$y = \sum_{i=1}^{N} \left(\alpha_i - \alpha_i^*\right) \cdot \langle \varphi(x_i), \varphi(x) \rangle + b$$

$$y = \sum_{i=1}^{N} \left(\alpha_i - \alpha_i^*\right) \cdot K(x_i, x) + b$$