SVM and SVR

Dr. Mohamed Elshenawy
mmelshenawy@gmail.com

1

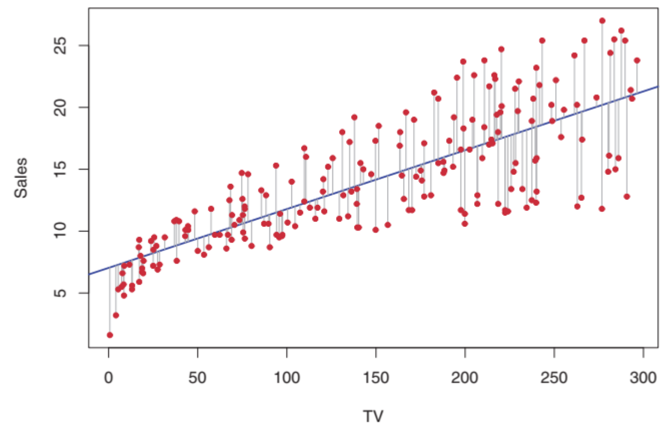
In Previous Lectures

- Linear Regression
- Polynomial Regression
- Training and test error
- Bias and Variance Tradeoff
- Classification
- KNN
- Logistic Regression

2

Review- Regression

- A simple, useful and widely used tool for predicting a quantitative response.
- It serves as a starting point for more complex approaches



An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

3

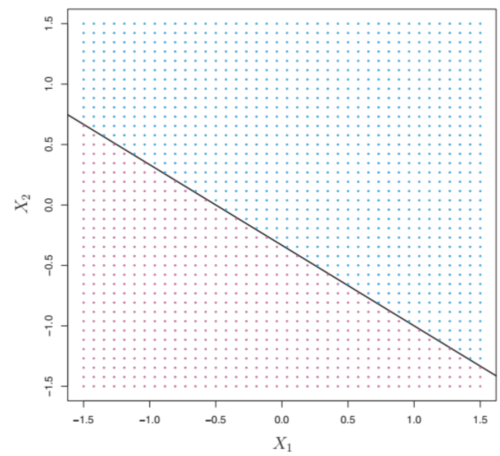
SVR - Key Idea

- Instead of minimizing the observed training error, Support Vector Regression (SVR) attempts to minimize the generalization error bound so as to achieve generalized performance. How?

4

Basic Notions - Hyperplane

- In a p-dimensional space, a hyperplane is a flat affine subspace of dimension p−1.
- In two dimensions, a hyperplane is a flat one-dimensional subspace (a line).
- In figure, the separating hyperplane is $1 + 2x_1 + 3x_2 = 0$

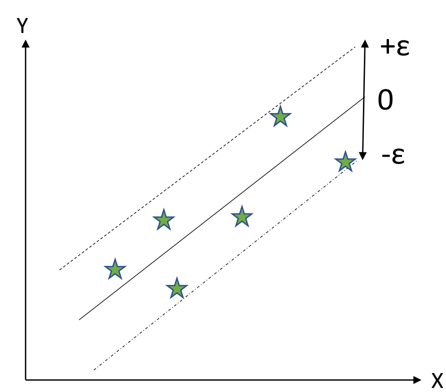


An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

5

Support Vector Regression (Epsilon-Support Vector Regression)

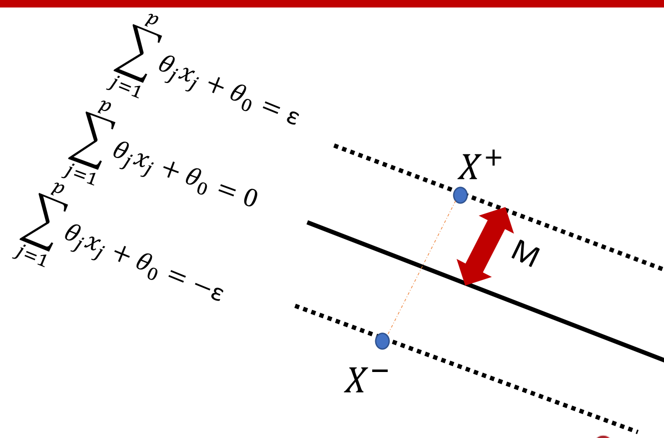
- Using a margin of tolerance.
- Our goal is to find a function $f(x)$ that has at most ε deviation from the targets y_i for all the training data, and at the same time is as flat as possible.
- In other words, we do not care about errors as long as they are less than ε but will not accept any deviation larger than this.
- $f(x)$ is the maximal margin hyperplane



6

Construction of the Maximal Margin Hyperplane

- Choose the values of the model parameters θ that maximize the margin M: $\max_{\theta_0,\theta_1,\ldots,\theta_p} M$
- The vector θ is orthogonal to the maximal margin hyperplane (why?)
- We can say
$$X^+ = \lambda\theta + X^-$$
- (i.e. we can get from X^- to X^+ by moving in the direction of θ for an unknown distance deterimined by λ)



7

- A=(-3,10), B = (b1, b2)
- -3b1+10b2 =0
- b2 =0.3 b1

8

Construction of the Maximal Margin Hyperplane (Cont.)

- We know that at X^+

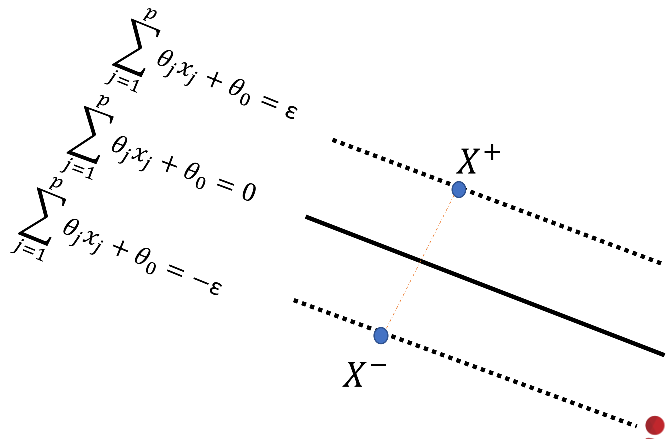
$$\sum_{j=1}^{p} \theta_j x_j + \theta_0 = \varepsilon$$

- Using the augmented form

$$\theta^T X^+ = \varepsilon$$

- But, $X^+ = \lambda\theta + X^-$
- $\theta^T(\lambda\theta + X^-) = \varepsilon$
- $\lambda\theta^T\theta + \theta^T X^- = \varepsilon$
- But we know $\theta^T X^- = -\varepsilon$
- $\lambda\theta^T\theta - \varepsilon = \varepsilon$

$$\lambda = \frac{2\varepsilon}{\theta^T\theta} = \frac{2\varepsilon}{\|\theta\|^2}$$



Construction of the Maximal Margin Hyperplane (Cont.)

- $M = \frac{1}{2}(X^+ - X^-) = \frac{1}{2}\|\lambda\theta\| = \frac{1}{2}\frac{2\|\theta\|}{\|\theta\|^2} = \frac{1}{\|\theta\|}$
- To maximize M, To maximize, we can maximize λ

- That is minimize $\frac{1}{2}\|\theta\|^2$, subject to constraints

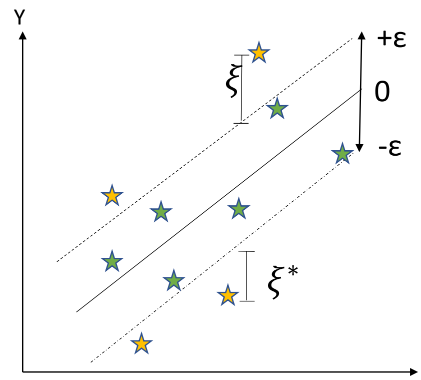
$$\sum_{j=1}^{p} \theta_j x_j + \theta_0 \leq \varepsilon$$

$$\sum_{j=1}^{p} \theta_j x_j + \theta_0 \geq -\varepsilon$$

- Constrained Optimization Problem (solved using Lagrange multipliers)
- Solution of this optimization equation is out of the scope of this course

Support Vector Regression – Using Soft Margin

- The assumption we had before it that there is a function that approximates all pairs (x_i, y_i) with ε precision.
- Sometimes, this is not the case.
- We may want to allow for some errors (How?).
- One can introduce slack variables ξ_i , $\xi_i^*$to cope with otherwise infeasible constraints of the optimization problem.



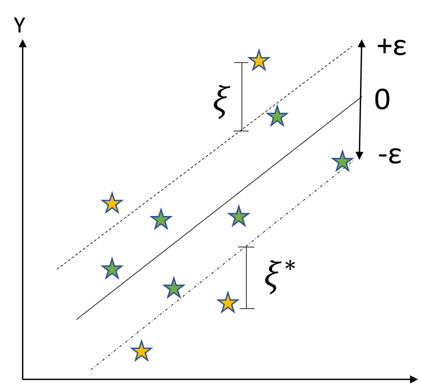
11

Support Vector Regression – Using Soft Margin (Cont.)

- minimize $\frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$
- N: number of examples in the dataset
- Subject to

$$\sum_{j=1}^{p}\theta_j x_j + \theta_0 \leq \varepsilon + \xi_i$$
$$\sum_{j=1}^{p}\theta_j x_j + \theta_0 \geq -(\varepsilon + \xi_i^*)$$
$$\xi_i, \xi_i^* \geq 0$$

- Again, a constrained optimization problem



12

Tuning the Hyperparameter C

- The constant $C > 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated.
- Increasing C, increases bias, reduces the variance (we become more tolerant of violations to the margin, margin will widen
- Decreasing C, reduces bias, increases variance (we become less tolerant of violations to the margin , margin narrows)
- C is generally chosen via cross-validation.

13

Solution of the constrained optimization problem

$$\max_{\alpha_i,\alpha_i^*}\left\{\frac{1}{2}\sum_{i,j=0}^{N}(\alpha_i-\alpha_i^*)(\alpha_i - \alpha_j^*)\langle X^{(i)},X^{(j)}\rangle-\varepsilon\sum_{i=1}^{N}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{N}y_i(\alpha_i-\alpha_i^*)\right\}$$

Subject to $\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)=0$ and $\alpha_i,\alpha_i^*\in[0,C]$

14

Solution of the constrained optimization problem

$$\theta = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)\, X^{(i)}$$

Linear SVR

$$f(X) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) . \langle X^{(i)}, X \rangle + b$$

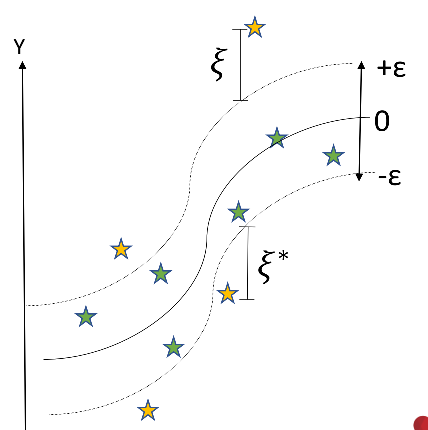
$\langle X^{(i)}, X^{(j)} \rangle$ is the dot product between the two observations $X^{(i)}, X^{(j)}$

- The solution has 2N parameters $\alpha_1, \alpha_1^* \alpha_2, \alpha_2^* \dots . \alpha_n, \alpha_n^*$
- The majority of these parameters will equal zero. Few (at support vectors) will have a value greater than 0
- The complexity of a function's representation by SVs is independent of the dimensionality of the input space X, and depends only on the number of SVs.

15

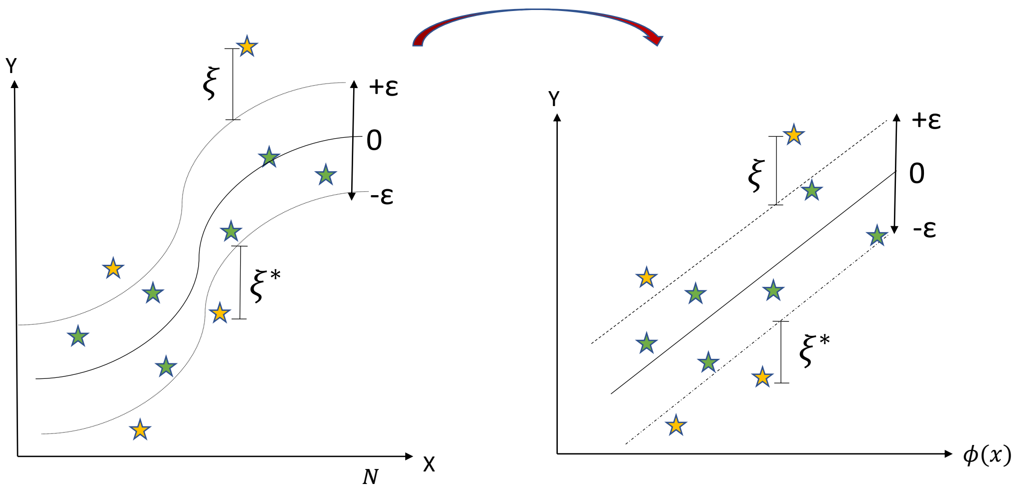
How about non-linear relationships

- Idea: we can transform the training data x_i by a map $\phi : \mathcal{X} \rightarrow \mathcal{F}$ into some feature space $\mathcal{F}$ in which we can do the linear separation.



16

Transformation



$$f(X) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) . \langle \phi(X^{(i)}), \phi(X) \rangle + b$$

17

Transformation - Implicit mapping via kernels

$$f(X) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) . \langle \phi(X^{(i)}), \phi(X) \rangle + b$$

- Knowing $\phi(X)$ is not easy.
- Do we need to know $\phi(X)$ explicitly?
- the SV algorithm only depends on dot products between patterns x_i. Hence it suffices to know the kernel function $k(x, x') := \langle \phi(x), \phi(x') \rangle$ rather than $\phi(X)$ explicitly
- Kernel: function that quantifies the similarity of two observations.
- There are conditions for kernel functions. The discussion of which is out of the scope of this course

18

Common Kernel

- Polynomial kernel of degree d (d>1)

$$K\left(X^{(i)},X^{(j)}\right)=\left(1+\sum_{l=1}^{p}x_{il}x_{jl}\right)^{d}$$

- Gaussian Radial Basis Function

$$K\left(X^{(i)},X^{(j)}\right)=\exp\left(-\gamma\sum_{l=1}^{p}(x_{il}-x_{jl})^{2}\right)$$

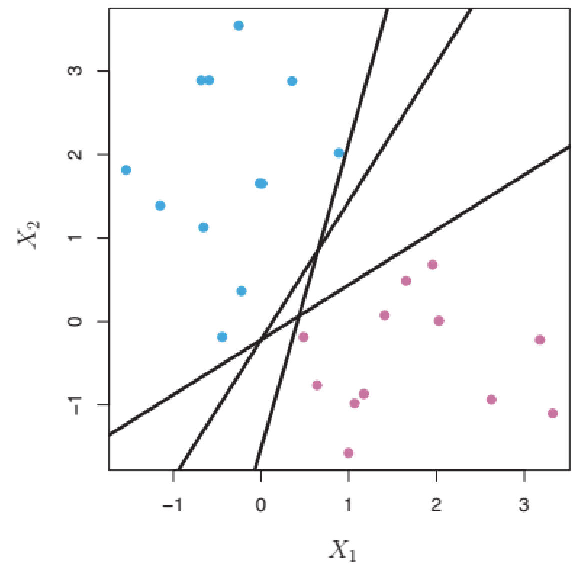
19

DEMO

20

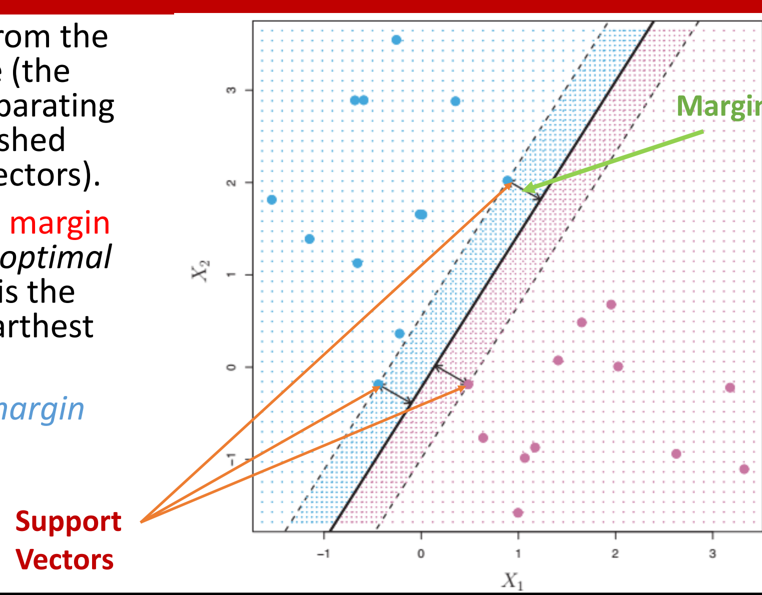
Classification

- Same dataset can be separated with many hyperplanes. Which one shall we choose?



The Maximal Margin Classifier

- Margin: the minimal distance from the observations to the hyperplane (the distance from the solid line (separating hyperplane) to either of the dashed lines (defined by the support vectors).
- A natural choice is the maximal margin hyperplane (also known as the *optimal separating hyperplane*), which is the separating hyperplane that is farthest from the training observations.
- This is known as the *maximal margin classifier.*



Review –Linear Algebra

- If you have a line $-3b_1 + 10b_2 = 0$ then you can say that vector [-3,10] is orthogonal to this line

23

Construction of the Maximal Margin Classifier (Cont.)

- The vector θ is orthogonal to the separating hyperplane
- We can say

$$X^+ = \lambda\theta + X^-$$

(i.e. we can get from X^- to X^+ by moving in the direction of θ for an unknown distance deterimined by λ)

Using the augmented form

$$\theta^T X^+ = 1$$

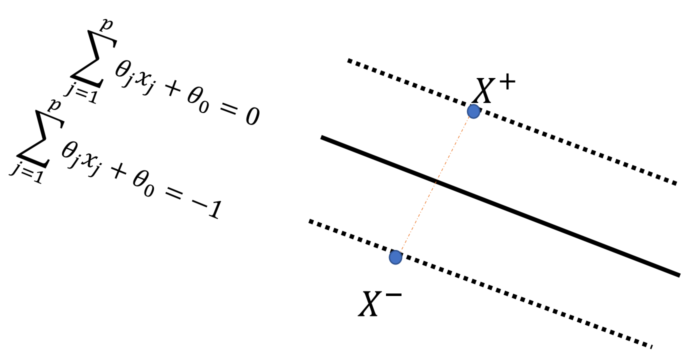
$$\theta^T(\lambda\theta + X^-) = 1$$

$$\lambda\theta^T\theta + \theta^T X^- = 1$$

But we know $\theta^T X^- = -1$

$$\lambda\theta^T\theta - 1 = 1$$

$$\lambda = \frac{2}{\theta^T\theta} = \frac{2}{\|\theta\|^2}$$



24

Construction of the Maximal Margin Classifier (Cont.)

However

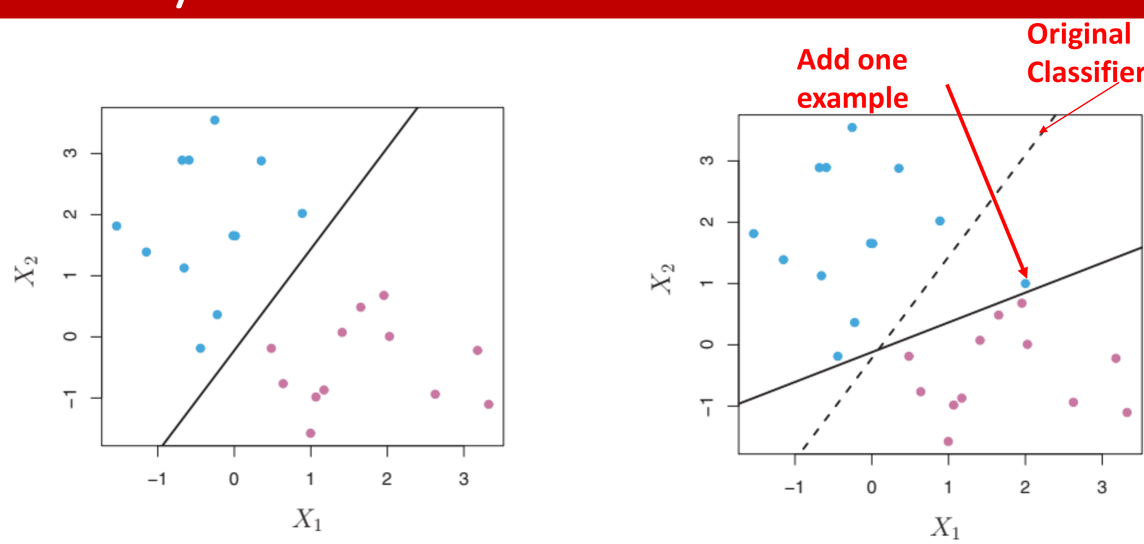
$$M = \frac{1}{2}(X^{+} - X^{-}) = \frac{1}{2}\|\lambda\theta\| = \frac{1}{2}\frac{2\|\theta\|}{\|\theta\|^2} = \frac{1}{\|\theta\|}$$

- To maximize M

$$\max_{\theta_0,\theta_1,\ldots,\theta_p} M$$

- We can minimize $\|\theta\|$subject to constraint $t_i\left(\sum_{j=1}^{p}\theta_j x_j + \theta_0\right) \geq 1$
- That is minimize $\|\theta\|^2$ such that $t_i\left(\sum_{j=1}^{p}\theta_j x_j + \theta_0\right) \geq 1$ for each observation (training example) i.
- The solution of this optimization problem is not in the scope of this course

The Need for Support Vector Classifiers- reduce sensitivity to individual observations



Support Vector Classifier

- Use a soft margin constraint
- Instead of

$$t_i \left(\sum_{j=1}^{p} \theta_j x_j + \theta_0 \right) \geq 1$$

- Let's use

$$t_i \left(\sum_{j=1}^{p} \theta_j x_j + \theta_0 \right) \geq (1 - \epsilon_i)$$

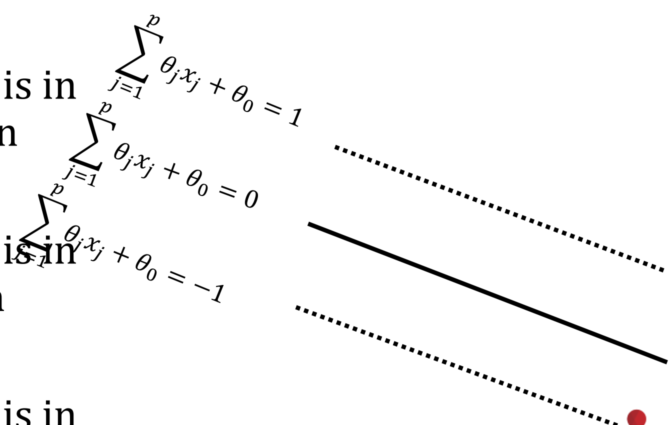
- $\epsilon_1, \epsilon_2, \ldots \epsilon_n$: Slack variables
- Where

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

27

Support Vector Classifier

- ϵ_i is a learnable parameter
- If $\epsilon_i = 0$, the ith observation is in the correct side of the margin
- If $\epsilon_i > 0$, the ith observation is in the wrong side of the margin
- If $\epsilon_i > 1$, the ith observation is in the wrong side of the hyperplane



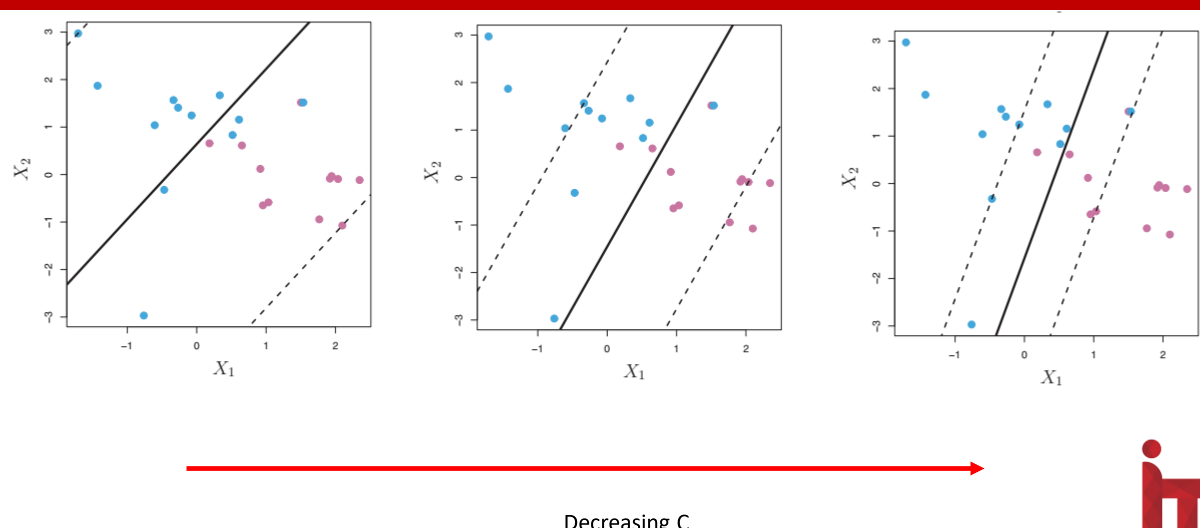
28

Tuning C hyperparameter

- C is a nonnegative tuning parameter.
- C bounds the sum of the ϵ_i 's, and so it determines the number and severity of the violations to the margin (and to the hyperplane) that we will tolerate.
- Increasing C, increases bias, reduces the variance (we become more tolerant of violations to the margin, margin will widen)
- Decreasing C, reduces bias, increases variance (we become less tolerant of violations to the margin, margin narrows)
- C is generally chosen via cross-validation.

29

Using different values of the tuning parameter C



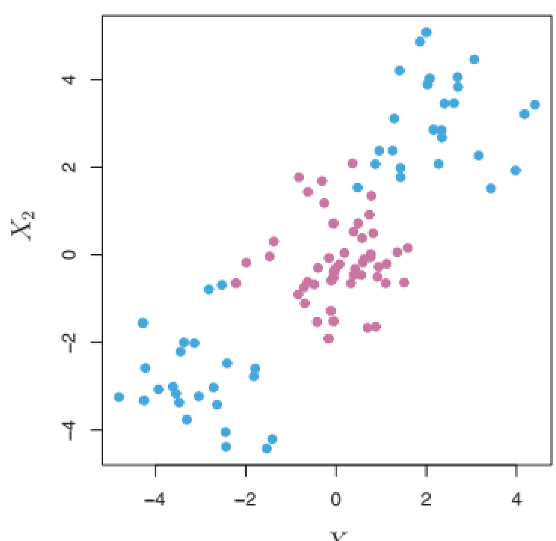
30

Using different values of the tuning parameter C

- For Unbalaced datasets, you way use a different C for each class.
- Assign a smaller C value for classes that have smaller number of examples (the model gives more importance to these classes)

31

Non-linear Decision Boundaries



32

Support Vector Machines (SVM)

- The support vector machine (SVM) is an extension of the support vector classifier that results from enlarging the feature space in a specific way, using kernels.
- The kernel approach allows us to accommodate a non-linear boundary between the classes
- Before discussing kernels, let's discuss the solution in case of the Maximal Margin Classifier

33

Solution to the optimization problem introduced earlier

- Using Lagrange multipliers, you get that

$$\theta = \sum_{i=1}^{n} \alpha_i t^{(i)} X^{(i)}$$

To calculate $\alpha_1, \alpha_2, \dots . \alpha_n, n: number\ of\ training\ examples$

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=0}^{n} t^{(i)} t^{(j)} \alpha_i \alpha_j \langle X^{(i)}, X^{(j)} \rangle$$

$\langle X^{(i)}, X^{(j)} \rangle$ is the dot product between the two observations $X^{(i)}, X^{(j)}$

Subject to

$$\alpha_i \geq 0, \sum_{i=0}^{n} \alpha_i t^{(i)} = 0$$

34

Solution to the optimization problem introduced earlier

- The solution has n parameters $\alpha_1, \alpha_2, \dots . \alpha_n$
- The majority of these parameters will equal zero. Few (at support vectors) will have a value greater than 0

35

Using Kernels

- Instead of the dot product $\langle X^{(i)}, X^{(j)} \rangle$, we use $K(X^{(i)}, X^{(j)})$
- $K\left(X^{(i)}, X^{(j)}\right)$, the kernel function: a function that quantifies the similarity of two observations.
- Take

$$K\left(X^{(i)}, X^{(j)}\right) = \sum_{l=1}^{p} x_{il} x_{jl}$$

- It gives us the support vector classifier (linear)

36

Common Kernels

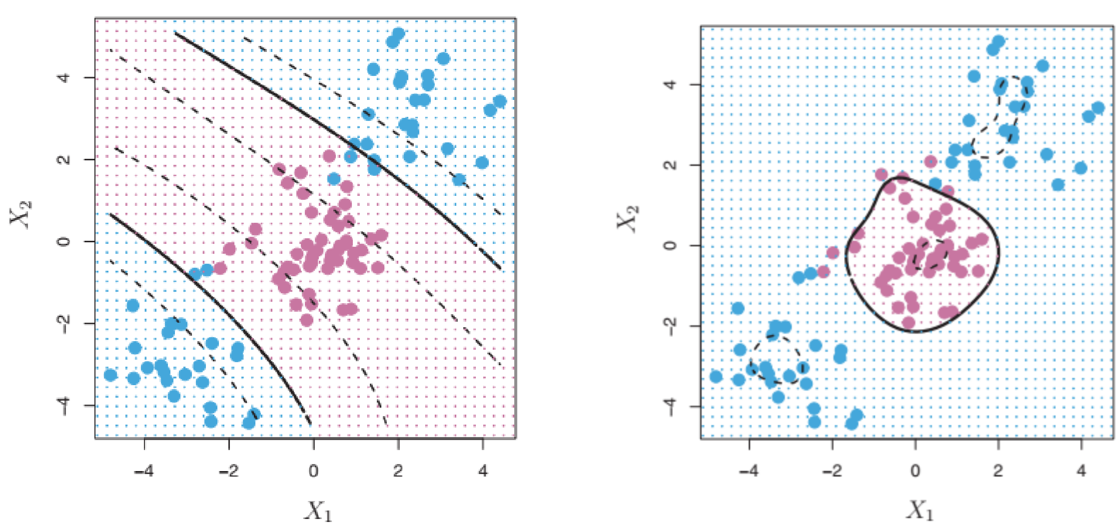
- Polynomial kernel of degree d (d>1)

$$K\left(X^{(i)}, X^{(j)}\right) = \left(1 + \sum_{l=1}^{p} x_{il} x_{jl}\right)^{d}$$

- Radial Kernel

$$K\left(X^{(i)}, X^{(j)}\right) = \exp\left(-\gamma \sum_{l=1}^{p} (x_{il} - x_{jl})^2\right)$$

37



Polynomial kernel of degree 3

Radial Kernel

38