Announcement

- Please submit TA FCQ: open 2/10-2/19
- Quiz 2 closes today 11:59pm
- Kaggle submission by Wed 11:59 pm
- Kaggle report due Thursday 11:59 pm

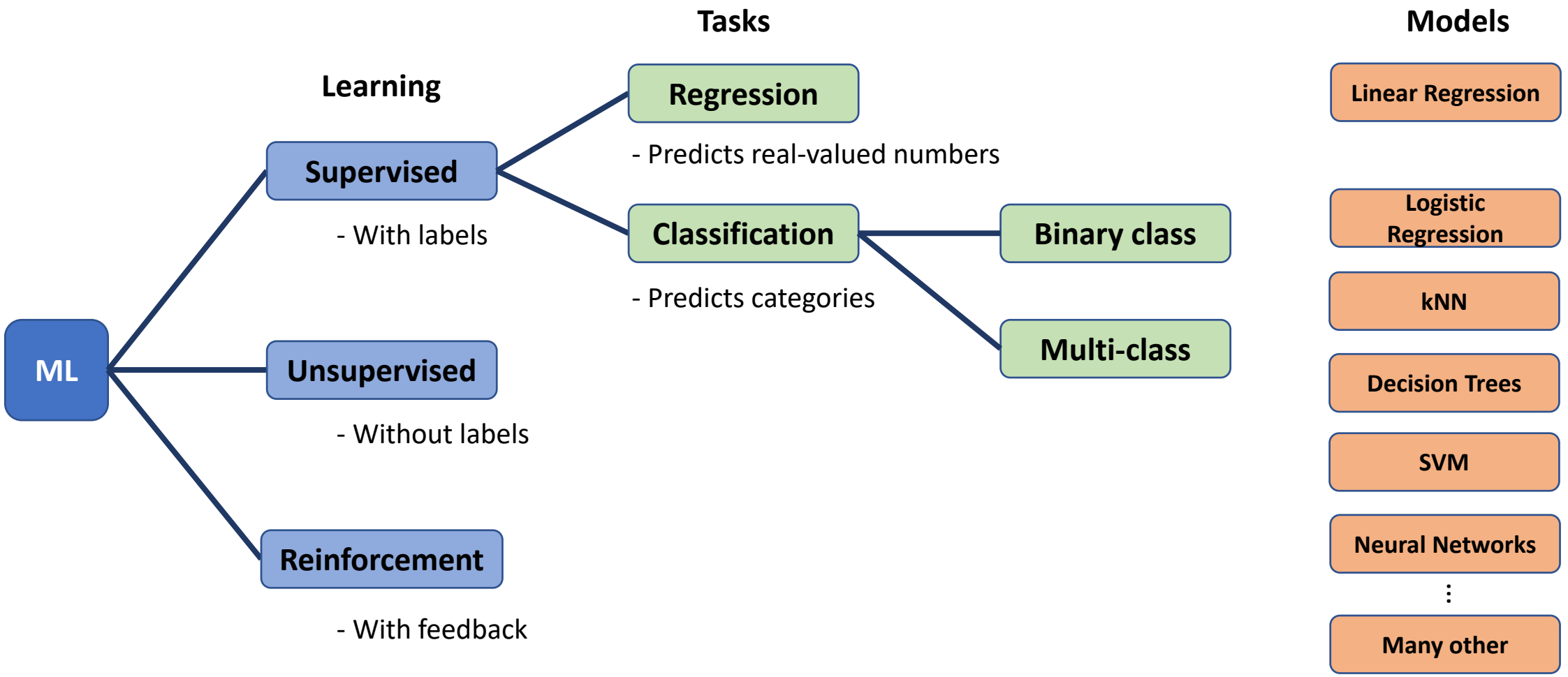
CSCI4622 Machine Learning 2020 Spring

Support Vector Machine

Geena Kim

*Slide contents adopted from ISLR material

Review: Types of machine learning problems



Review: Types of machine learning problems

Models	Hyperparameters	Parameters	Loss or Criteria
Linear Regression			
Logistic Regression			
kNN			
Decision Trees			
SVM			
Neural Networks			

Review: Binary Classification

Yes or No problem

- Creditcard Default
- Fradulant Insurance Claim
- Spam Filtering
- Medical Diagnosis
- Survival Prediction
- Customer Retention
- Image Recognition

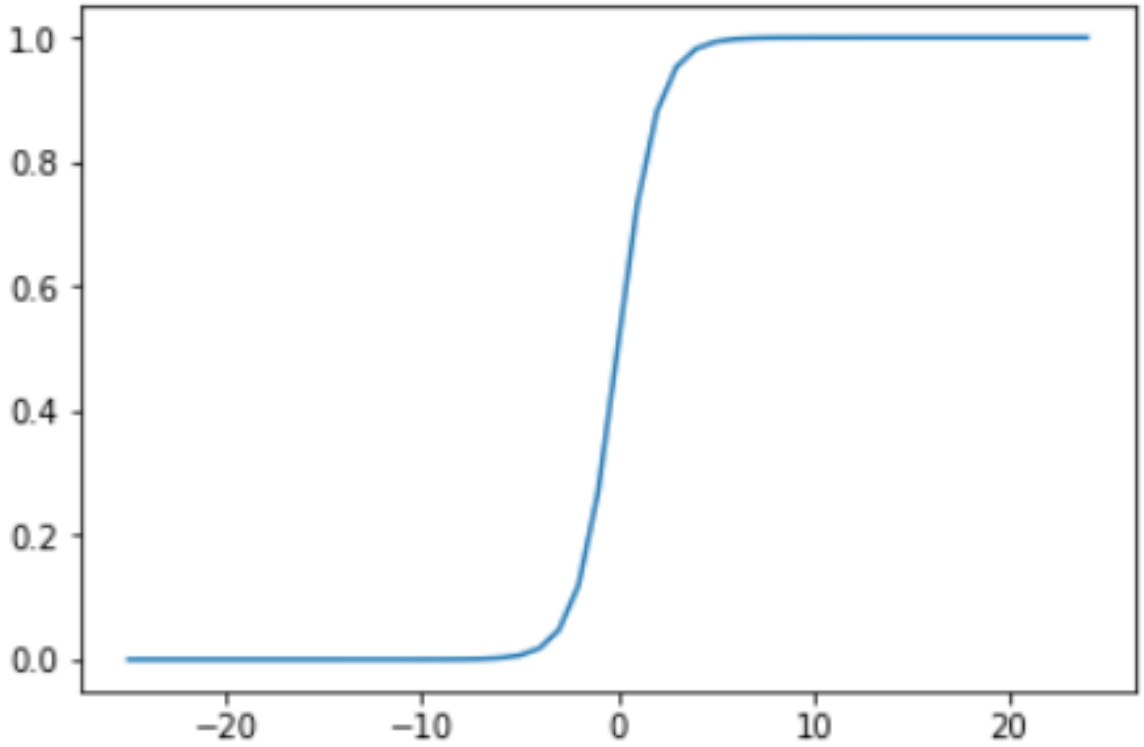
Review: Logistic Function

$$P^{(i)} \in \mathbb{R}[0, 1]$$

$$P^{(i)} = \sigma(z^{(i)})$$

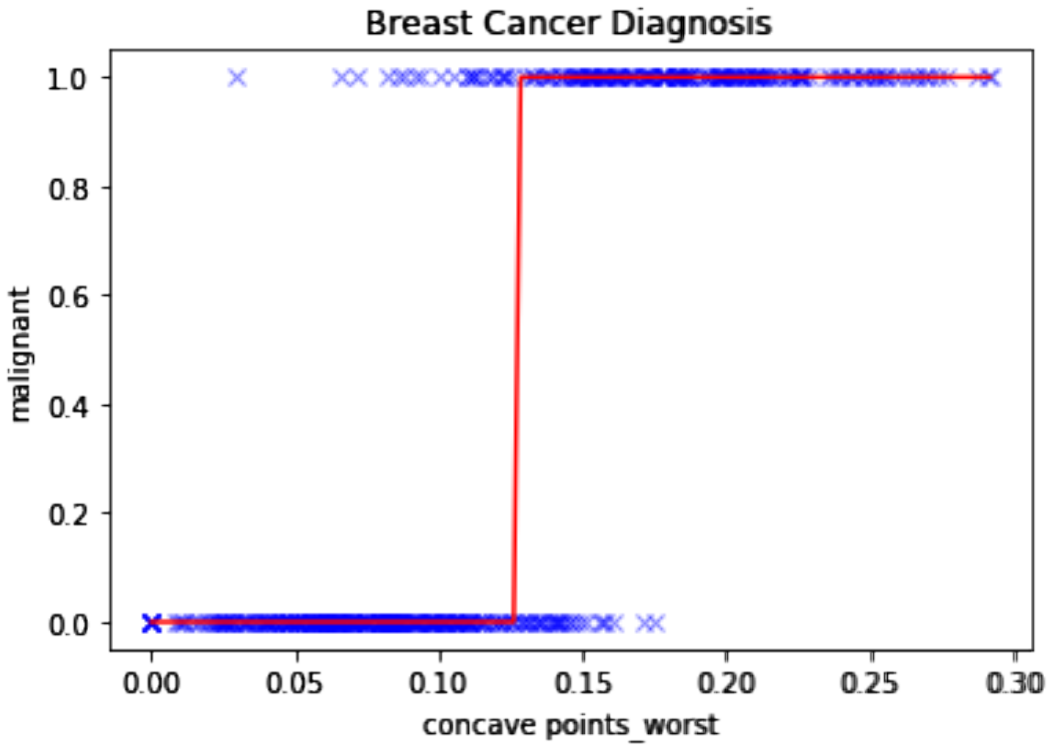
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

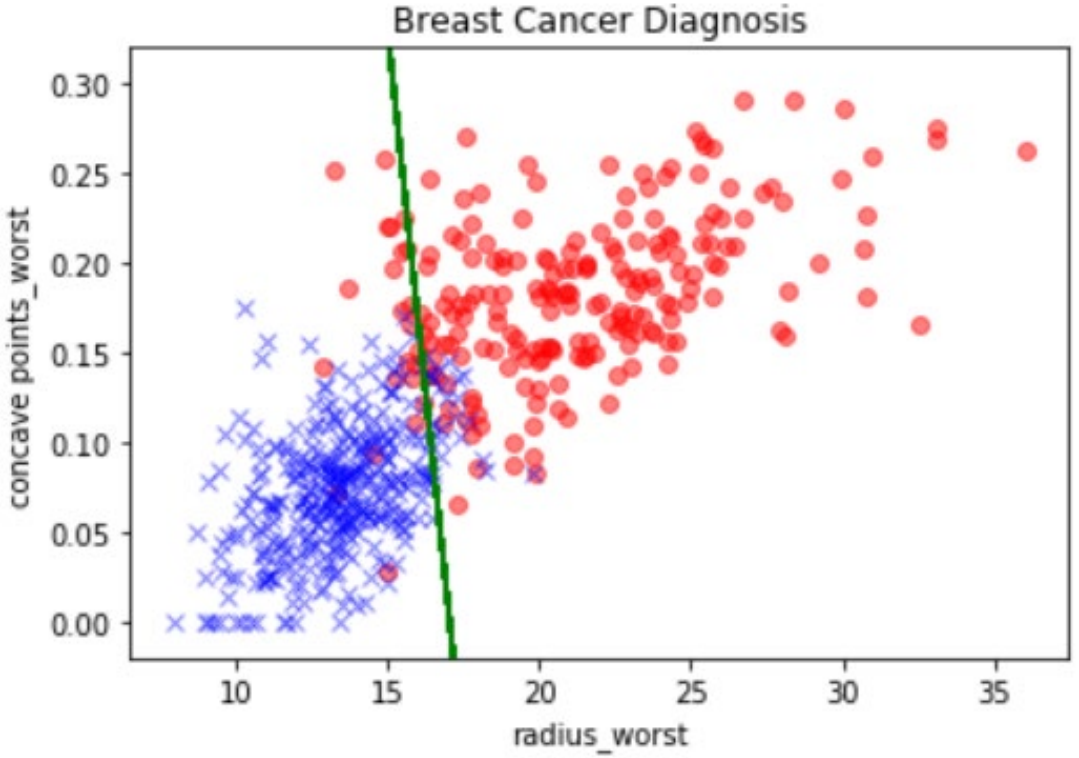
$$z^{(i)} = \boldsymbol{W} \cdot \boldsymbol{X} + b$$



Called “logit” and is related to the decision boundary

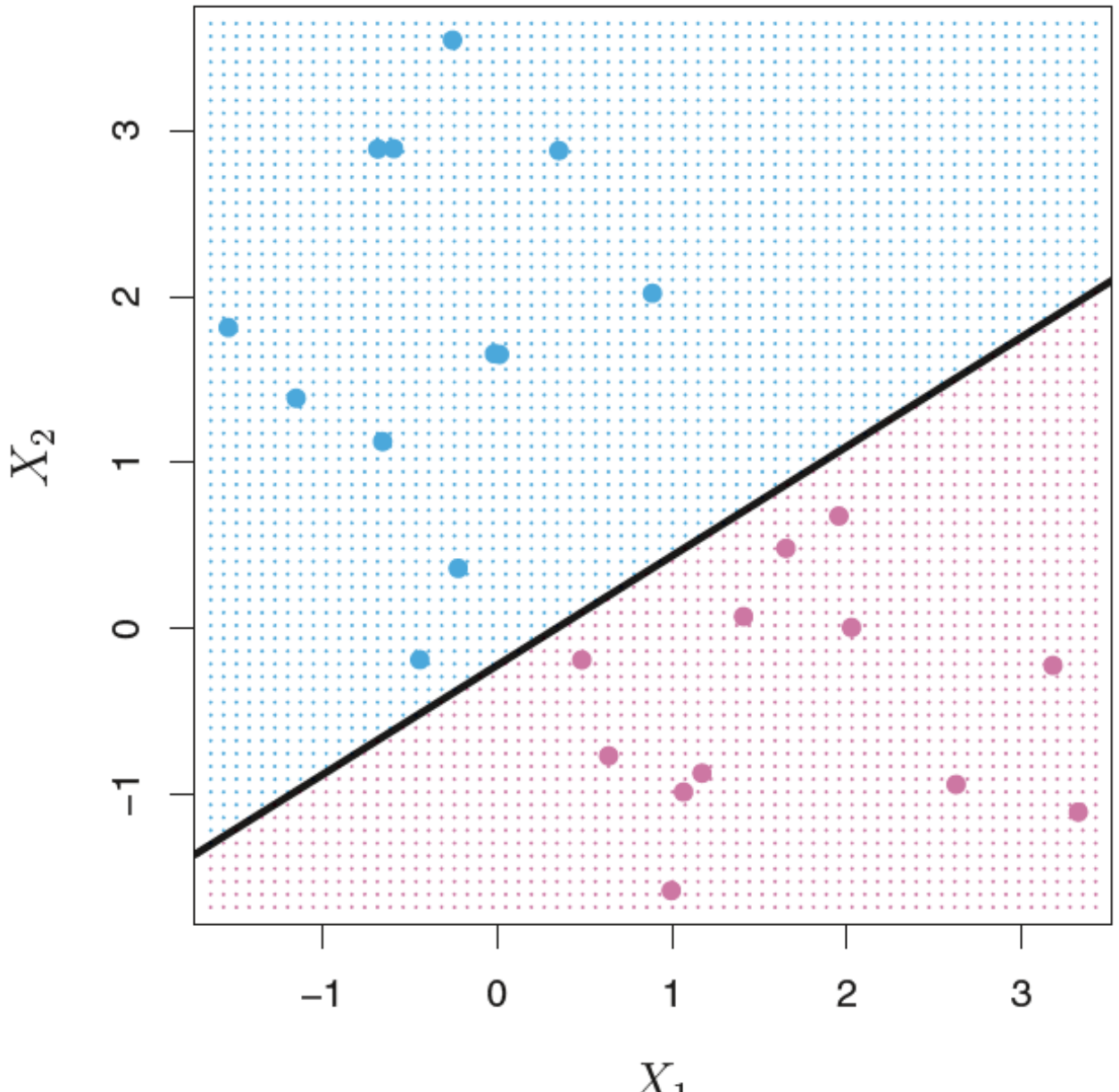
Review: Logistic Regression Decision Boundary





$z = 0.443\ x1 + 2.76\ x2 - 7.57 = 0$

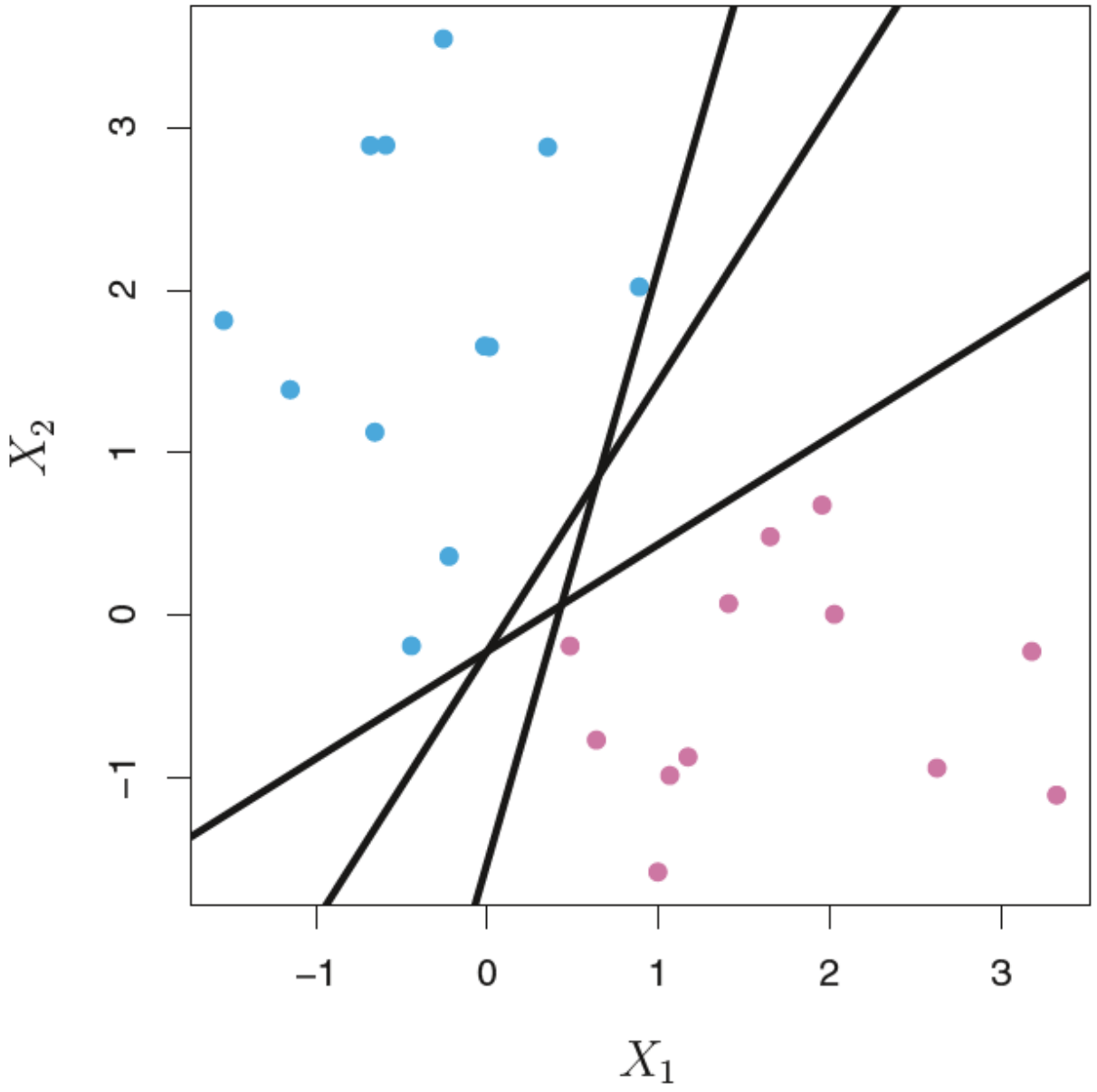
Hyperplane as a Decision Boundary



We can separate the two classes using a hyper plane!

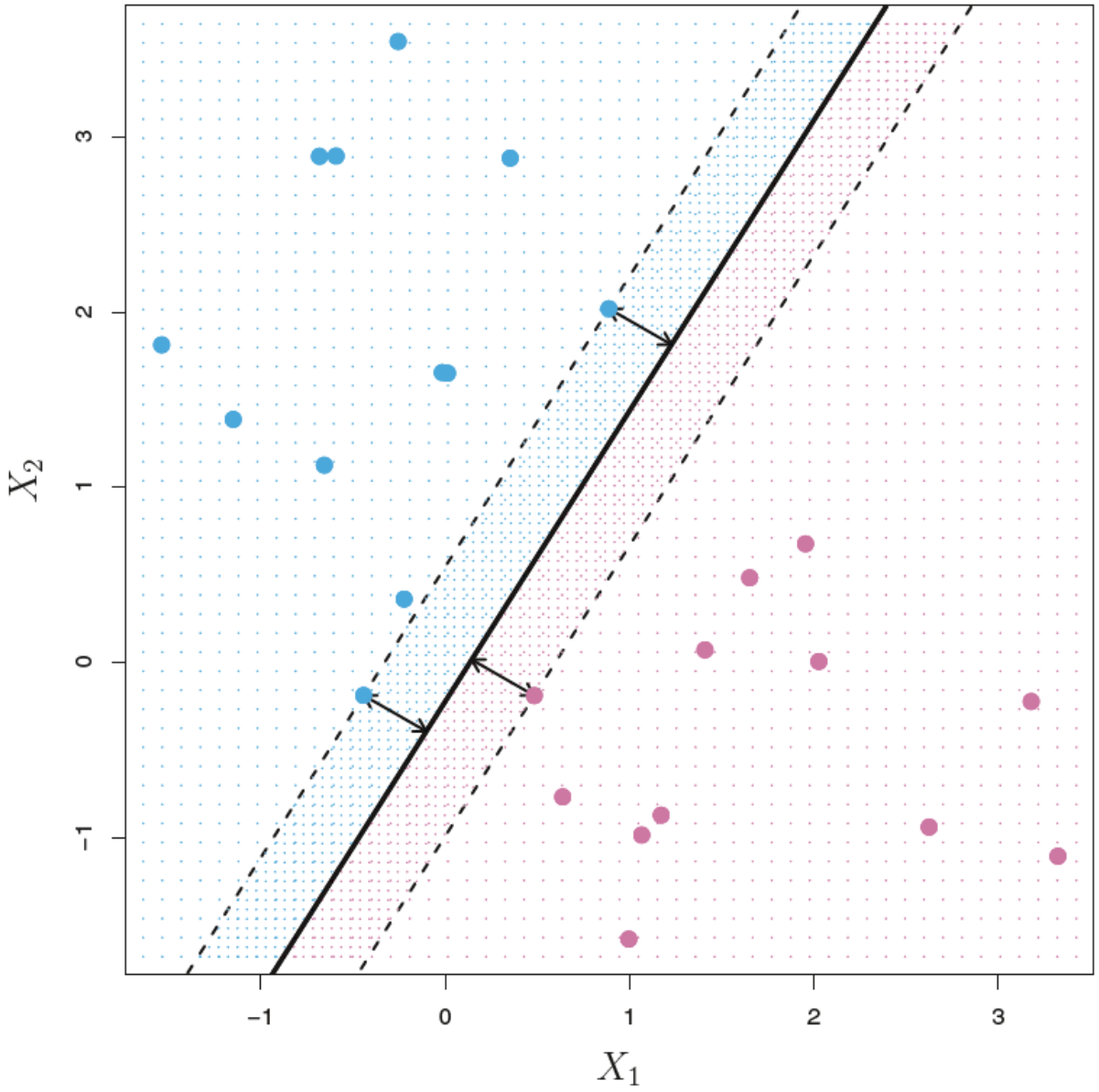
This hyperplane is called "separating hyperplane"

Hyperplane as a Decision Boundary



But which hyperplane should we choose?

Maximum margin classifier

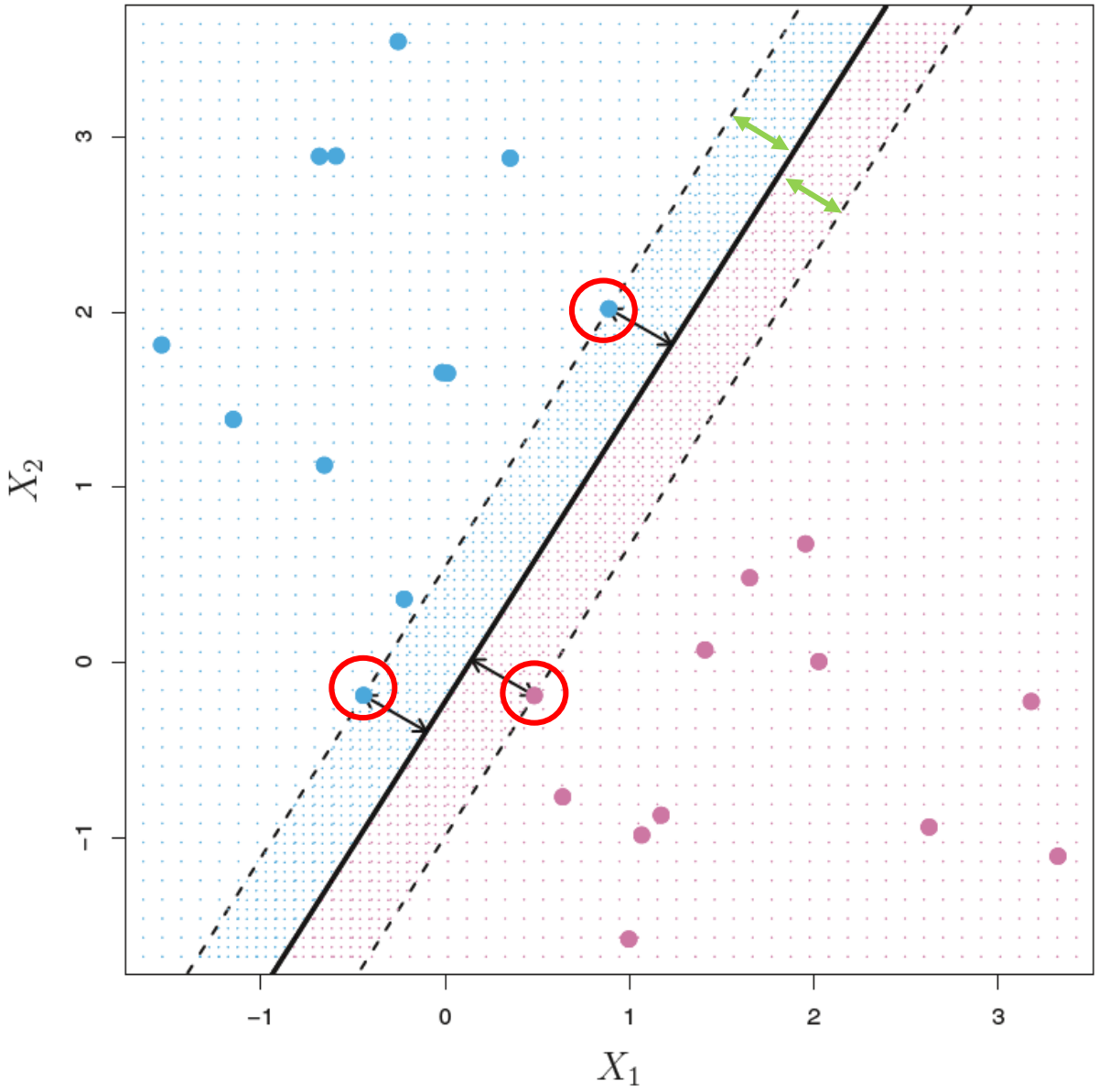


Which hyperplane should we choose?

The one with the least likely to misclassify the test data

= The one with the biggest margin

Maximum margin classifier

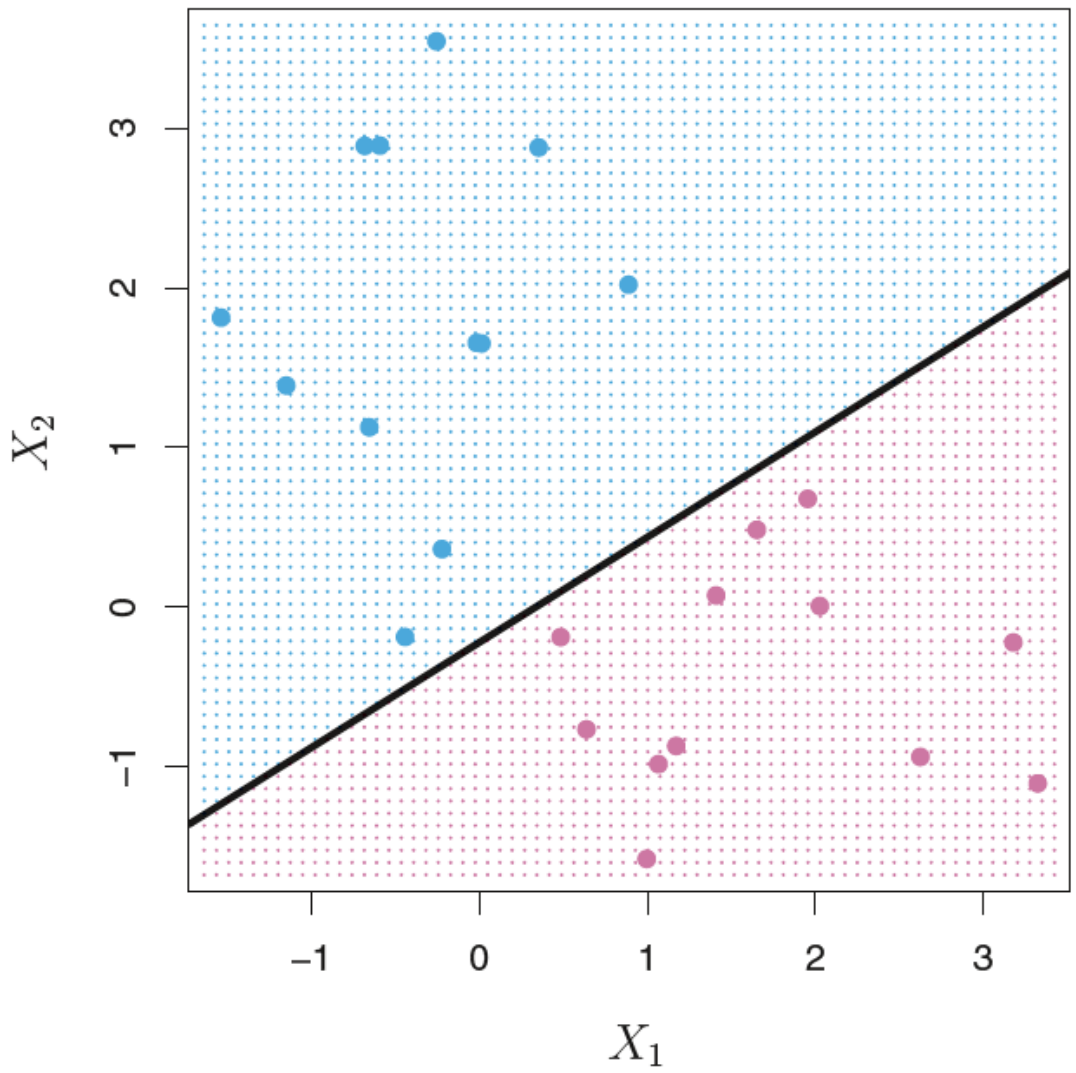


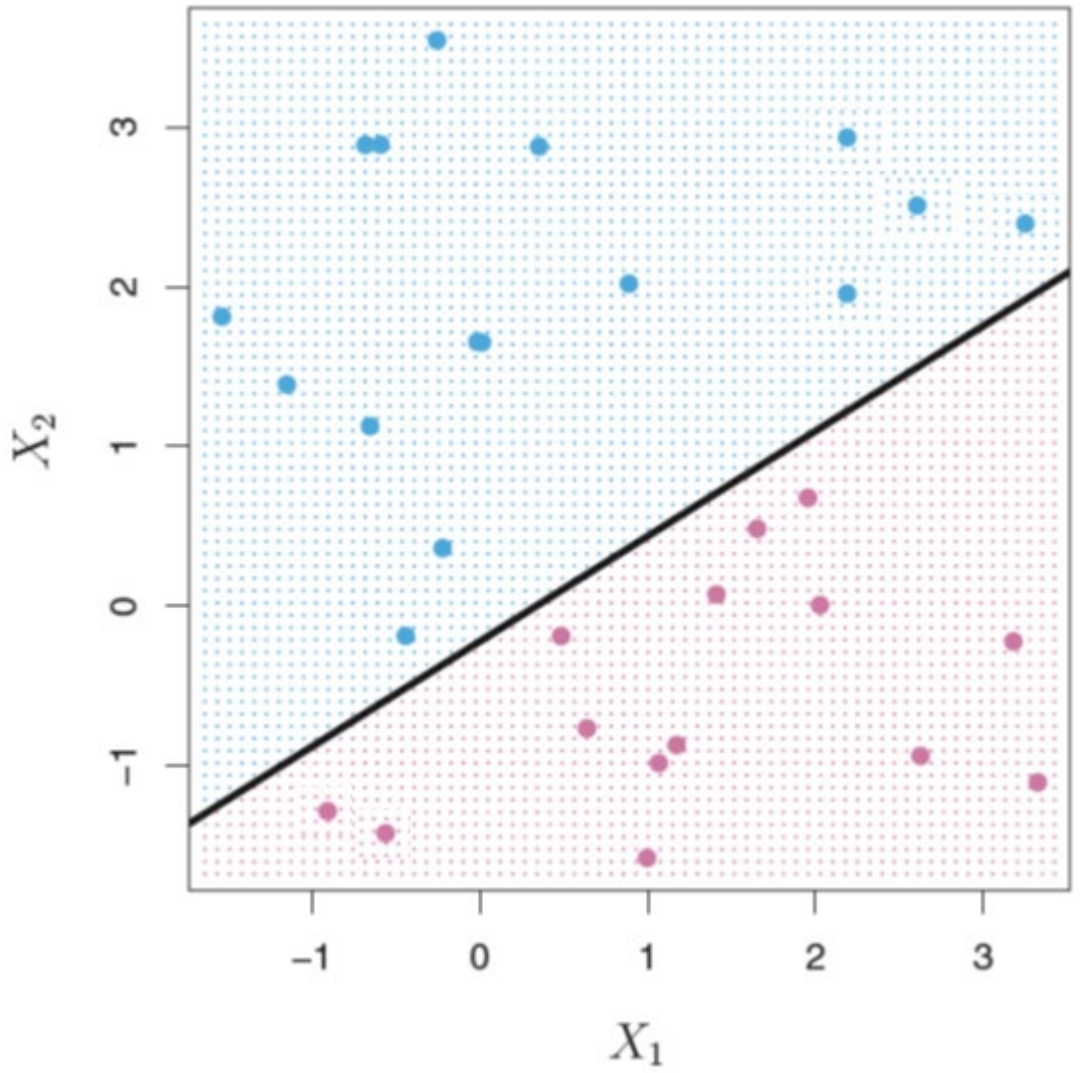
Support

Margin

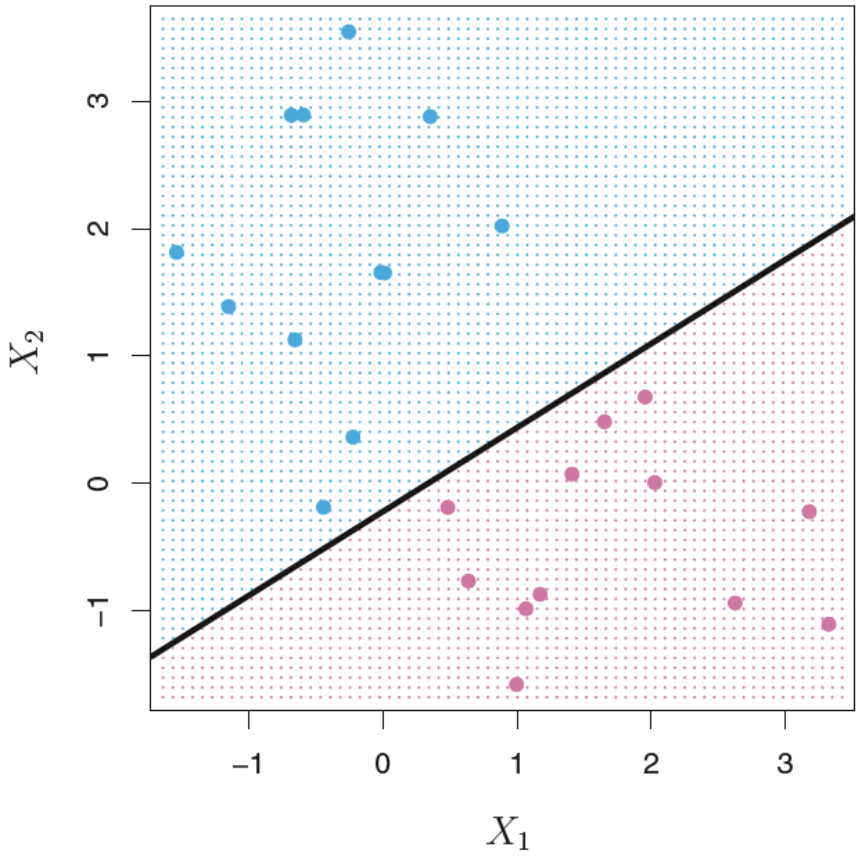
Quiz

What happens to the separating hyperplane when adding new train data points?





How to find the maximal margin hyperplane?



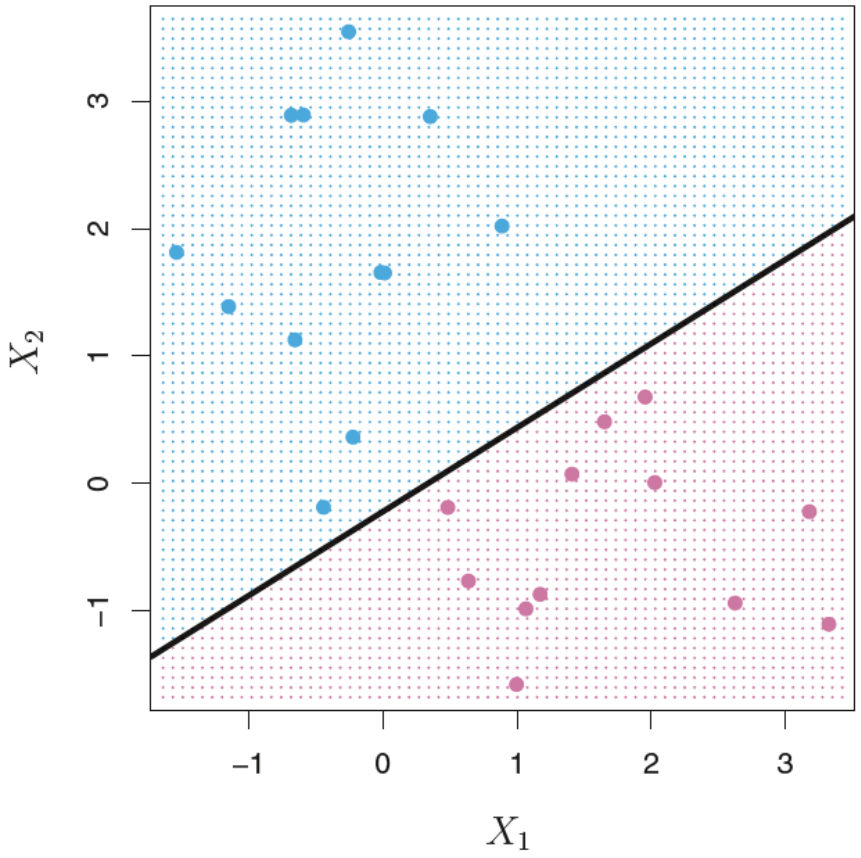
$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} > 0 \text{ if } y_i = 1$$

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} < 0 \text{ if } y_i = -1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) > 0$$

$$\text{for all } i = 1, \ldots, n$$

How to find the maximal margin hyperplane?



$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) > 0$$

for all $i = 1, \ldots, n$

$y_1, \ldots, y_n \in \{-1, 1\}$

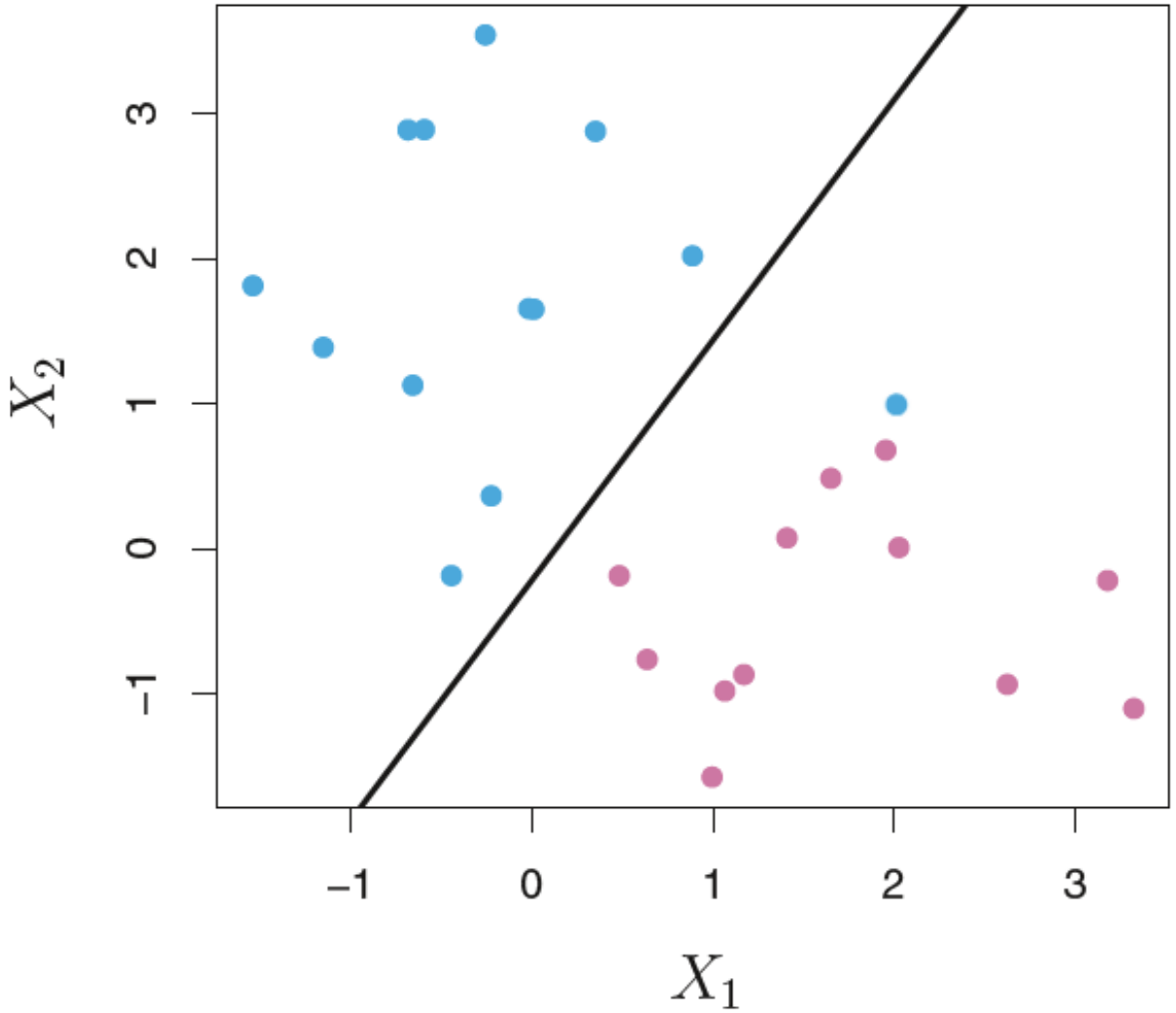
$$\underset{\beta_0, \beta_1, \ldots, \beta_p, M}{\text{maximize}} \; M$$

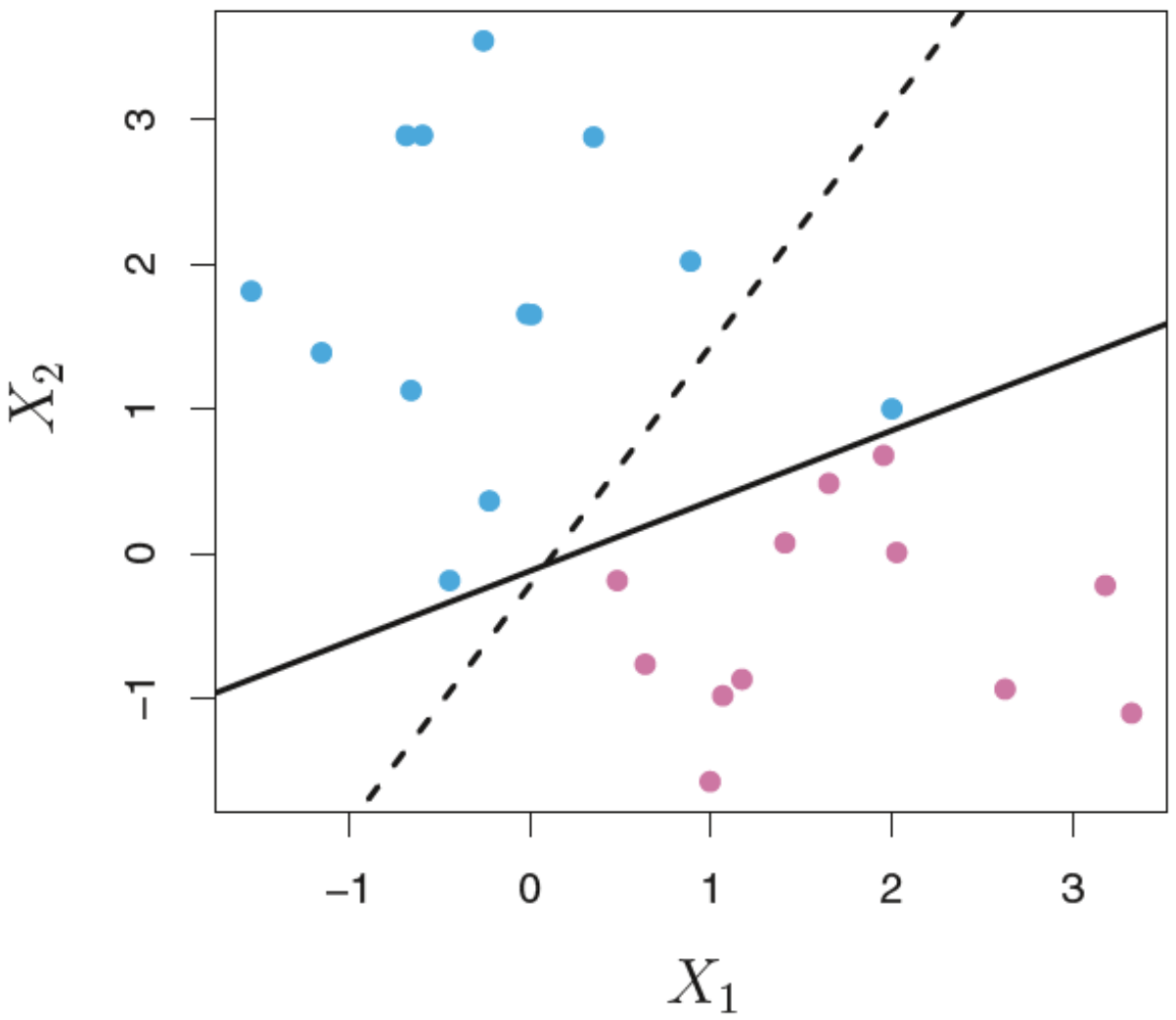
$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \;\; \forall\, i = 1, \ldots, n$$

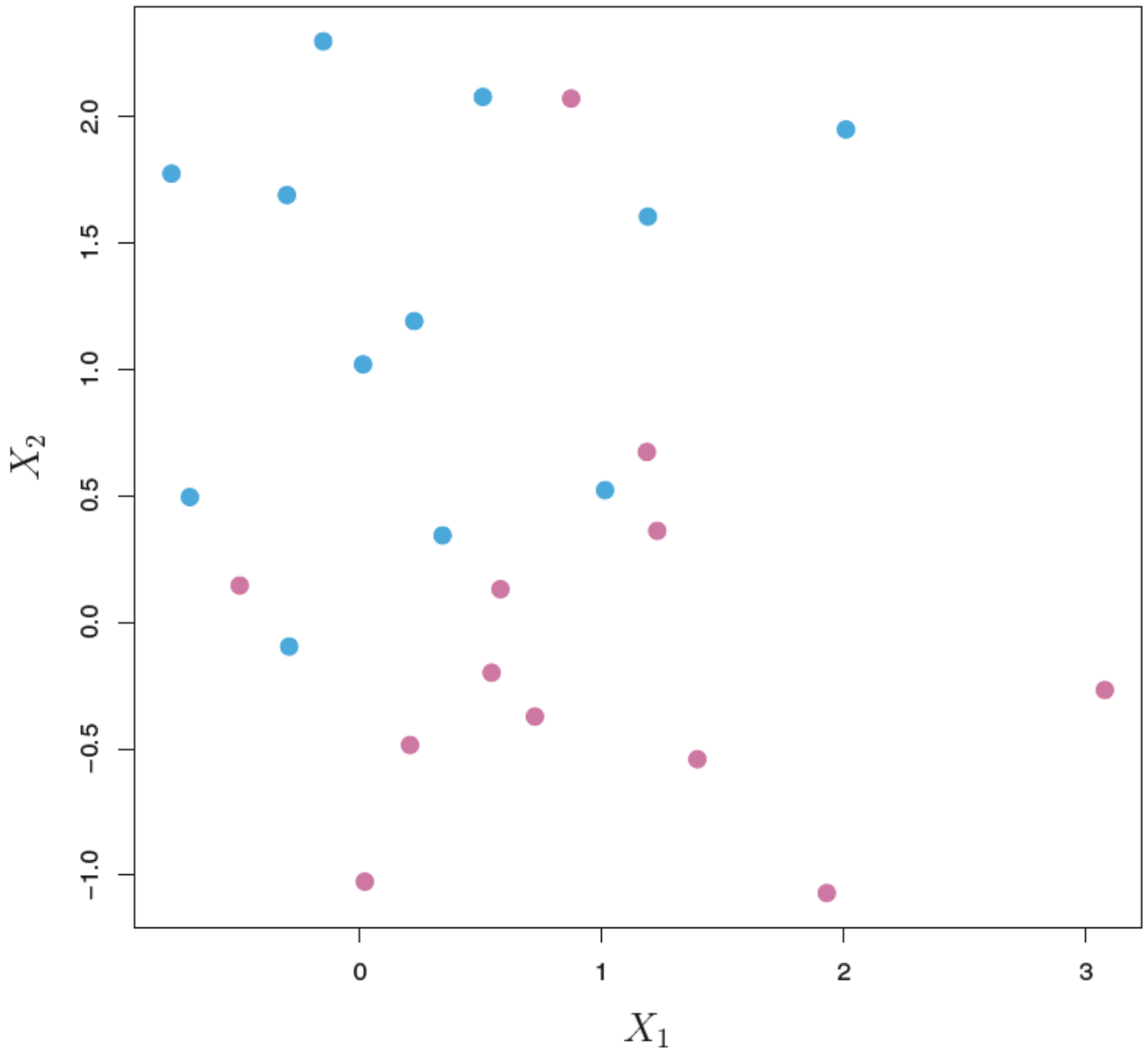
Drawback of the hard margin

What happens to the separating hyperplane?



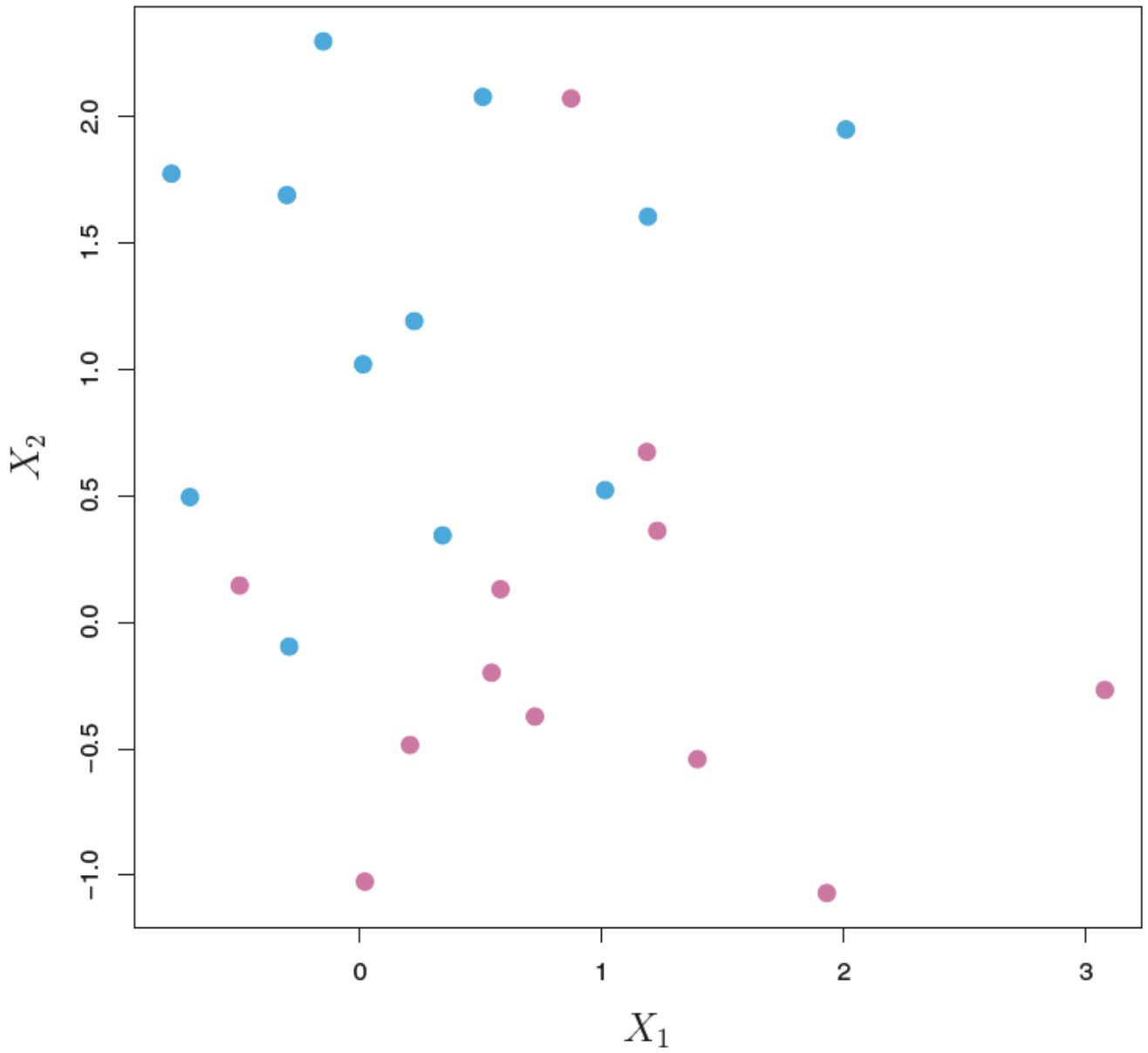


The impossible case…



Can you separate this with a hyperplane?

How to deal with an inseparable case

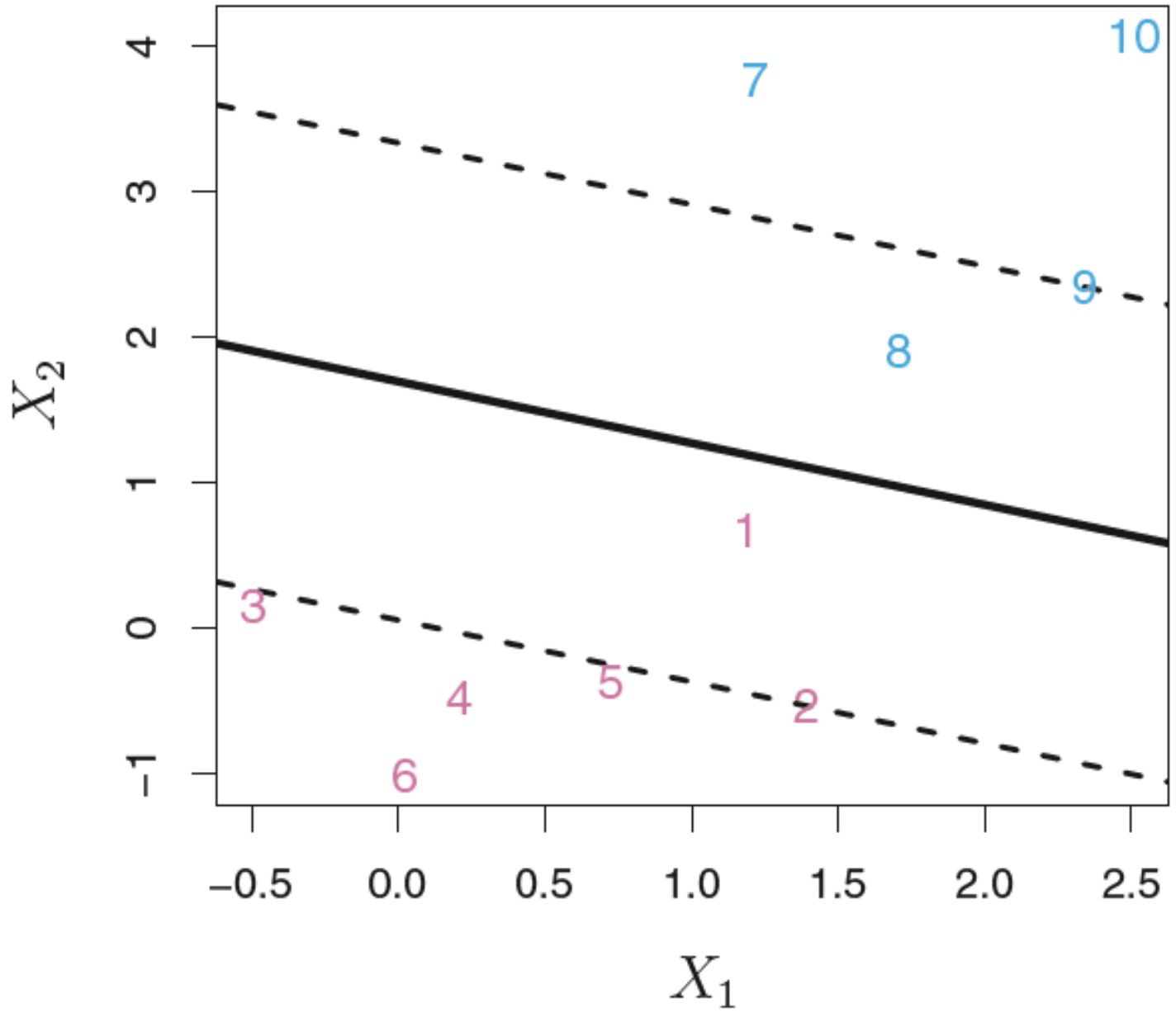


We'll have to accept some errors by softening the margin

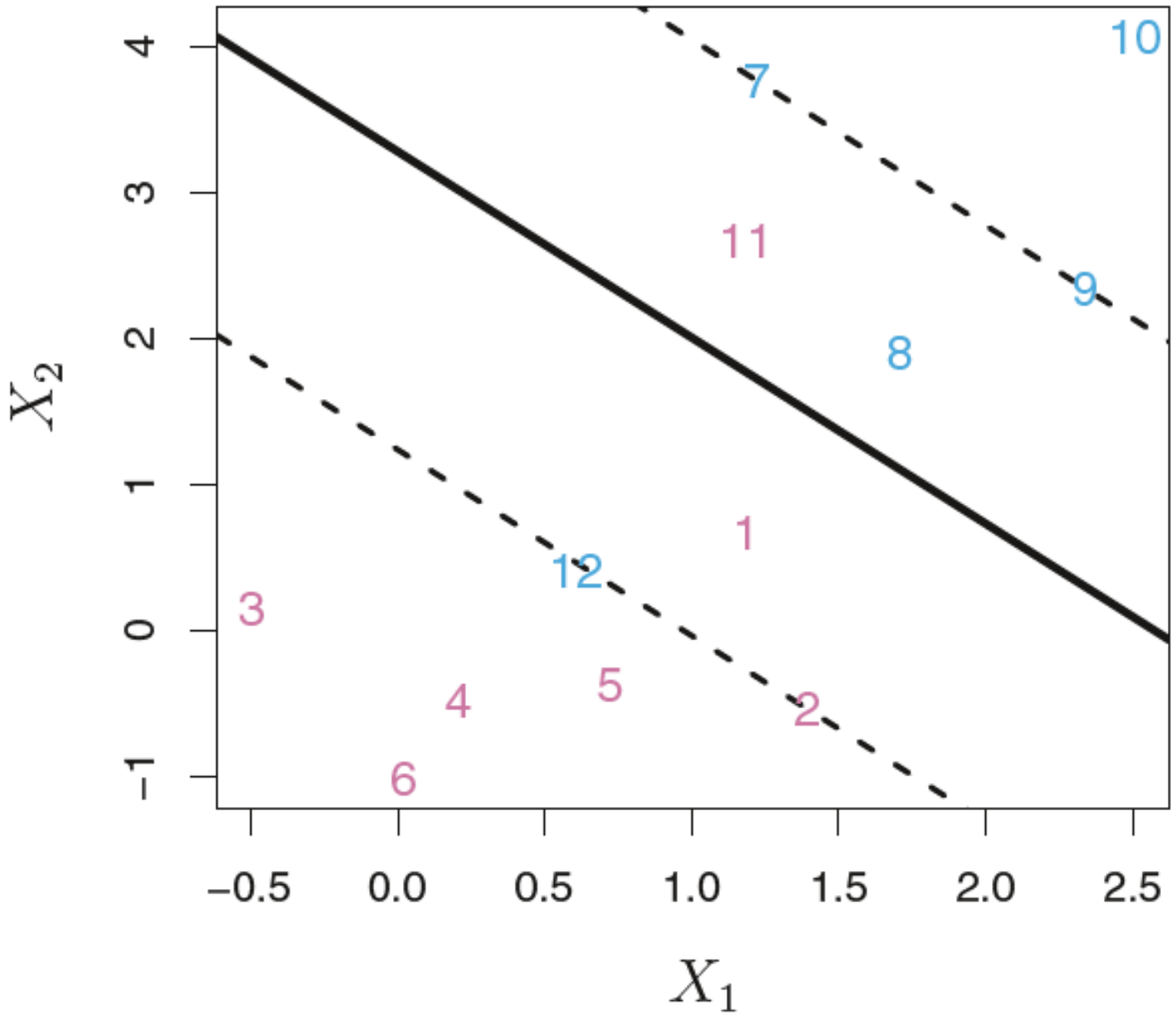
"soft margin classifier"

or called
"support vector classifier"

Soft margin classifier



Soft margin classifier



Soft margin classifier

Formulating support vector classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,M}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1$$

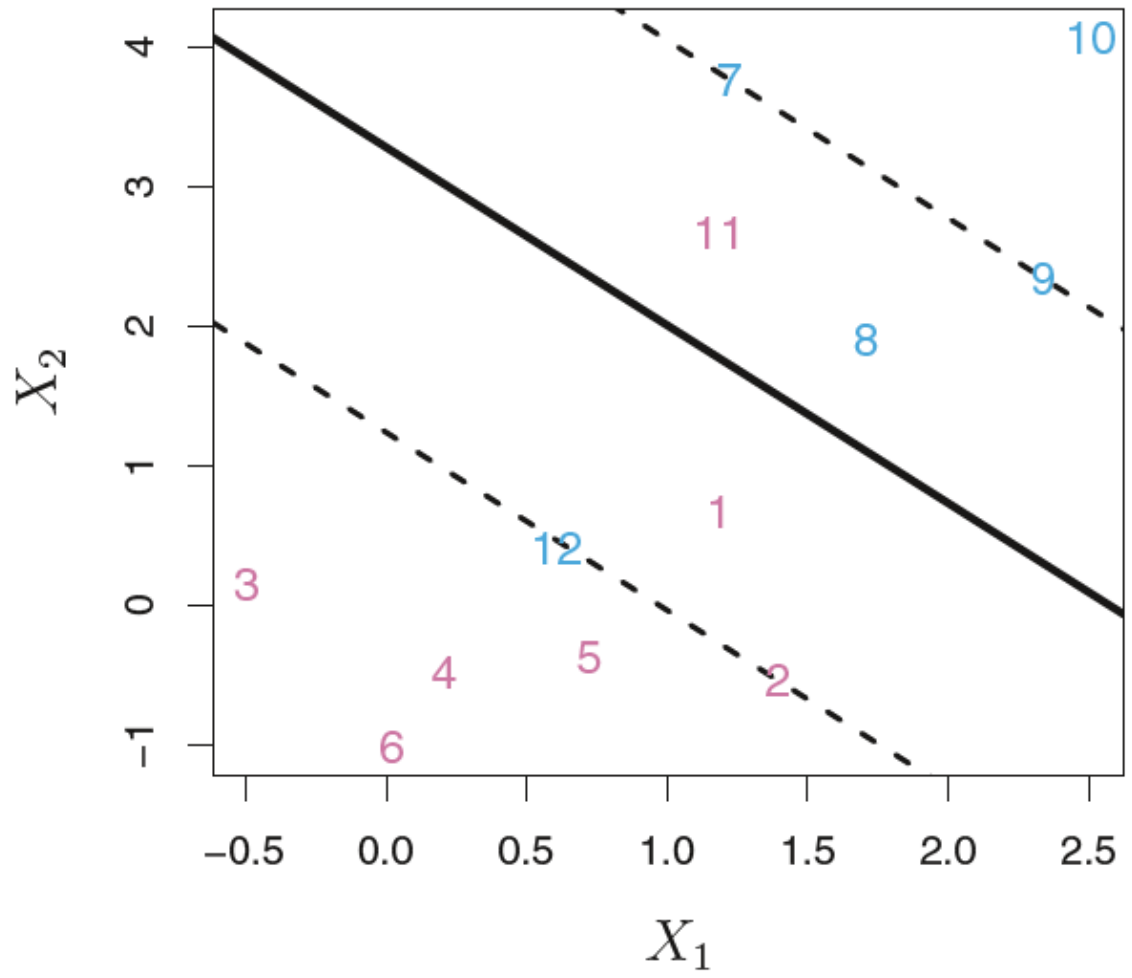
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0$$

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

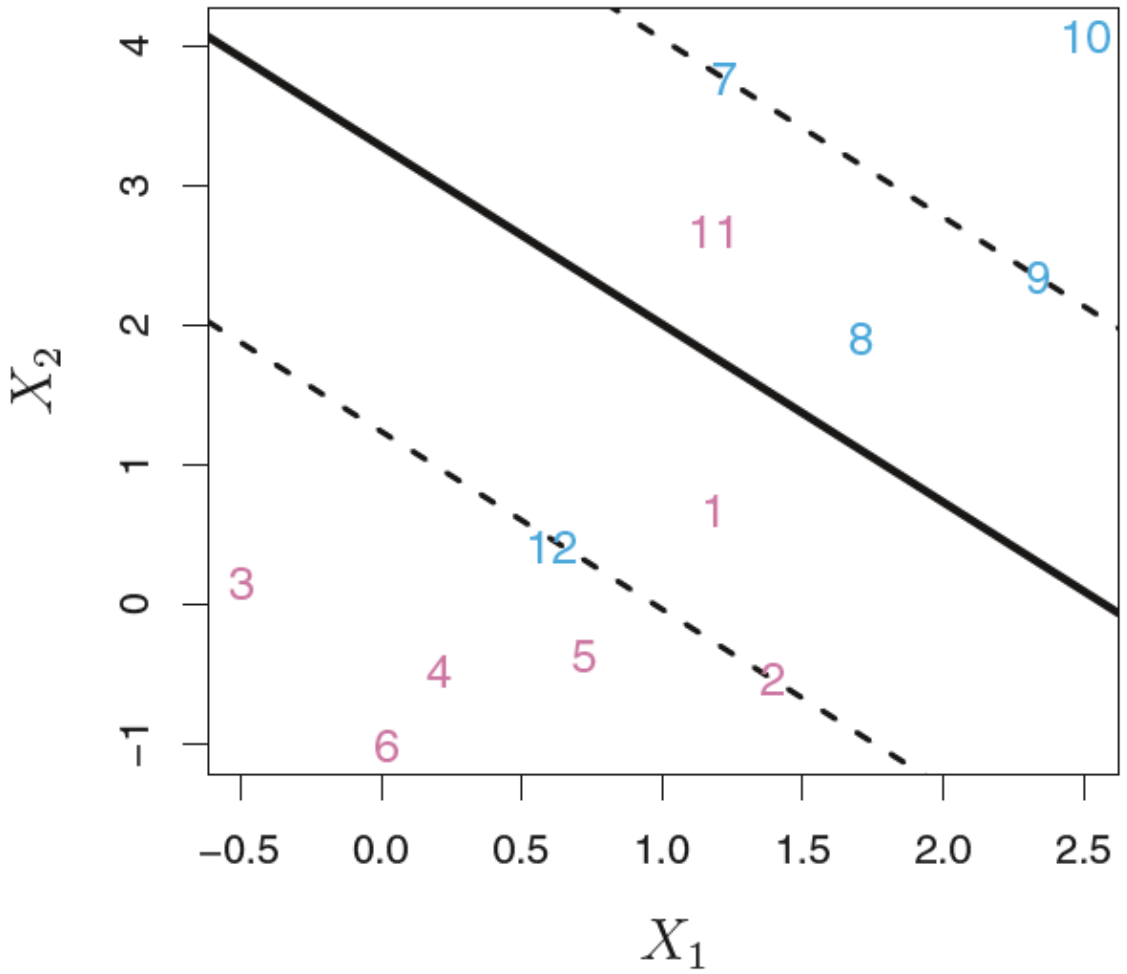
Soft margin classifier

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i) \qquad \epsilon_i \geq 0$$



Soft margin classifier

Q. What should we minimize?



(a) Number of support vectors
(b) Number of supports on the wrong side of the hyperplane
(c) The sum of distance of support vectors to the hyperplane
(d) The sum of distance of the support vectors to the correct-side margin M
(e) The sum of distance of the support vectors to the wrong-side margin –M

The role of C parameter

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0$$

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

C bounds both number and severity of violations

C is an error budget

C is a hyperparameter

The role of C parameter

Q1. What's the maximum number of supports on the wrong side the hyperplane given C?

Q2. What happens to the margin M when C decreases?

Q3. What happens to the bias and variance when C is small?

The role of C parameter

Q1. What's the maximum number of supports on the wrong side the hyperplane given C?

ANS: C

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

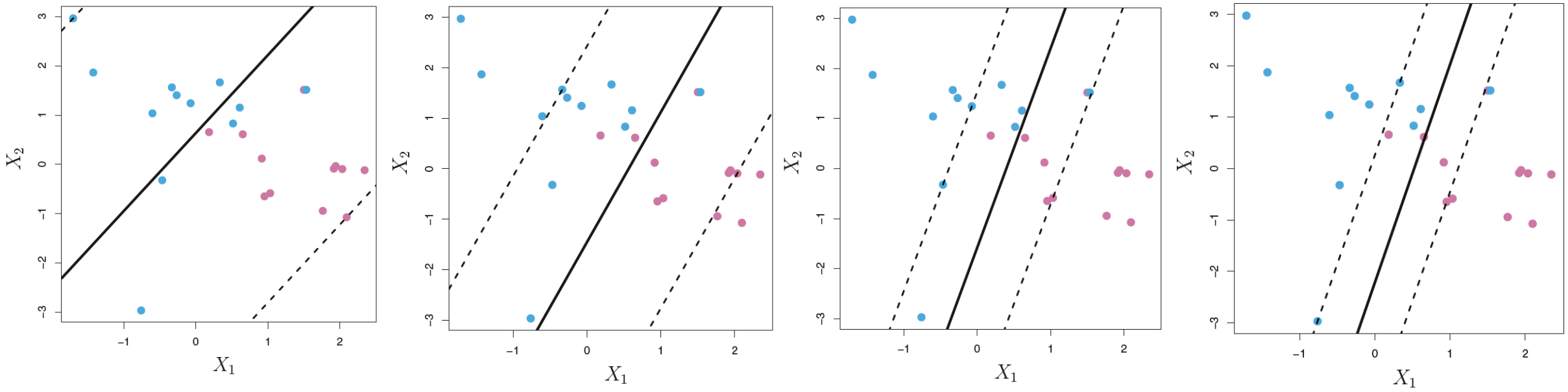
$$\epsilon_i \geq 0$$

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

The role of C parameter

Q2. What happens to the margin when C decreases?

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



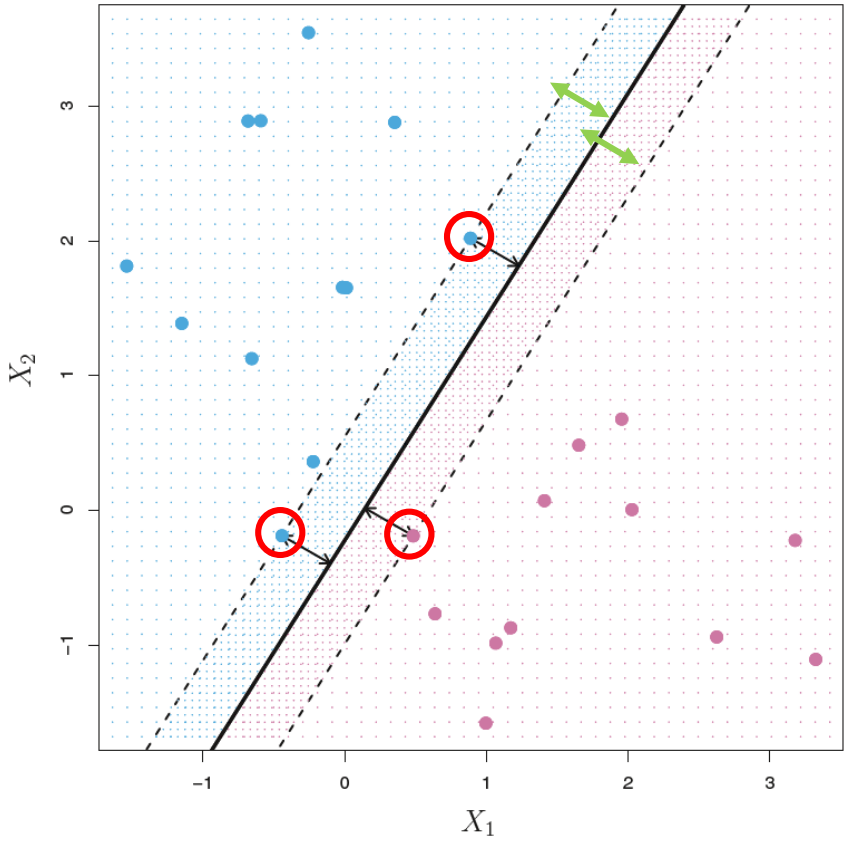
ANS: the magin becomes narrower

The role of C parameter

Q3. What happens to the bias and variance when C is small?

ANS: small C gives lower bias and higher variance

Recap



Support

Margin

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,M}{\text{maximize}} \; M$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \;\; \forall\, i = 1, \ldots, n$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0$$

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

Next time: Beyond linearly separable data

How can we separate this kind of data with SVC?