Welcome to CME 250 Introduction to Machine Learning!

Spring 2020 – Online version

April 28th 2020

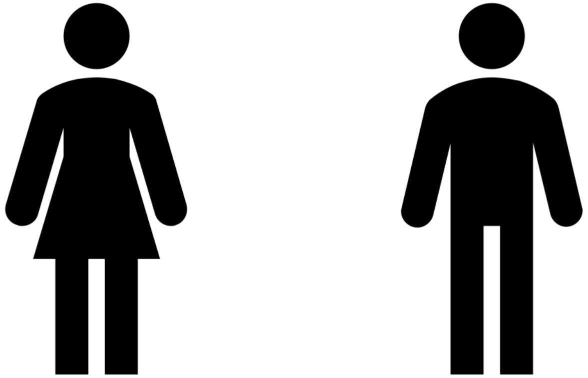
Stanford | Institute for Computational & Mathematical Engineering

Today's schedule: Classification

- Why does the distinction between regression and classification matters?
- Classification looking at Y as a random variable:
  - Logistic regression as a Generalized linear model
- Classification finding boundaries:
  - Support Vector Machines
- How to measure classification success?
  - Confusion Matrix

Let's get to know each other…

Breakout room



You

Another student

Name

Location

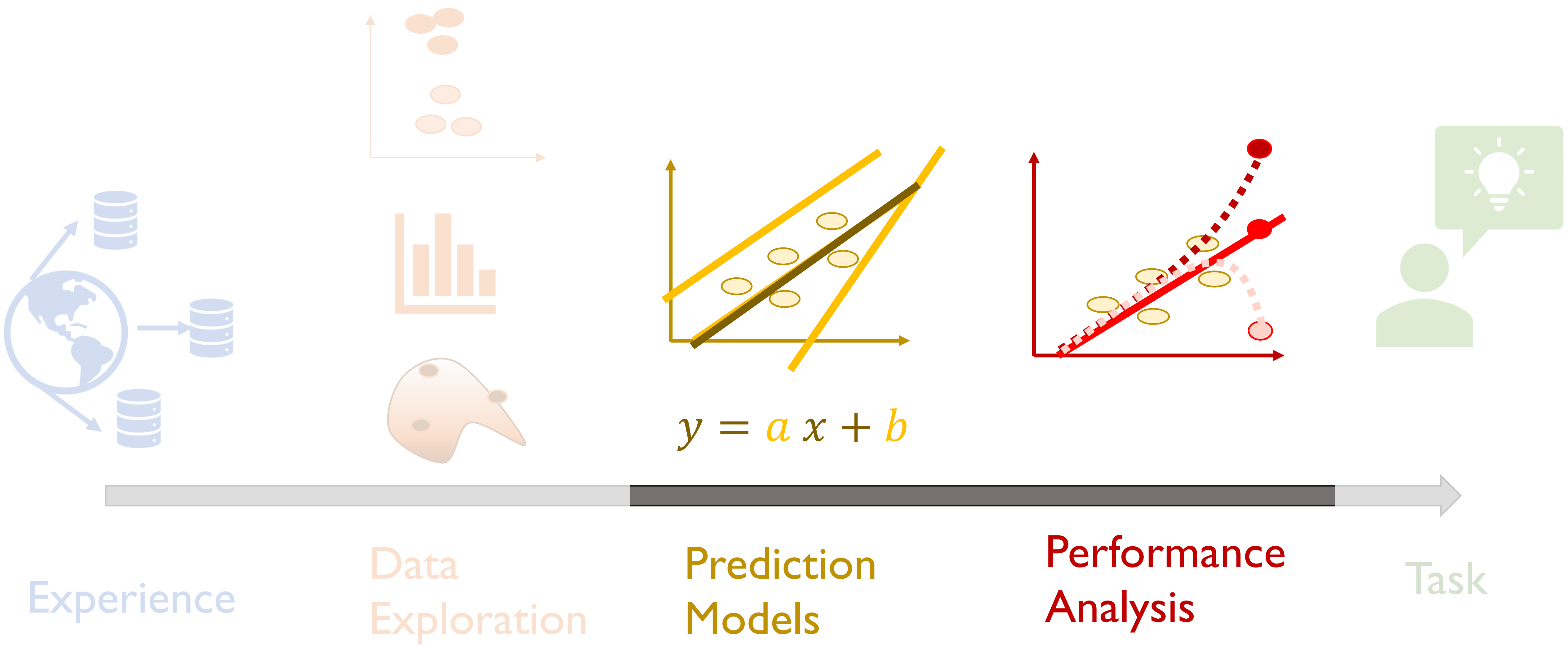
Department

Year

How was Part I Project? Interesting/unexpected/ unforgettable lessons or insights.

3 mins

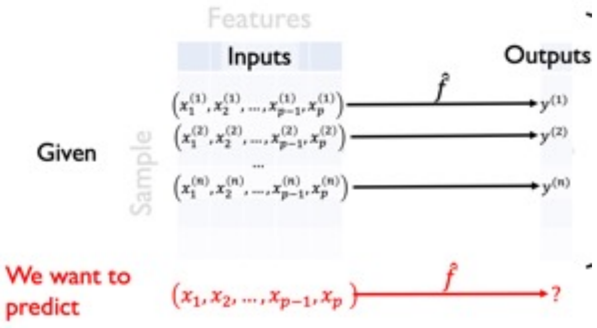
Chat/Audio/Video


$y = a\,x + b$
Experience
Data Exploration
Prediction Models
Performance Analysis
Task

Last week recap

Supervised Learning

Learn from examples



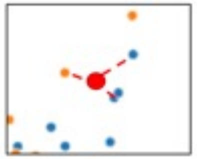
Regression

Y is quantitative

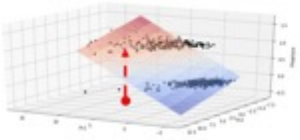
Classification

Y is categorical

KNN

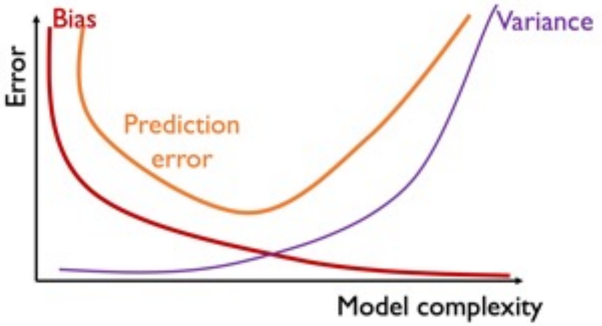


Linear Regression



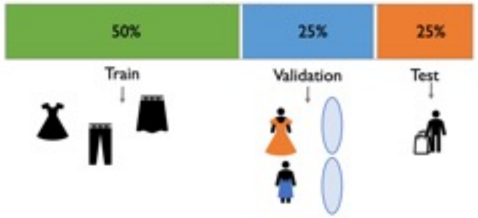
Using dummy variables?

Model Selection



Cross-Validation

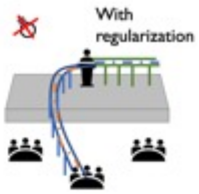
Estimate prediction error



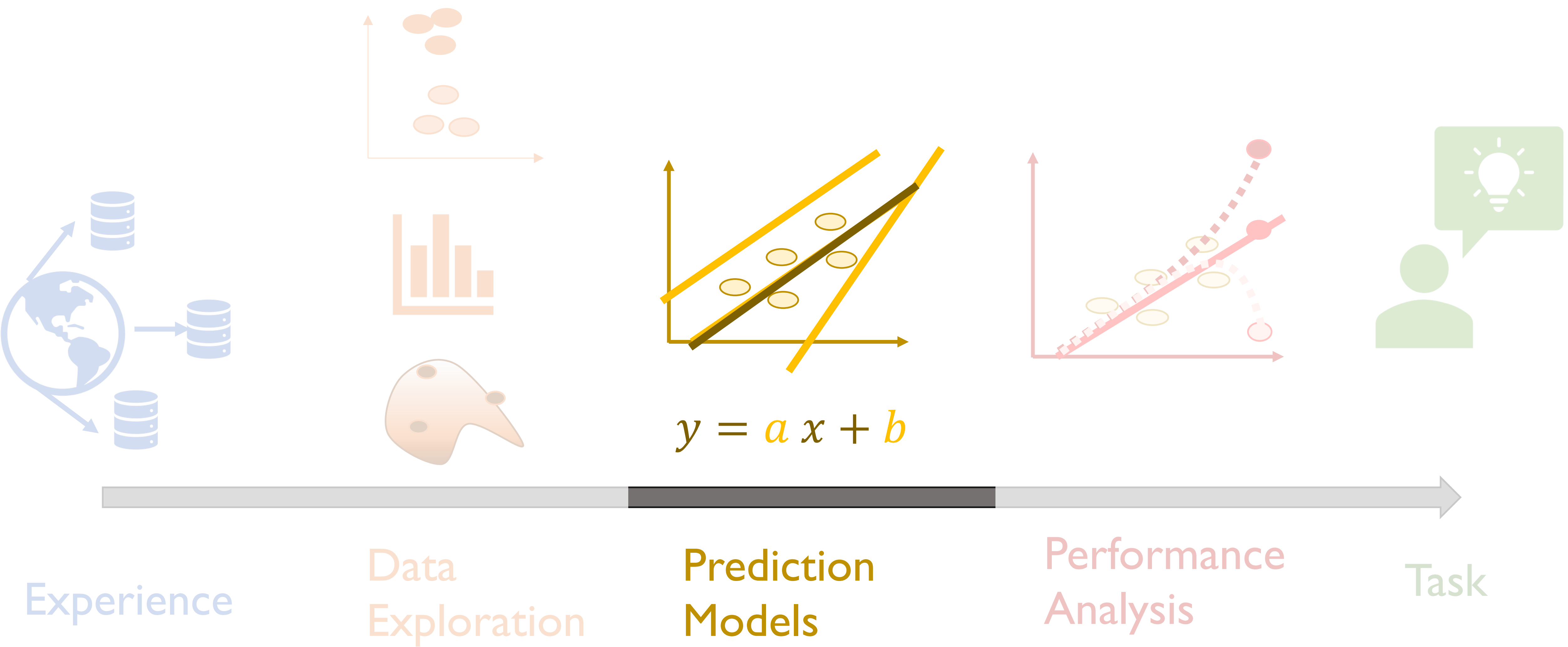
K-fold CV, LOOCV

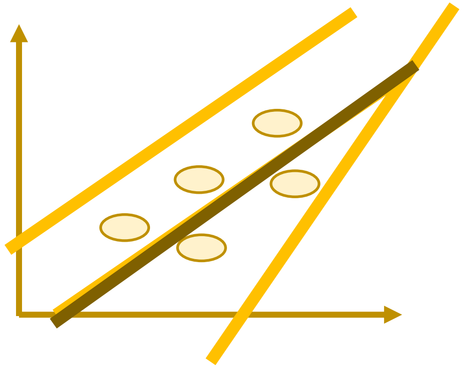
Regularization

Control complexity: Hyperparameters



Ridge, Lasso


$y = a\,x + b$
Experience
Data Exploration
Prediction Models
Performance Analysis
Task



$$y = a\,x + b$$

Prediction Models

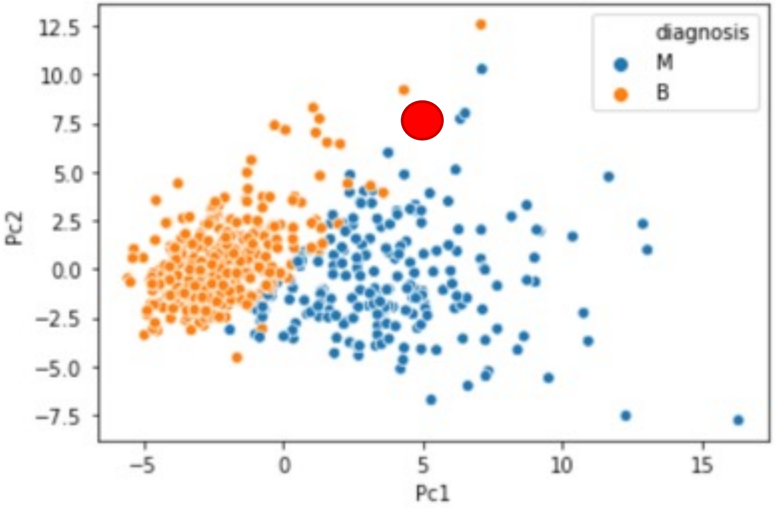
Supervised Learning Part II: Prediction Models for Classification

Introduction to Statistical Learning
Chapter 4: Classification
Chapter 9: Support Vector Machines

Elements Statistical Learning
Chapter 3.2: Linear Methods for Classification
Chapter 12: Support Vector Machines

More on Generalized Linear Models
Bayesian and Frequentist Regression Methods.
Jon Wakefield, 2013
Chapter 6.3: Generalized Linear Models

Breast Cancer Wisconsin (Diagnostic) Dataset



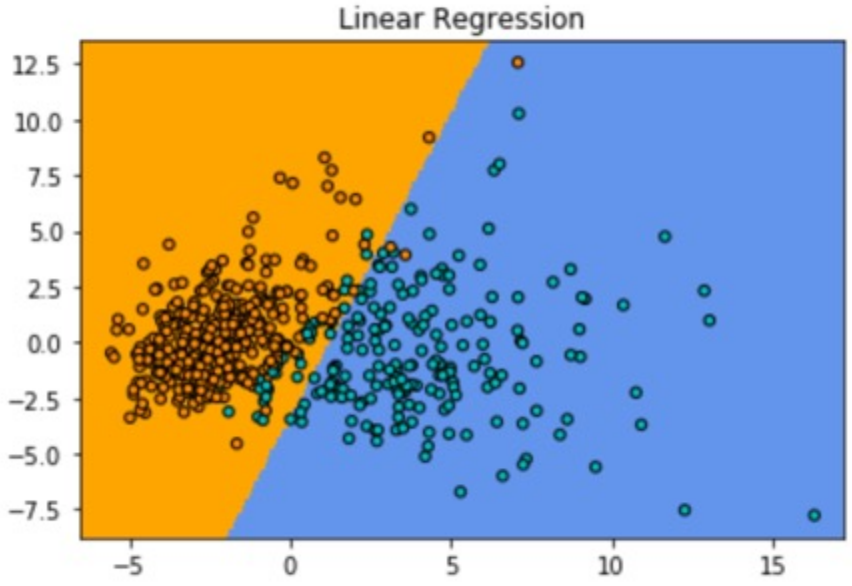
What is the diagnosis for this sample?

Use dummy variable for Y

M → 1

- - - - - - 0.5

B → 0



$Y \approx f(x) = \beta_0 + \beta_1 X_{PC1} + \beta_2 X_{PC2}$

What if we have more than 2 categories?

Linear regression with more than 2 categories

Option 1)

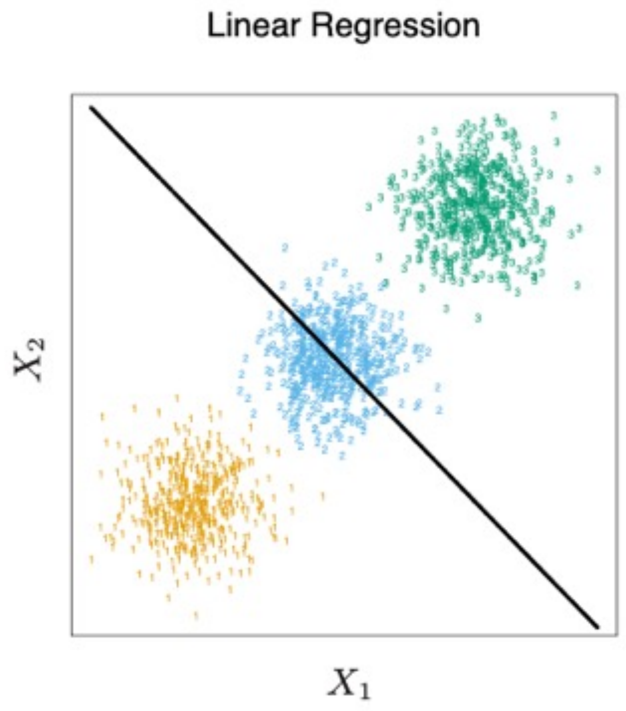
	Y
Cat. 1	0
	- - - - - 0.5
Cat.2	1
	- - - - - 1.5
Cat. 3	2

✗ Imposes order in categories

Option 2)

	Y_1	Y_2	Y_3
Cat. 1	1	0	0
Cat.2	0	1	0
Cat. 3	0	0	1

✗ Ignores category 2

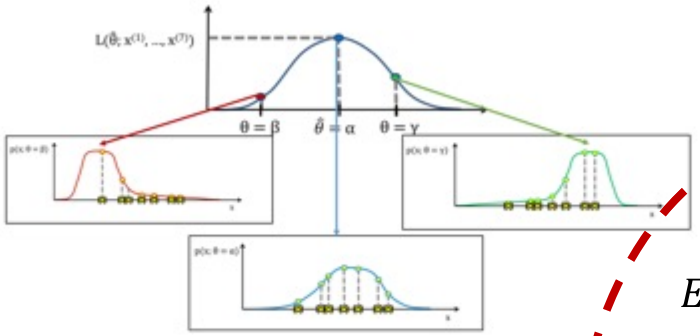


ESL Fig 4.2

We need a different approach!

How can we extend Linear Regression?

LR is Maximum Likelihood estimator



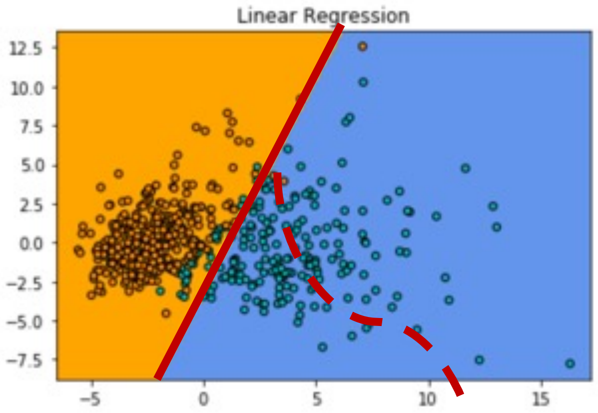
$$\boxed{Y \sim Normal}$$

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Find a better distribution for Y
categorical

Logistic Regression

LR creates separating hyperplanes



Y

M $\longrightarrow$ 1

- - - - - - 0.5

B $\longrightarrow$ 0

Optimize the hyperplane

Support Vector Machines

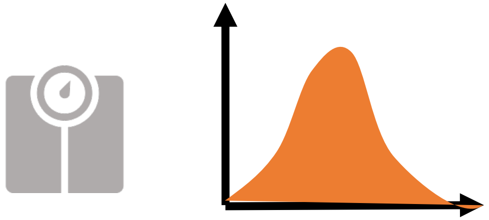
Extend LR: Generalized Linear Models

Exponential Family

$$\boxed{p(y|\,\theta,\alpha) = \exp\left(\frac{y\theta - b(\theta)}{\alpha} + c(y,\alpha)\right)}$$

Normal distribution

$N(\mu,\sigma^2)$



$$\boxed{p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{(y-\mu)^2}{2\sigma^2}\right)}$$

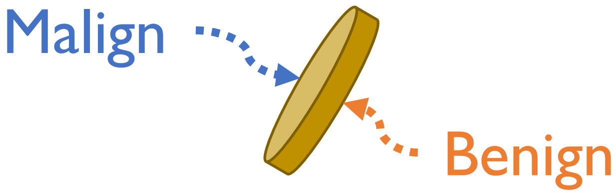
Poisson distribution

$Poisson(\lambda)$

$$\boxed{p(y|\lambda\,) = \frac{\lambda^y \exp(-\lambda)}{y!}}$$

Bernoulli distribution

$Bernoulli(p)$



$$\boxed{p(y|\lambda\,) = p^y(1-p)^{1-y}}$$

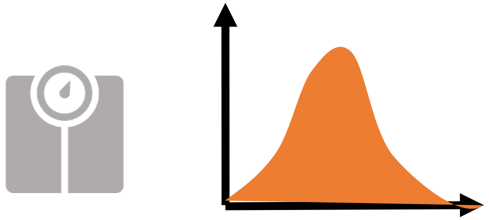
Extend LR: Generalized Linear Models

Exponential Family

$\boxed{E(y|\ \theta, \alpha) = b'(\theta)}$

Normal distribution

$N(\mu, \sigma^2)$



$b(\theta) = \theta^2/2$

$\mu = E[Y|\theta, \alpha] = \theta$

Poisson distribution

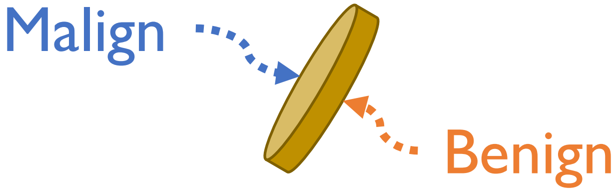
$Poisson(\lambda)$

$b(\theta) = \exp(\theta)$

$\lambda = E[Y|\theta, \alpha] = \exp(\theta)$

Bernoulli distribution

$Bernoulli(p)$



$b(\theta) = \log(1 + \exp(\theta))$

$p = E[Y|\theta, \alpha] = \dfrac{\exp(\theta)}{1 + \exp(\theta)}$

Calculate a linear regression of $\theta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

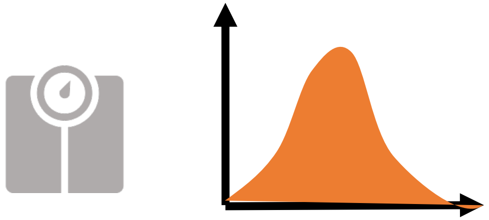
Extend LR: Generalized Linear Models

Exponential Family

$\boxed{E(y|\,\theta,\alpha) = b'(\theta)}$

Normal distribution

$N(\mu,\sigma^2)$



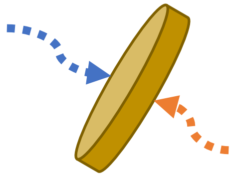
Poisson distribution

$Poisson(\lambda)$

Bernoulli distribution

$Bernoulli(p)$

Malign



Benign

Linear Regression

$y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Log-Linear Regression

$\log(y) \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

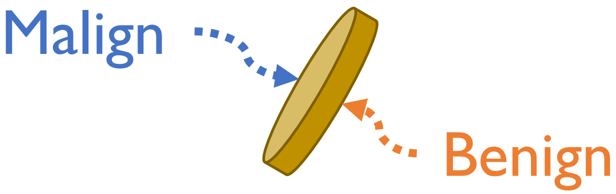
Logistic Regression

$\log\left(\frac{p}{1-p}\right) \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Logistic Regression

Bernoulli distribution

$Bernoulli(p)$

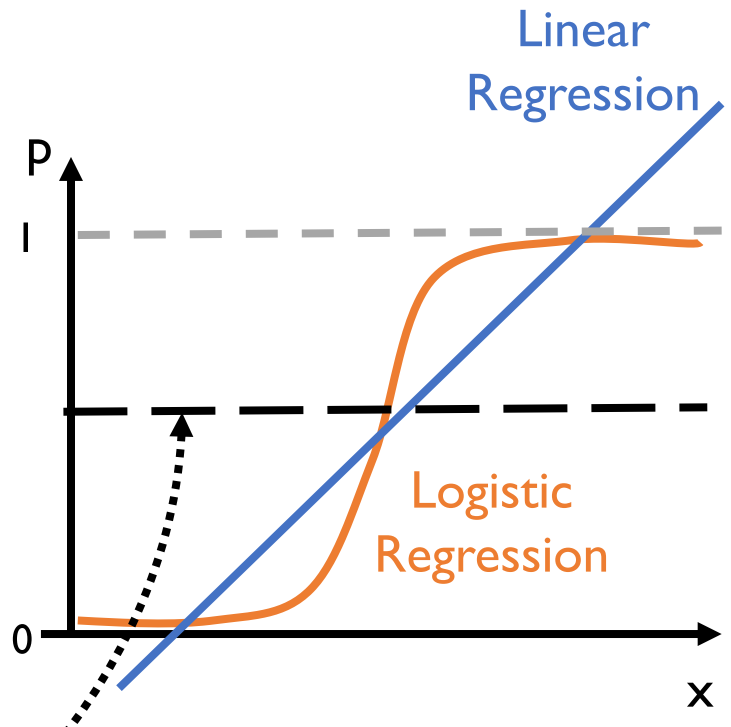


Log-odds / Logit

$$\boxed{\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$

$$\boxed{p = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\ )}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\ )}}$$

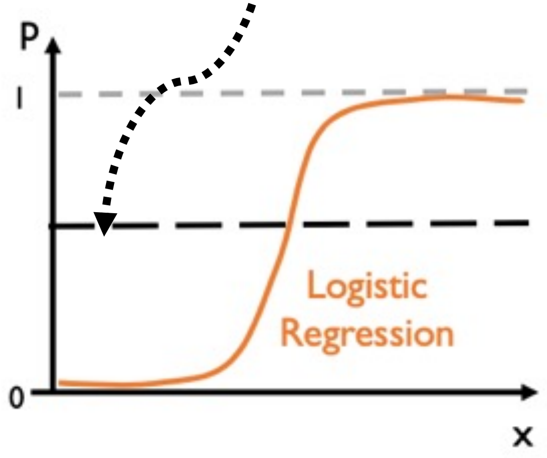
Sigmoid



To classify, we specify probability threshold

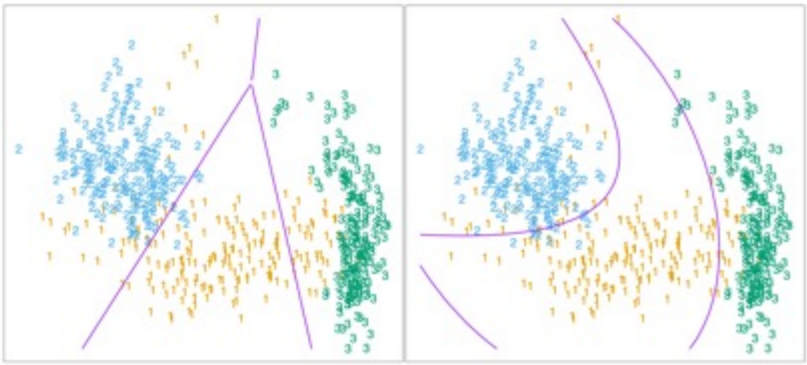
Challenges of Logistic Regression

Probability Threshold



Hyperparameter
Usually 0.5 (not always)

Linear Decision Boundary



ESL Fig 4.1

As in Linear regression:
Add additional features
$X_1, X_2, X_1X_2, X_1^2, X_2^2$

Multiple Categories

$$p_1 = \frac{\exp(\beta_{01} + \beta_1^T X)}{1 + \sum_{l=1}^{K-1} \exp\left(\beta_{0l} + \beta_l^T X\right)}$$

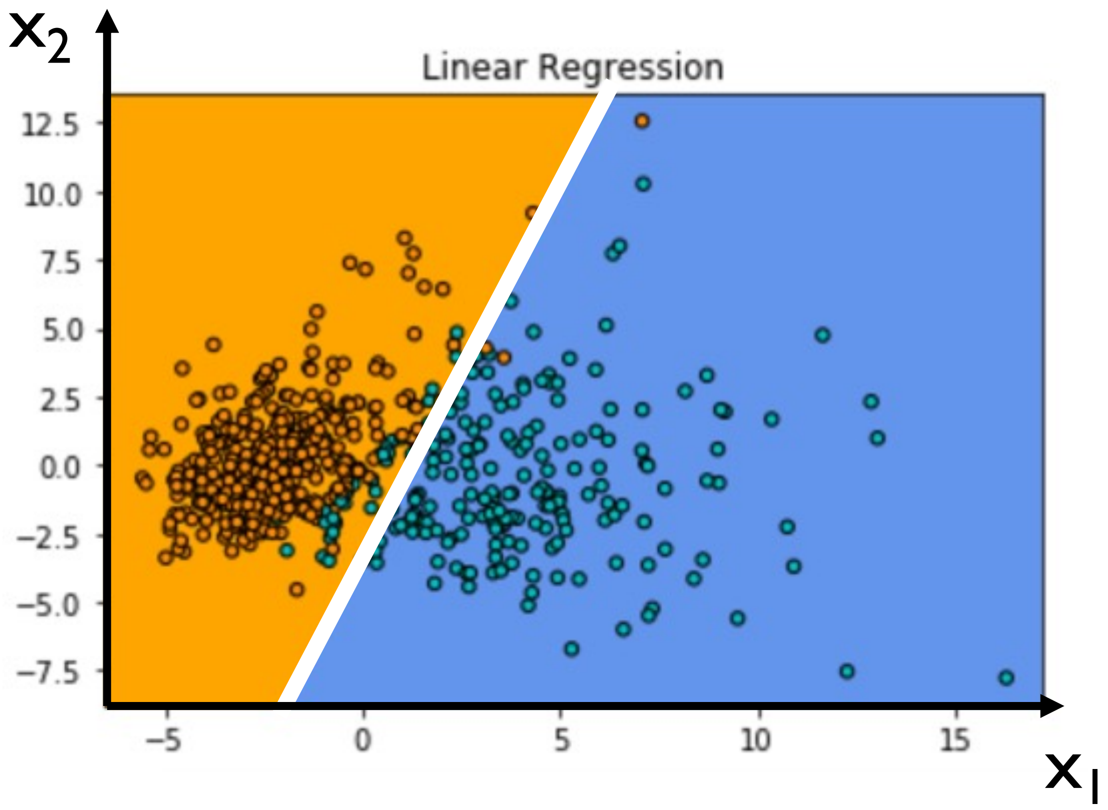
…

$$p_{K-1} = \frac{\exp(\beta_{0K-1} + \beta_K^T X)}{1 + \sum_{l=1}^{K-1} \exp\left(\beta_{0l} + \beta_l^T X\right)}$$

$$p_K = \frac{1}{1 + \sum_{l=1}^{K-1} \exp\left(\beta_{0l} + \beta_l^T X\right)}$$

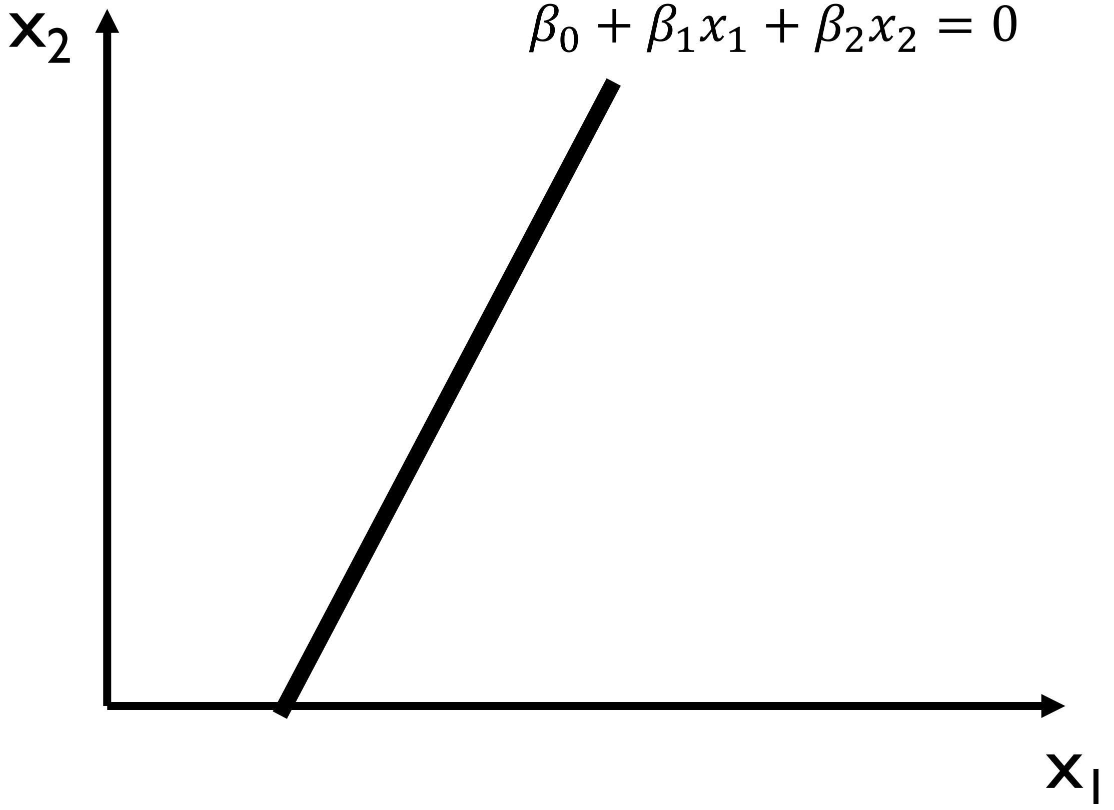
Do we need probabilities or decision boundaries?

Find “optimal” boundary to separate classes



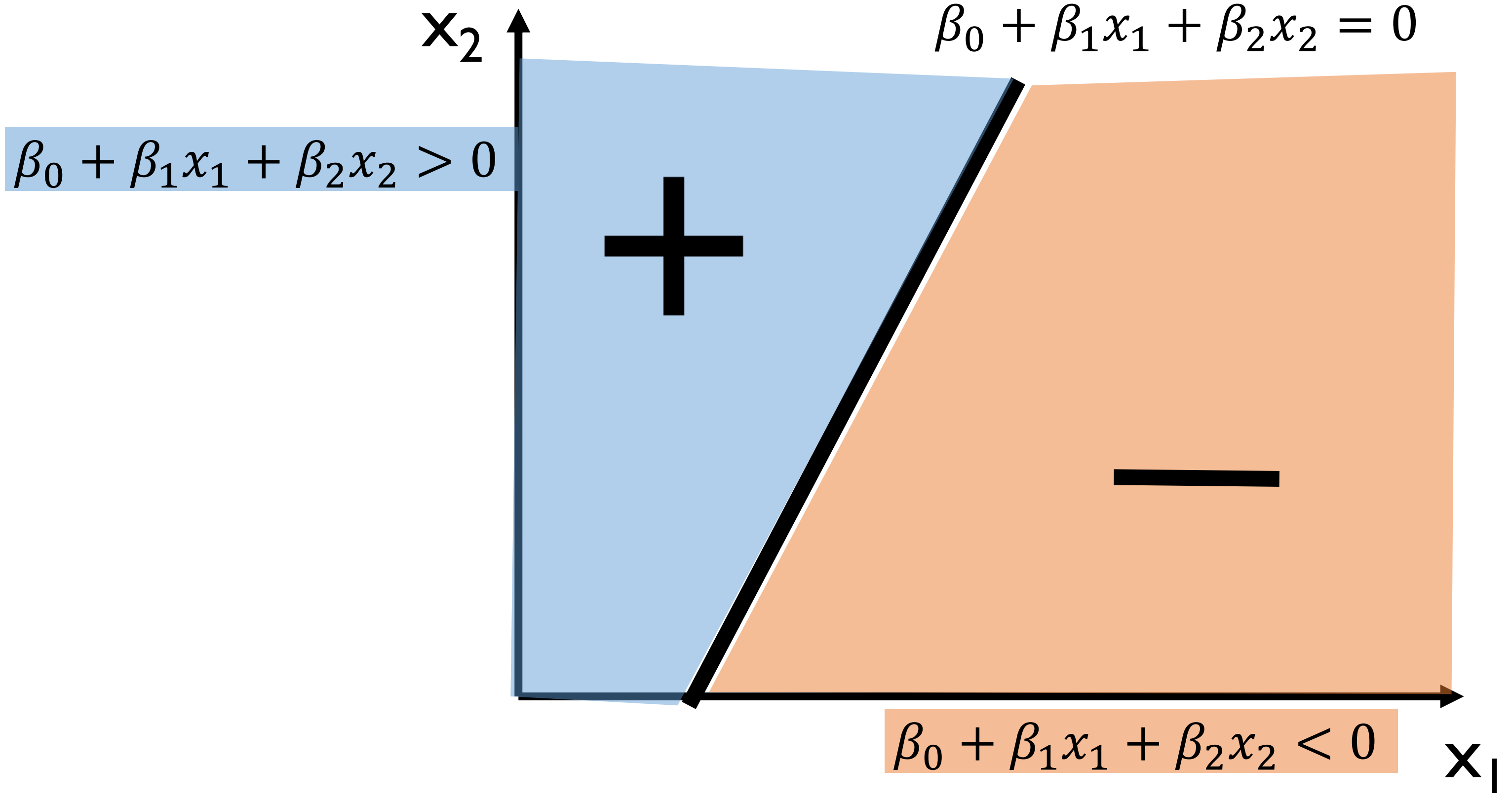
Do we need probabilities or decision boundaries?

Find "optimal" boundary to separate classes



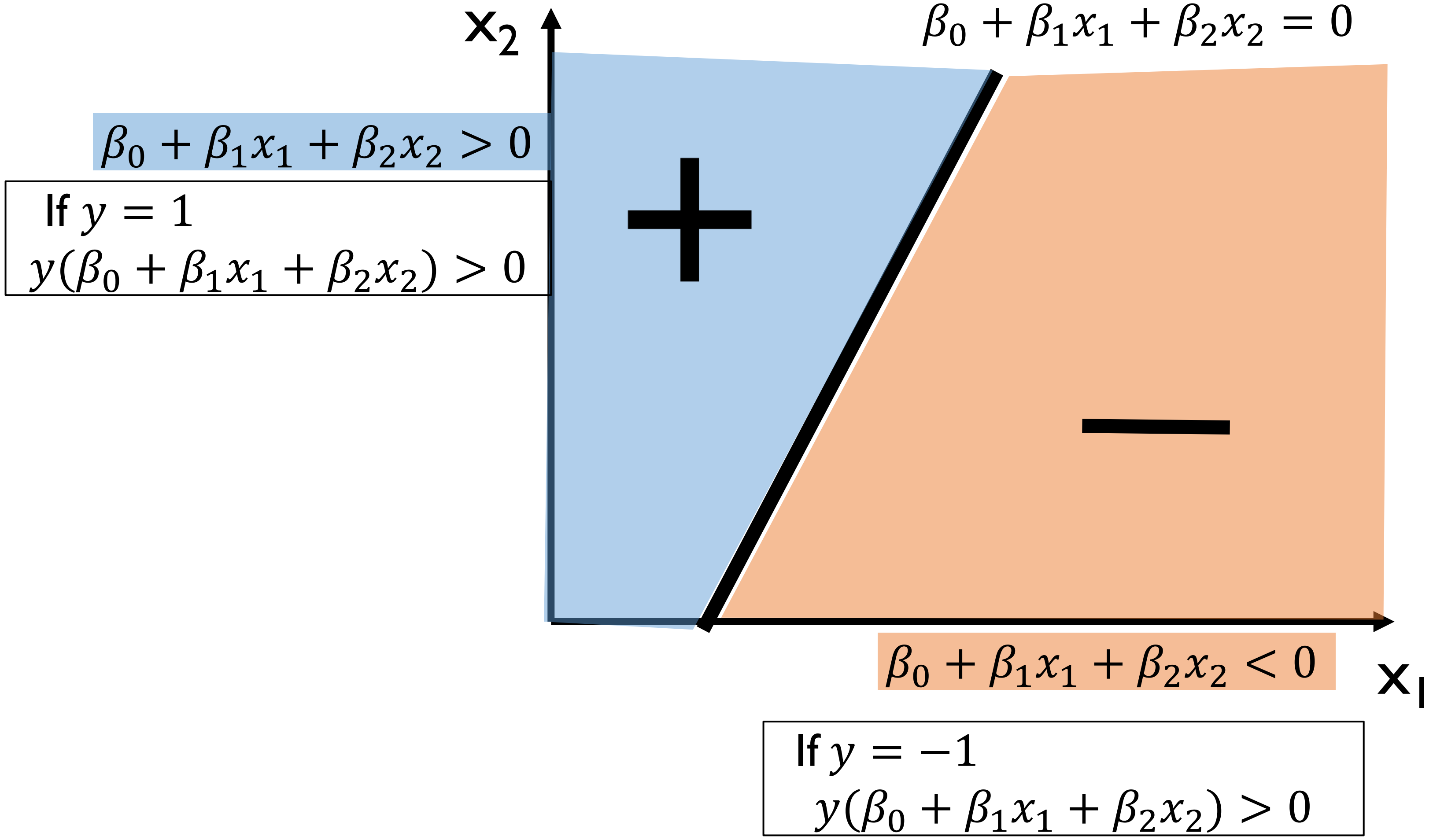
Do we need probabilities or decision boundaries?

Find "optimal" boundary to separate classes



Do we need probabilities or decision boundaries?

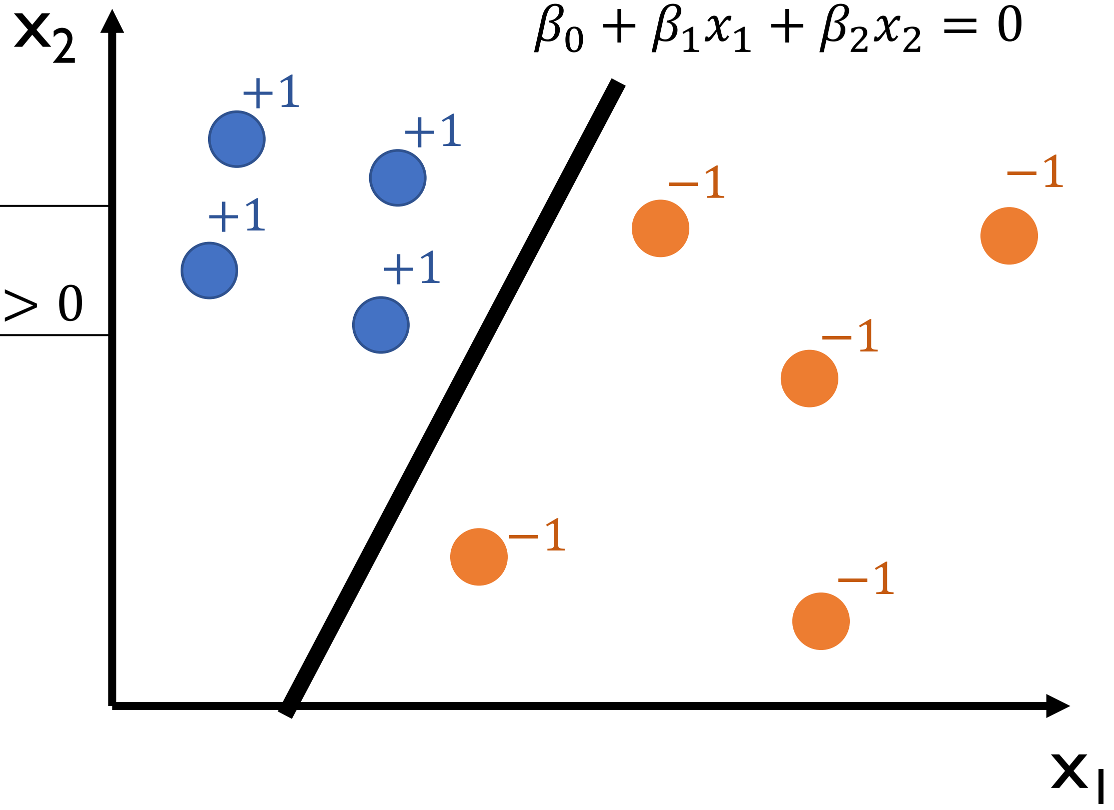
Find "optimal" boundary to separate classes



Do we need probabilities or decision boundaries?

Find "optimal" boundary to separate classes

If $y = 1$
$y(\beta_0 + \beta_1 x_1 + \beta_2 x_2) > 0$

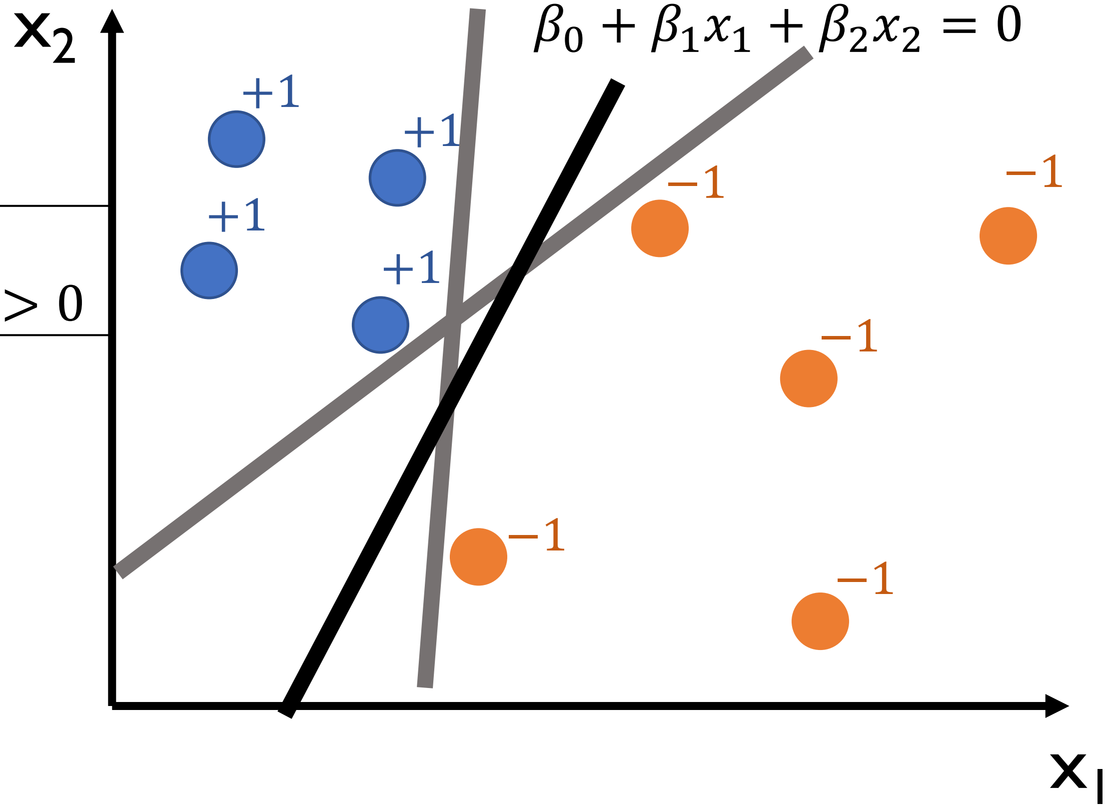


If $y = -1$
$y(\beta_0 + \beta_1 x_1 + \beta_2 x_2) > 0$

Do we need probabilities or decision boundaries?

Find **"optimal"** boundary to separate classes

If $y = 1$
$y(\beta_0 + \beta_1 x_1 + \beta_2 x_2) > 0$

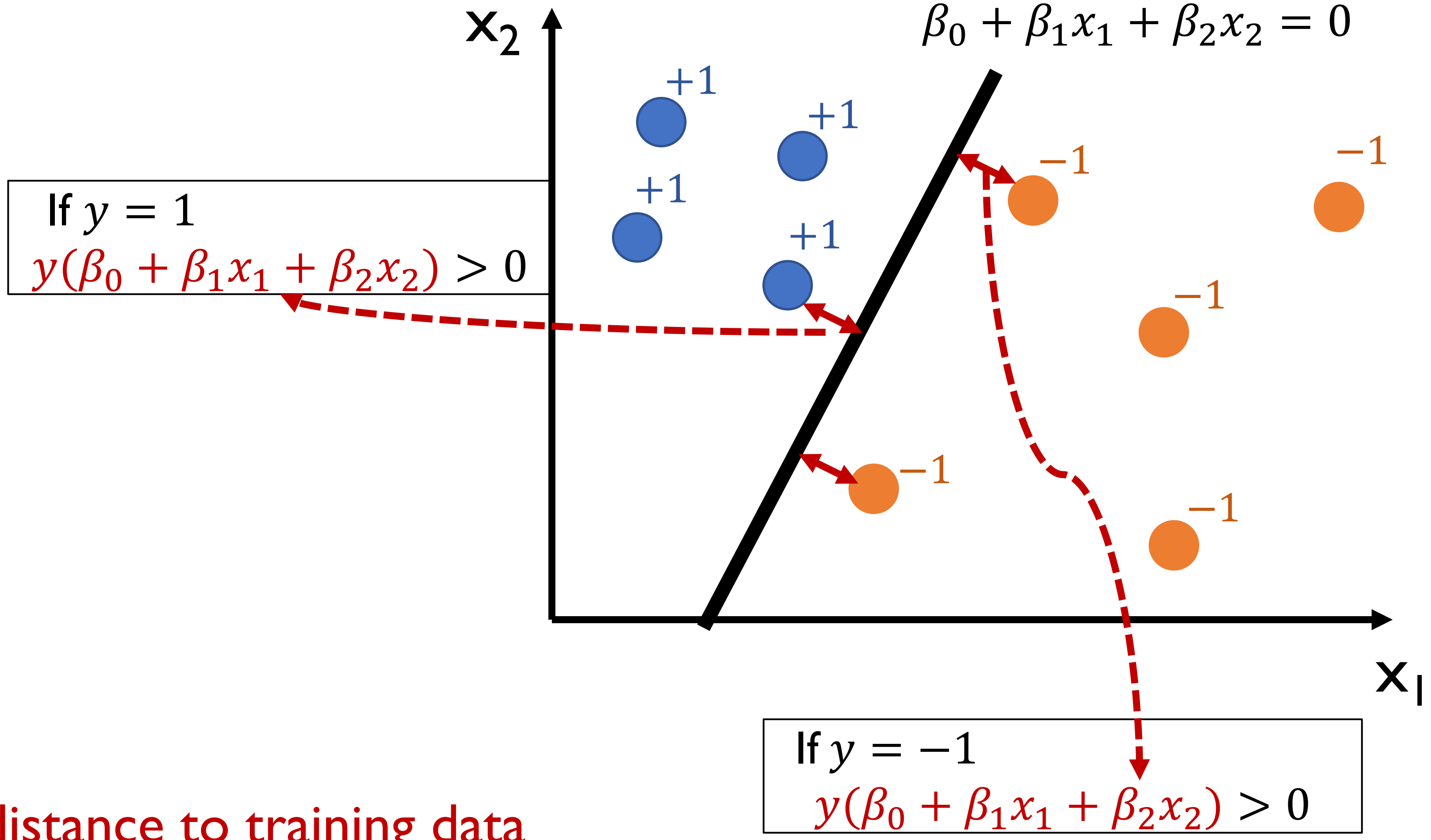


If $y = -1$
$y(\beta_0 + \beta_1 x_1 + \beta_2 x_2) > 0$

Optimal: Farthest distance to training data

Maximal Margin Classifier

Find **“optimal”** boundary to separate classes



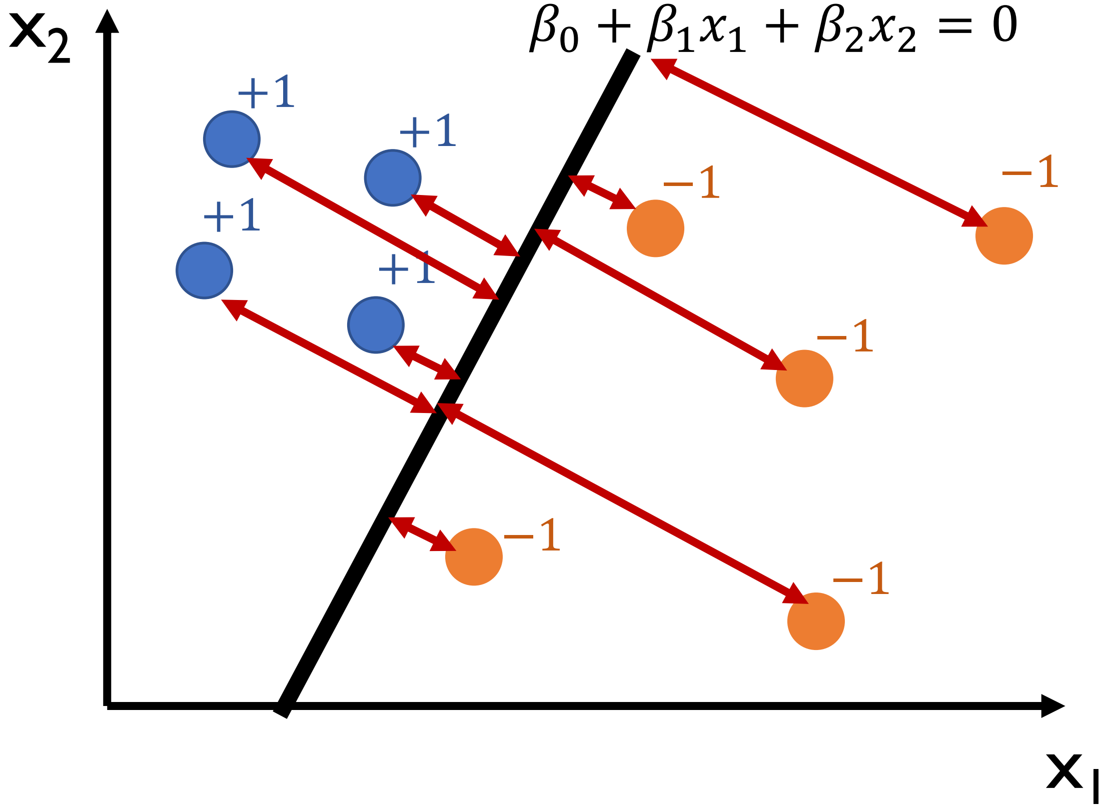
Optimal: Farthest distance to training data

Maximal Margin Classifier

$$\max_{\beta_0,\beta_1,\beta_2} M$$

such that $\beta_0^2 + \beta_1^2 + \beta_2^2 = 1$

For all training data
$$y^{(i)}\left(\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)}\right) \geq M$$



Optimal: Farthest distance to training data

Maximal Margin Classifier

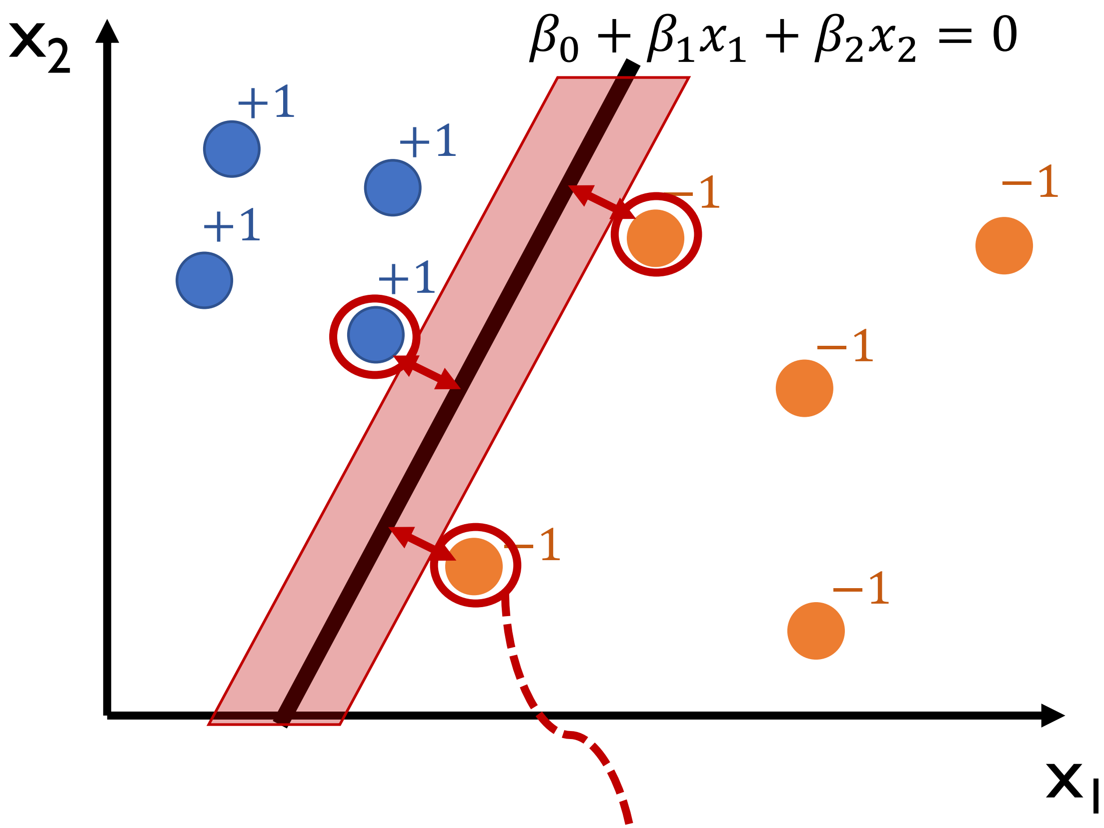
$$\max_{\beta_0,\beta_1,\beta_2} \; M$$

such that $\beta_0^2 + \beta_1^2 + \beta_2^2 = 1$

For all training data

$$y^{(i)}\left(\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)}\right) \geq M$$

Only points with equality matter



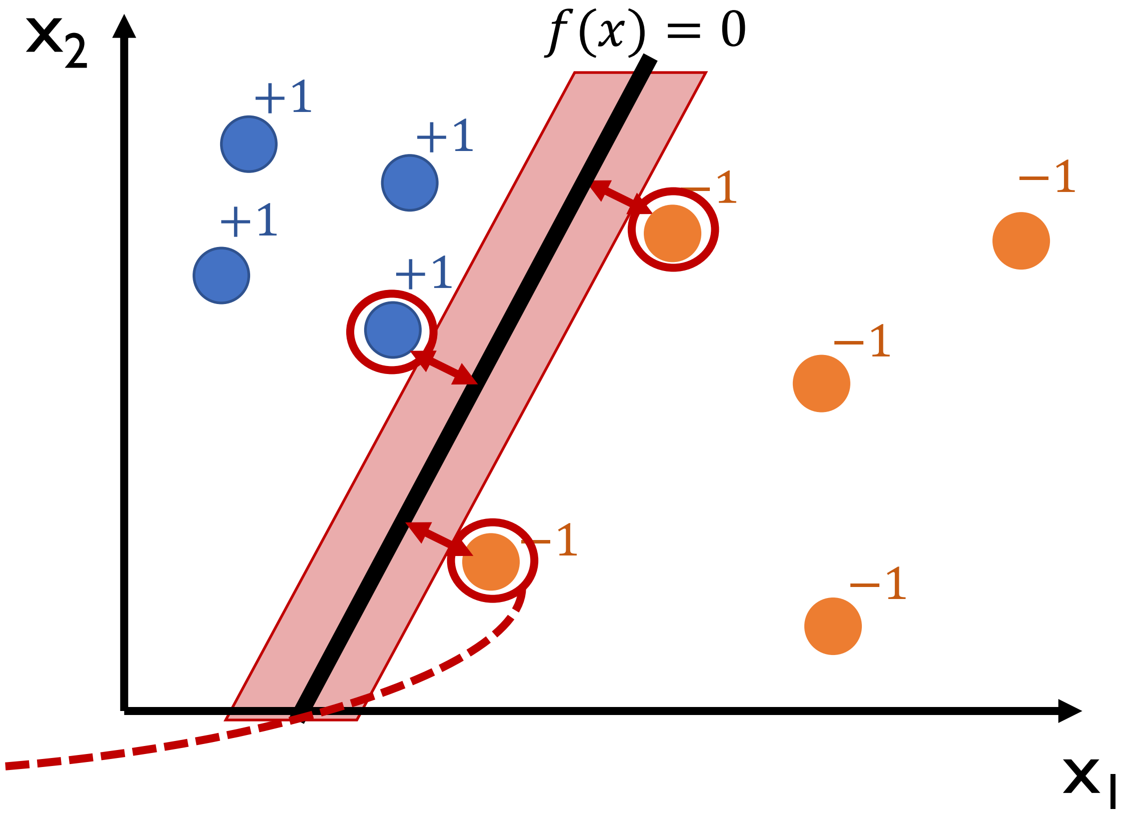
Support Vectors

Maximal Margin Classifier

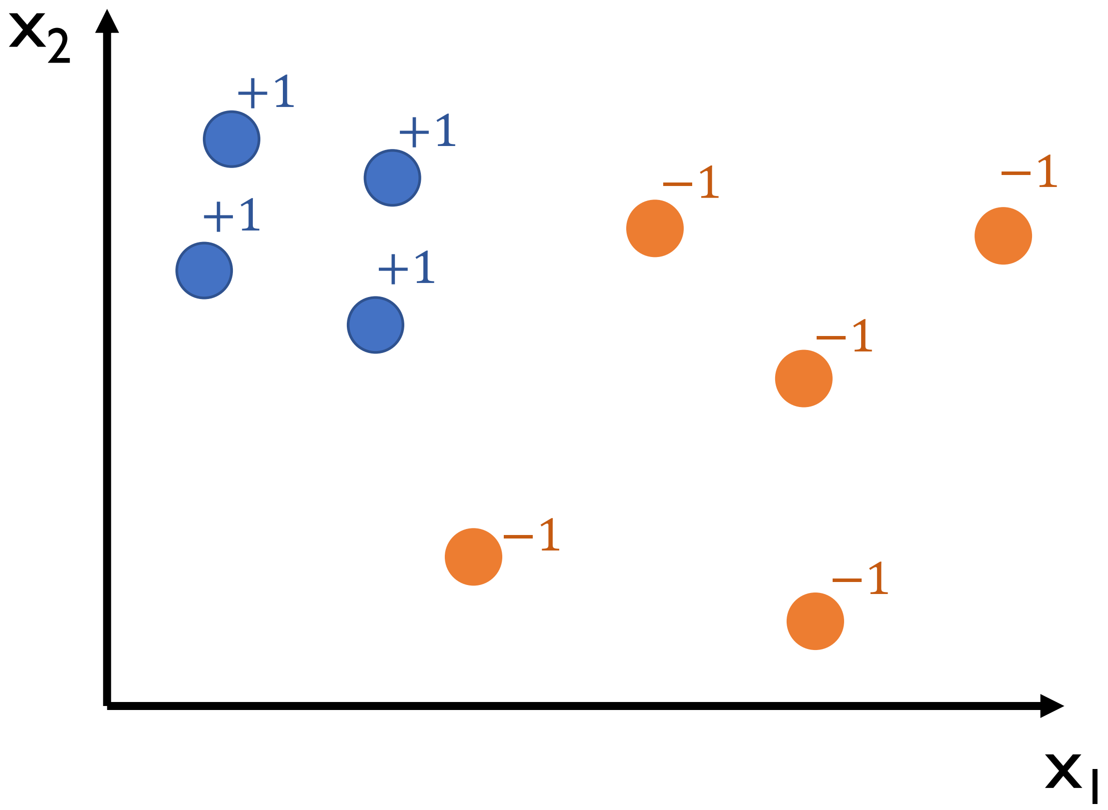
Solving optimization problem we find $\beta_0, \alpha_1, \ldots, \alpha_N$ such that hyperplane:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \, x^T x^{(i)}$$

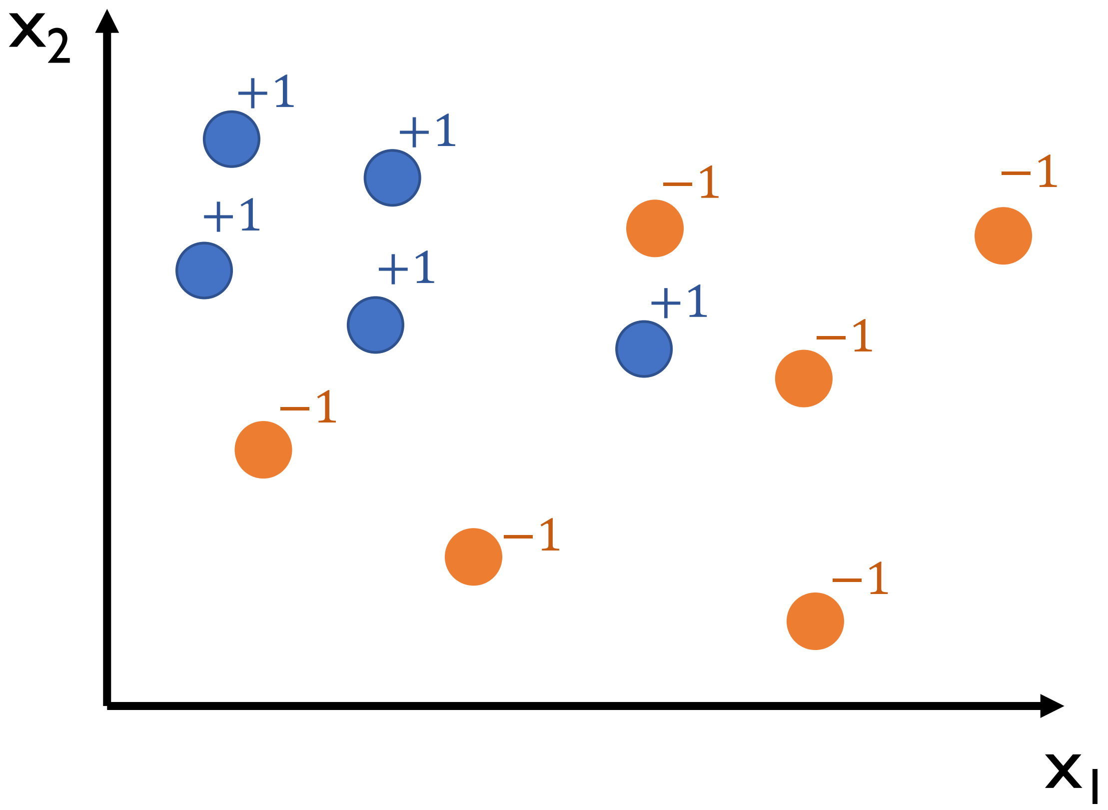
Support Vectors



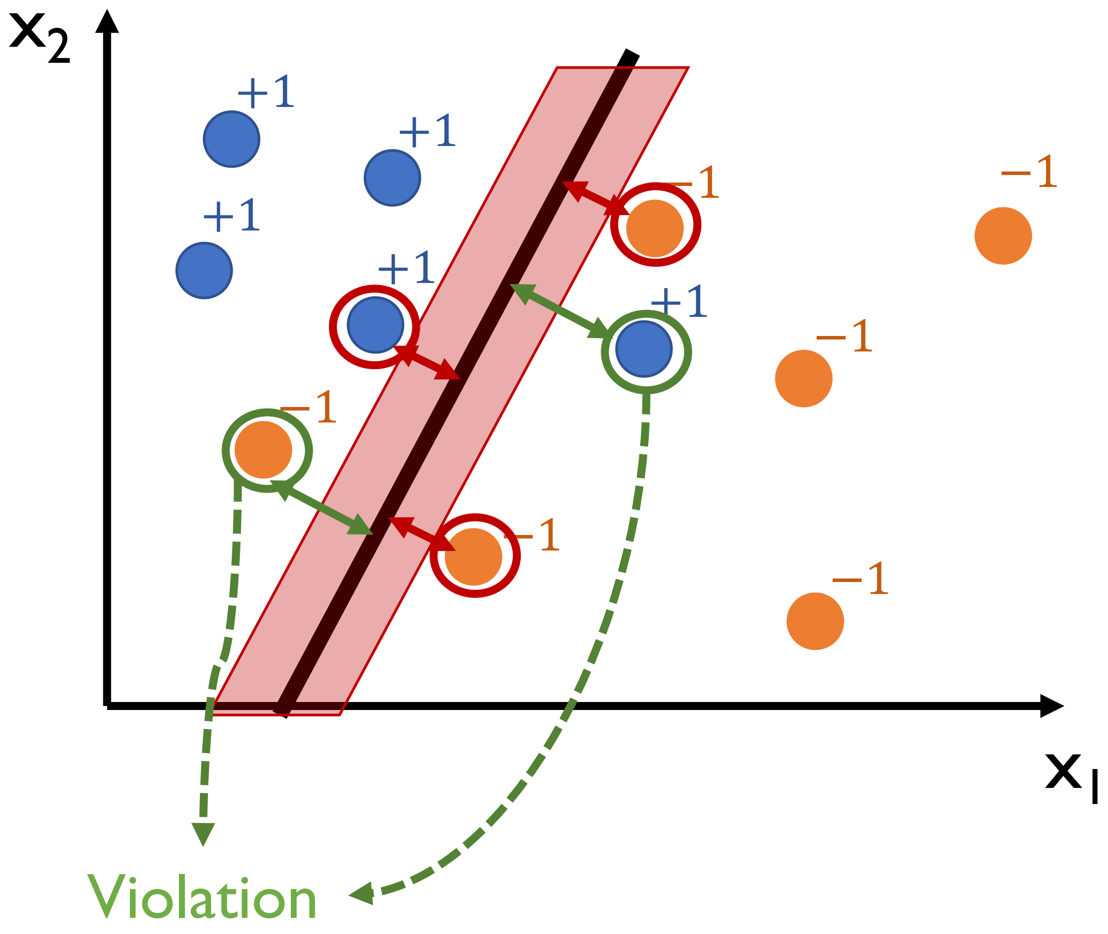
What if there is no separating hyperplane?



What if there is no separating hyperplane?



Support Vector Classifier



Support Vector Classifier

$$\max_{\beta_0,\beta_1,\beta_2} M$$

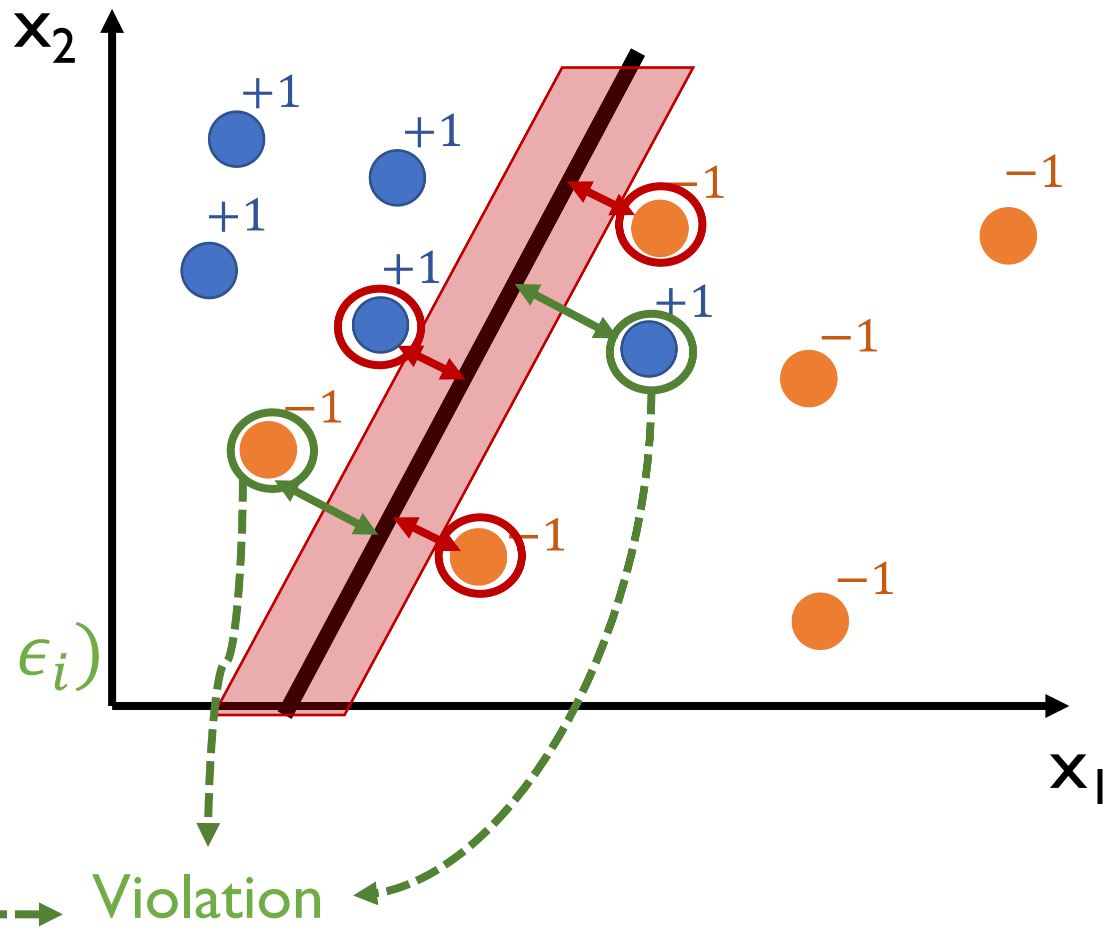
such that $\beta_0^2 + \beta_1^2 + \beta_2^2 = 1$

For all training data

$$y^{(i)}\left(\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)}\right) \geq M(1 - \epsilon_i)$$

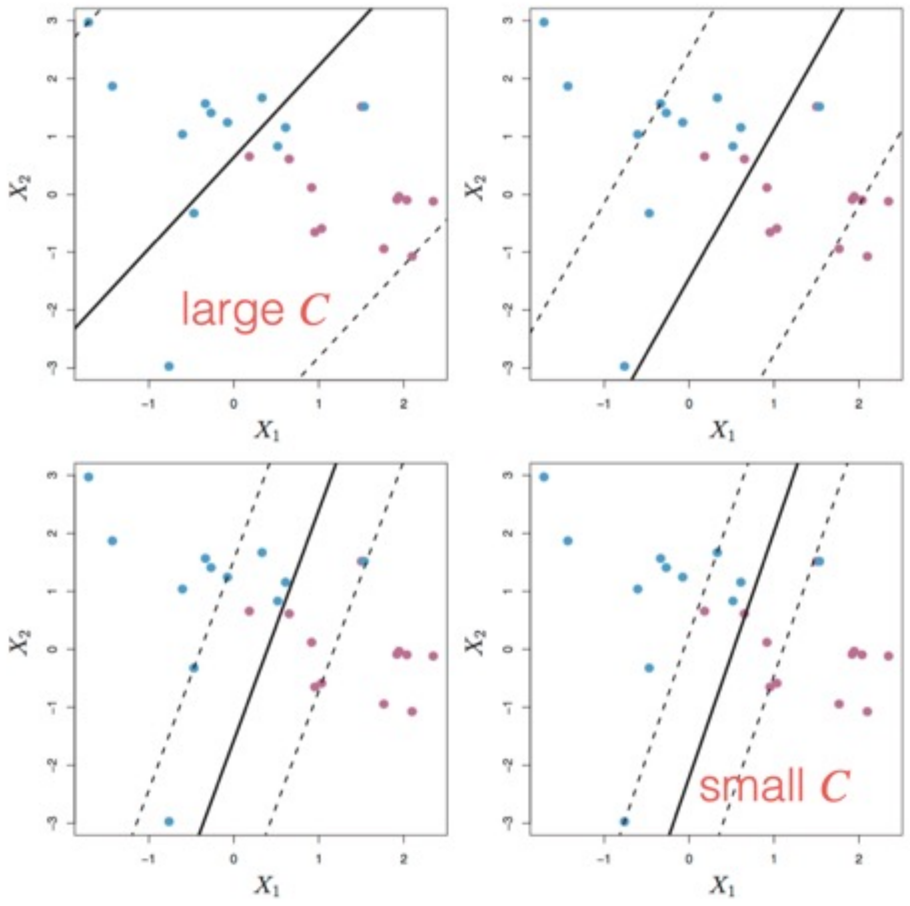
such that

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



Support Vector Classifier

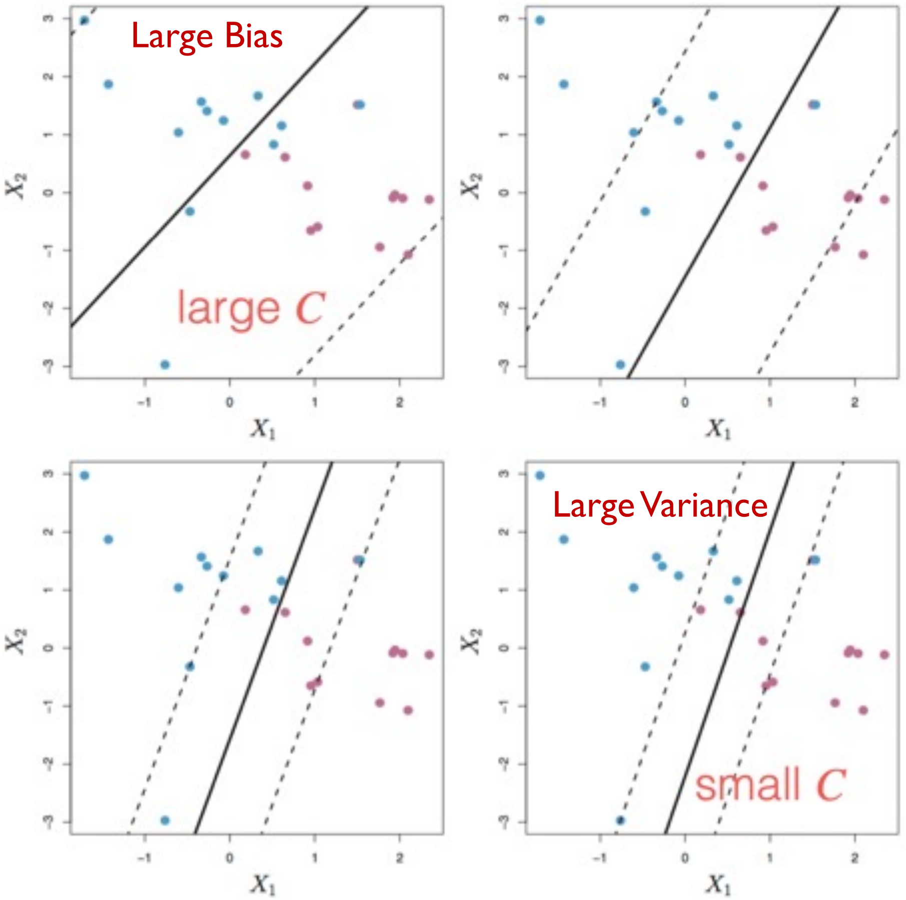
C is the "budget" for violations



ISL Fig 9.7

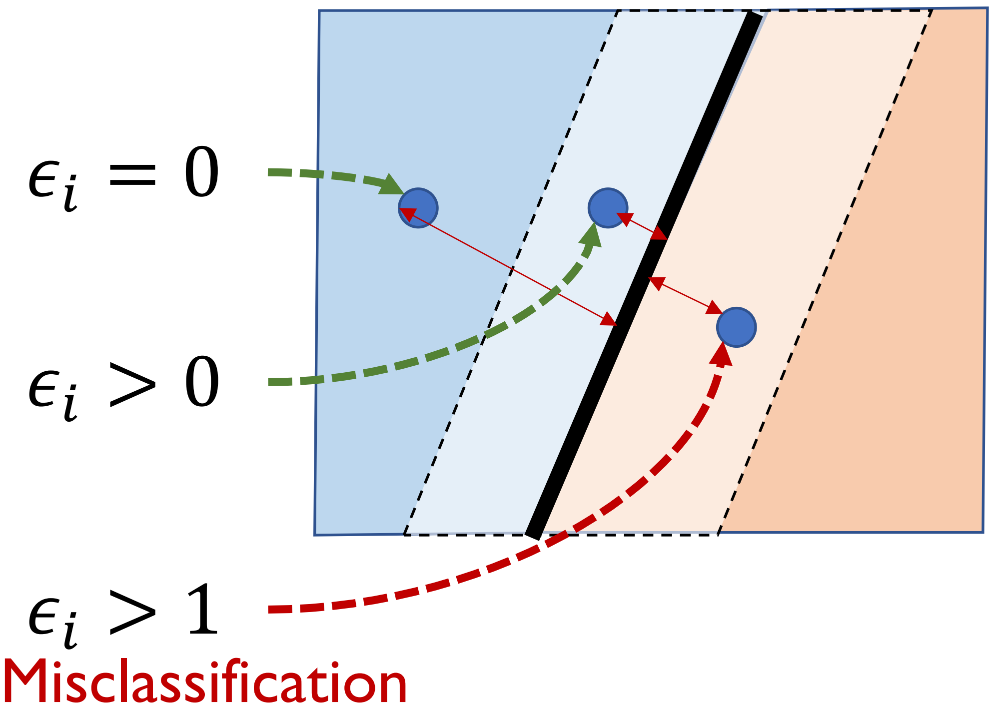
Support Vector Classifier

C is the "budget" for violations



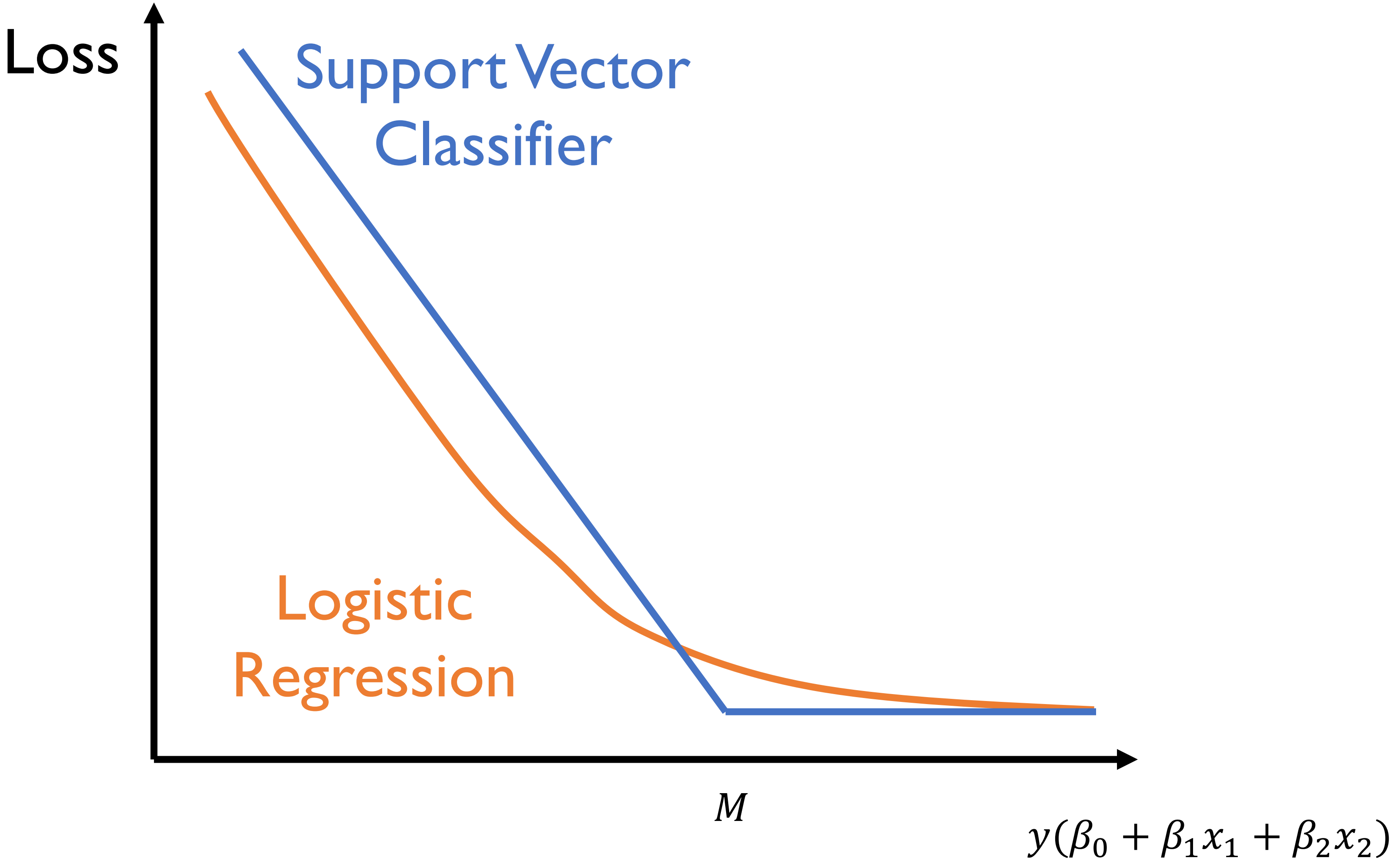
ISL Fig 9.7

Slack variables ϵ_i



Support Vector Classifier vs Logistic Regression

Similar Behavior

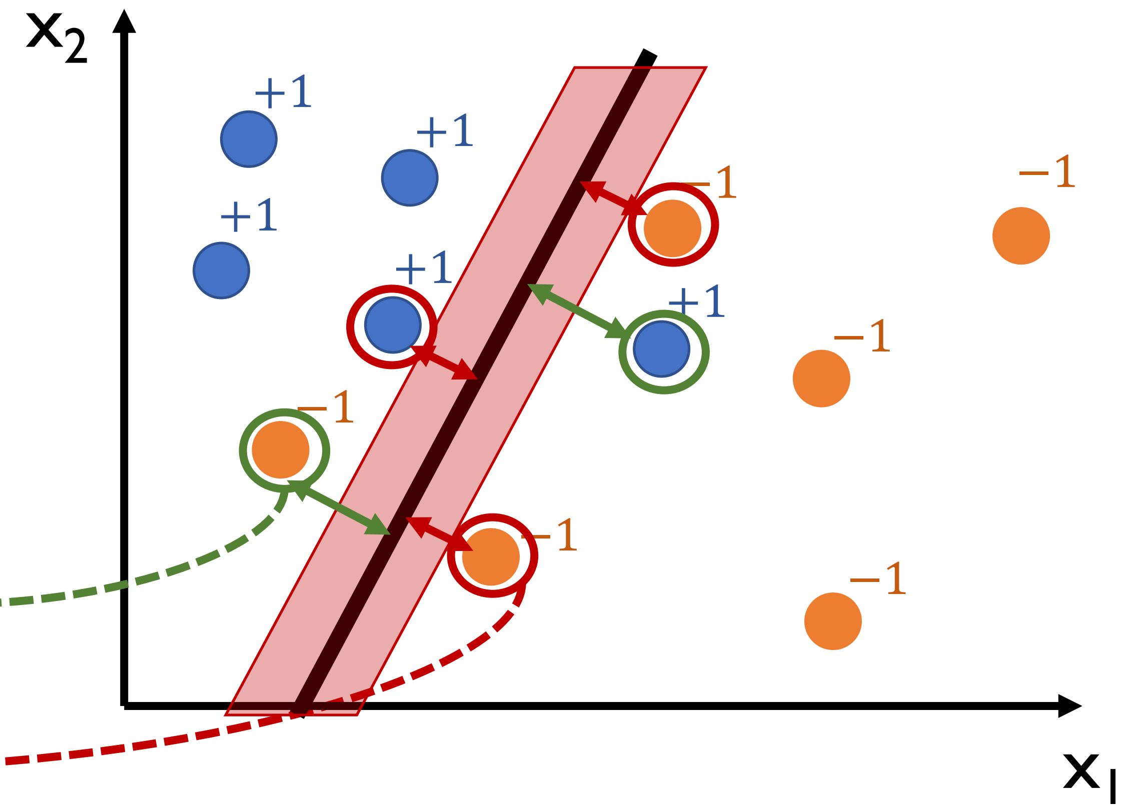


Support Vector Classifier

Solving optimization problem we find $\beta_0, \alpha_1, \ldots, \alpha_N$ such that hyperplane:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \, x^T x^{(i)}$$

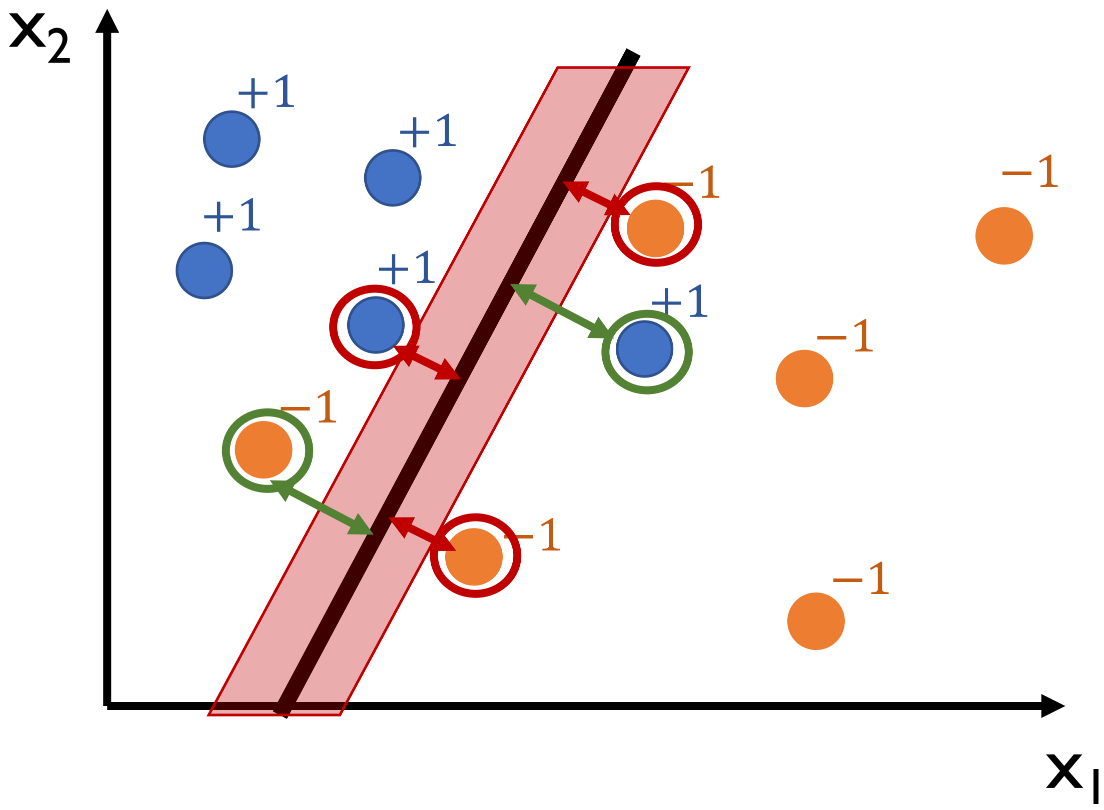
Support Vectors



Support Vector Classifier

Solving optimization problem we find $\beta_0, \alpha_1, \ldots, \alpha_N$ such that hyperplane:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \boxed{x^T x^{(i)}}$$



We only get linear boundaries

Beyond linear decision boundaries

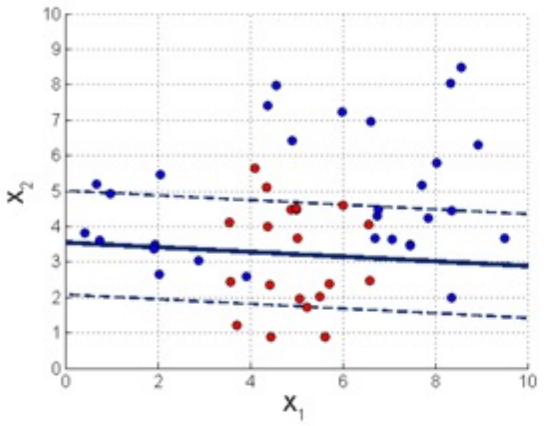
$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \boxed{x^T x^{(i)}}$$

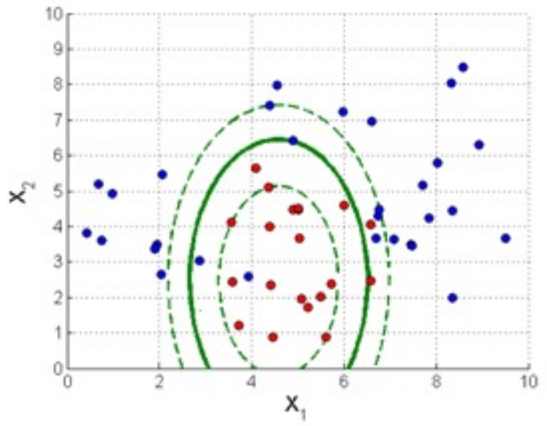
Option 1)
Add additional features

$$X_1, X_2, X_1X_2, X_1^2, X_2^2, \ldots$$

Option 2)
Generalize inner product

Kernels $\quad x^T x^{(i)} \rightarrow K(x, x^{(i)})$





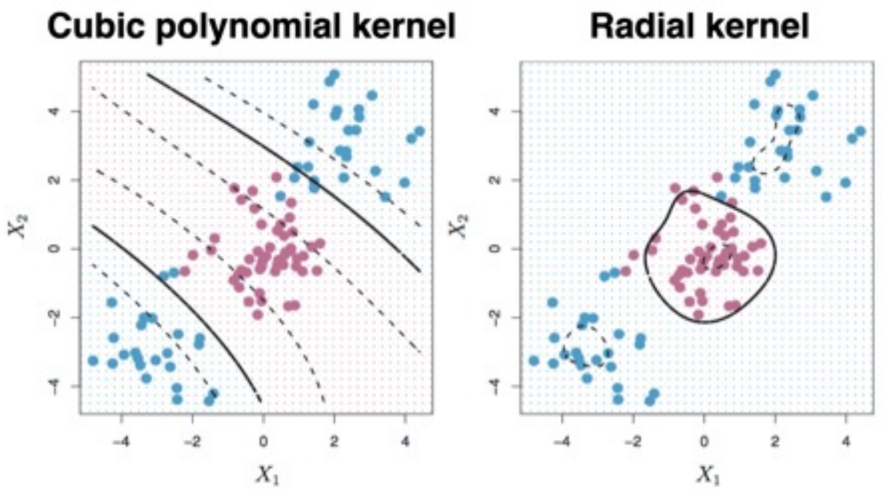


FIGURE 9.9, ISL (8th printing 2017)

Support Vector Machines

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \boxed{K(x, x^{(i)})}$$

Linear Kernel $K\left(x, x^{(i)}\right) = x^T x^{(i)}$

Polynomial Kernel $K\left(x, x^{(i)}\right) = \left(1 + x^T x^{(i)}\right)^p$ Includes $x_1^k x_2^l$, $k + l \le p$

Radial Basis Kernel $K\left(x, x^{(i)}\right) = \exp\left(-\gamma \|x - x^{(i)}\|^2\right)$ Includes infinite # features

Kernel ≈ Similarity Measure

Challenges Support Vector Machines

Hyperparameters: C and Kernel

✓ Flexibility

✗ Delicate tuning

✗ Difficult to Interpret

✓ Convex Optimization (easy)

Extend to more than 2 classes

One vs one
1 SVM per pair

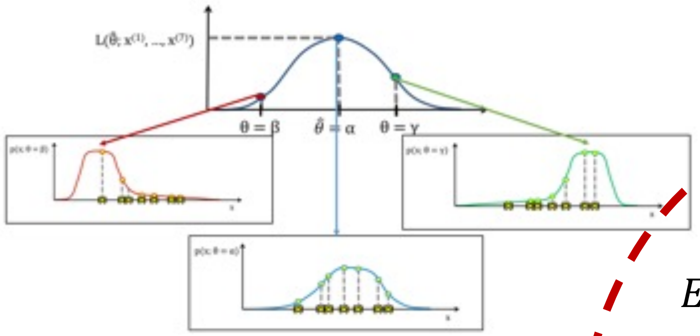
	C1	C2	C3	C4
C1				
C2				
C3				
C4				

One vs all
1 SVM per class

C1	C2 C3 C4
C2	C1 C3 C4
C3	C1 C2 C4
C4	C1 C2 C3

How can we extend Linear Regression?

LR is Maximum Likelihood estimator



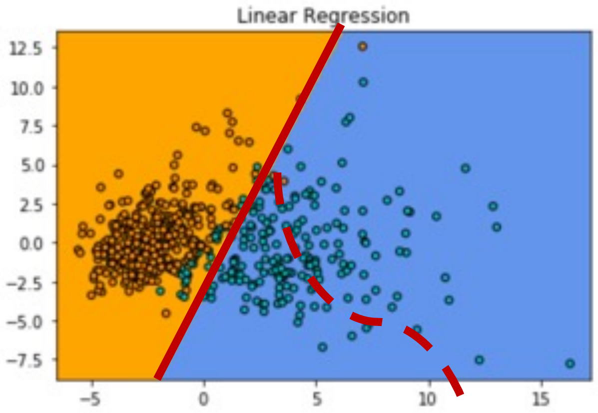
$$\boxed{Y \sim Normal}$$

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Find a better distribution for Y categorical

Logistic Regression

LR creates separating hyperplanes



Y

M $\longrightarrow$ 1

0.5

B $\longrightarrow$ 0

Optimize the hyperplane

Support Vector Machines

How do we measure the error?

Most common approach: Error rate

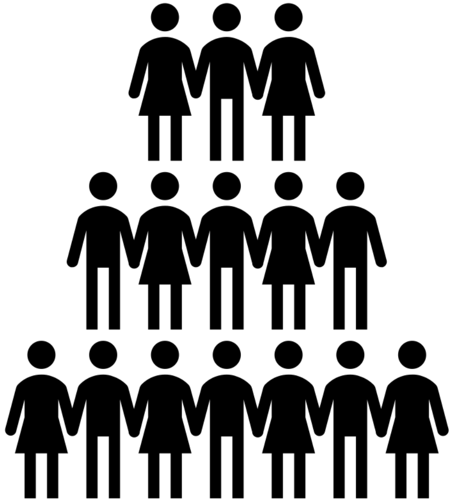
$$error = \frac{1}{N}\sum_{i=1}^{N} I\left(\hat{y}^{(i)} \neq y^{(i)}\right)$$

1 if $\hat{y}^{(i)} \neq y^{(i)}$
0 if $\hat{y}^{(i)} = y^{(i)}$

Most common approach: Error rate

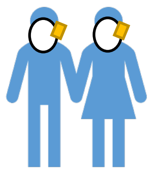
$$error = \frac{1}{N}\sum_{i=1}^{N} I\left(\hat{y}^{(i)} \neq y^{(i)}\right)$$

1 if $\hat{y}^{(i)} \neq y^{(i)}$
0 if $\hat{y}^{(i)} = y^{(i)}$



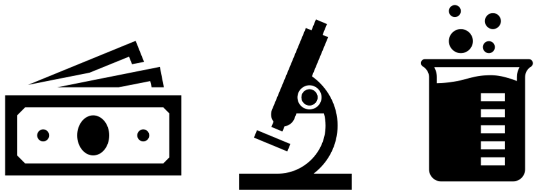
998 healthy

+



2 sick

Design test with
$error \leq 0.2\%$

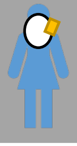
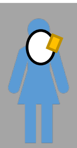
Option 1 

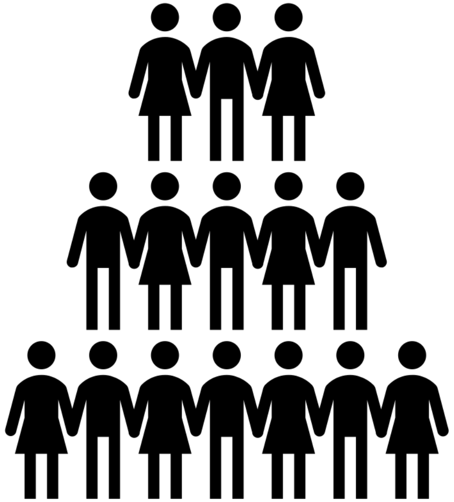
Option 2 Assign healthy to everyone

Useful approach: Confusion Matrix

Predicted Labels

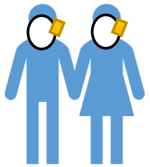
True Labels

		
	0	2
	0	998



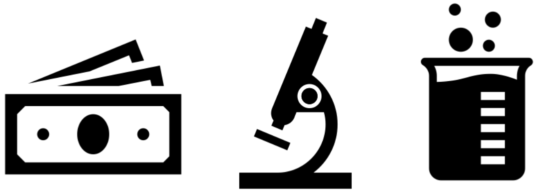
998 healthy

\+



2 sick

Design test with
$error \leq 0.2\%$

Option 1 

Option 2 Assign healthy to everyone

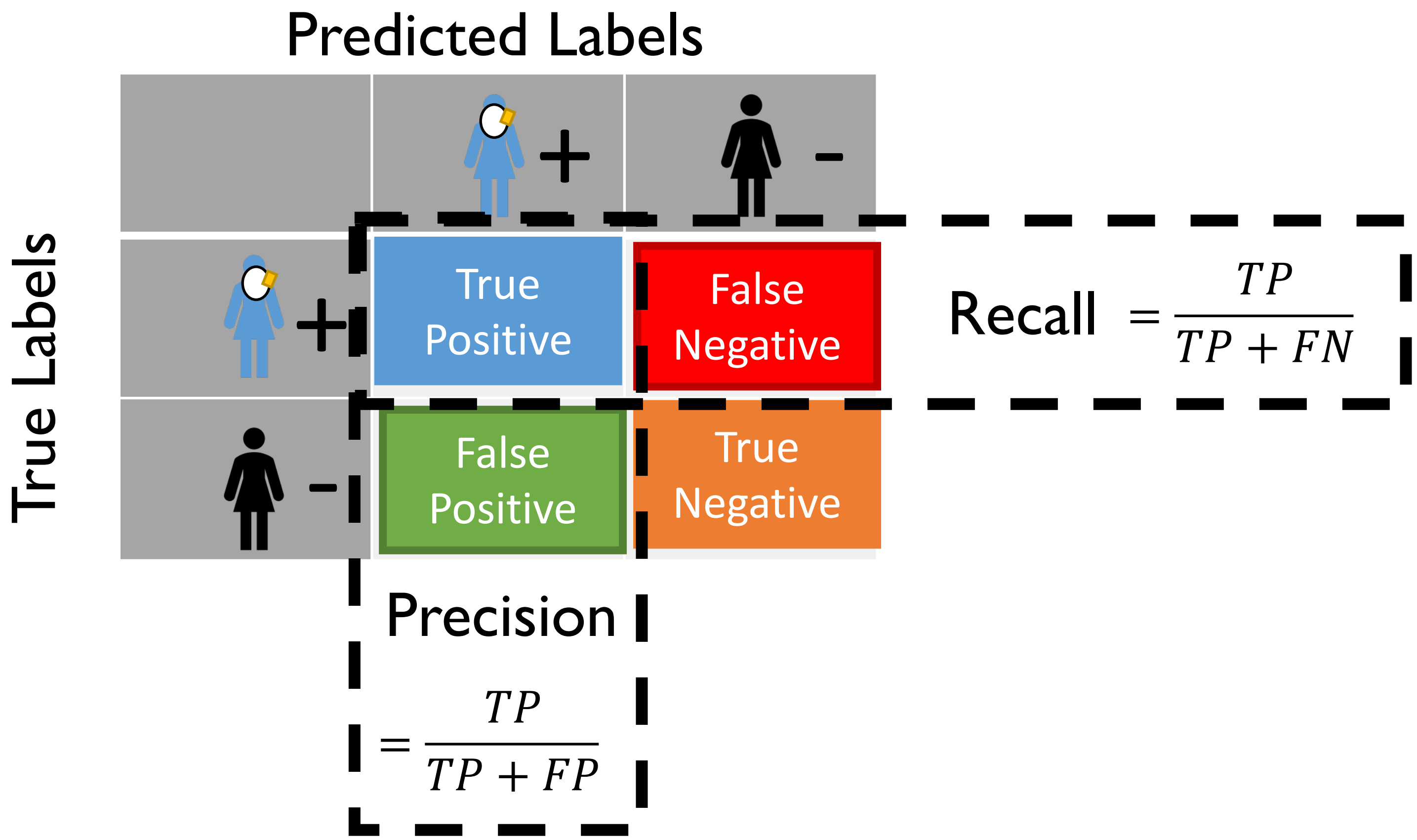
Useful approach: Confusion Matrix

Predicted Labels

True Labels

	+	-
+	True Positive	False Negative
-	False Positive	True Negative

Useful approach: Precision vs Recall



(+) samples correctly classified

(+) predictions that are truly(+)

Useful approach: Precision vs Recall

Predicted Labels

True Labels	+	-
+	0	2
-	0	998

Recall $= 0$

(+) samples correctly classified

Precision

$=?$

(+) predictions that are truly(+)

Useful approach: Precision vs Recall

Predicted Labels

True Labels	+	-
+	2	0
-	0	998

Recall $= 1$

(+) samples correctly classified

Precision $= 1$

(+) predictions that are truly(+)

Perfect Classifier

Useful approach: Precision vs Recall

Predicted Labels

True Labels	+	-
+	1	1
-	499	499

Recall $= 0.5$

(+) samples correctly classified

Precision $= 0.002$

(+) predictions that are truly(+)

Random Guessing 50/50

Useful approach: ROC curve

Predicted Labels

True Labels	+	-
+	True Positive	False Negative
-	False Positive	True Negative

$$\text{TP Rate} = \frac{TP}{TP + FN}$$

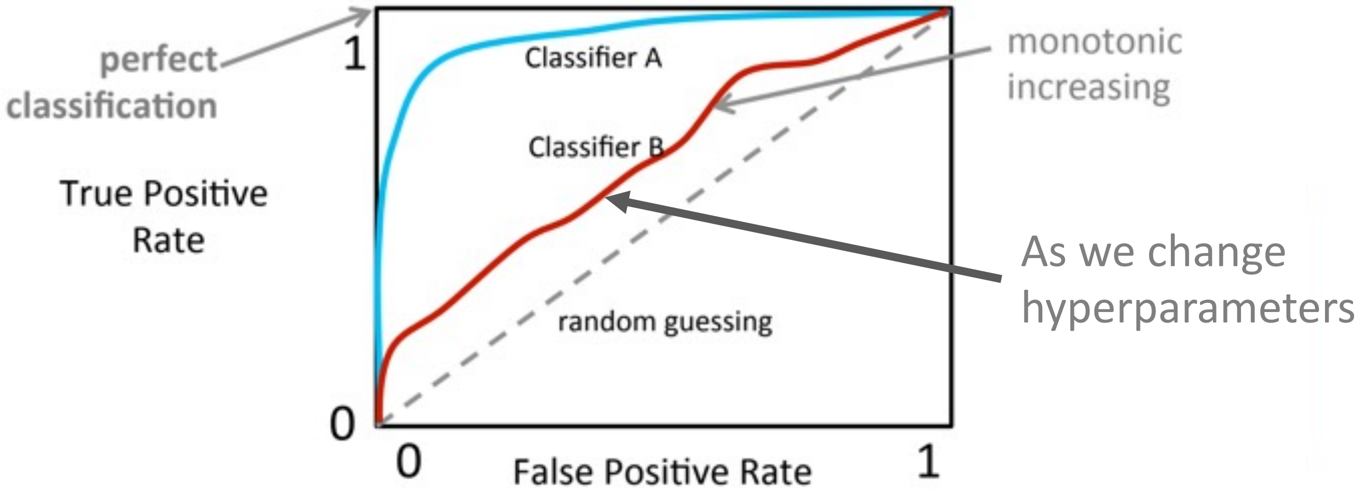
$$\text{FP Rate} = \frac{FP}{FP + TN}$$

Useful approach: ROC curve

True Labels \ Predicted Labels	+	-
+	True Positive	False Negative
-	False Positive	True Negative

$$\text{TP Rate} = \frac{TP}{TP + FN}$$

$$\text{FP Rate} = \frac{FP}{FP + TN}$$

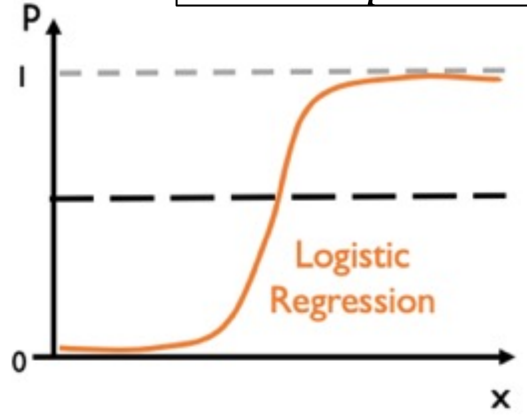


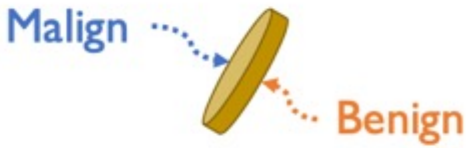
Today's Recap

Classification

Logistic Regression

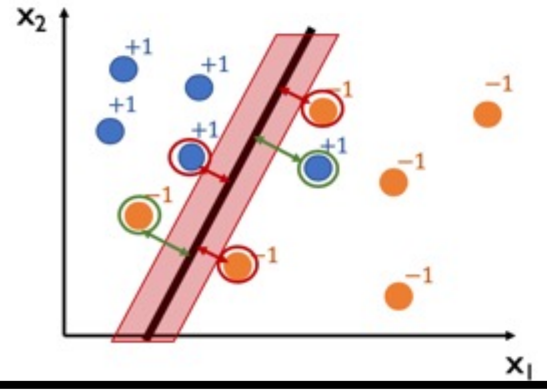
$$\boxed{\log\left(\frac{p}{1-p}\right) \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$





Support Vector Machines

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x^{(i)})$$



Evaluation:
Confusion Matrix