NYU Center for Data Science

DS-GA 1003
Machine Learning

Week 2: Lecture 2
Model Selection – Classification and Loss Functions

CC BY NC ND

NYU Center for Data Science

How can we put machine learning into practice?

DS-GA 1003
Machine Learning

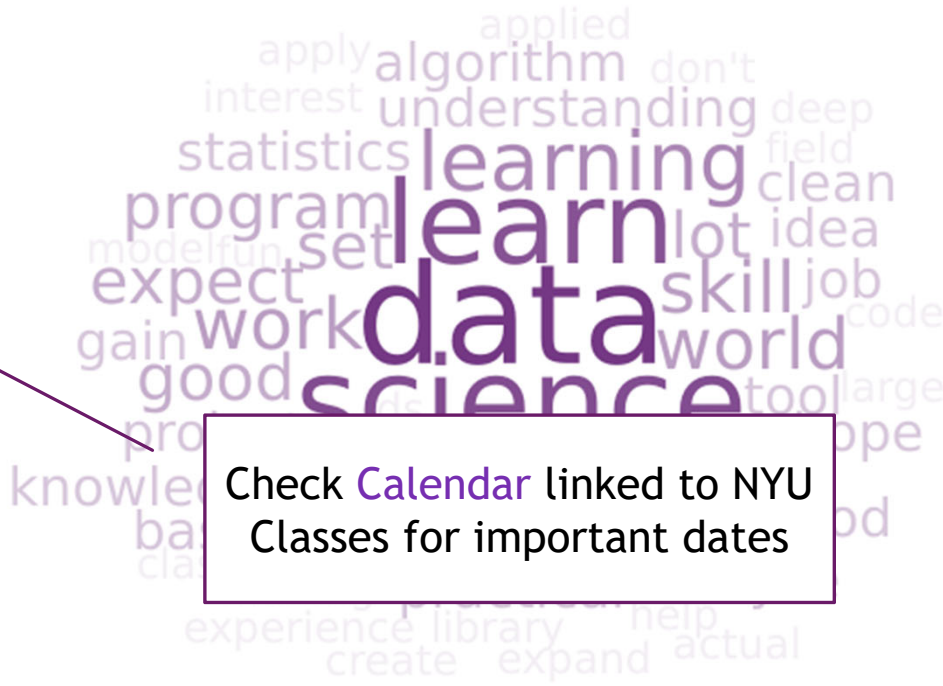
Week 2: Lecture 2
Model Selection – Classification and Loss Functions

Adapted from Rosenberg, Abu-Mostafa, Rangan, Shewchuk

CC BY NC ND

Announcements

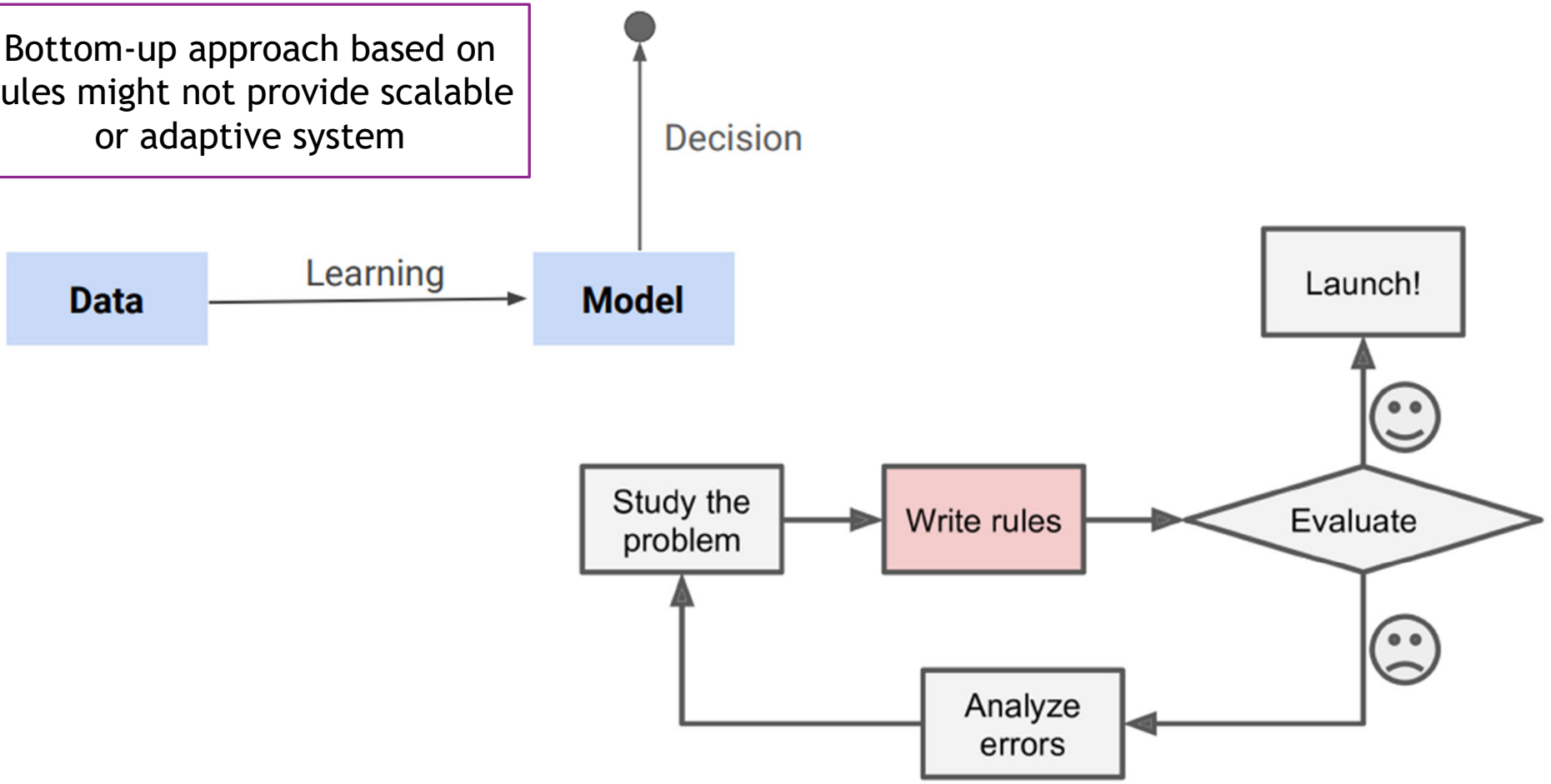
- Please check Week 2 agenda on NYU Classes
  - Homework 1
  - Survey 1
  - Section, Tutoring Session, Office Hours
- Remember to post to Piazza



CC BY NC ND

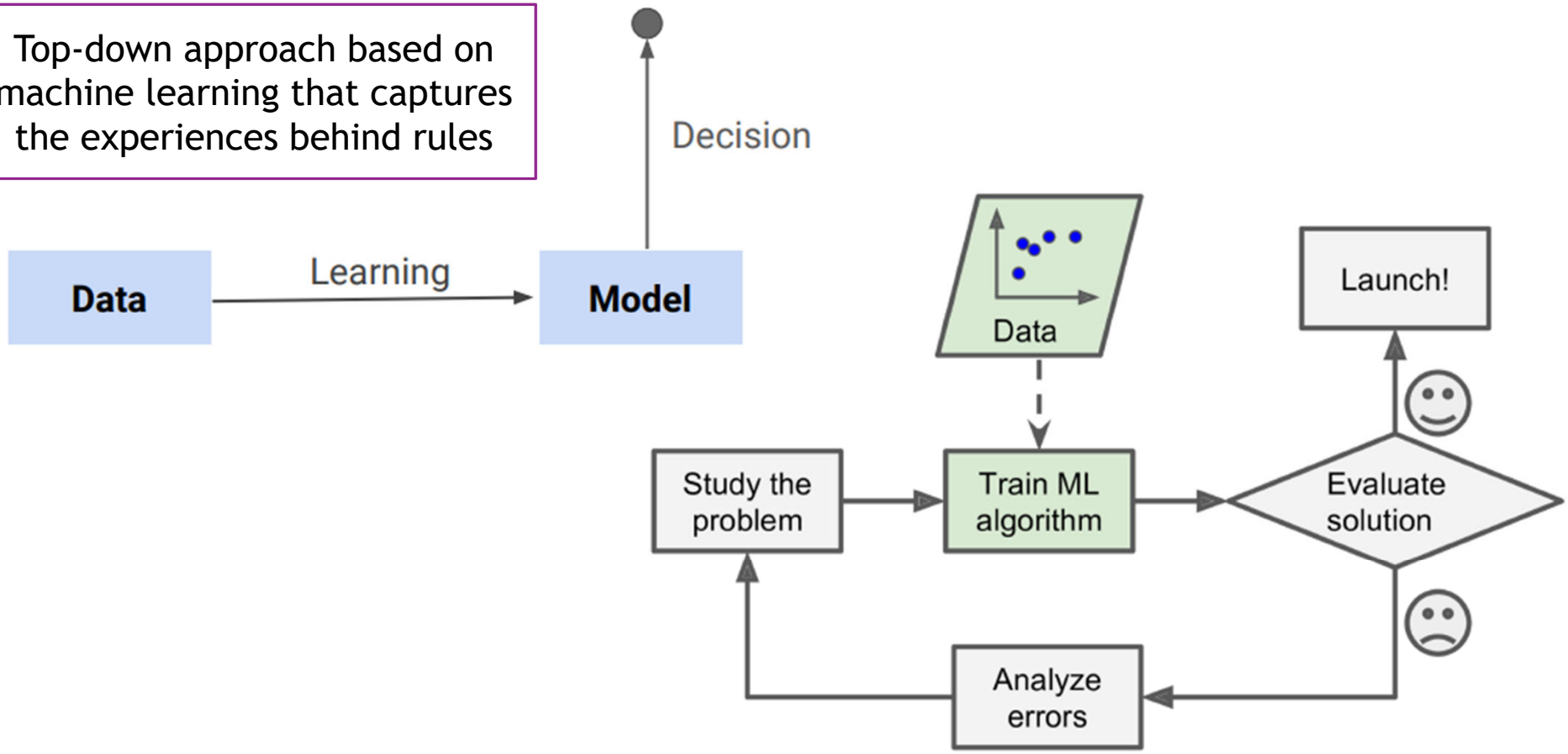
Review: What is machine learning?

Bottom-up approach based on rules might not provide scalable or adaptive system

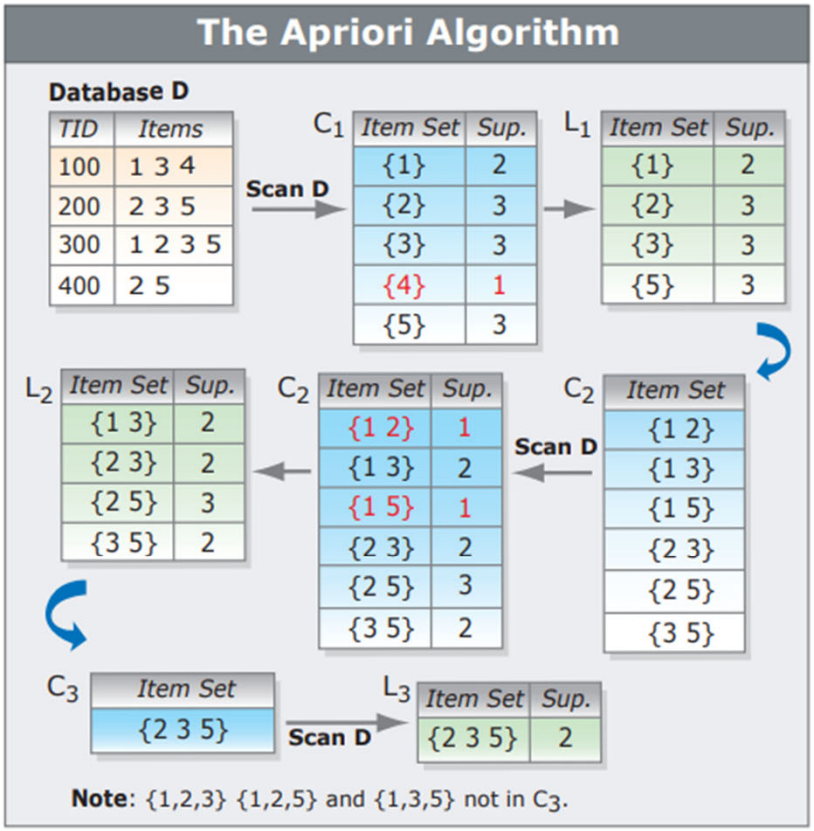


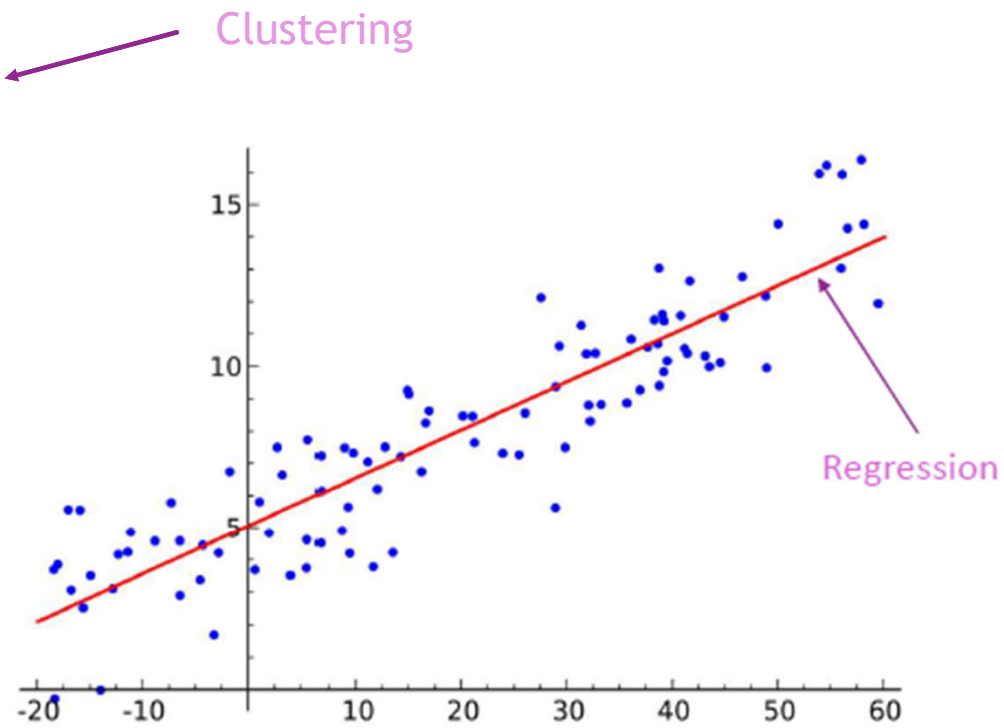
Review: What is machine learning?

Top-down approach based on machine learning that captures the experiences behind rules

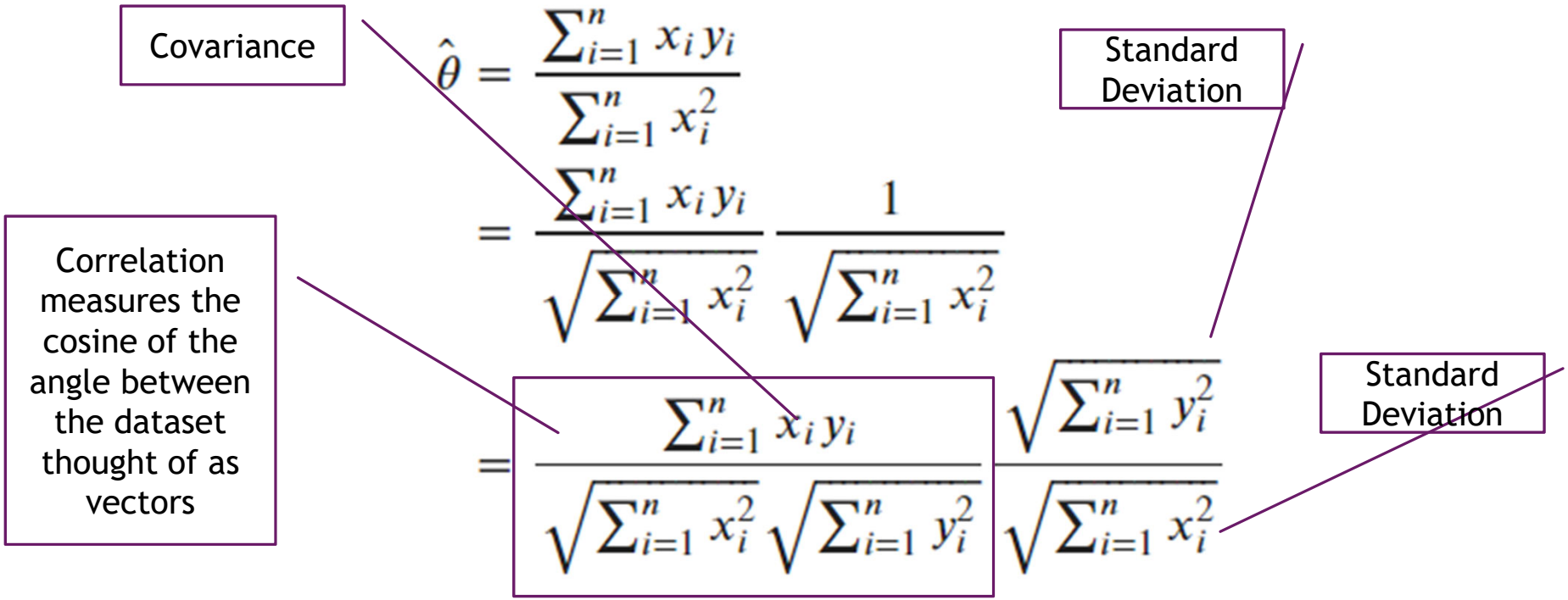


Review: Regression and Clustering





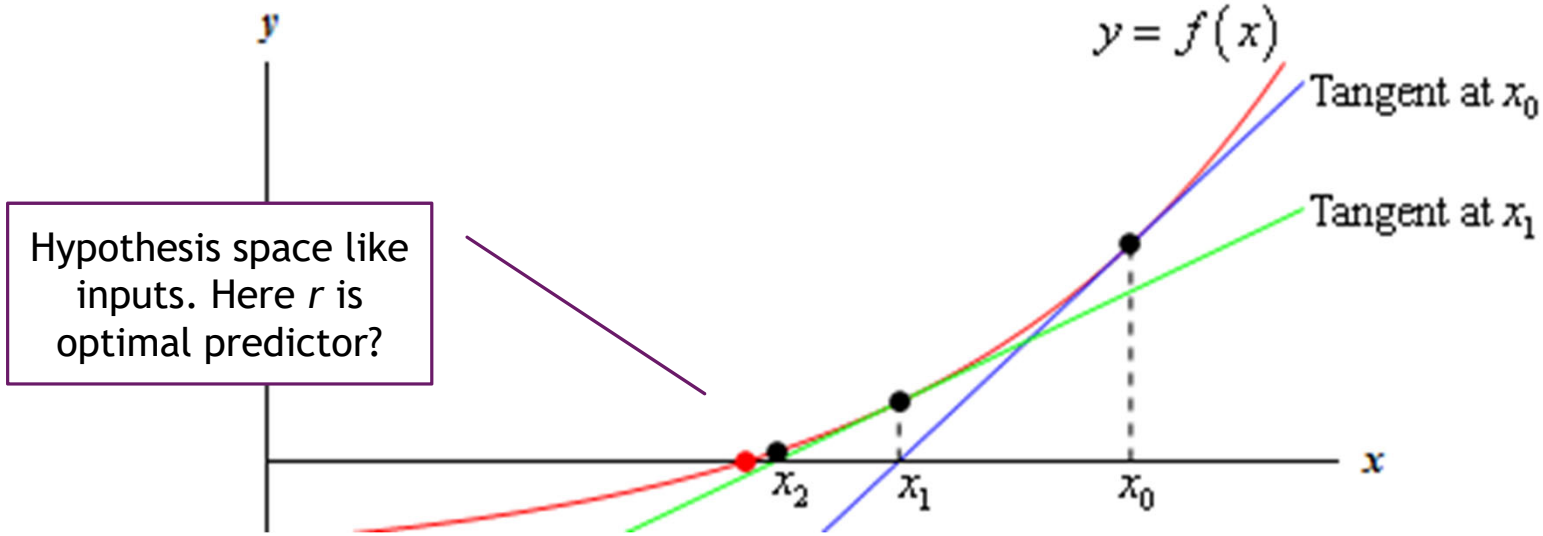
Review: Regression and Clustering



Steps for Machine Learning

- Loss Function l(x)
  - Set f(x) = l'(x) derivative
  - Find r such that f(r) = 0

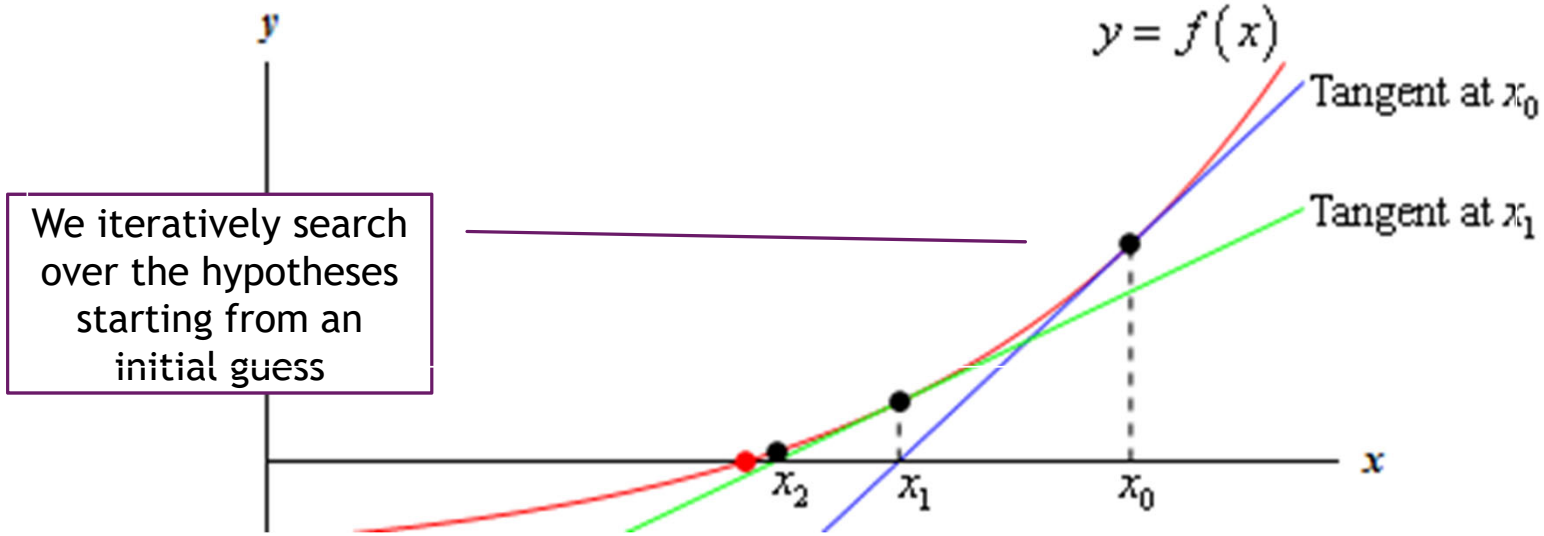
Analogy



Steps for Machine Learning

Analogy

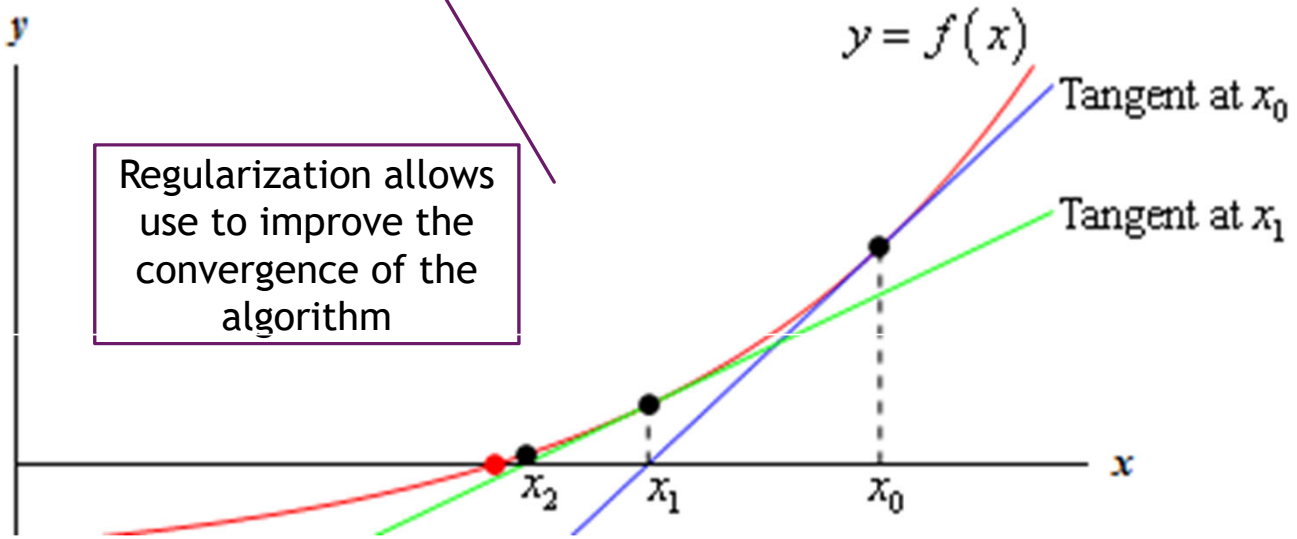
- Optimization
  - Set g(x) = x – (f(x) / f'(x))
  - Take $x_{t+1} = g(x_t)$



Steps for Machine Learning

- Regularization
  - Set g(x) = x – m (f(x) / f'(x))
  - Take $x_{t+1} = g(x_t)$

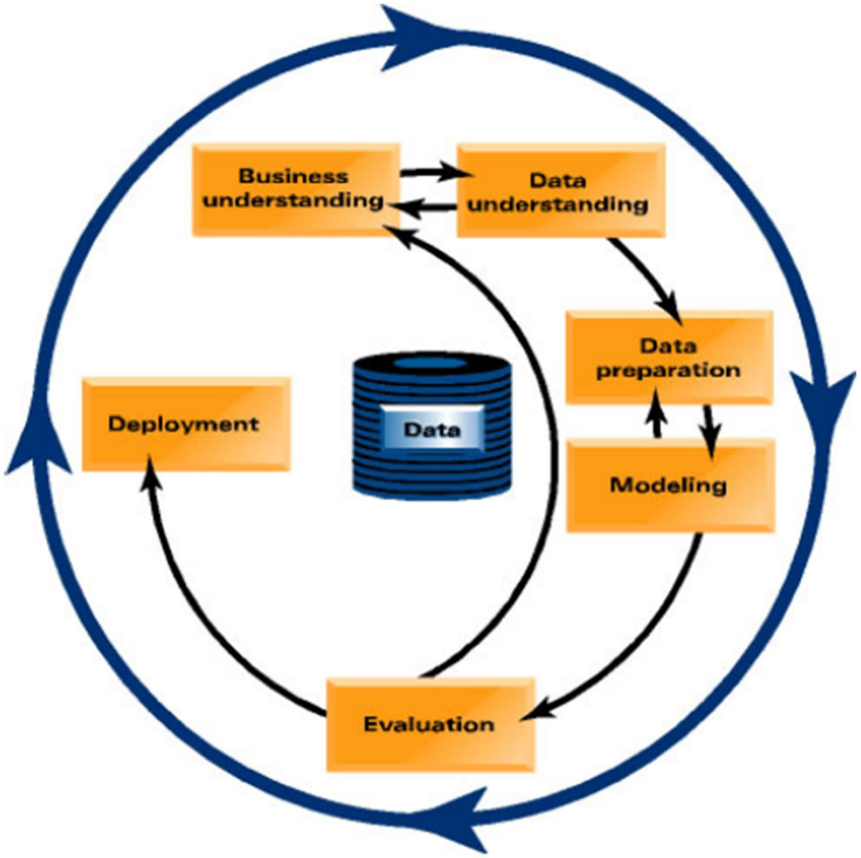
Analogy



Regularization allows use to improve the convergence of the algorithm

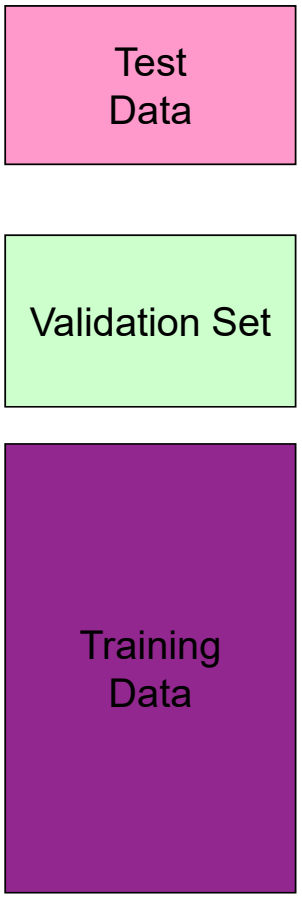
Agenda

- Steps for Machine Learning
  - Hypothesis Space
  - Loss Functions
  - Optimization
  - Regularization
- Putting Steps into Practice
  - Data
  - Features and Labels
  - Experimentation
  - Evaluation



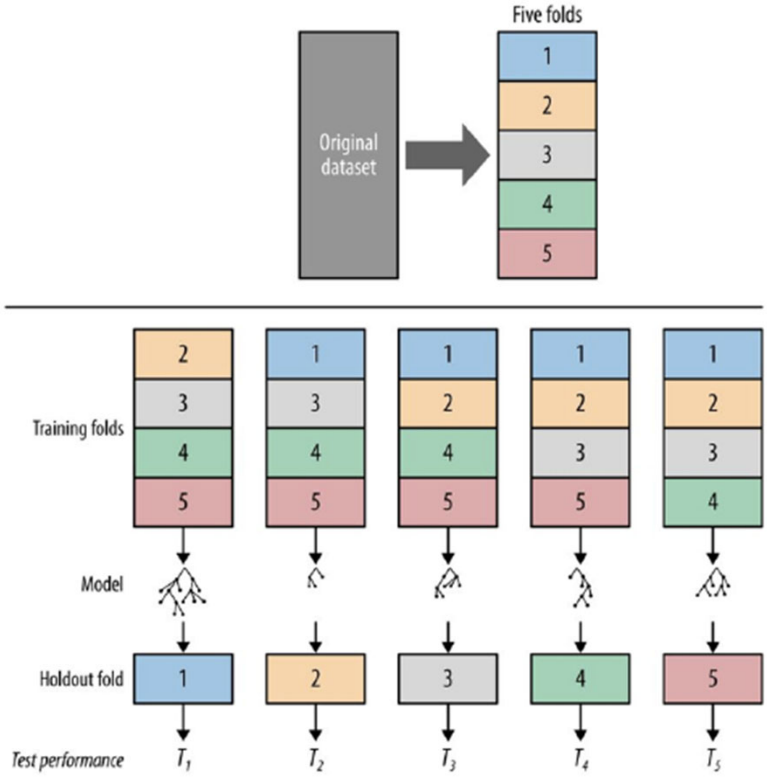
Data

- Data
  - Split
    - Training set
    - Held out set (sometimes call Validation set)
    - Test set
  - Randomly allocate to these three, e.g. 60/20/20



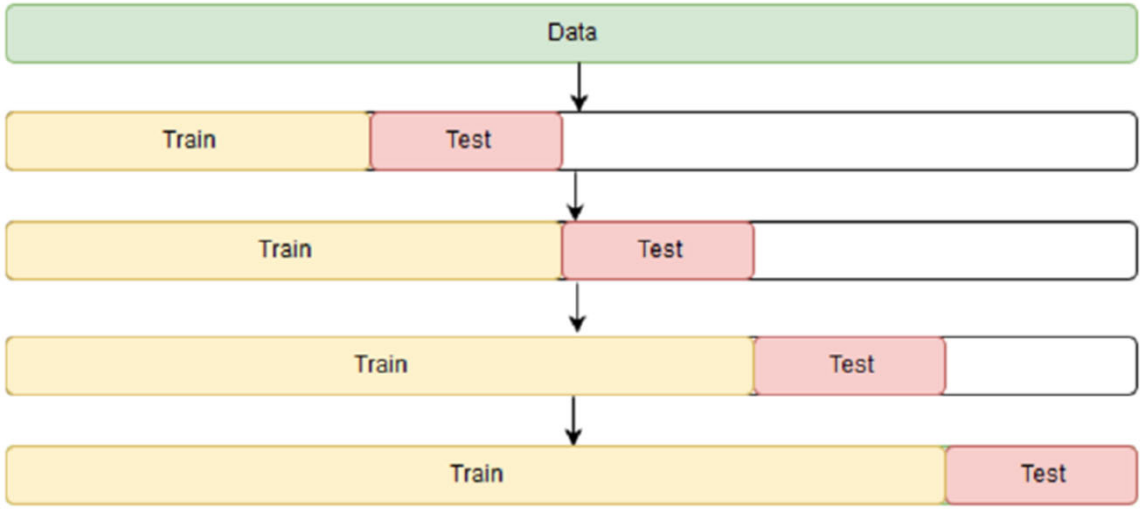
Data

- Data
  - Split
    - Training set
    - Held out set (sometimes call Validation set)
    - Test set
  - Randomly allocate to these three, e.g. 60/20/20



Data

- Data
  - Split
    - Training set
    - Held out set (sometimes call Validation set)
    - Test set
  - Randomly allocate to these three, e.g. 60/20/20



Features and Labels

- Features
  - Data comes in different forms
    - Text
    - Recordings
    - Images
  - Translate into features amenable to model
- Labels
  - For supervised learning, we have labels
    - Spam or Not Spam

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Features and Labels

	id	subject	email	spam
0	0	Subject: A&L Daily to be auctioned in bankrupt...	url: http://boingboing.net/#85534171\n date: n...	0
1	1	Subject: Wired: "Stronger ties between ISPs an...	url: http://scriptingnews.userland.com/backiss...	0
2	2	Subject: It's just too small ...	<html>\n <head>\n </head>\n <body>\n <font siz...	1
3	3	Subject: liberal defnitions\n	depends on how much over spending vs. how much...	0
4	4	Subject: RE: [ILUG] Newbie seeks advice - Suse...	hehe sorry but if you hit caps lock twice the ...	0

```
dear ricardo1 ,

<html>
<body>
<center>
<b><font color = "red" size = "+2.5">cost effective direct email advertising</font><br>
<font color = "blue" size = "+2">promote your business for as low as </font><br>
<font color = "red" size = "+2">$50</font> <font color = "blue" size = "+2">per
<font color = "red" size = "+2">1 million</font>
<font color = "blue" size = "+2"> email addresses</font></font><p>
<b><font color = "#44c300" size ="+2">maximize your marketing dollars!<p></font></b>
<font size = "+2">complete and fax this information form to 309-407-7378.<br>
a consultant will contact you to discuss your marketing needs.<br>
```

Experimentation

- Experimentation
  - Select a hypothesis f
    - Usually depends on numbers called parameters
    - Fit parameters to model on training set. Compute accuracy of test set.
  - Tune hyperparameters on validation set
    - Usually arise from regularization
    - For example early stopping
  - Data Snooping

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Experimentation

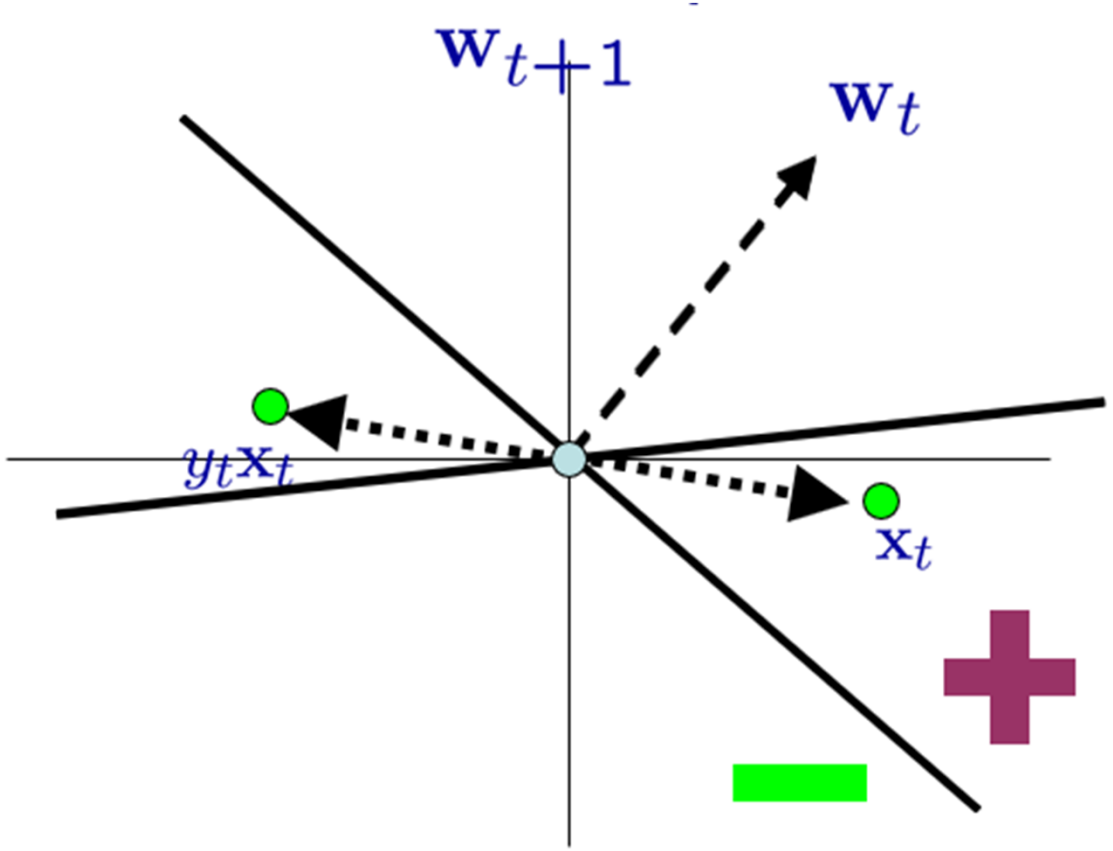
- Experimentation
  - Select a hypothesis f
    - Usually depends on numbers called parameters
    - Fit parameters to model on training set. Compute accuracy of test set.
  - Tune hyperparameters on validation set
    - Usually arise from regularization
    - For example early stopping
  - Data Snooping

x_1	x_2	x_3	x_4	y
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	?
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

Rule	Counterexample
$\Rightarrow y$	1
$x_1 \Rightarrow y$	3
$x_2 \Rightarrow y$	2
$x_3 \Rightarrow y$	1
$x_4 \Rightarrow y$	7
$x_1 \wedge x_2 \Rightarrow y$	3
$x_1 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \Rightarrow y$	3
$x_2 \wedge x_4 \Rightarrow y$	3
$x_3 \wedge x_4 \Rightarrow y$	4
$x_1 \wedge x_2 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3

Fitting Parameters

$\langle \mathbf{w}, \mathbf{x} \rangle$ \ y	$+1$	-1
$+1$	$+1$	-1
-1	-1	$+1$



Fitting Parameters

x

“free money”

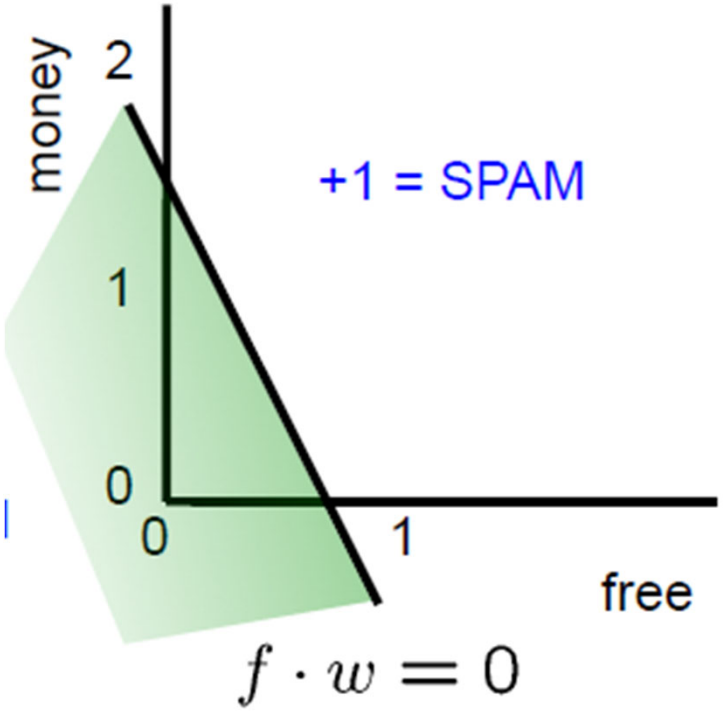
$f(x)$

BIAS	:	1
free	:	1
money	:	1
...		

w

BIAS	:	-3
free	:	4
money	:	2
...		

$$(1)(-3) + (1)(4) + (1)(2) + \ldots = 3$$



Fitting Parameters

- Step 1

$$\text{sign}\left(\langle \mathbf{w}, \mathbf{f}(x)\rangle - \text{threshold}\right) = \begin{cases} 1 & \text{then spam} \\ -1 & \text{then not spam} \end{cases}$$

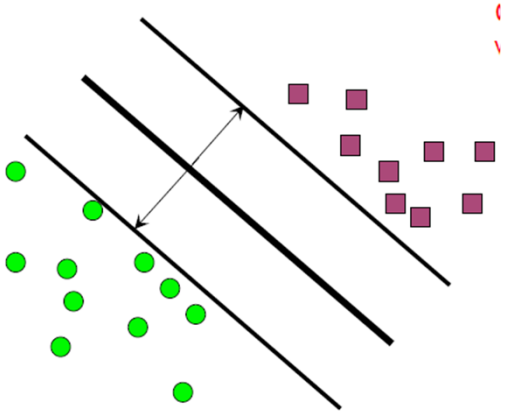
- Step 2 (Combine)

$$f_{N+1}(x) \equiv 1$$

$$w_{N+1} = -\text{threshold}$$

- Step 3 (Output)

$$\text{sign}\left(\langle \mathbf{w}, \mathbf{f}(x)\rangle\right) = \begin{cases} 1 & \text{then spam} \\ -1 & \text{then not spam} \end{cases}$$



Fitting Parameters

Perceptron Algorithm

input: A training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$

initialize: $\mathbf{w}^{(1)} = (0, \ldots, 0)$

for $t = 1, 2, \ldots$

if $(\exists\, i \text{ s.t. } y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0)$ **then**

$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$

else

output $\mathbf{w}^{(t)}$

Fitting Parameters

Perceptron Algorithm

input: A training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$

initialize: $\mathbf{w}^{(1)} = (0, \ldots, 0)$

for $t = 1, 2, \ldots$

if $(\exists\, i \text{ s.t. } y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0)$ **then**

$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$

else

output $\mathbf{w}^{(t)}$

$$y \langle w_t, x \rangle$$

$$y \langle w_{t+1}, x \rangle = y \langle w_t, x \rangle + ||x||^2$$

Exercise

Fix x (vector), w (vector) and b (number). Assume absolute value of w is 1. Determine v that minimizes

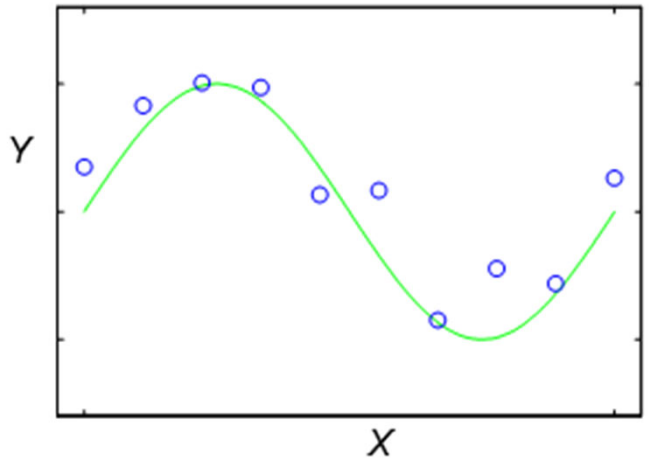
$$\min\{\|\mathbf{x}-\mathbf{v}\| : \langle \mathbf{w}, \mathbf{v}\rangle + b = 0\}$$

Hint: Consider

$$\mathbf{v} = \mathbf{x} - (\langle \mathbf{w}, \mathbf{x}\rangle + b)\mathbf{w}$$

Tuning Hyperparameters

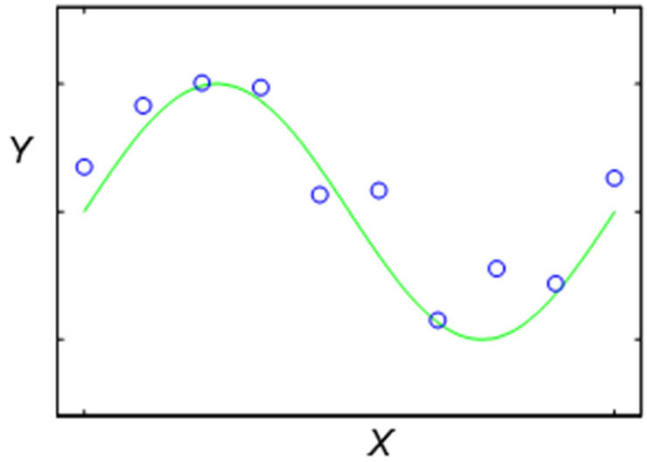
Dataset: 10 (X,Y) points generated from a sin function, with noise



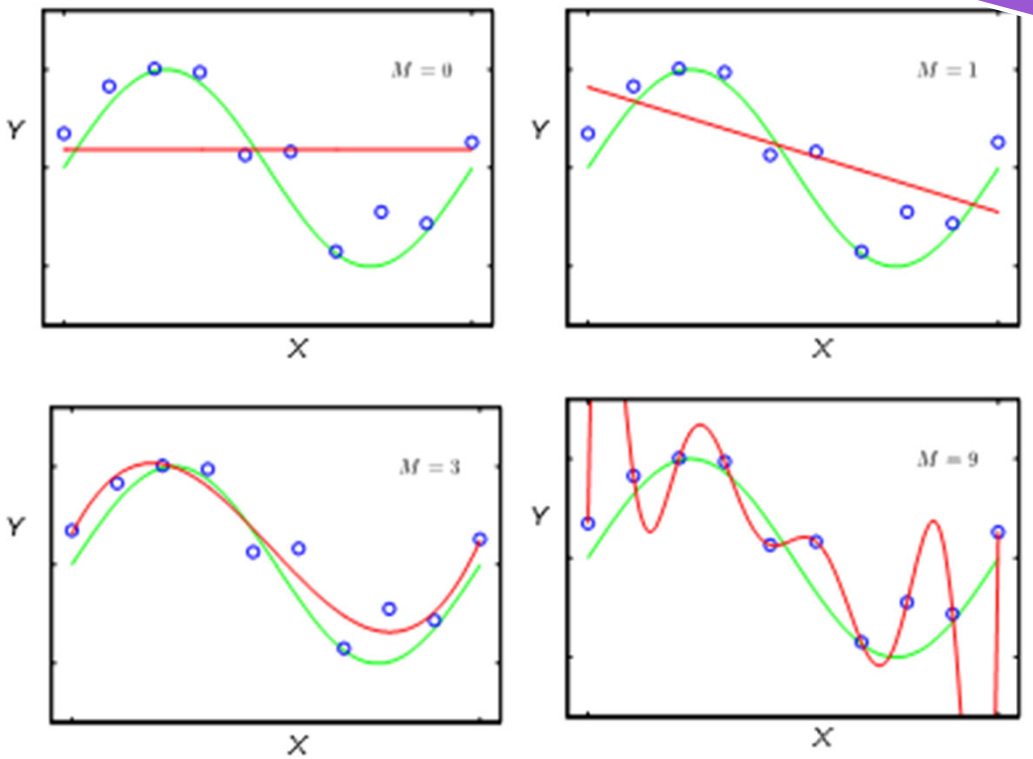
Think about the nonlinear transformation that allowed us to model an inverse relationship between mpg and horsepower using linear regression

Tuning Hyperparameters

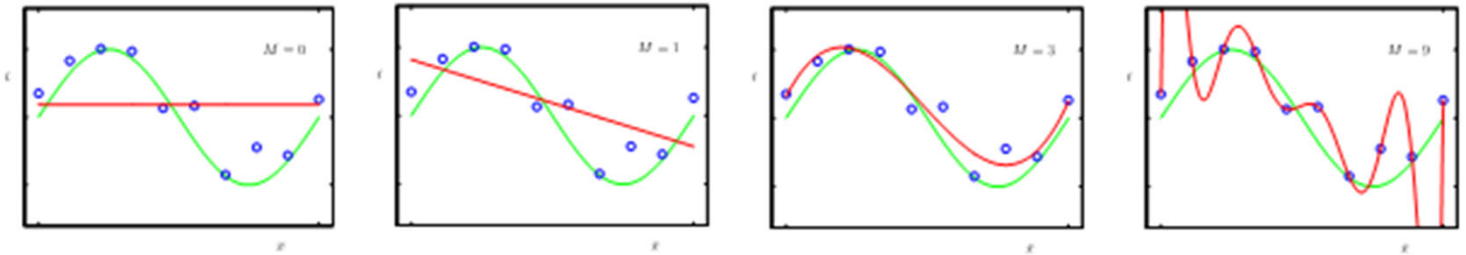
Dataset: 10 (X,Y) points generated from a sin function, with noise



Which to choose?



Fitting Parameters



We measure error using a *loss function* $L(y, \hat{y})$

For regression, a common choice is squared loss:

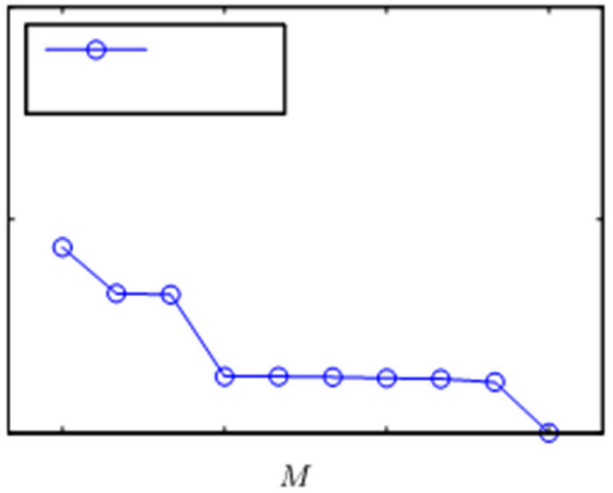
$$L(y_i, f(x_i)) = (y_i - f(x_i))^2$$

Squared error

The *empirical loss* of the function *f* applied to the training data is then:

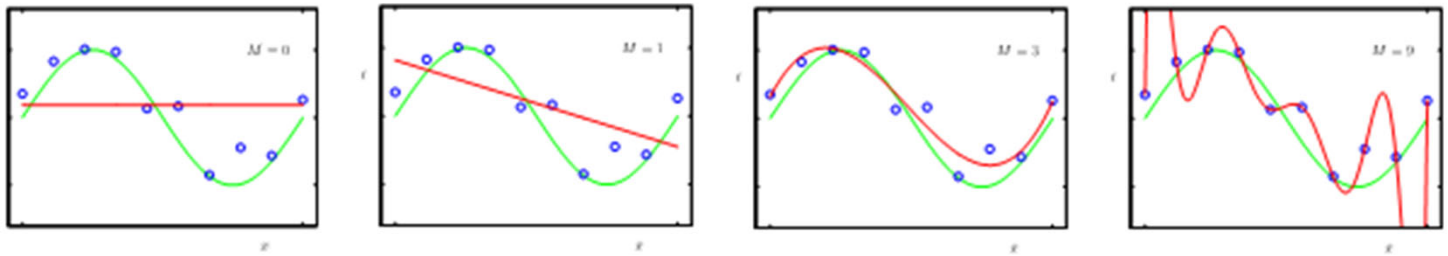
$$\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) = \frac{1}{N}\sum_{i=1}^{N} (y_i - f(x_i))^2$$

Learning curve



Measure of model complexity

Fitting Parameters



We measure error using a *loss function* $L(y, \hat{y})$

For regression, a common choice is squared loss:

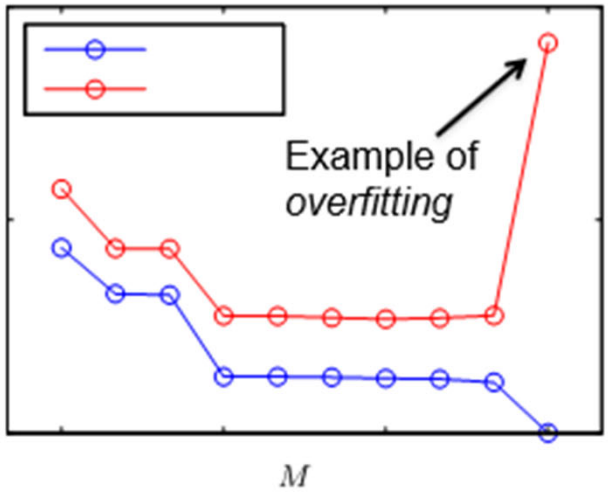
$$L(y_i, f(x_i)) = (y_i - f(x_i))^2$$

Squared error

The *empirical loss* of the function *f* applied to the training data is then:

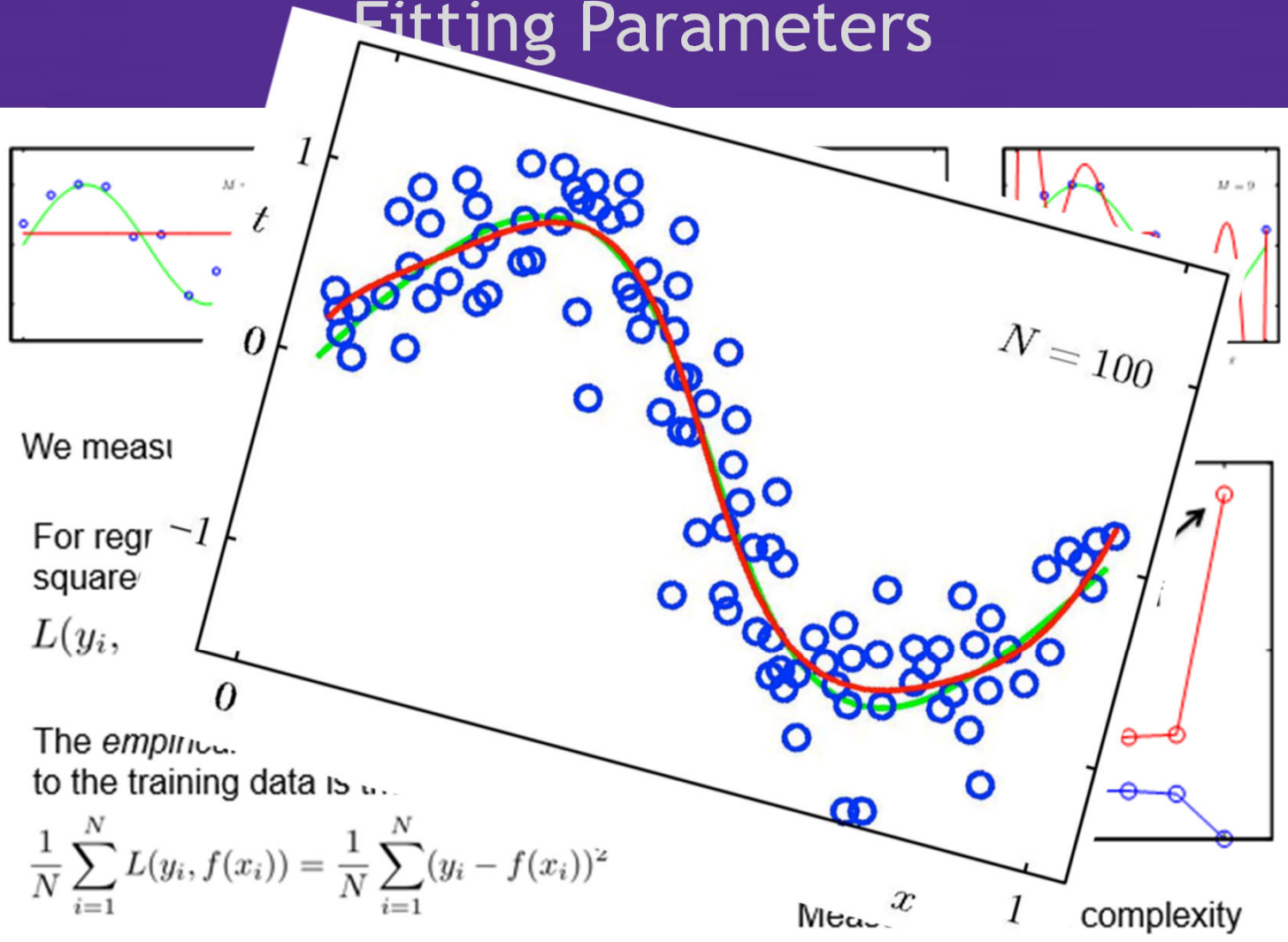
$$\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

Learning curve

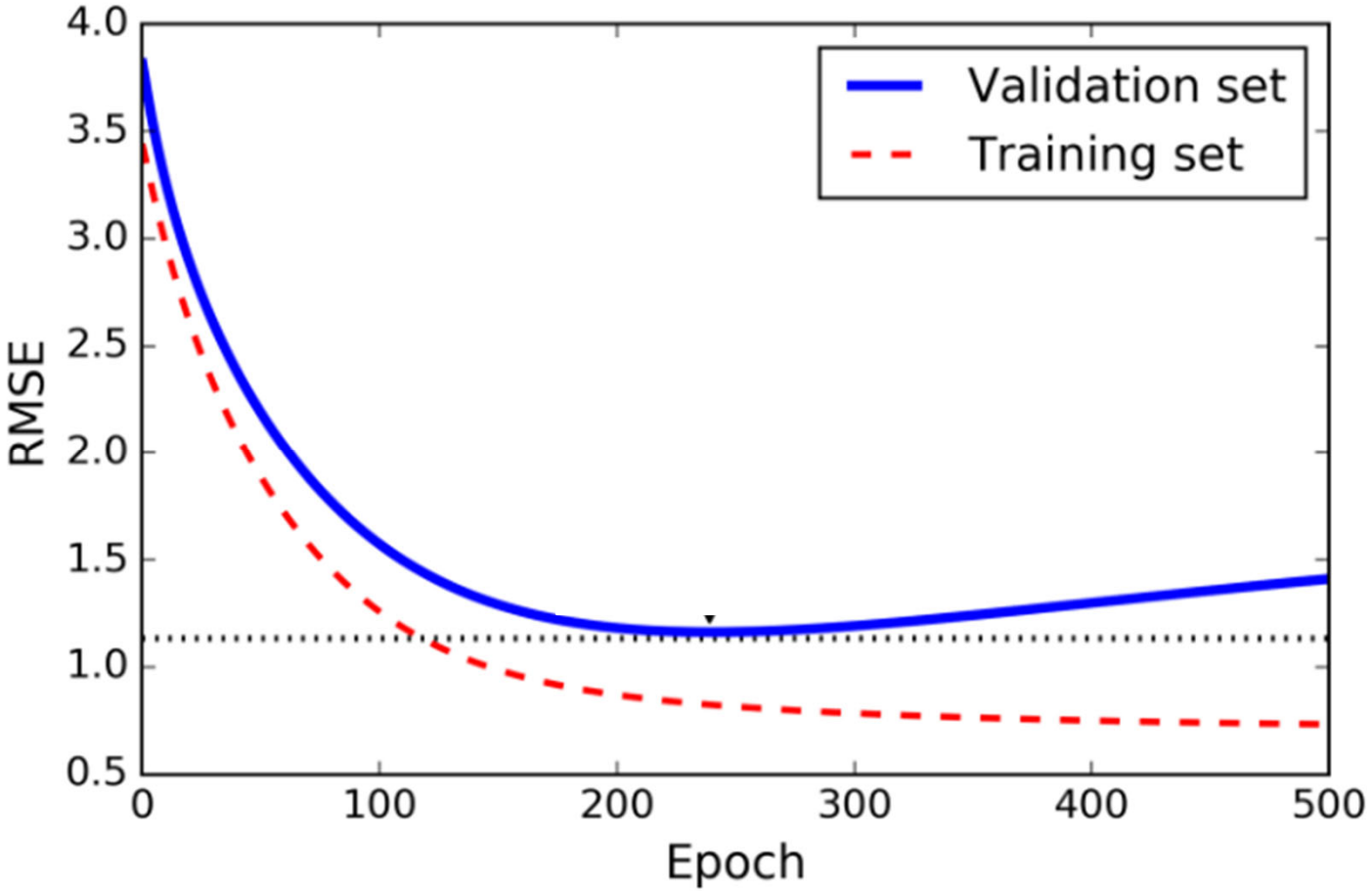


Measure of model complexity

Fitting Parameters

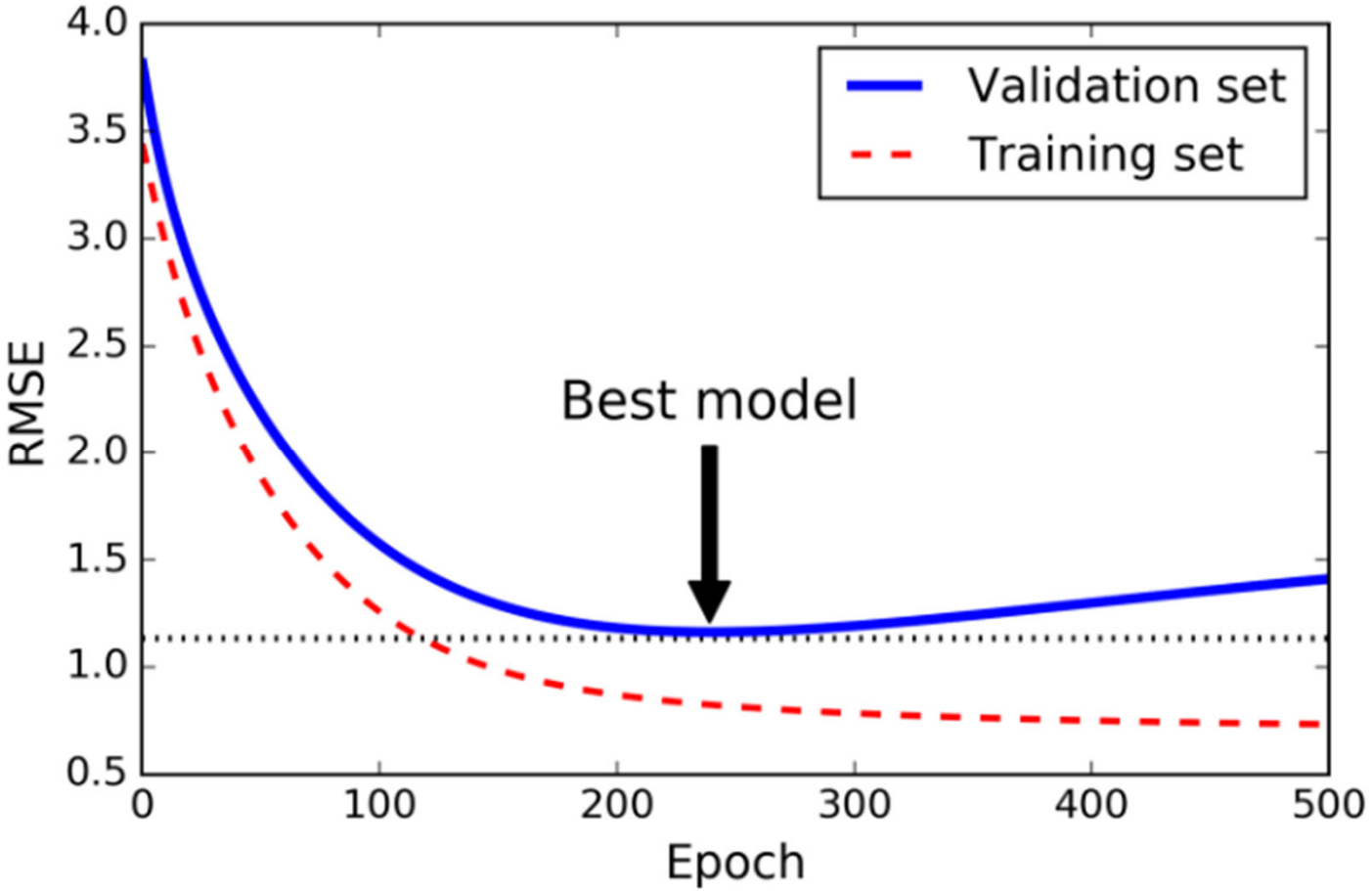


Regularization



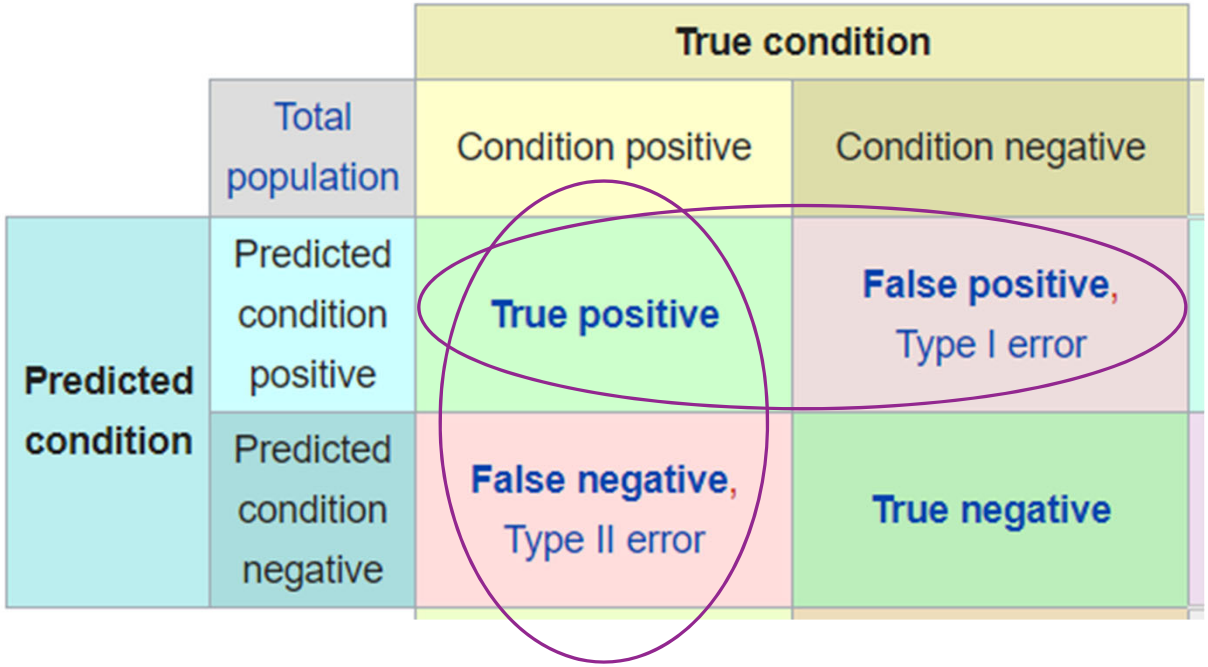
Parsimony

Regularization



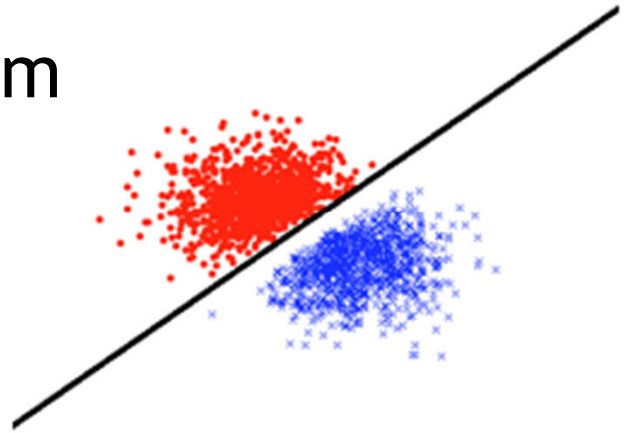
Evaluation

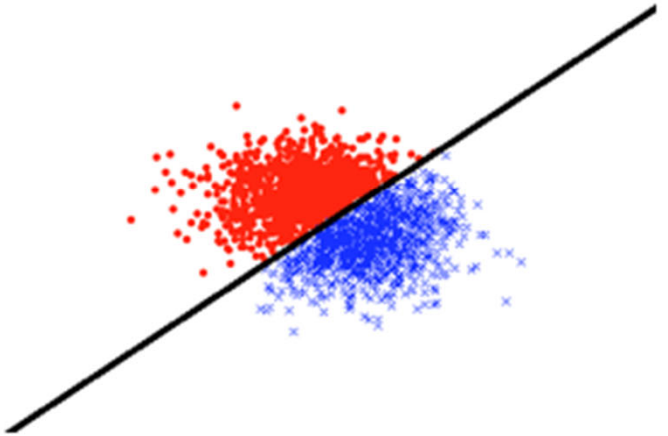
- Evaluation
  - Metrics measure the accuracy according to different criteria
  - Often accuracy refers to fraction of instances predicted correctly
    - 0-1 loss function



Perceptron

- Advantages
  - Error Bound
  - Online Algorithm
- Disadvantages
  - Many Decision Boundaries
    - Overfitting
  - Separable Data





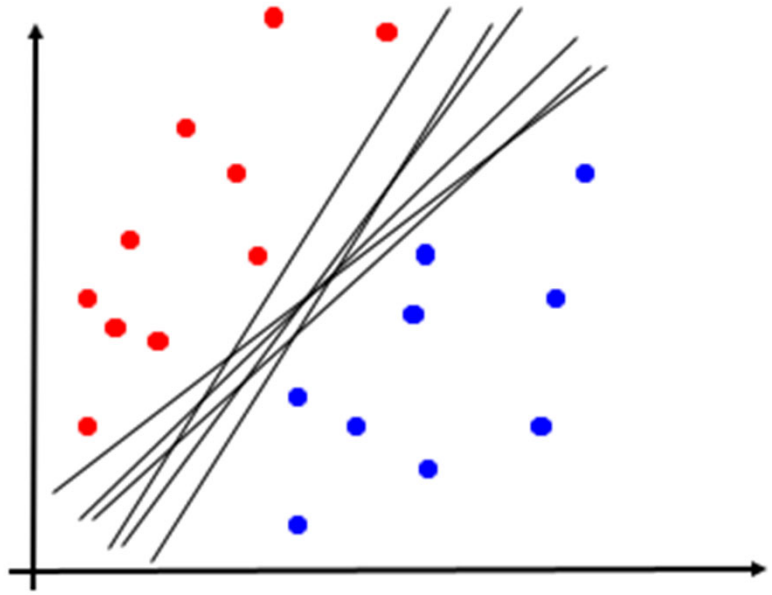
Perceptron

- Advantages
  - Error Bound
  - Online Algorithm
- Disadvantages
  - Many Decision Boundaries
    - Overfitting
  - Separable Data

1 $\mathbf{w}_1 \leftarrow \mathbf{w}_0$ ▷ typically $\mathbf{w}_0 = \mathbf{0}$
2 **for** $t \leftarrow 1$ **to** T **do**
3 RECEIVE($\mathbf{x}_t$)
4 $\widehat{y}_t \leftarrow \text{sgn}(\mathbf{w}_t \cdot \mathbf{x}_t)$
5 RECEIVE(y_t)
6 **if** $(\widehat{y}_t \neq y_t)$ **then**
7 $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_t \mathbf{x}_t$ ▷ more generally $\eta y_t \mathbf{x}_t, \eta > 0$.
8 **else** $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t$
9 **return** $\mathbf{w}_{T+1}$

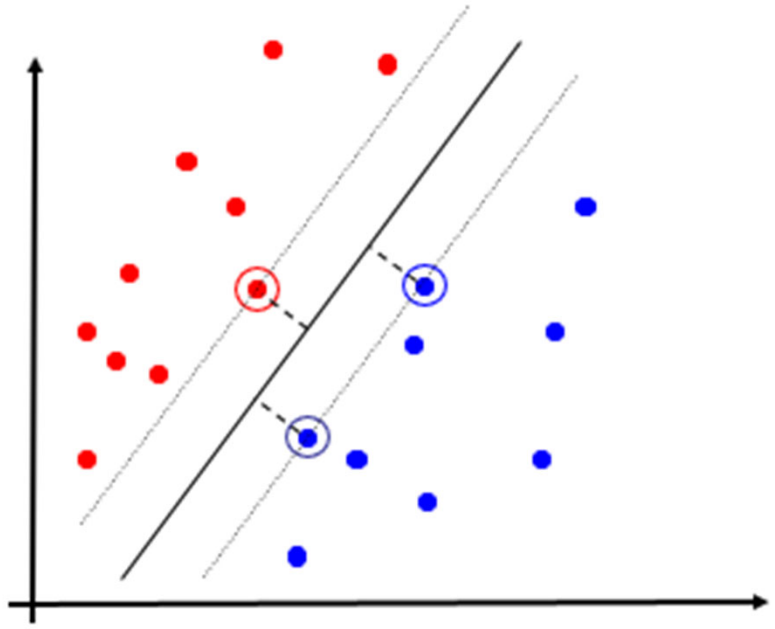
Perceptron

- Advantages
  - Error Bound
  - Online Algorithm
- Disadvantages
  - Many Decision Boundaries
    - Overfitting
  - Separable Data



Perceptron

- Advantages
  - Error Bound
  - Online Algorithm
- Disadvantages
  - Many Decision Boundaries
    - Overfitting
  - Separable Data



Perceptron

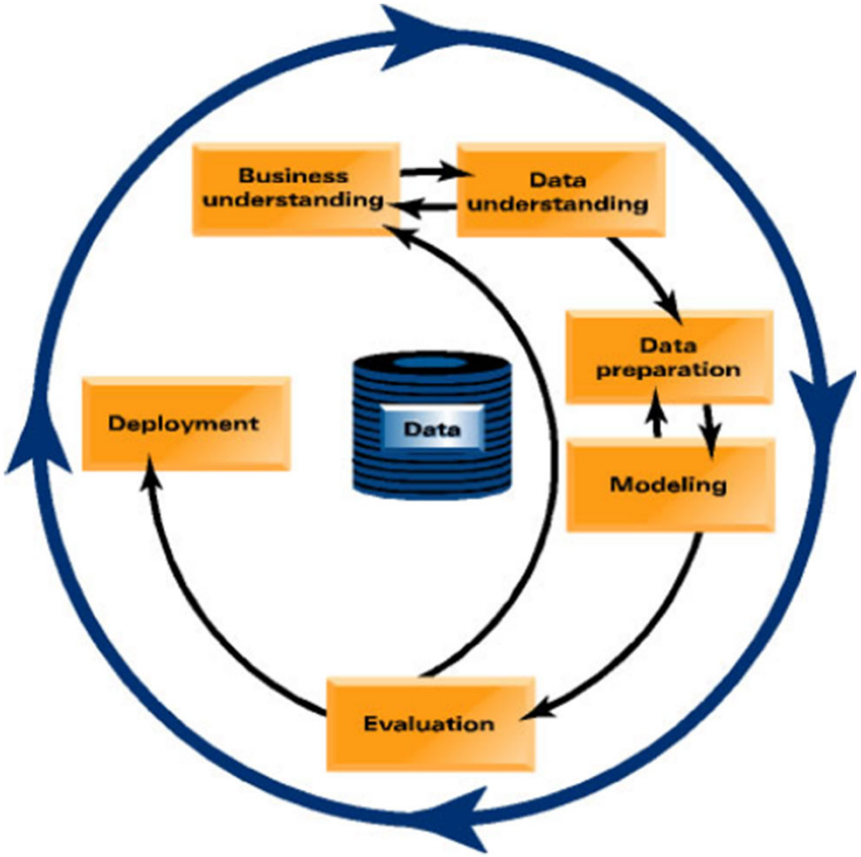
- Advantages
  - Error Bound
  - Online Algorithm
- Disadvantages
  - Many Decision Boundaries
    - Overfitting
  - Separable Data

Perceptron

- Advantages
  - Error Bound
  - Online Algorithm
- Disadvantages
  - Many Decision Boundaries
    - Overfitting
  - Separable Data

Summary

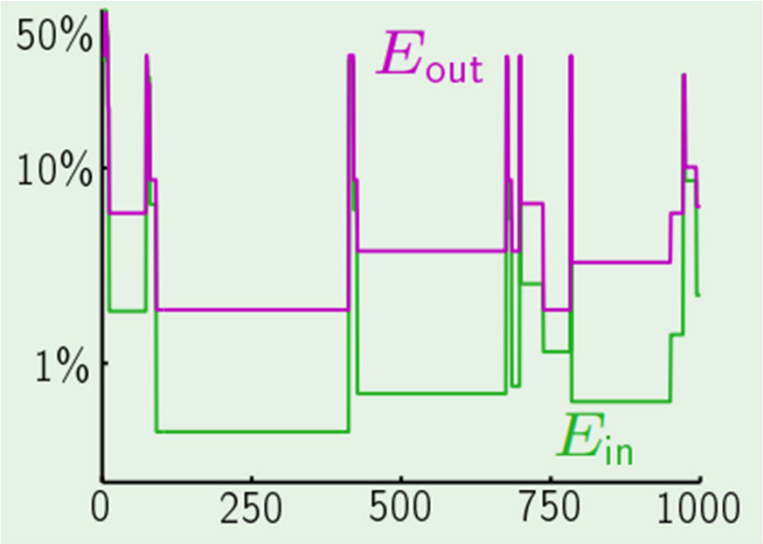
- Steps for Machine Learning
  - Hypothesis Space
  - Loss Functions
  - Optimization
  - Regularization
- Putting Steps into Practice
  - Data
  - Features and Labels
  - Experimentation
  - Evaluation



Questions

Pocket Algorithm

- Questions on Piazza?
- Question for You!
  - Can you think of another way to use Perceptron for non-separable data

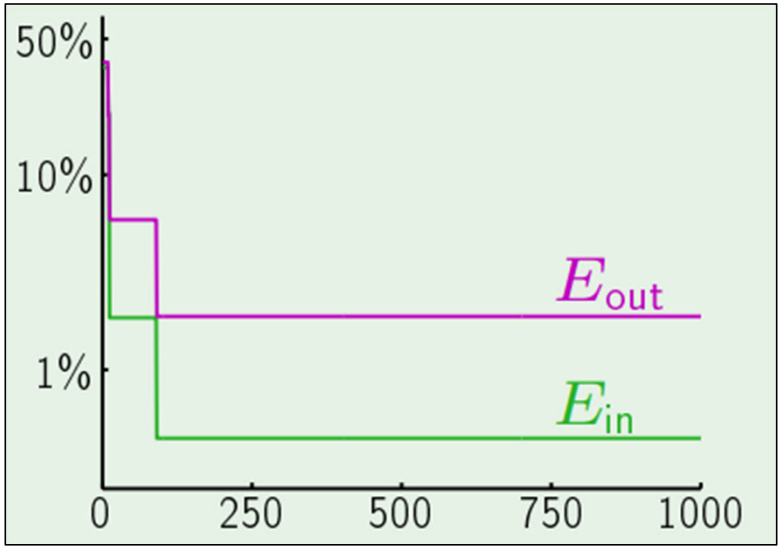


cc BY NC ND

Questions

Pocket Algorithm

- Questions on Piazza?
- Question for You!
  - Can you think of another way to use Perceptron for non-separable data



cc BY NC ND