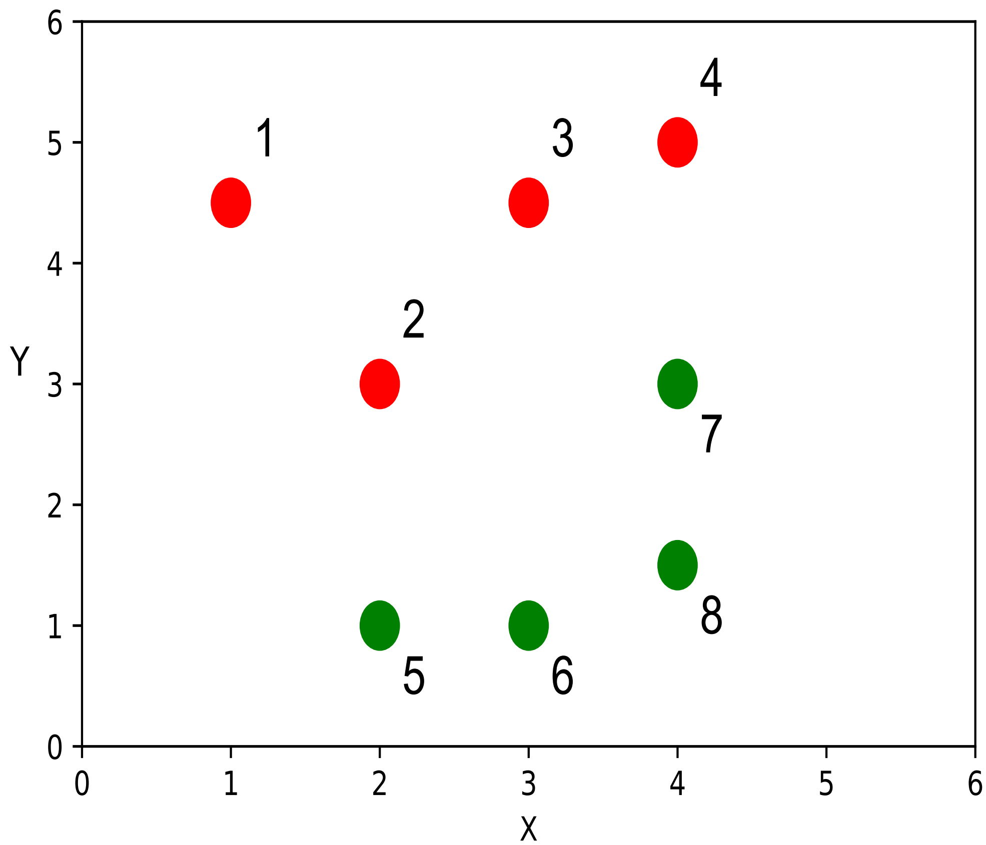
HISTORICAL

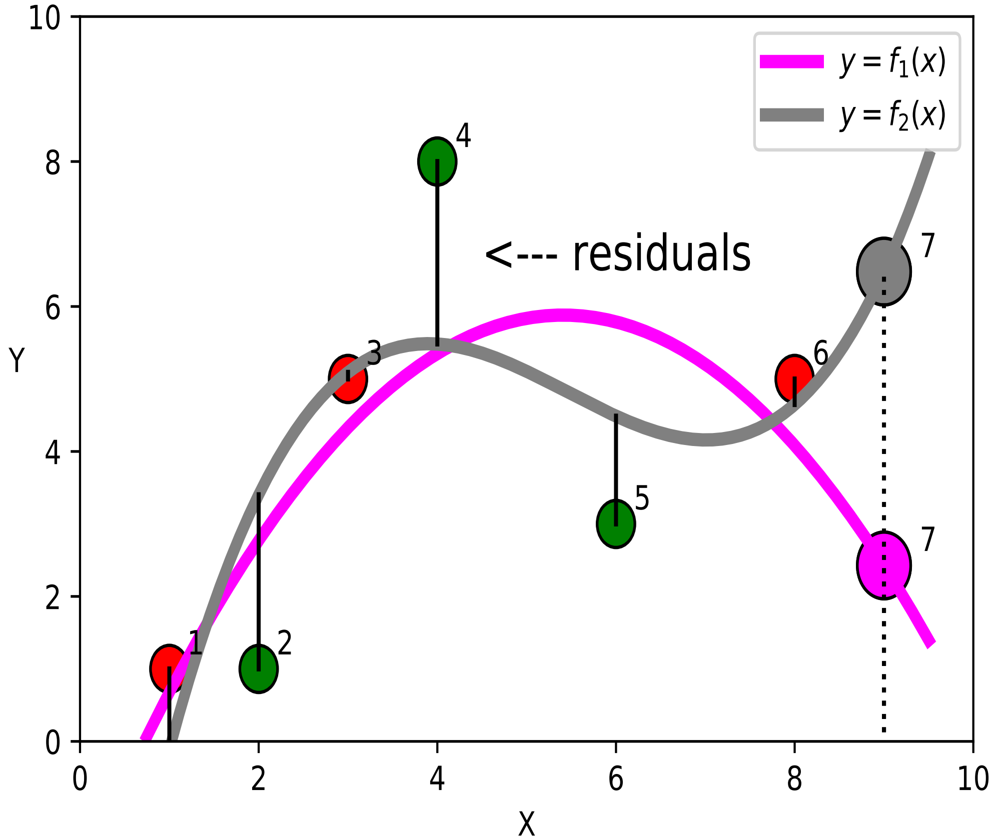
EXAMPLE

Problem Statement



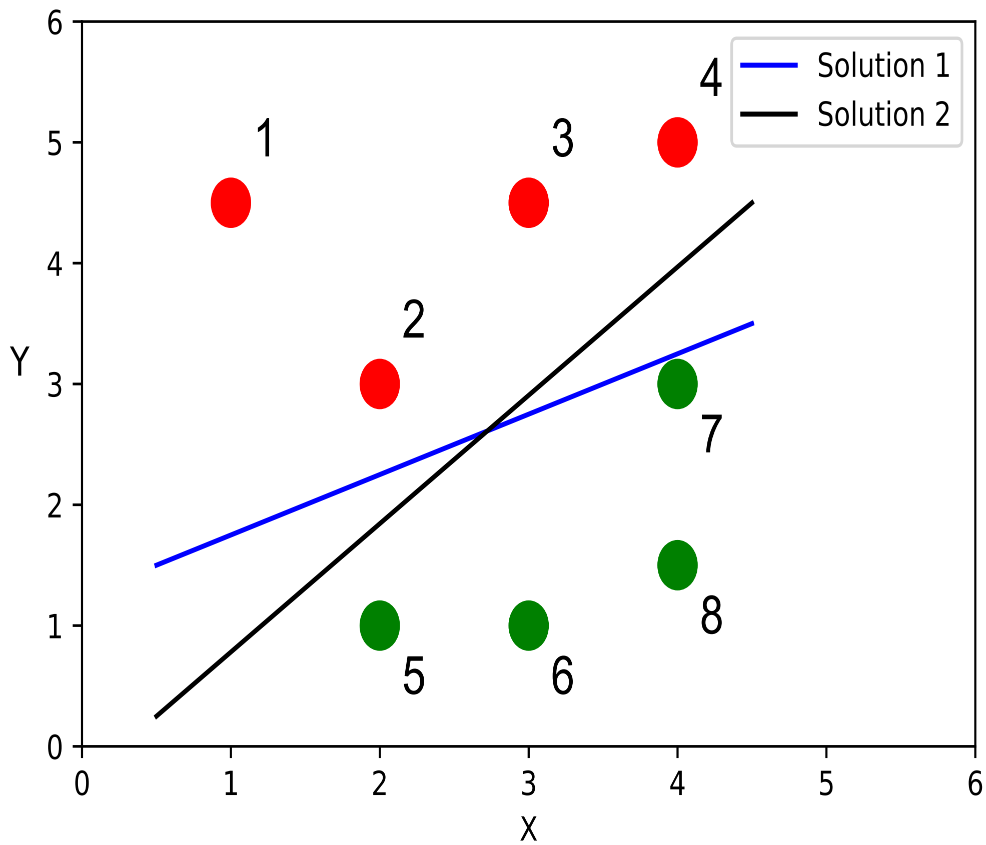
- choose a function $y = f(x)$ for classification or prediction

Prediction Problem



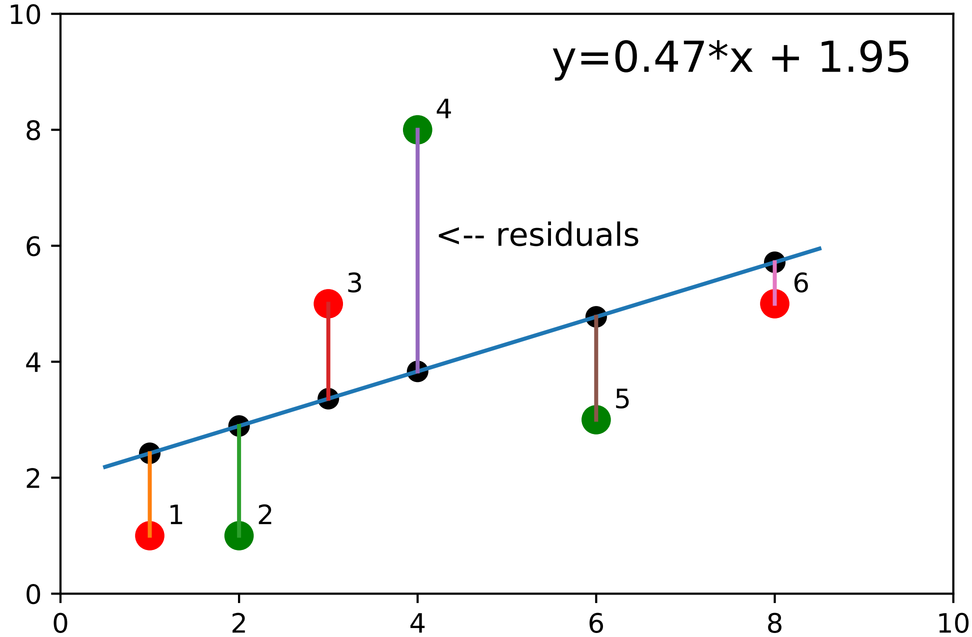
- find $f(x)$ to match data and give good prediction

Classification Problem



- find $f(x)$ to assign labels

Ex: Linear Regression



- assume $y = ax + b$
- choose line to minimize "loss'

$$Q = \sum_{i=1}^{n} e_i^2$$

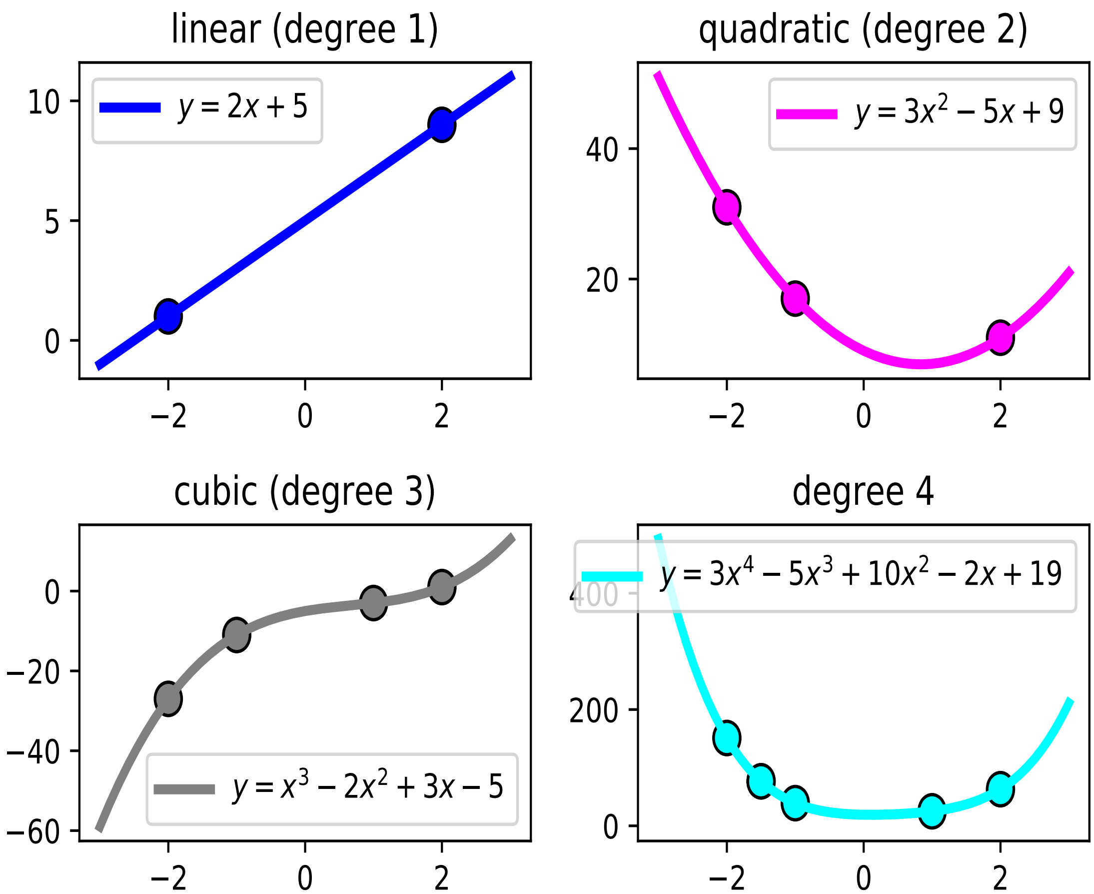
Generalized Linear Models

- ideal case: $y = ax + b$
- most models do not fit
- solution: use *link* functions $f(\cdot), g(\cdot)$
- our model:

$$f(y) = ag(x) + b$$

- typical link functions: polynomials, $\exp(\cdot)$, $\log(\cdot)$

Example: Polynomials



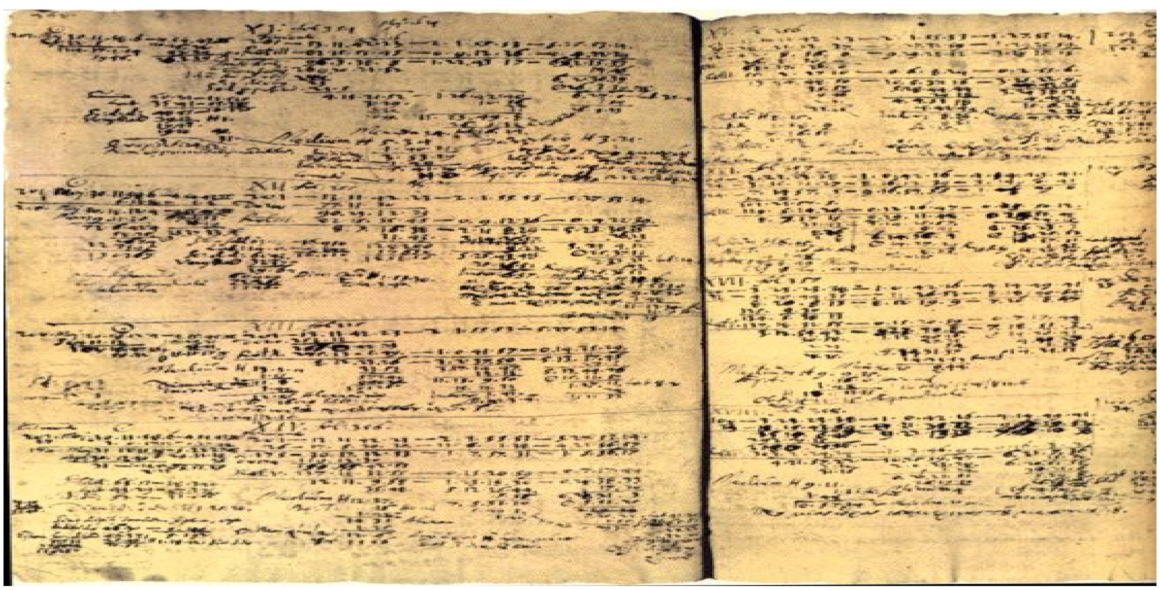
- underfitting/overfitting
- model complexity

General Approach

- want $f(y) = ag(x) + b$
- collect and clean data
- split data into training and testing
- choose link functions f and g
- use training set to compute parameters for f and g
- use testing to choose between models

Example: Kepler's Laws

- Johannes Kepler worked for Tycho Brahe
- Brahe compiled detailed observations (especially Mars)



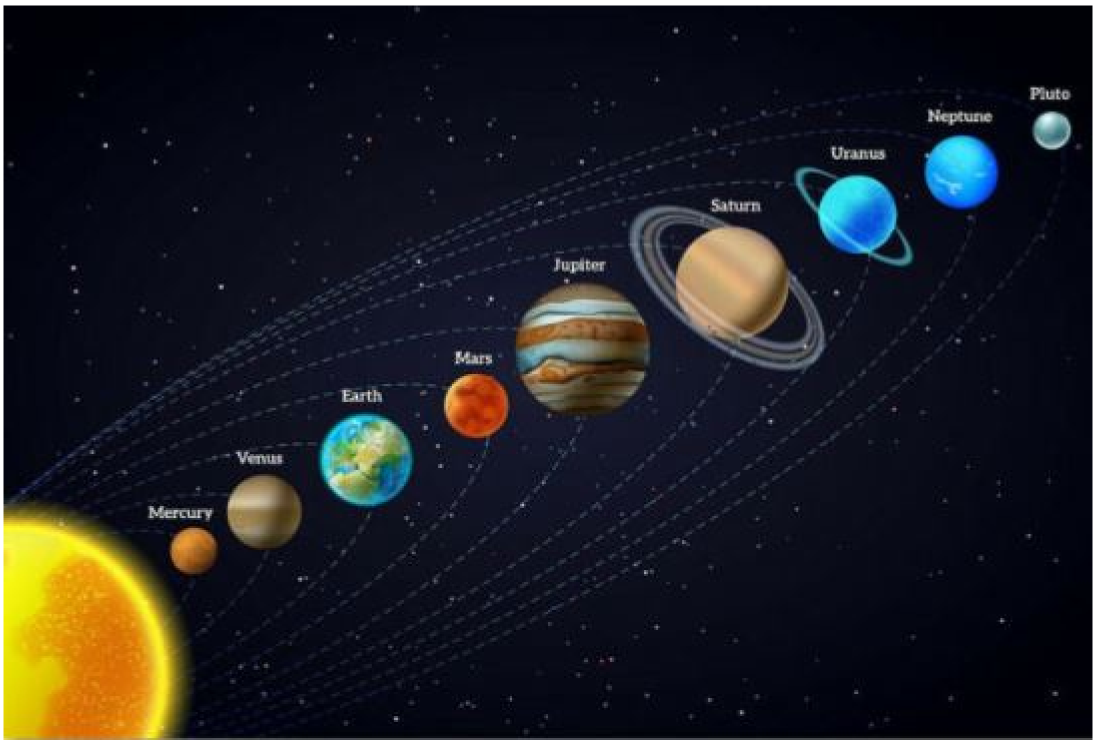
Distance, Periods Data

- Kepler "cleaned" data:

Planet	Period (days) T	Distance (AU) to Sun R
Mercury	87.77	0.389
Venus	224.70	0.724
Earth	365.25	1
Mars	686.95	1.524
Jupiter	4332.62	5.2
Saturn	10759.2	9.510

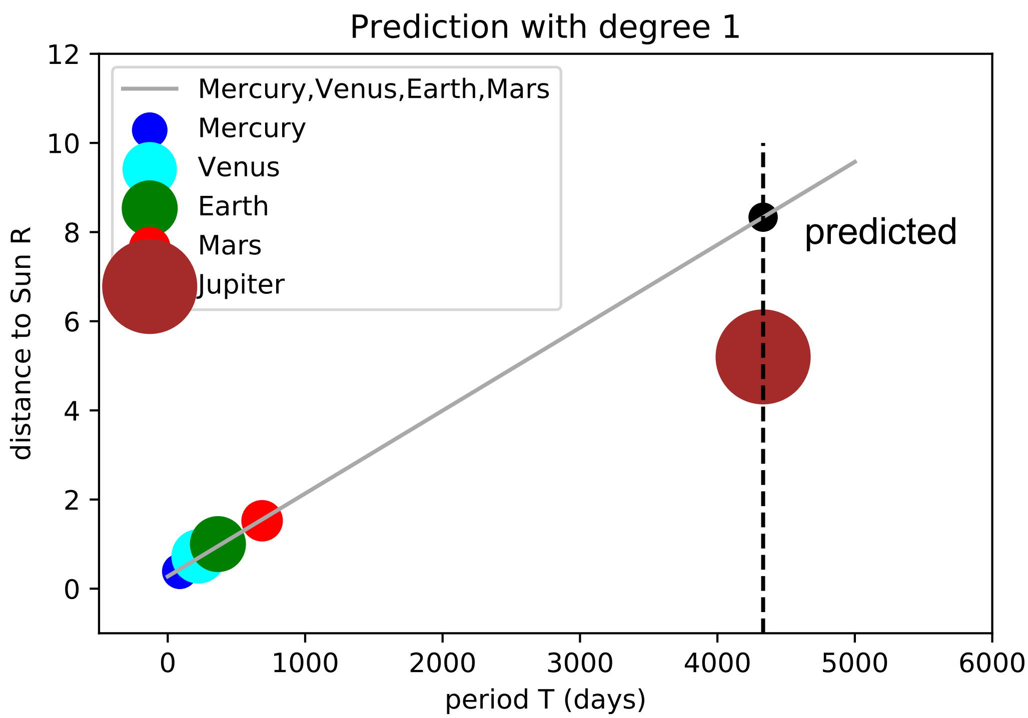
- what is R vs. T?
- Kepler discovered $R^3 = aT^2$

Periods and Orbits




```
"I first believed I was dreaming
But it is absolutely certain and exact
that the ratio which exists between
the period times of any two planets
is precisely the ratio of the 3/2th
power of the mean distance."
translated from Harmonies of the World by Kepler (1619)"
```

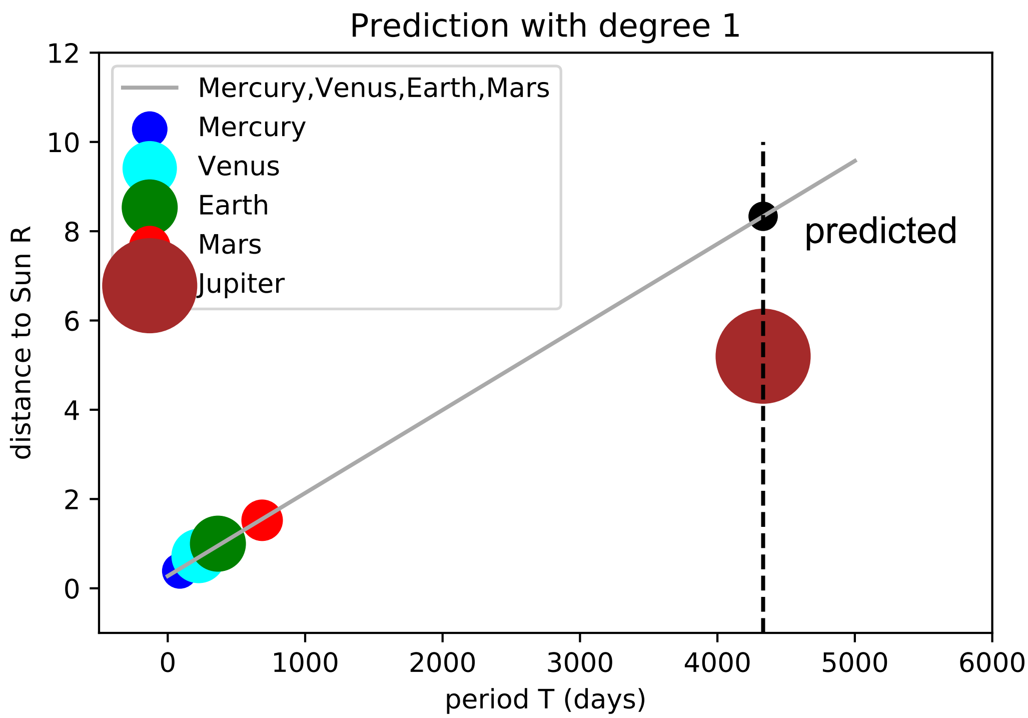
Kepler's Question



$$f(R) = ag(T) + b$$

- how are R and T related?
- what link functions $f(R)$ and $g(T)$ match the data?

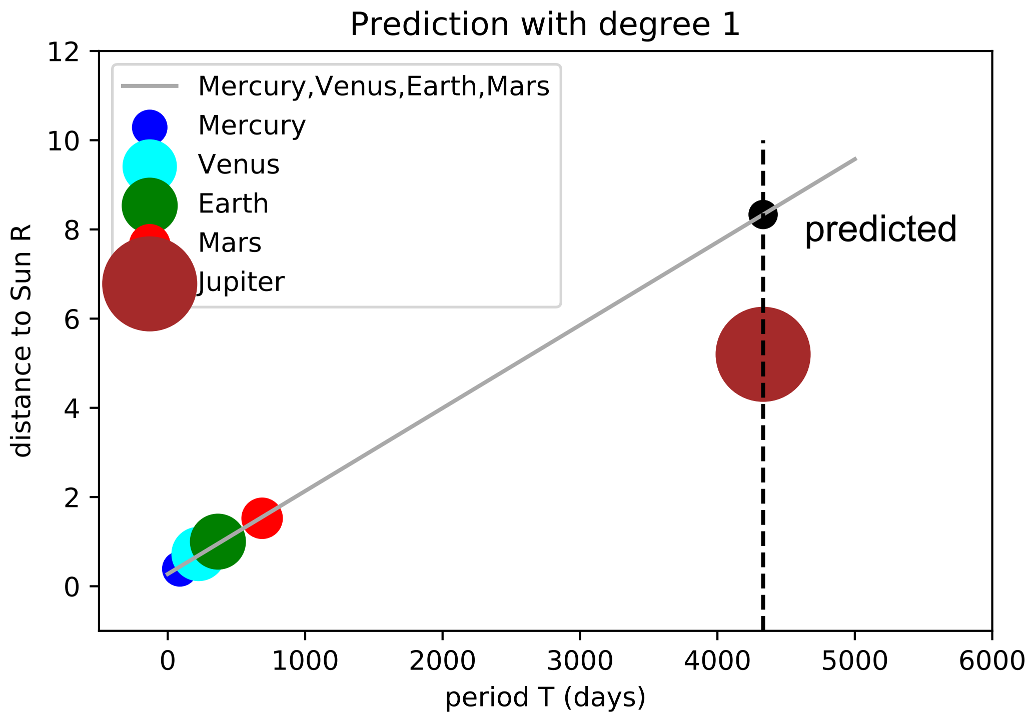
Linear Prediction



$$R = aT + b$$

- training set: Mercury, Venus, Earth, Mars
- testing set: Jupiter

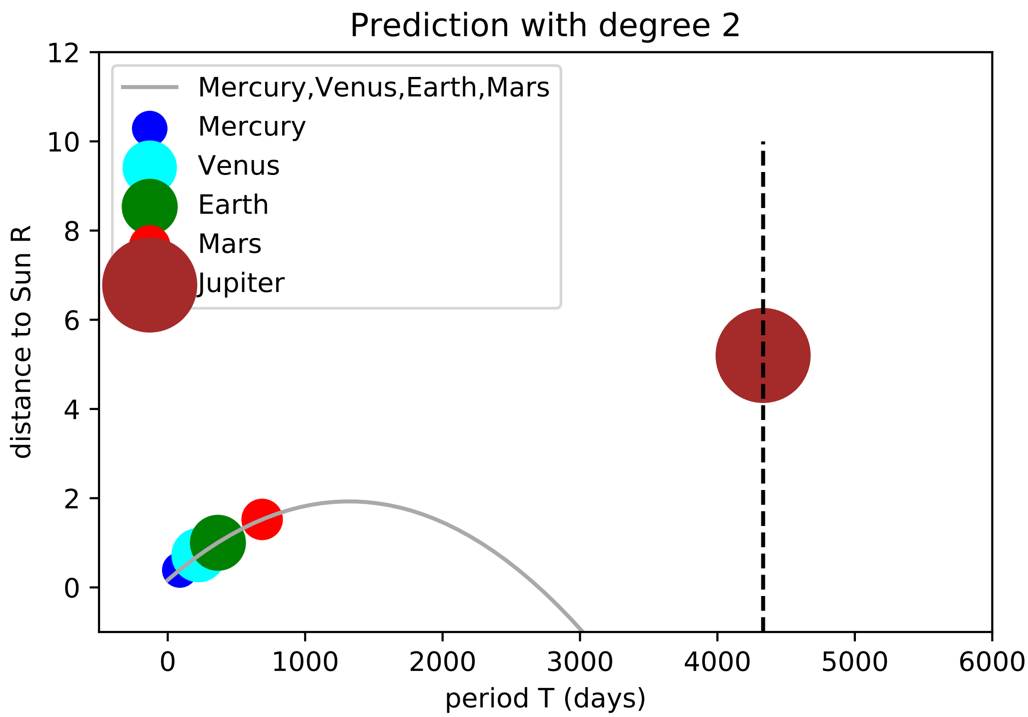
Linear Prediction



$$R = (1.86 \times 10^{-3}) \cdot T + 0.27$$

- predicted $R(\text{Jupiter}) = 8.33$
- "exact": 5.2 (rel. error 60%)

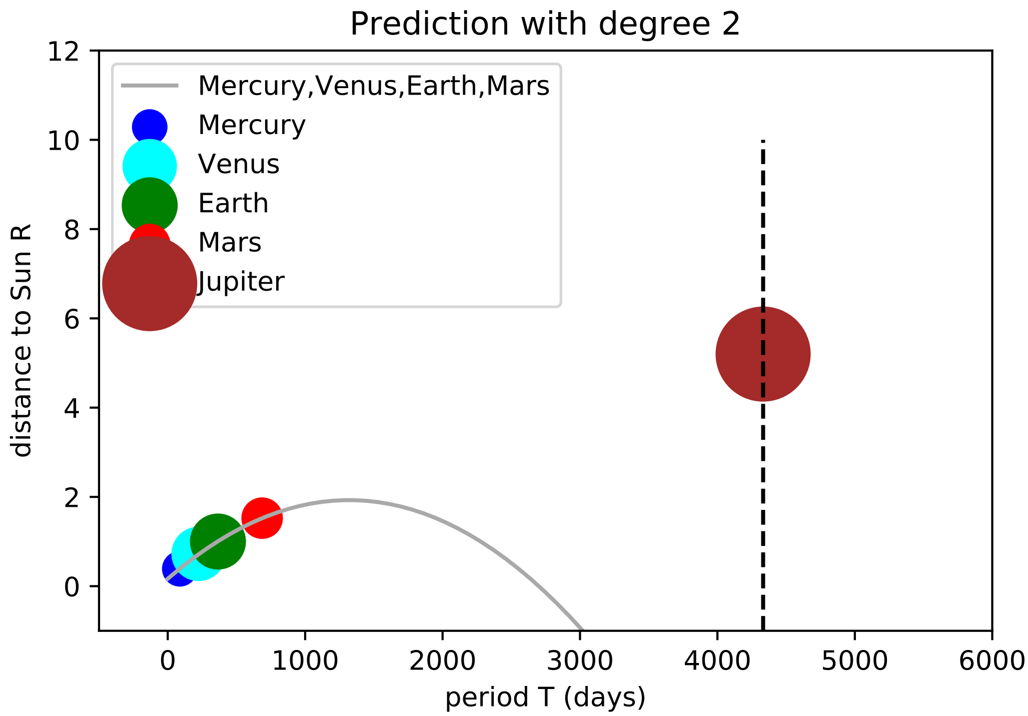
Quadratic Prediction



$$R = aT^2 + bT + c$$

- polynomial link for $g(T)$
- training set: Mercury, Venus, Earth, Mars

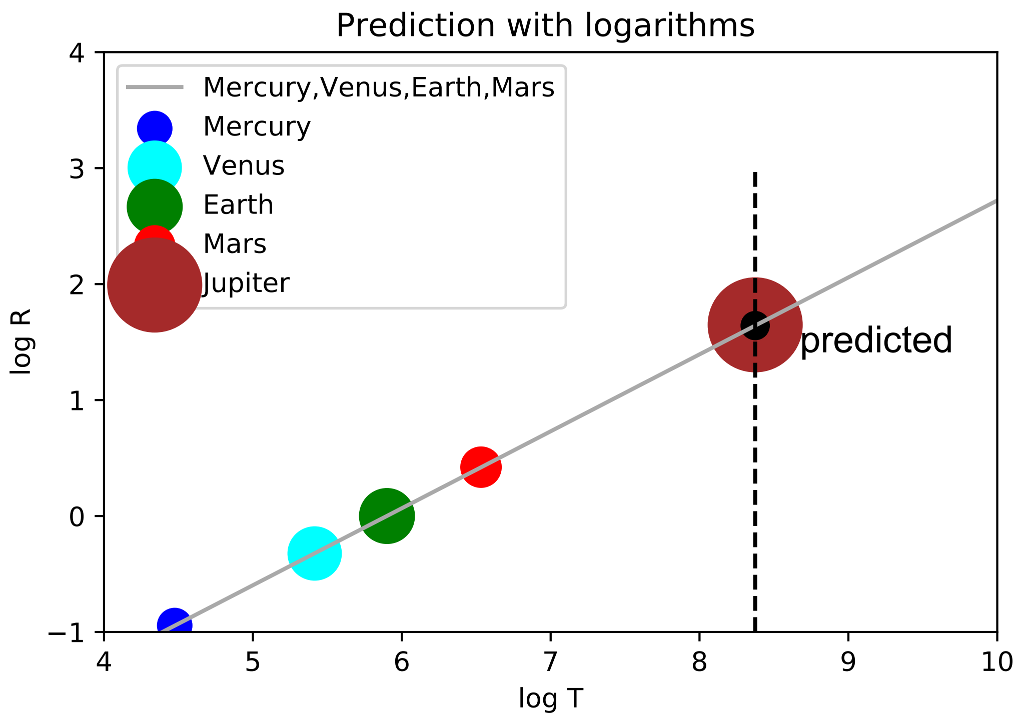
Quadratic Prediction



$$R = -(1.01 \times 10^{-6}) \cdot T^2 + (2.67 \times 10^{-3}) \cdot T + 0.17$$

- predicted $R(\text{Jupiter}) < 0$!!!

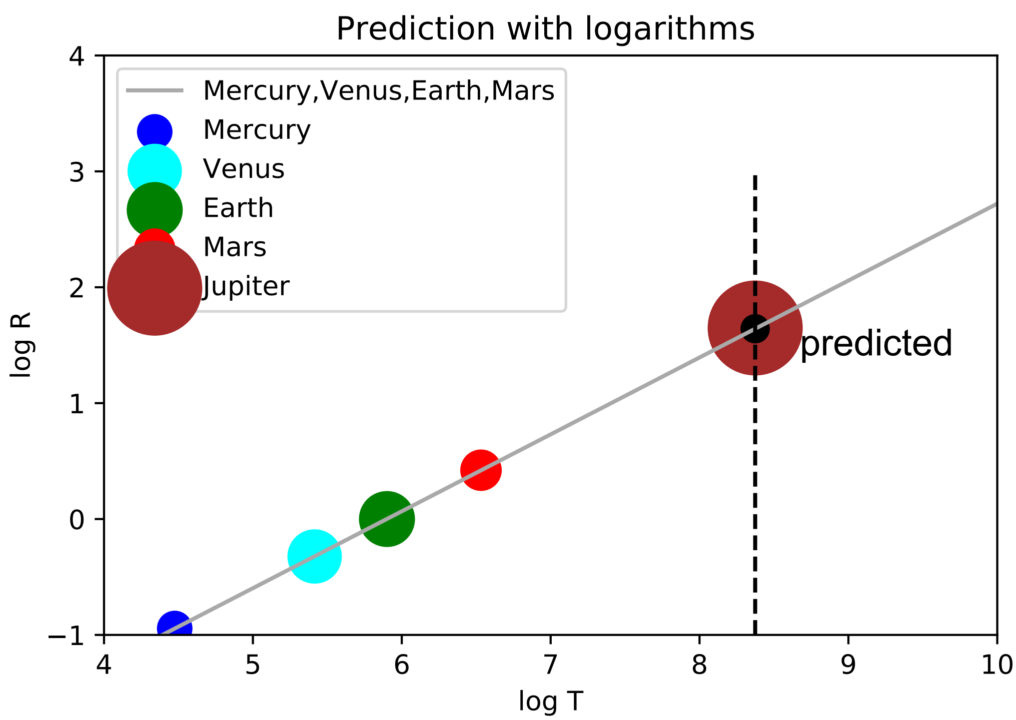
log-log Prediction



$$\log R = a \log T + b$$

- logarithmic link for f and g
- log() function was invented by John Napier in 1614

log-log Prediction



$$\log R = 0.66 \cdot \log T - 3.91$$

$$3 \cdot \log R = 2 \cdot \log T - 11.73$$

$$\Rightarrow \; R^3 \propto T^2$$

Kepler Result

Planet	Period (days) T	Distance (AU) to Sun R	R^3/T^2 $10^{-6}\mathrm{AU}^3/\mathrm{day}^2$
Mercury	87.77	0.389	7.64
Venus	224.70	0.724	7.52
Earth	365.25	1	7.50
Mars	686.95	1.524	7.50
Jupiter	4332.62	5.2	7.49
Saturn	10759.2	9.510	7.43

- Kepler's third law $T^2 \propto R^3$

square of orbital period of a planet is proportional to the cube of the semi-major axis of its orbit

Modern Data

Planet	Period (days) T	Distance to Sun R	R^3/T^2 $10^{-6}\text{AU}^3/\text{day}^2$
Mercury	87.9693	0.38710	7.496
Venus	224.7008	0.72333	7.496
Earth	365.2564	1	7.496
Mars	686.9796	1.52366	7.495
Jupiter	4332.8201	5.20336	7.504
Saturn	10775.599	9.53707	7.498
Uranus	30687.153	19.1913	7.506
Neptune	60190.03	30.0690	7.504

- ”close” match to Kepler’s
- Earth: 7.5 vs. 7.496

Modern Derivation

- Newton's gravitational law
- for circular orbits centripetal force equals gravitational force

$$mR\left(\frac{2\pi}{T}\right)^2 = G\frac{mM}{R^2}$$

$$\Rightarrow\ T^2 = \left(\frac{4\pi^2}{GM}\right)R^3$$

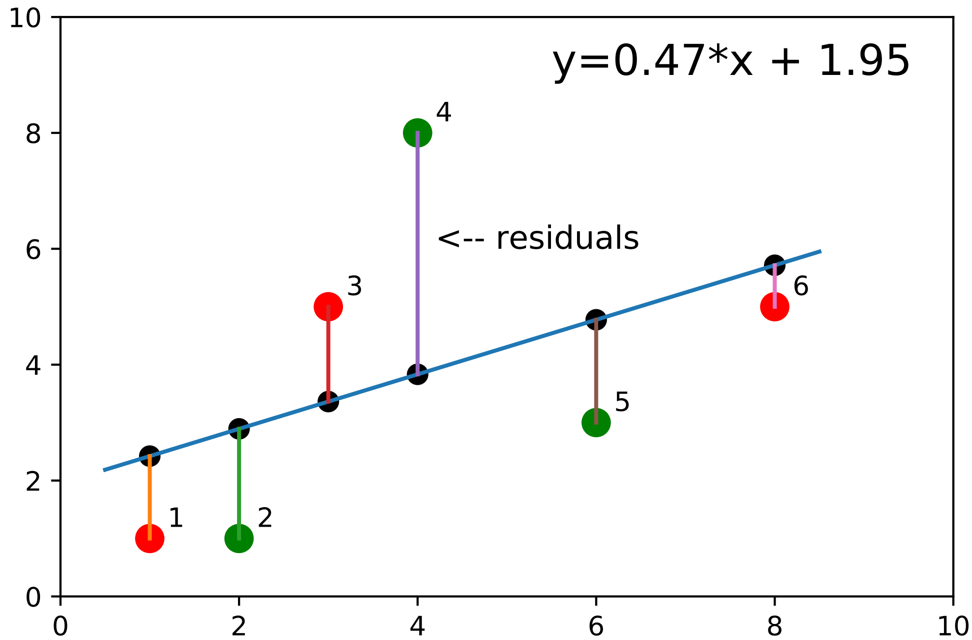
$$\Rightarrow\ T^2 \propto R^3$$

General Approach Revisited

- want: $f(y) = ag(x) + b$
- choose f and g
- use training to compute f, g
- use testing to choose model
- many methods in data science are in this category:

1. linear regression
2. logistic regression
3. support vector machines

Ex: Linear Regression



- $y = ax + b$
- no link functions for x or y
- simple interpretation for a, b

Ex: Logistic Regression

- assume probability P
- define *odds* as

$$\text{odds}(P) = \frac{P}{1 - P}$$

- $P = 0.25 \mapsto \text{odds}(P) = 1/3$
- $P = 0.50 \mapsto \text{odds}(P) = 1$
- $P = 0.75 \mapsto \text{odds}(P) = 3/1$
- want to model

$$f(\text{odds}(P)) = ag(x) + b$$

Main Idea

- use $f = \log()$ as *link* for the *odds*
- use regression

$$\log\left(\frac{P}{1-P}\right) = b_0 + b_1 x$$

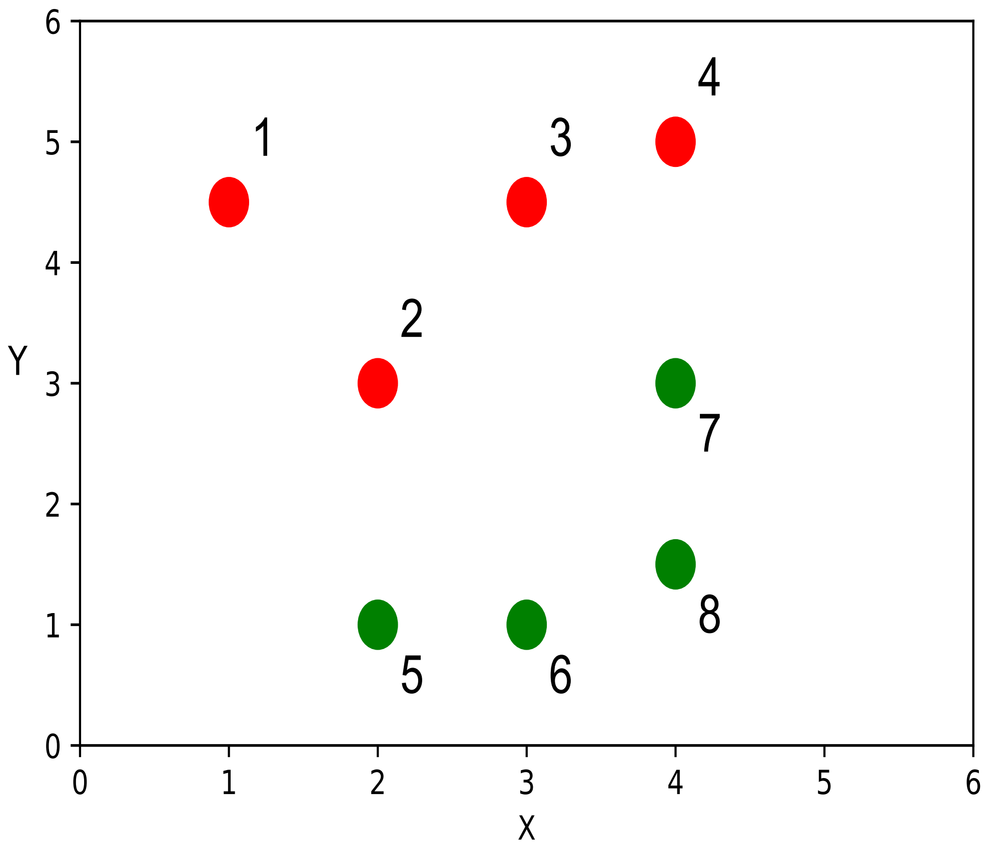
$$\frac{P}{1-P} = \exp(b_0 + b_1 x)$$

$$P = \frac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)}$$

- note:

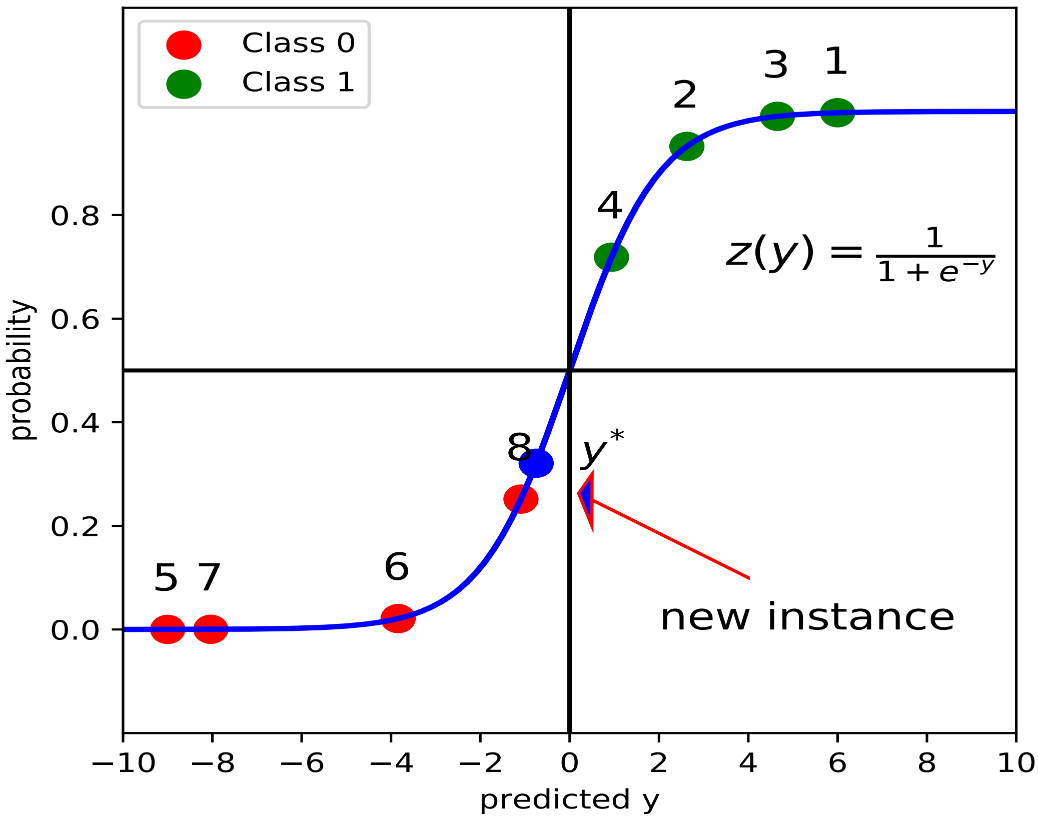
$$\frac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)} = \frac{1}{1 + \exp(-(b_0 + b_1 x))}$$

Original Dataset



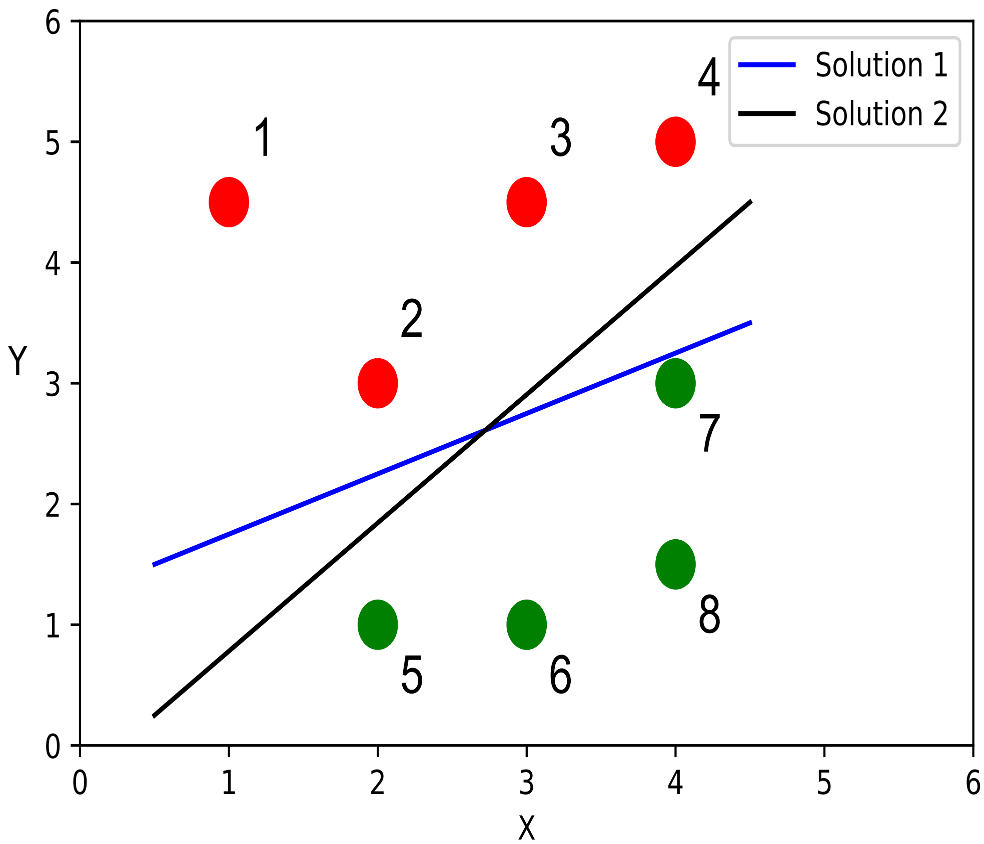
- use *logit*() function

Computing Class Labels



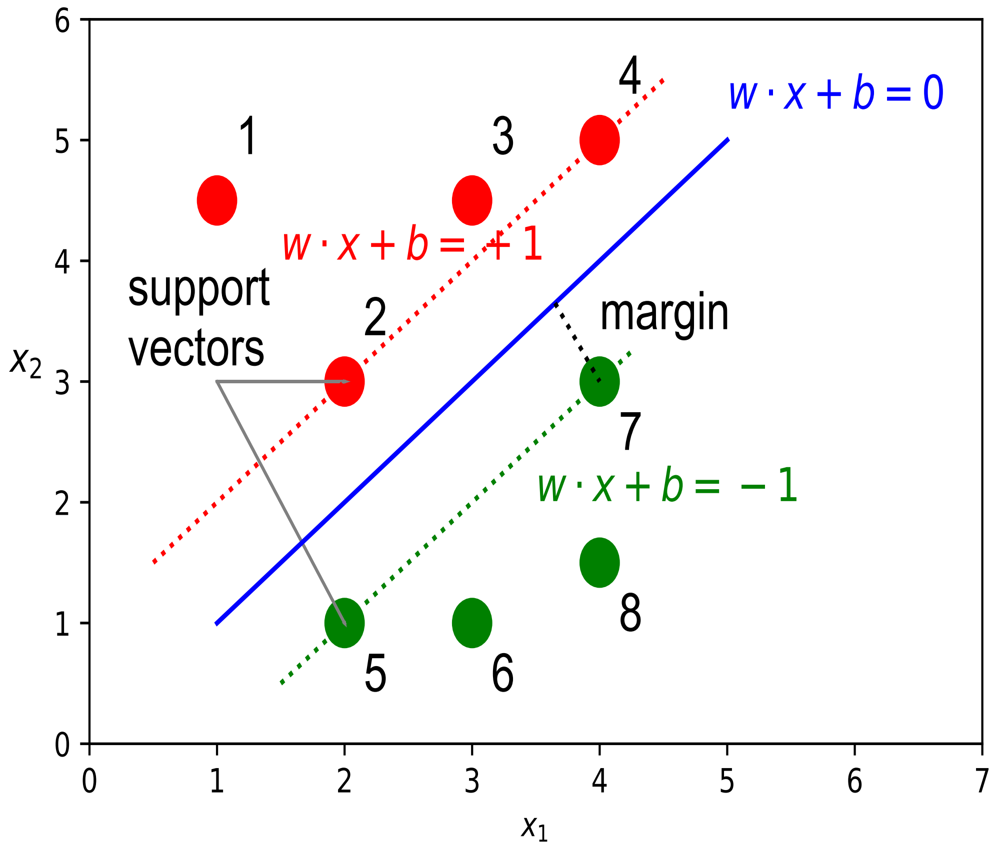
- $z(y^*) < 0.5$ - "red" (class 0)

SVM: How to Separate?



- many possibilities

SVM Intuition



- use "thickest" line
- maximize margins

Summary

- linear functions for predictions and/or classification are simple and easy to explain
- most models are not linear
- idea: use *link* functions to transform models
- look for linear functions in the transformed space
- use these functions to solve our original problem