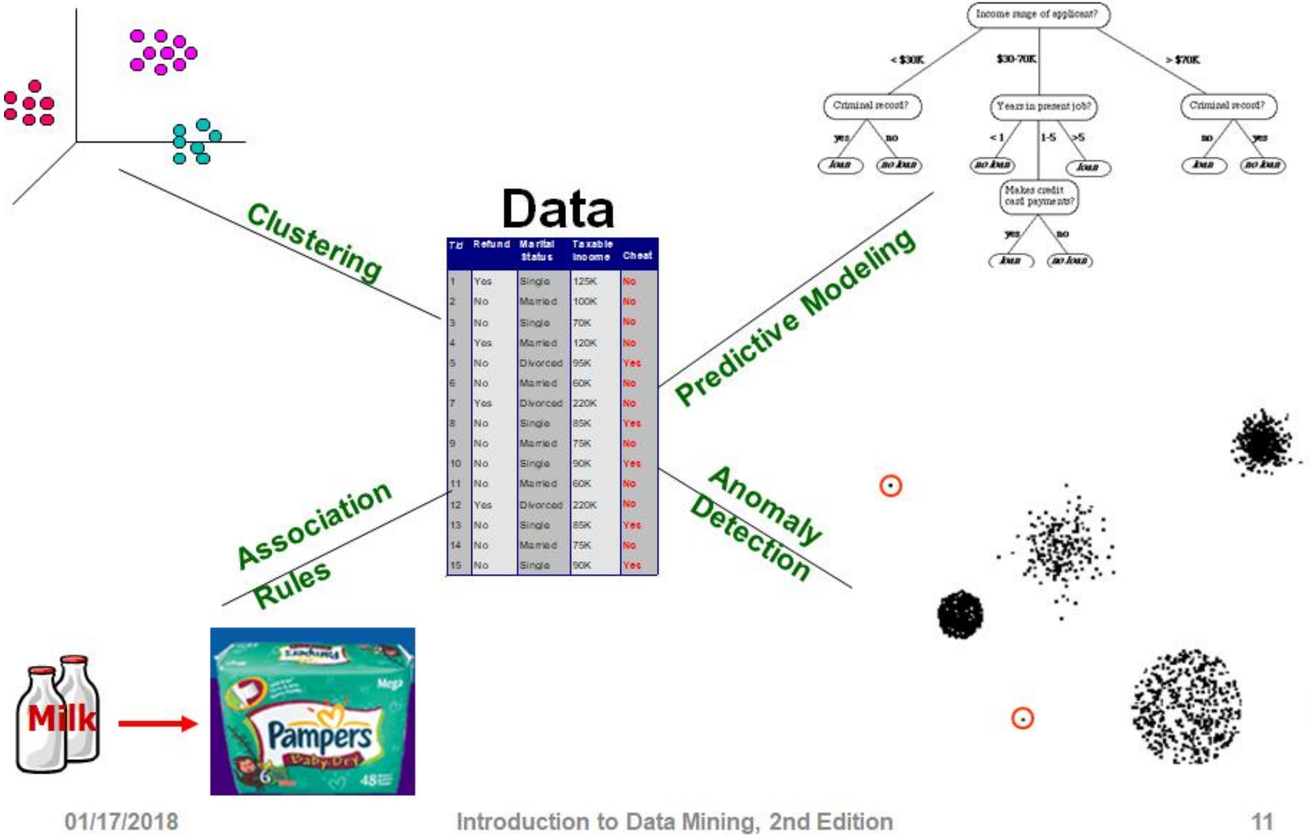
Foundations of Data Science & Analytics: Regression

Ezgi Siir Kibris

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar

Tasks



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Regression

Predictive Modeling

	Output
Classification:	Classes / Categories
Regression:	Continuous Values

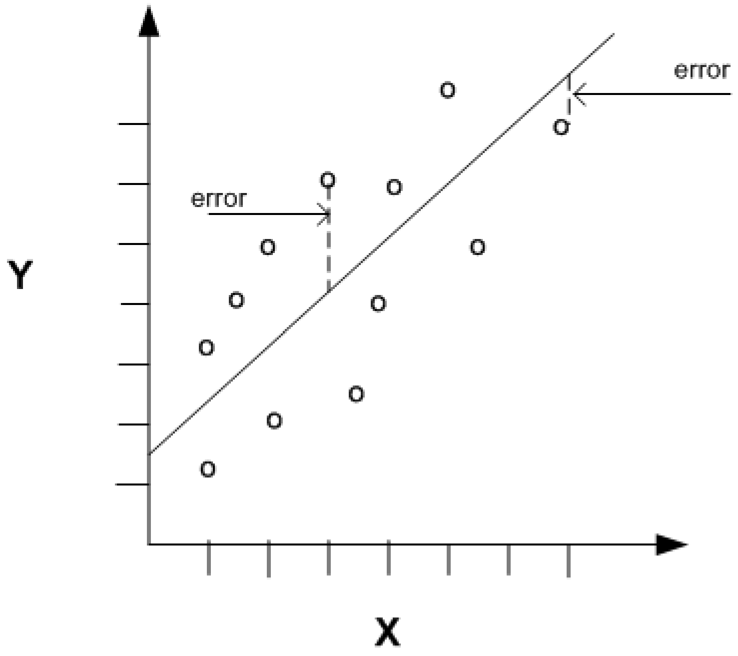
Regression: Definition

- Given a set of training data (x, y), where **x** = (x1, x2, x3, …, xn) are the independent variables and **y** is the dependent variable,

- Fit a mapping function **f(x, w)** from **x** to **y** so that for an unseen **x**, the function can correctly predict its **y** value.

- y = f(x, w) + e
- The learning process is to find the optimal coefficient **w** so that the error **e**, also called the residual is minimized in the training data.

Linear Regression



$$f(x, w) = w_1 x + w_0$$

$$\min L(w) = \sum_i (y^{(i)} - f(x^{(i)}, w))^2$$

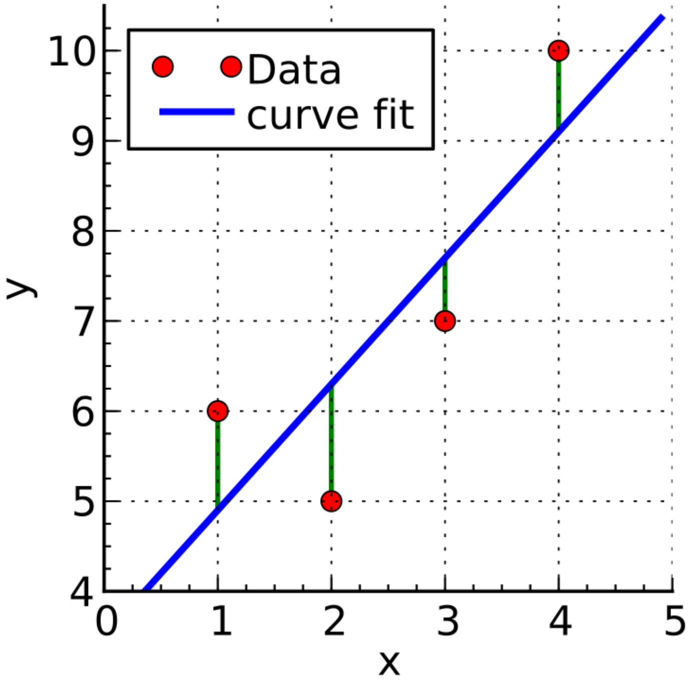
$$\frac{\partial L}{\partial w} = 2 \sum_i \left[(f(x^{(i)}, w) - y^{(i)}) \frac{\partial f(x^{(i)}, w)}{\partial w} \right]$$

$$\sum_i (w_1 x^{(i)} + w_0 - y^{(i)}) x^{(i)} = 0$$

$$\sum_i (w_1 x^{(i)} + w_0 - y^{(i)}) = 0$$

Has unique solution as long as Size of training data N > dimensionality of x

Linear Regression



$$(1,6)(2,5)(3,7)(4,10)$$

$$f(x,w) = w_1 x + w_0$$

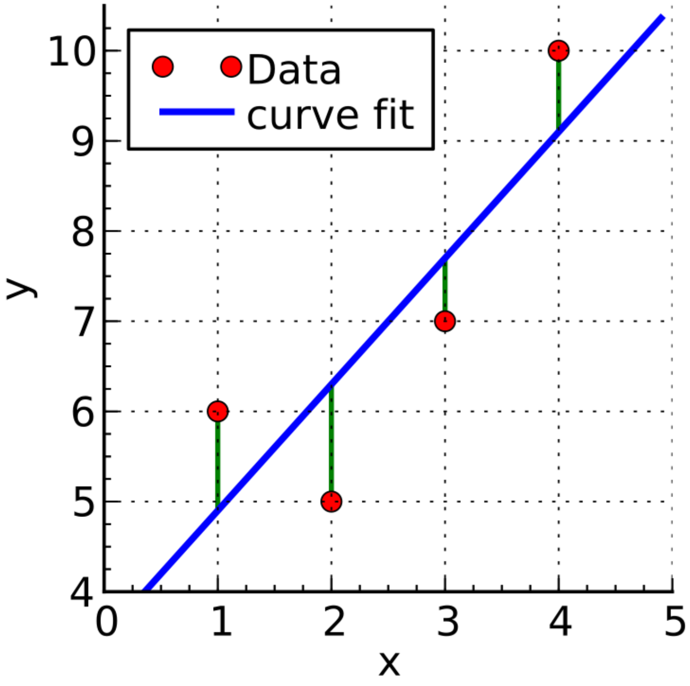
$$\min L(w) = \sum_i (y^{(i)} - f(x^{(i)}, w))^2$$

$$\frac{\partial L}{\partial w} = 2 \sum_i \left[\left(f(x^{(i)}, w) - y^{(i)} \right) \frac{\partial f(x^{(i)}, w)}{\partial w} \right]$$

$$\sum_i (w_1 x^{(i)} + w_0 - y^{(i)}) x^{(i)} = 0$$

$$\sum_i (w_1 x^{(i)} + w_0 - y^{(i)}) = 0$$

Linear Regression



$$(1,6)(2,5)(3,7)(4,10)$$

$$\begin{aligned}&(w_1 + w_0 - 6) + (2w_1 + w_0 - 5)\\&+ (3w_1 + w_0 - 7) + (4w_1 + w_0 - 10)\\&= 10w_1 + 4w_0 - 28 = 0\end{aligned}$$

$$\begin{aligned}&(w_1 + w_0 - 6) + 2(2w_1 + w_0 - 5)\\&+ 3(3w_1 + w_0 - 7) + 4(4w_1 + w_0 - 10)\\&= 30w_1 + 10w_0 - 77 = 0\end{aligned}$$

$$f(x, w) = 1.4x + 3.5$$

Linear Regression

Linear regression imposes two key restrictions:

We assume the relationship between the response Y and the predictors X1, . . . , Xp is:

1 **Linear**

2 **Additive**

The truth is almost never linear; but often the linearity and additivity assumptions are *good enough*

Linear Regression

“Essentially, all models are wrong, but some are useful.”

—George Box

- As an analyst, you can make your models more useful by
    1 Making sure you’re solving useful problems
    2 Carefully interpreting your models in meaningful, practical terms

- So that just leaves one question…
    How can we make our models less wrong?

Polynomial Regression

- Start with a variable X

- Create new variables:

$$X_1 = X,\ X_2 = X^2,\ \cdots,\ X_k = X^k$$

$$\Phi(X) = (X,\ X^2,\ \cdots,\ X^k)$$

- Fit linear regression model with new variables.

Polynomial Regression

$$f(x, w) = w_2 x^2 + w_1 x + w_0$$

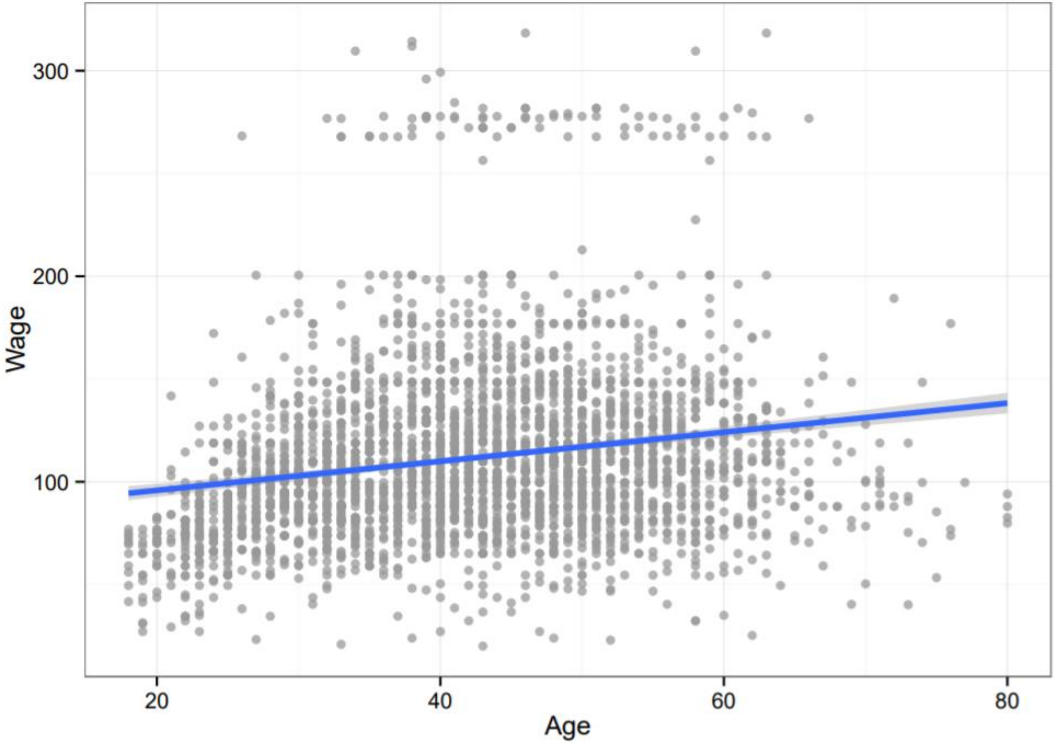
$$\min L(w) = \sum_i (y^{(i)} - f(x^{(i)}, w))^2$$

$$\frac{\partial L}{\partial w} = 2 \sum_i \left[\left(f(x^{(i)}, w) - y^{(i)} \right) \frac{\partial f(x^{(i)}, w)}{\partial w} \right]$$

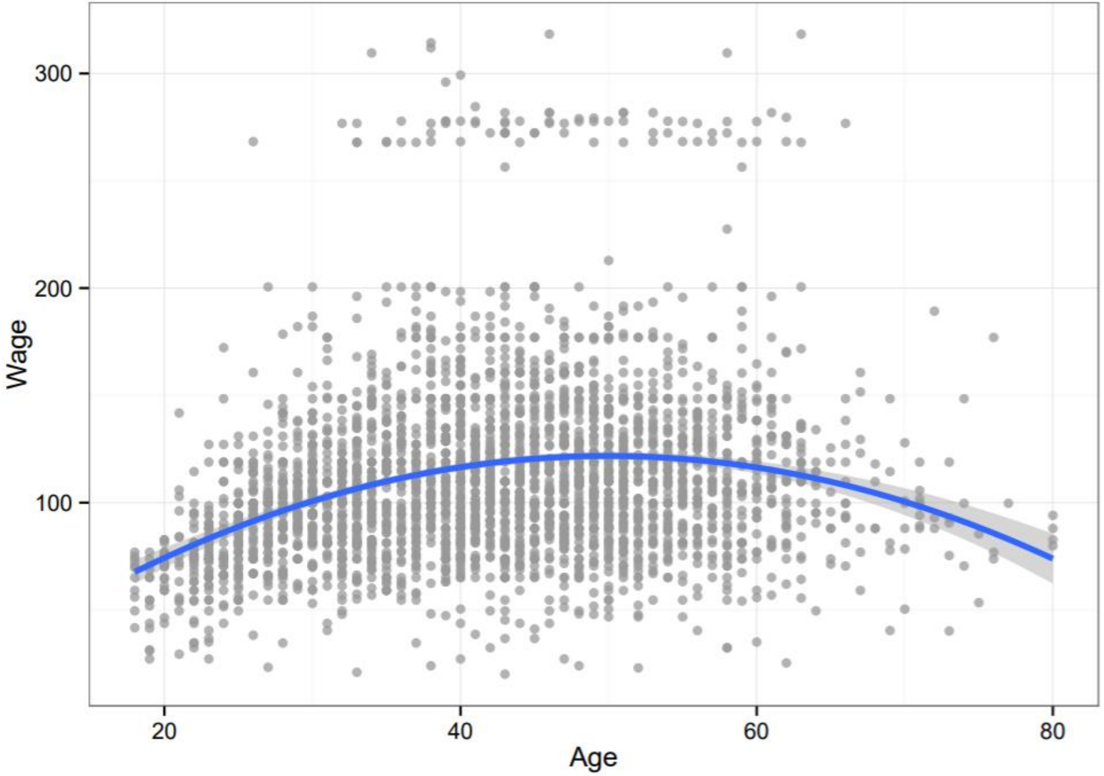
$$\sum_i (w_2 x^{(i)2} + w_1 x^{(i)} + w_0 - y^{(i)}) x^{(i)2} = 0$$

$$\sum_i (w_2 x^{(i)2} + w_1 x^{(i)} + w_0 - y^{(i)}) x^{(i)} = 0$$

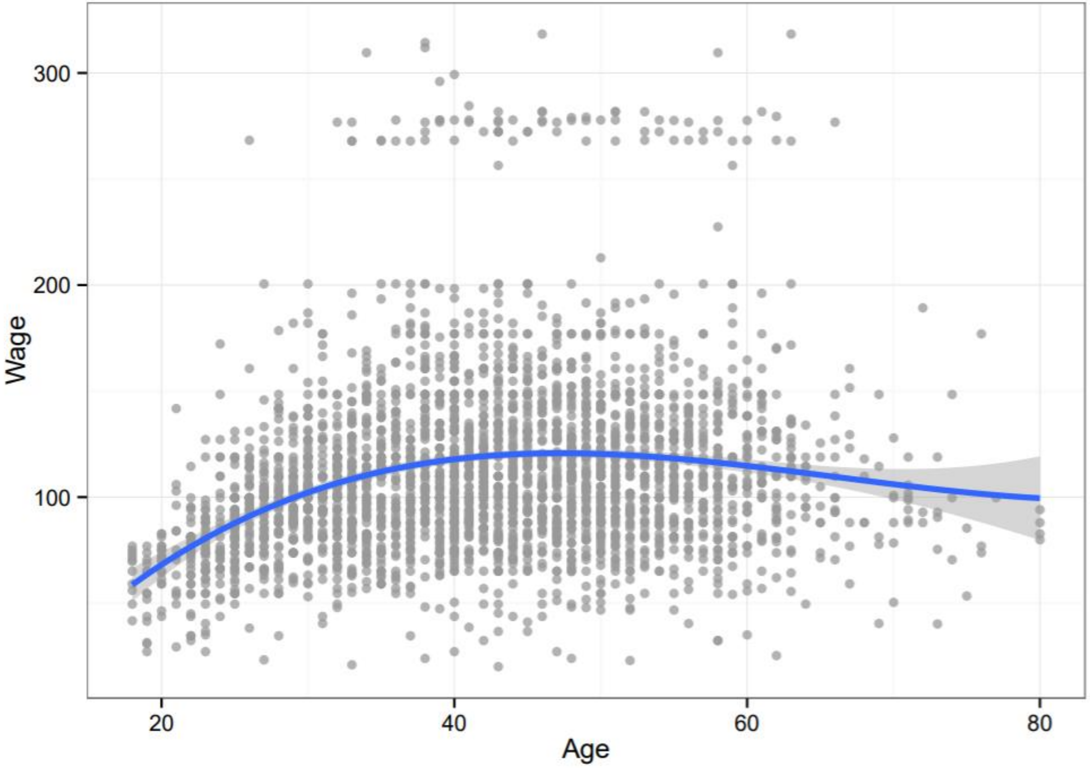
$$\sum_i (w_2 x^{(i)2} + w_1 x^{(i)} + w_0 - y^{(i)}) = 0$$




```
lm(wage ∼ age, data = Wage)
```

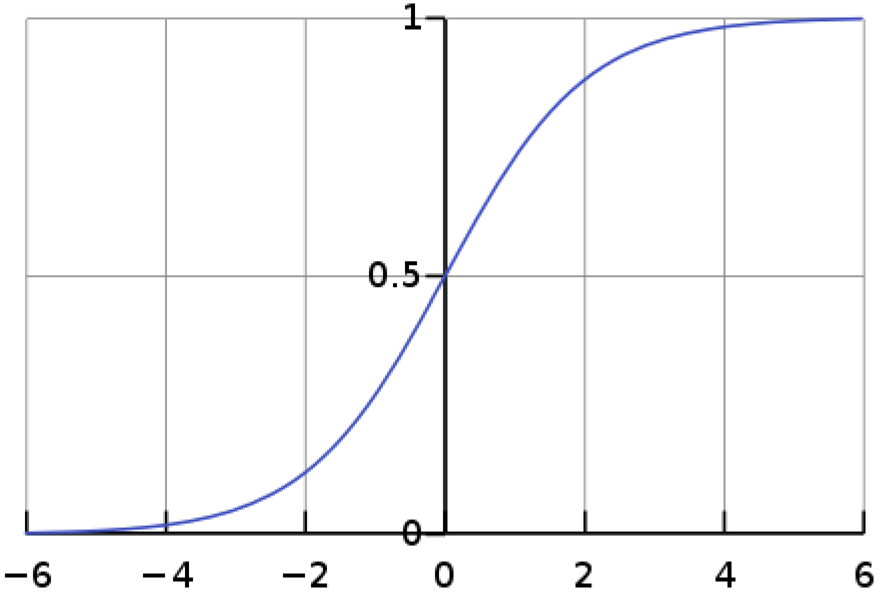



```
lm(wage ~ poly(age, 2), data = Wage)
```




```
lm(wage ~ poly(age, 3), data = Wage)
```

Logistic Regression

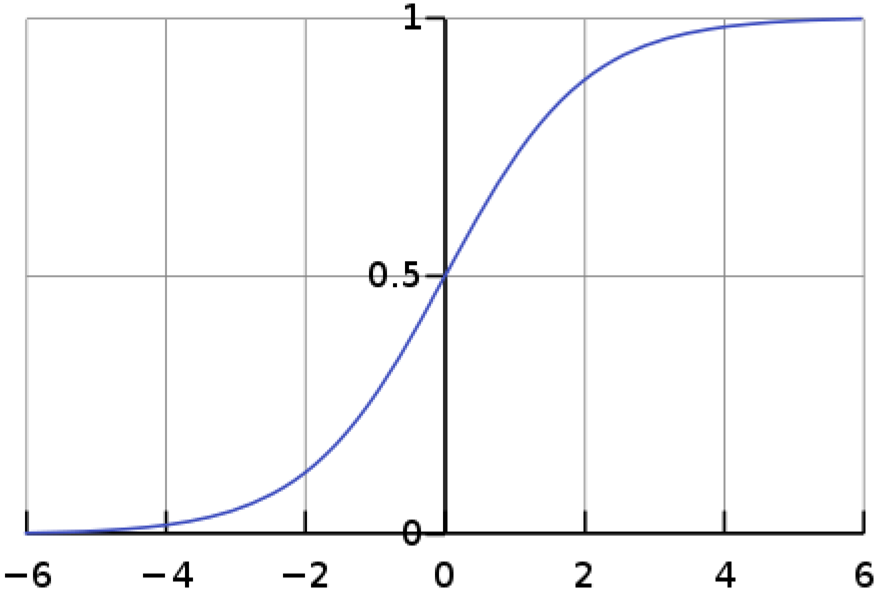


Fit the logistic (sigmoid) function:

$$f(x, w) = \frac{1}{1+e^{-(w_0+w_1 x)}}$$

$$f(-\infty, w) = 0$$
$$f(\infty, w) = 1$$

Logistic Regression



Learns like a regression model, but solves classification problems:

$$y = 0 \ or \ 1$$

The same to a perceptron with logistic activation function.

Loss Functions

$$MSE = \frac{1}{n}\sum_{i}^{n}(y^{(i)} - f(x^{(i)}))^2$$

Convex, easy to solve

$$MAE = \frac{1}{n}\sum_{i}^{n}|y^{(i)} - f(x^{(i)})|$$

Robust to outliers, hard to solve (constant gradient)