* Regularization

$x \rightarrow$ features

$w \rightarrow$ weights

$\hat{y} \rightarrow$

Polynomial regression: High order polynomials

loss will be less

but leads to overfitting

add penalty term.

$$\hat{y} = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_K x^K$$

→ High order polynomials give better fit and lower data loss.

→ However complicated hypotheses leads to overfitting

→ Idea:

· Change the loss function to penalize hypothesis

* Ridge Regression

MSE + λ (norm of wt vectors)

$$L_{ridge} = \frac{1}{2n} \sum_{j=1}^{n} (\hat{y_j} - y_j)^2 + \lambda \frac{1}{2} \sum_{i=1}^{p} w_i^2$$

λ → regularization coefficient

$$y = \underset{0.1}{w_1} x + \underset{0.2}{w_2} x^2$$

$$\begin{matrix} 0.1 & & 0.2 \\ 10 & & 20 \end{matrix}$$

if $x = 2$; $y = 1$

$x = 2.1$;

Forcing my weight parameters to be small.

→ Also known as L_2 or squared value regularization

→ Tries to reduce the length ||w|| of the parameter vector, promoting lesser dependency on $\hat{y}$ on predictors (lower model complexity).

→ λ controls the regularization penalty. $\lambda = 0$ results in regular MSE

LASSO Regression (feature selection)

$$L_{LASSO} = \frac{1}{2n} \sum_{j=1}^{n} (\hat{y}_j - y_j)^2 + \lambda \sum_{i=1}^{p} |\omega_i|$$

$$\omega_1 + \omega_2 \leq C$$

→ Least Absolute Shrinkage and Selection Operator (LASSO), also known as L_1 or absolute value regularization.

plot: $\frac{1}{1+e^{-x}}$

change $x \rightarrow 10$ to 20

classmate
Date
Page

Logistic Regression

* Linear classifier:

$$z = w_0 + w_1 x_1 + \cdots + w_d x_d$$

- The output above is unbounded, whereas it should be between 0 and 1, ie;

$$z = [0 \leq p(x) \leq 1]$$

- Changing p by the same amount requires a bigger changes in x when p is already large (or small) than when p is close to $1/2$.

$$\log \frac{p(x)}{1-p(x)} = w_0 + w_1 x_1 + \cdots w_d x_d$$

$$p(x) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \rightarrow \text{sigmoid curve.}$$

Understanding Sigmoid Curve:

$$y = \frac{1}{1+e^{-(wx+b)}}$$

* Loss function

→ In binary classification, the cross-entropy loss is defined as:

$$L(y, p) = -(y \log p + (1-y) \log(1-p))$$

- It is a measure of the difference between the predicted