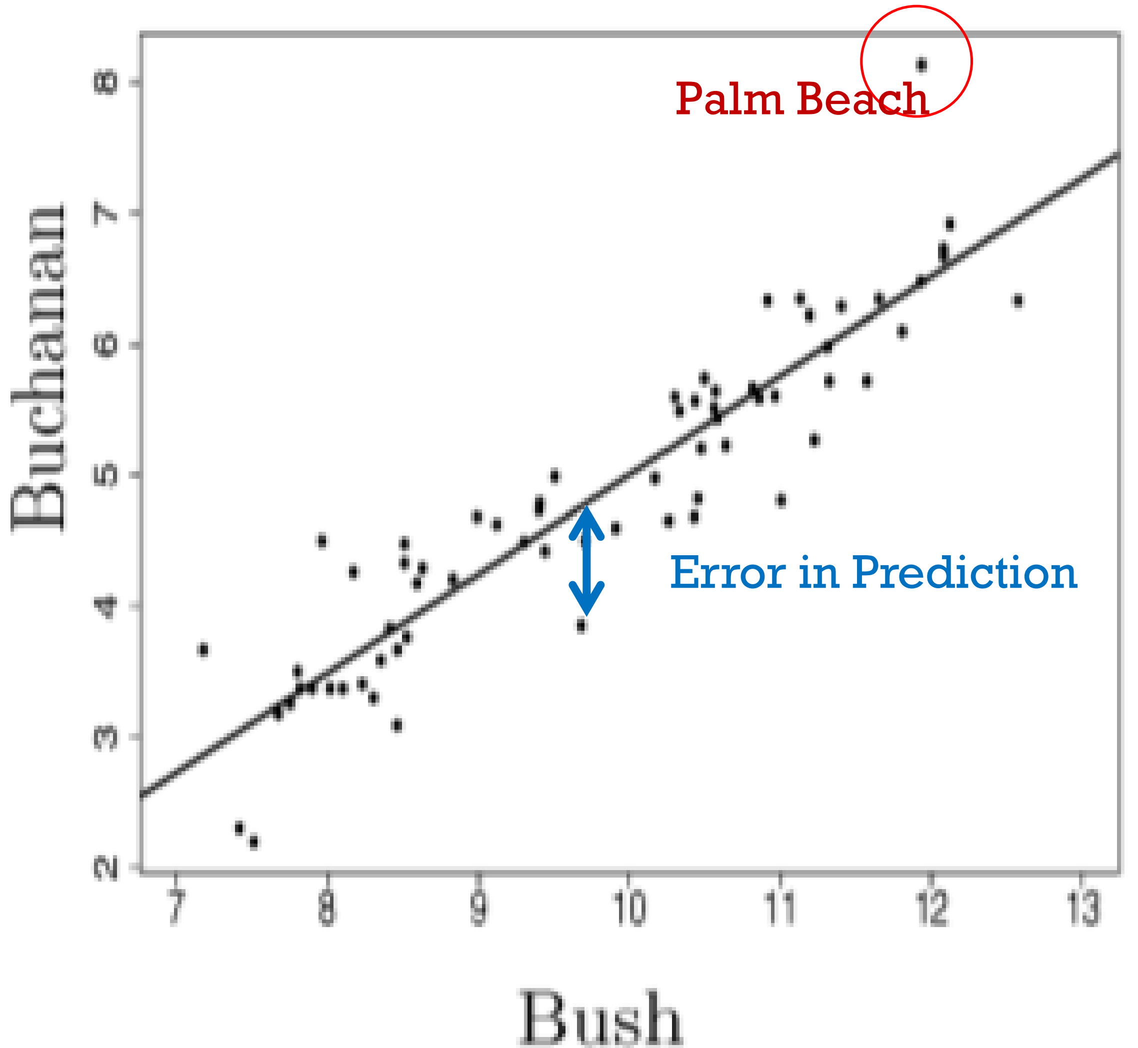
REGRESSION

BUSH VS BUCHANAN



REGRESSION

Training Data. $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \dots, \left(x^{(n)}, y^{(n)}\right)$

- Features $x^{(t)} \in \mathbb{R}^d$
- Response $y^{(t)} \in \mathbb{R}$

We want to learn functions $f: \mathbb{R}^d \to \mathbb{R}$ such that for all data (x, y)

$$y \approx f(x; \theta).$$

We want $y - f(x; \theta)$ to be small.

SQUARED ERROR

Risk = "Expected Loss"
Empirical = "of the Data"

Loss Function.

$$\text{Loss}(z) = \frac{1}{2} z^2$$

CONVEX!!

Empirical Risk.

$$\mathcal{L}_n(\theta) = \frac{1}{n} \sum_{\text{data } (x,y)} \text{Loss}\big(y - f(x;\theta)\big)$$

$$= \frac{1}{n} \sum_{\text{data } (x,y)} \frac{1}{2} \big(y - f(x;\theta)\big)^2$$

- Big errors are penalized more heavily
- Want to apply convex optimization

LINEAR REGRESSION

Model. Set of linear functions

$$f(x;\theta,\theta_0) = \theta_1 x_1 + \cdots + \theta_d x_d + \theta_0 = \theta^\top x + \theta_0$$

Model Parameters.

$$\theta \in \mathbb{R}^d, \theta_0 \in \mathbb{R}$$

Training Data.

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

Learning Objective.

$$\mathcal{L}_n(\theta,\theta_0) = \frac{1}{n} \sum\nolimits_{\text{data }(x,y)} \frac{1}{2}\left(y - (\theta^\top x + \theta_0)\right)^2$$

ALGORITHMS

STOCHASTIC GRADIENT DESCENT

Ignore θ_0 for now

Gradient.

$$\nabla_\theta \mathcal{L}_n(\theta) = \frac{1}{n} \sum_{\text{data } (x,y)} \nabla_\theta \left(\frac{1}{2} (y - \theta^\top x)^2 \right)$$

$$\nabla_\theta \left(\frac{1}{2} (y - \theta^\top x)^2 \right) = \left(y - \theta^T x \right) \nabla_\theta (y - \theta^\top x)$$

$$= -(y - \theta^\top x)\, x$$

Algorithm.

Set $\theta = 0$

Randomly select data (x, y)

$$\theta \longleftarrow \theta + \eta_k \,(y - \theta^\top x)\, x$$

EXACT SOLUTION

$$\nabla_\theta \mathcal{L}_n(\theta) = \frac{1}{n}\sum_{\text{data }(x,y)} -(y - \theta^\top x)\, x$$

$$= \frac{1}{n}\sum_{\text{data }(x,y)} -xy + (\theta^\top x)\, x$$

$$= \frac{1}{n}\sum_{\text{data }(x,y)} -xy + x(x^\top \theta) \; = -B + A\theta$$

where

$$B = \frac{1}{\mathrm{n}}\sum_{t=1}^{n} x^{(t)} y^{(t)} = \frac{1}{n}\left[x^{(1)}, \dots, x^{(n)}\right]\left[y^{(1)}, \dots, y^{(n)}\right]^\top = \frac{1}{n} X^\top Y$$

$$\mathrm{A} = \frac{1}{\mathrm{n}}\sum_{t=1}^{n} x^{(t)} x^{(t)\top} = \frac{1}{n}\left[x^{(1)}, \dots, x^{(n)}\right]\left[x^{(1)}, \dots, x^{(n)}\right]^\top = \frac{1}{n} X^\top X$$

$$X = \left[x^{(1)}, \dots, x^{(n)}\right]^\top, \;\; Y = \left[y^{(1)}, \dots, y^{(n)}\right]^\top$$

EXACT SOLUTION

Optimization problem is convex, so the minimum is attained when the gradient is zero.

$$
\begin{aligned}
\nabla_\theta \mathcal{L}_n\big(\hat\theta\big) = 0 \quad &\Leftrightarrow \quad \frac{1}{n}(X^\top X)\,\hat\theta = \frac{1}{n}X^\top Y \\
&\Leftrightarrow \quad \hat\theta = (X^\top X)^{-1}X^\top Y
\end{aligned}
$$

Issues.

1. Need $X^\top X$ to be invertible
   - Feature vectors $x^{(1)}, \dots, x^{(n)}$ must span $\mathbb{R}^d$
   - Must have more data than features, $n > d$.
2. What if $X^\top X \in \mathbb{R}^{d\times d}$ is a large matrix?
   - Takes long time to invert
   - Use stochastic gradient descent

REGULARIZATION

REGULARIZATION

Temp. on Mars

Weight Age

Height $y \approx \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d + \theta_0$

How do we ensure that $\theta_k = 0$ when feature x_k is irrelevant?

Pick simplest model that explains data ➔ **generalization**

Add a penalty.

$$\mathcal{L}_{n,\lambda}(\theta) = \frac{1}{n} \sum_{\text{data }(x,y)} \frac{1}{2}(y - \theta^\top x)^2 + \frac{\lambda}{2} \|\theta\|^2$$

Ridge Regression

Regularization parameter $\lambda \geq 0$

ALGORITHMS FOR REGULARIZATION

Stochastic Gradient Descent

Gradient. $\nabla_\theta \mathcal{L}_{n,\lambda}(\theta) = \lambda\theta - (y - \theta^\top x)\, x$

Update rule. $\theta \longleftarrow (1 - \eta_k \lambda)\, \theta + \eta_k\, (y - \theta^\top x)\, x$

Exact Solution

$$\begin{aligned} \nabla_\theta \mathcal{L}_{n,\lambda}(\hat{\theta}) = 0 \quad &\Leftrightarrow \quad \lambda\hat{\theta} + \frac{1}{n}(X^\top X)\,\hat{\theta} = \frac{1}{n} X^\top Y \\ &\Leftrightarrow \quad \hat{\theta} = (n\lambda I + X^\top X)^{-1} X^\top Y \end{aligned}$$

Performance Metrics

Learning Objective

$$\mathcal{L}_{n,\lambda}(\theta) = \frac{1}{n} \sum_{\boxed{\text{trg data } (x,y)}} \frac{1}{2}(y - \theta^\top x)^2 + \boxed{\frac{\lambda}{2} \|\theta\|^2}$$

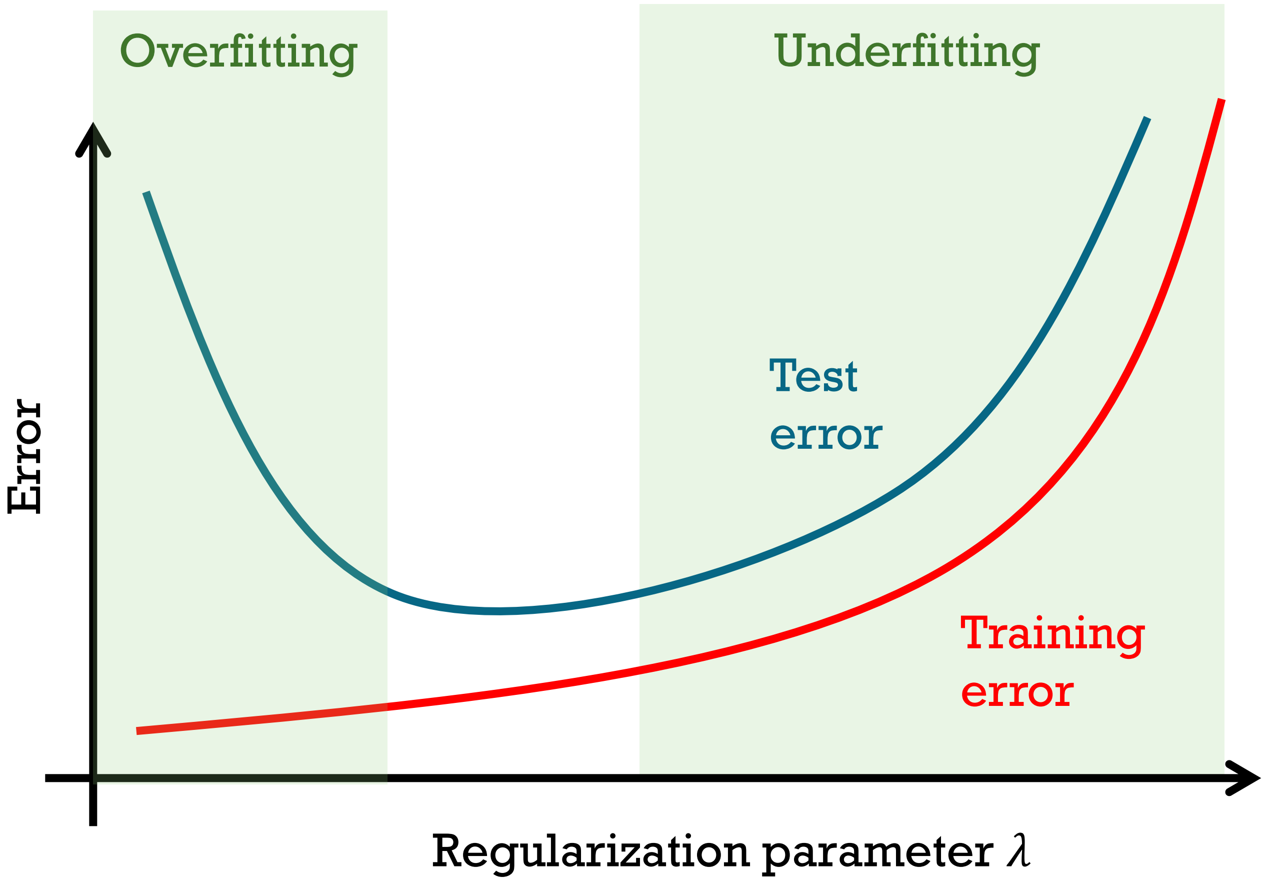
Training Error

$$\mathcal{L}_{n}(\theta) = \frac{1}{n} \sum_{\boxed{\text{trg data } (x,y)}} \frac{1}{2}(y - \theta^\top x)^2$$

Test Error

$$\mathcal{R}(\theta) = \frac{1}{n} \sum_{\boxed{\text{test data } (x,y)}} \frac{1}{2}(y - \theta^\top x)^2$$

EFFECT OF REGULARIZATION



DISCUSSION

WHY LINEAR?

Expressive power is in the features.

e.g. polynomial regression

$$f(x;\theta) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_d x^d$$

feature vector (computed from x)

$$(1, x, x^2, \dots, x^d)$$

SOURCES OF ERROR

Estimation Error (variance)

- Noisy data
- Too few data

Structural Error (bias)

- Data is not linear
- Too many parameters