Linear Regression and Bias-Variance Tradeoff

Machine Learning - 10601

Geoff Gordon, Miroslav Dudík

http://www.cs.cmu.edu/~ggordon/10601/
September 21, 2009

Last time: linear regression

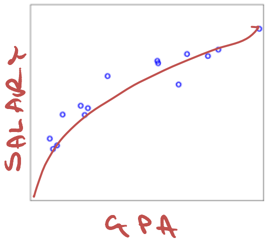
Goal: TASK

predict a continuous response
from continuous/categorical inputs

EXPERIENCE

Input: (GPA, GENDER...)

$(x_1,y_1), \ldots, (x_N,y_N)$ ← SALARY "RESPONSE"



Model: LINEAR

$y \approx \sum_{j=1}^{M} w_j \phi_j(x)$ ← FEATURES BASIS FCTS

Performance measure:

$\sum_n (y_n - w \cdot \phi(x_n))^2$ ← MEAN SQUARED ERROR

Linear regression

Input: $(x_1,y_1), \dots, (x_N,y_N)$

Assume: $y \approx \sum_i w_j\phi_j(x)$

Goal: $\min_{w} \sum_n \left(y_n - w \cdot \phi(x_n)\right)^2$

feature vecor

inner product

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \approx \begin{pmatrix} \phi(x_1)\cdot w \\ \vdots \\ \phi(x_N)\cdot w \end{pmatrix} = \underbrace{\begin{pmatrix} \phi(x_1) \\ \vdots \\ \phi(x_N) \end{pmatrix}}_{\Phi \text{ MATRIX}} \cdot w$$

ROW VECTOR OF FEATURES

COLUMN VECTOR OF WEIGHTS

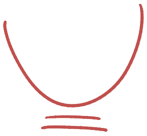
LIN. IN w

$$\min_w \| y - \Phi w \|^2$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

vec. of predictions

CONVEX IN w



looking for $\hat{y}$ in column space of Φ that is closest to y

$$\nabla_w = -2\,\Phi^T(y - \Phi w) = 0$$

$$2\,\Phi^T\Phi\, w = 2\,\Phi^T y$$

IF INVERTIBLE

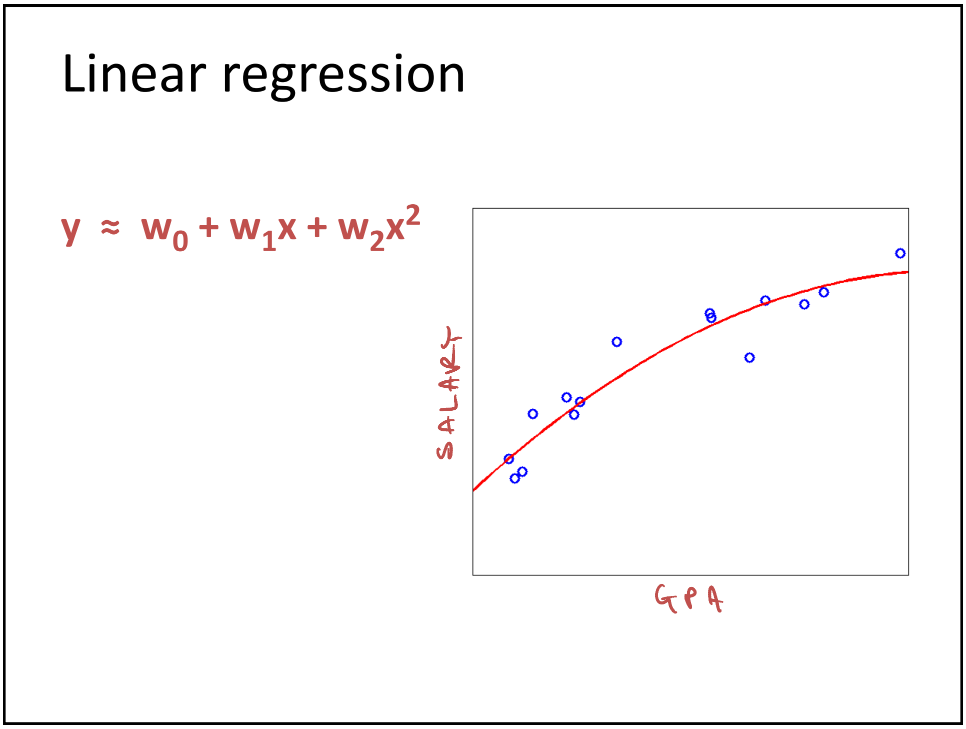
$$w = (\Phi^T\Phi)^{-1}\Phi^T y$$

INVERSION NOT POSSIBLE IF FEATURES LINEARLY DEPENDENT

PROBLEMS

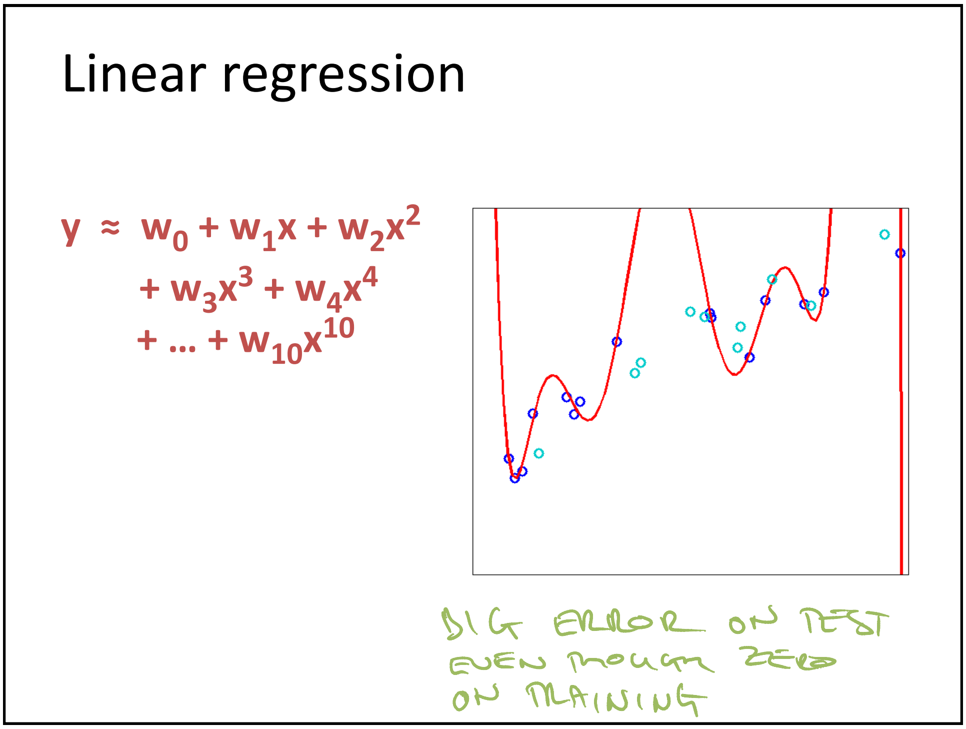
I. COLINEARITY AMONG FEATURES

I. TOO MANY FEATURES CAN FIT ANYTHING



Linear regression

$y \approx w_0 + w_1x + w_2x^2$

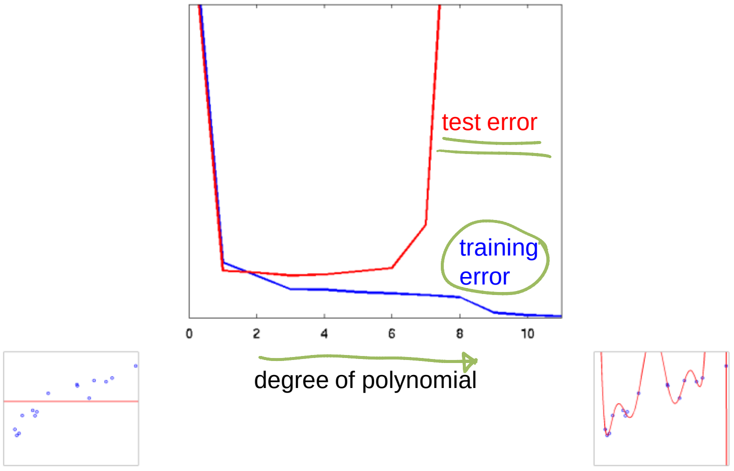


Linear regression

$y \approx w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \ldots + w_{10}x^{10}$

BIG ERROR ON TEST
EVEN THOUGH ZERO
ON TRAINING

Training & Test Error Don’t Match



Training & Test Error Don’t Match: Why?

unknown dist. on (x,y)

$$\text{error}_{\text{train}} = \boxed{\frac{1}{N}\sum_n \left(y_n - \phi(x_n)\cdot w\right)^2}$$

ALWAYS TOO OPTIMISTIC

$$\text{error}_{\text{true}} = E_{x,y}\left[\left(y - \phi(x)\cdot w\right)^2\right]$$

$$= \int \left(y - \phi(x)\cdot w\right)^2 \underline{p(x,y)}\, dx\, dy$$

samples $(x'_1, y'_1), \ldots, (x'_K, y'_K)$

MONTE CARLO SAMPLING

$$\approx \boxed{\frac{1}{K}\sum_k \left(y'_k - \phi(x'_k)\cdot w\right)^2} = \boxed{\text{error}_{\text{test}}}$$

indep. of $\hat{w}$

Training & Test Error

Training error:

overly optimistic

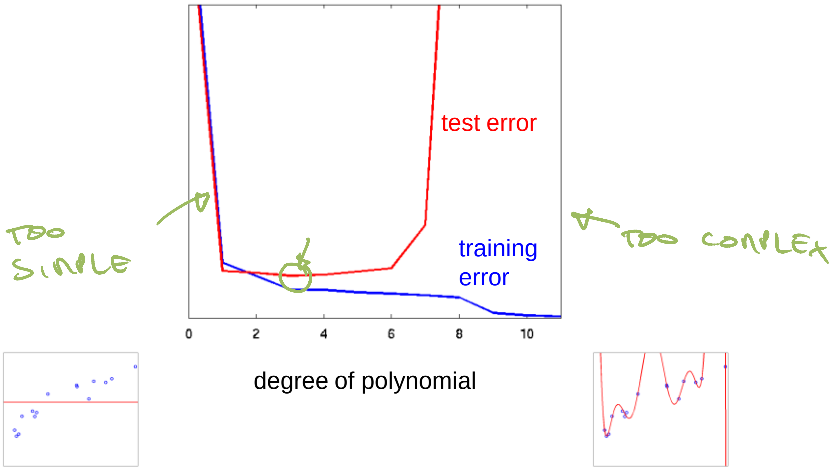
Test error:

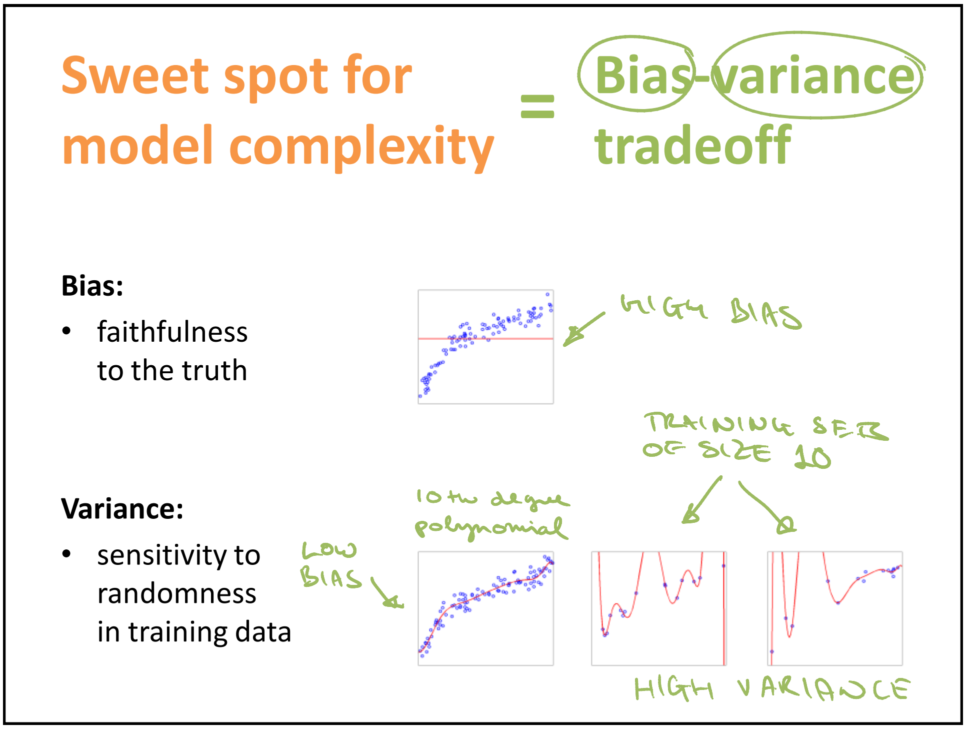
approximation of **prediction error**

as long as

test data set never touched during training

Sweet spot for model complexity





Sweet spot for model complexity = Bias-variance tradeoff

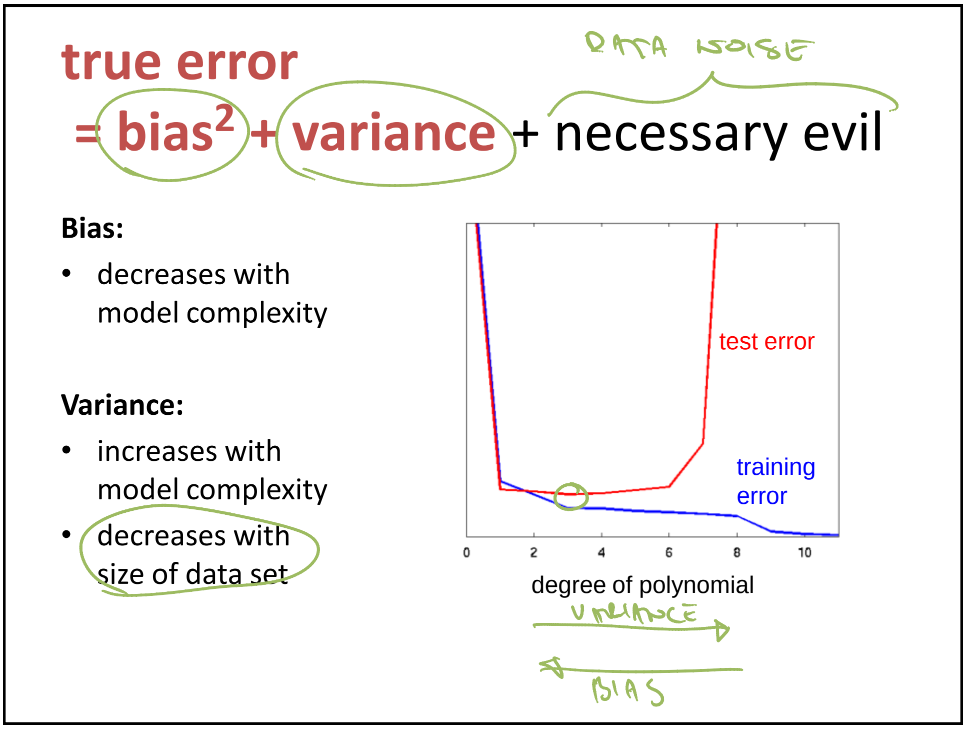
Bias:

- faithfulness to the truth

Variance:

- sensitivity to randomness in training data

HIGH BIAS
TRAINING SET OF SIZE 10
10th degree polynomial
LOW BIAS
HIGH VARIANCE



true error

$$= \text{bias}^2 + \text{variance} + \overbrace{\text{necessary evil}}^{\text{DATA NOISE}}$$

Bias:

- decreases with model complexity

Variance:

- increases with model complexity
- decreases with size of data set

test error
training
error
0
2
4
6
8
10
degree of polynomial
VARIANCE
BIAS

true error

= bias² + variance + necessary evil

$X \sim p(x)$

$Y = f(X) + \varepsilon$ $\varepsilon \sim N(0, \sigma^2)$

prediction at x_0: $\underset{\text{RANDOM}}{\text{data}} \rightarrow \hat{w}$ RANDOM VAR

$x_0 \mapsto \phi(x_0)\cdot\hat{w}$ → RANDOM VAR: $\hat{Y}(x_0)$

error(x_0) = $E_{Y,\mathcal{D}}\left[(Y - \hat{Y}(x_0))^2 \mid X = x_0\right]$

$\parallel$

$$E\Big[\Big(\underbrace{(Y - E_Y[Y \mid x_0])}_{\text{EXPECT}=0} + (E_Y[Y \mid x_0] - \overset{\text{CONST.}}{E_{\mathcal{D}}[\hat{Y}(x_0)]}) + \underbrace{(E_{\mathcal{D}}[\hat{Y}(x_0)] - \hat{Y})}_{\text{EXPECTATION}=0}\Big)^2 \Big| X = x_0\Big]$$

$f(x)$ $f(x_0)$

true error

= bias² + variance + necessary evil

$$\text{error}(x_0) = E\Big[\Big(\overbrace{(\underset{\varepsilon}{Y - f(x)})}^{\text{EXP}=0} + \overbrace{(f(x) - E_{\mathcal{D}}[\hat{Y}(x_0)])}^{\text{CONST.}} + \underbrace{(E_{\mathcal{D}}[\hat{Y}(x_0)] - \hat{Y}(x_0))}_{\text{EXP}=0}\Big)^2 \Big| x_0\Big]$$

$$(A+B+C)^2 = A^2 + B^2 + C^2 + 2AB + 2AC + 2BC$$

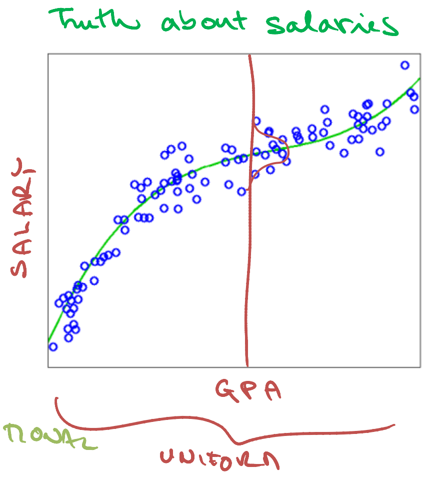
$$\text{error}(x_0) = E[\underbrace{(Y - f(x_0))^2}_{\sigma^2}] + \underbrace{(f(x_0) - E_{\mathcal{D}}[\hat{Y}(x_0)])^2}_{\text{BIAS}^2} + \underbrace{E\left[(\hat{Y}(x_0) - E[\hat{Y}(x_0)])^2\right]}_{\text{VARIANCE}}$$

$$\text{error} = \int \text{error}(x)\,p(x)\,dx$$

Announcements

- project proposals due
  this Wednesday at 10:30am
- HW #3 OUT SOON
  (DUE OCT 7)

Least squares fit = max likelihood for Gaussians



$X \sim p(x)$

$N(0, \sigma^2)$

$Y \approx w \cdot \phi(x) + \epsilon$

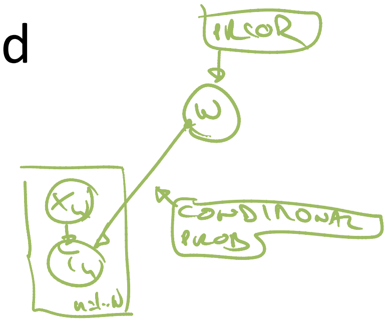
$p(y|x) = N(\underline{w} \cdot \phi(x), \sigma^2)$

$\max_w \prod_n p(y_n|x_n)$ ← CONDITIONAL MLE

$\max_w \sum_n \left(\cancel{\text{const.}} - \frac{(y_n - w \cdot \phi(x_n))^2}{2\sigma^2} \right)$

$\min_w \sum_n (y_n - w \cdot \phi(x_n))^2$

Beyond max likelihood for Gaussians



- add a prior over **w**

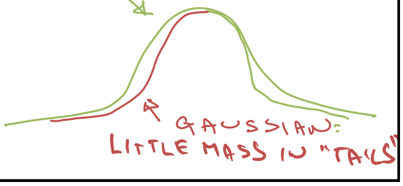
MAP, prediction

- replace the Gaussian by a different model
  - different noise model
  - different support for **Y** $\in \mathbb{R}$

$Y \in \{0,1\}$

FATTER TAILS ≈ ALLOW OUTLIERS

$p(\varepsilon) \propto \exp\{-|\varepsilon|\}$



Regression with a Gaussian prior

$Y \sim N(\phi(x)\cdot w, \sigma^2)$

$w_j \sim N(0, \tau^2)$

MAP:

$$\min_w \left[-\log p(w|\text{data})\right]$$

$$// \qquad // \|w\|^2$$

$$\text{const} + \underbrace{\frac{1}{2\sigma^2}\sum_n (y_n - w\cdot\phi(x_n))^2}_{\text{NEG. LOG. LIKELIHOOD}} + \underbrace{\frac{1}{2\tau^2}\sum_j w_j^2}_{\text{NEG. LOG. PRIOR}} \qquad / \cdot \sigma^2$$

$$\min_w \underbrace{\frac{1}{2}\|y - \Phi w\|^2}_{\text{ERROR}} + \frac{1}{2}\boxed{\frac{\sigma^2}{\tau^2}}\|w\|^2$$

λ

CONTROL OVER COMPLEXITY

PUSHING TOWARDS $w=0$

Regression with a Gaussian prior

truth

$Y \sim \underline{\phi(x)\cdot \boxed{w}} + \text{noise}$

$w \sim \overline{N(0,\tau^2)}$

ALWAYS INVERTIBLE:

$$\hat{\omega} = (\Phi^T\Phi + \underline{\underline{\lambda I}})^{-1}\Phi^T y$$

OPTIMIZATION BETTER DEFINED

"REGULARIZED"

=

"SHRINKAGE"

=

NEG. LOG. PRIOR

$$E[\hat{\omega}] = (\Phi^t\Phi + \lambda I)^{-1}\underbrace{\Phi^T\Phi}_{A}\cdot w$$

A^{-1} $\lambda = 0$ A

$E[\hat{\omega}] = w$ if $\lambda = 0$: $\boxed{\text{NO BIAS}}$

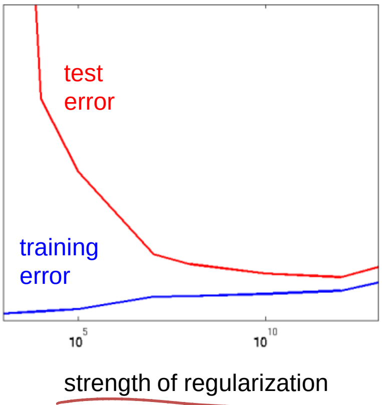
$E[\hat{\omega}]$ is "shrinking" towards zero as λ increases $\boxed{\text{INCREASING BIAS}}$

Bias-variance tradeoff for ridge regression

REGULARIZATION

$\frac{\lambda}{2}\|\omega\|^2$

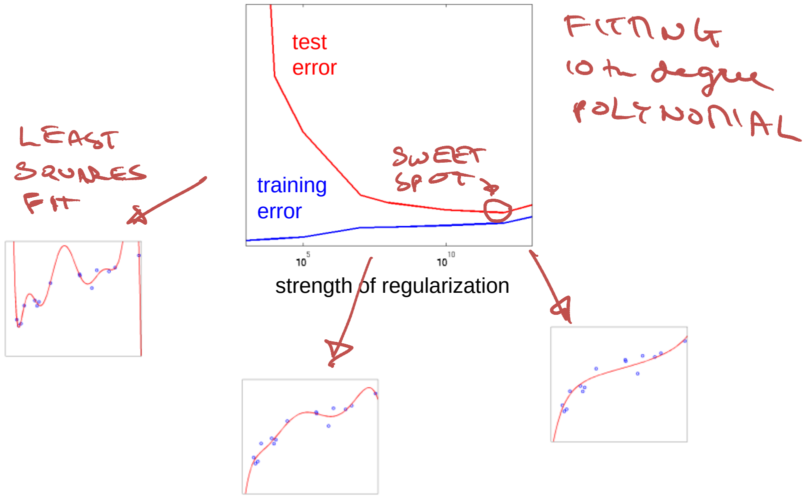
POLY OF DEGREE 10



BIAS

VARIANCE

Bias-variance tradeoff for ridge regression



LEAST SQUARES FIT

SWEET SPOT

FITTING 10th degree POLYNOMIAL