# Machine Learning 4771

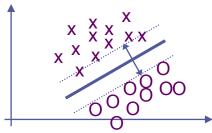
Instructor: Tony Jebara

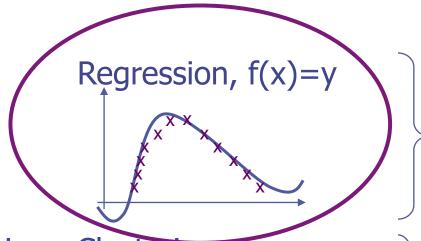
#### Topic 2

- Regression
- Empirical Risk Minimization
- Least Squares
- Higher Order Polynomials
- Under-fitting / Over-fitting
- Cross-Validation

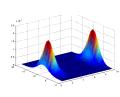
#### Regression

#### Classification

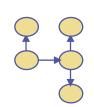




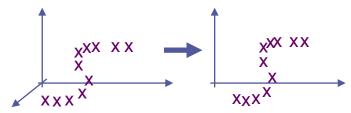
Density/Structure Estimation



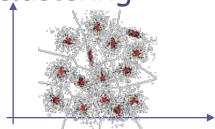




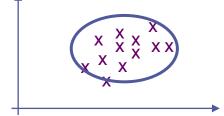
**Feature Selection** 







**Anomaly Detection** 



Unsupervised

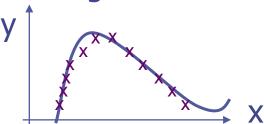
Supervised

#### **Function Approximation**

Start with training dataset

$$\mathcal{X} = \left\{\!\!\left(\boldsymbol{x}_{\!\scriptscriptstyle 1}, \boldsymbol{y}_{\!\scriptscriptstyle 1}\right),\!\!\left(\boldsymbol{x}_{\!\scriptscriptstyle 2}, \boldsymbol{y}_{\!\scriptscriptstyle 2}\right),\!\ldots,\!\!\left(\boldsymbol{x}_{\!\scriptscriptstyle N}, \boldsymbol{y}_{\!\scriptscriptstyle N}\right)\!\!\right\} \quad \boldsymbol{x} \in \mathbb{R}^{\scriptscriptstyle D} = \right|$$

- Have N (input, output ) pairs
- •Find a function f(x) to predict y from x That fits the training data well



- •Example: predict the price of house in dollars y using x = [#rooms; latitude; longitude; ...]
- Need: a) Way to evaluate how good a fit we have
  - b) Class of functions in which to search for f(x)

#### **Empirical Risk Minimization**

- •Idea: minimize 'loss' on the training data set
- •Empirical = use the training set to find the best fit
- •Define a loss function of how good we fit a single point: L(y,f(x))•Empirical Risk = the average loss over the dataset

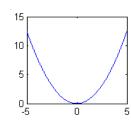
$$R = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

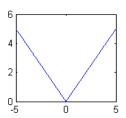
•Simplest loss: squared error from y value

$$L\!\left(\boldsymbol{y}_{\boldsymbol{i}}, f\!\left(\boldsymbol{x}_{\boldsymbol{i}}\right)\right) = \frac{1}{2}\!\left(\boldsymbol{y}_{\boldsymbol{i}} - f\!\left(\boldsymbol{x}_{\boldsymbol{i}}\right)\right)^{2}$$



$$L(y_{i}, f(x_{i})) = |y_{i} - f(x_{i})|$$





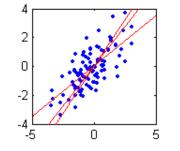
#### **Linear Function Classes**

Linear is simplest class of functions to search over:

$$f(x;\theta) = \theta^T x + \theta_0 = \sum_{d=1}^D \theta_d x(d) + \theta_0$$

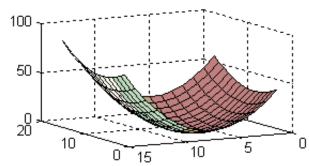
•Start with x being 1-dimensional (D=1):

$$f(x;\theta) = \theta_1 x + \theta_0$$



ullet Plug in the above & minimize empirical risk over  $\theta$ 

$$R(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \left( y_i - \theta_1 x_i - \theta_0 \right)^2$$



- •Note: minimum occurs when  $R(\theta)$  gets flat (not always!)
- •Note: when R( $\theta$ ) is flat, gradient  $\nabla_{\theta} R = 0$

Min by Gradient=0

•Gradient=0 means the partial 
$$\nabla_{\theta} R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
•Take partials of empirical risk:

•Take partials of empirical risk:

$$R(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \left( y_i - \theta_1 x_i - \theta_0 \right)^2$$

 $\begin{array}{l} \text{Min by Gradient=0} \\ \text{•Gradient=0 means the partial} \\ \nabla_{\theta} R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{array}$ 

•Take partials of empirical risk:

$$\begin{split} R\left(\theta\right) &= \tfrac{1}{2N} \sum\nolimits_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right)^2 \\ &\tfrac{\partial R}{\partial \theta_0} = \tfrac{1}{N} \sum\nolimits_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right) \left(-1\right) = 0 \end{split}$$

 $\begin{array}{l} \text{Min by Gradient=0} \\ \text{•Gradient=0 means the partial} \\ \text{•derivatives are all 0} \end{array} \quad \nabla_{\theta} R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ 

•Take partials of empirical risk:

$$\begin{split} R\left(\theta\right) &= \frac{1}{2N} \sum\nolimits_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right)^2 \\ &\frac{\partial R}{\partial \theta_0} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right) \left(-1\right) = 0 \\ &\frac{\partial R}{\partial \theta_1} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right) \left(-x_i\right) = 0 \end{split}$$

## Min by Gradient=0

•Gradient=0 means the partial  $\nabla_{\theta}R=\begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta} \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ 

$$abla_{\scriptscriptstyle{0}}R = \left[ egin{array}{c} rac{\partial R}{\partial heta_{\scriptscriptstyle{0}}} \ rac{\partial R}{\partial heta_{\scriptscriptstyle{0}}} \end{array} 
ight] = \left[ egin{array}{c} 0 \ 0 \end{array} 
ight]$$

• Take partials of empirical risk:

$$\begin{split} R\left(\theta\right) &= \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right)^2 \\ \frac{\partial R}{\partial \theta_0} &= \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right) \left(-1\right) = 0 \\ \frac{\partial R}{\partial \theta_1} &= \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right) \left(-x_i\right) = 0 \\ \theta_0 &= \frac{1}{N} \sum y_i - \theta_1 \frac{1}{N} \sum x_i \end{split}$$

## Min by Gradient=0

•Gradient=0 means the partial derivatives are all 0 
$$\nabla_{\theta} R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_0} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

• Take partials of empirical risk:

$$\begin{split} R\left(\theta\right) &= \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right)^2 \\ \frac{\partial R}{\partial \theta_0} &= \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right) \left(-1\right) = 0 \\ \frac{\partial R}{\partial \theta_1} &= \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \theta_1 x_i - \theta_0\right) \left(-x_i\right) = 0 \\ \theta_0 &= \frac{1}{N} \sum y_i - \theta_1 \frac{1}{N} \sum x_i \\ \theta_1 \sum x_i^2 &= \sum y_i x_i - \theta_0 \sum x_i \end{split}$$

## Min by Gradient=0

•Gradient=0 means the partial derivatives are all 0 
$$\nabla_{\theta} R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_0} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

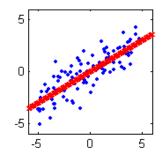
• Take partials of empirical risk:

$$\begin{split} R\left(\theta\right) &= \frac{1}{2N} \sum\nolimits_{i=1}^{N} \left(y_{i} - \theta_{1}x_{i} - \theta_{0}\right)^{2} \\ &\frac{\partial R}{\partial \theta_{0}} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \left(y_{i} - \theta_{1}x_{i} - \theta_{0}\right) \left(-1\right) = 0 \\ &\frac{\partial R}{\partial \theta_{1}} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \left(y_{i} - \theta_{1}x_{i} - \theta_{0}\right) \left(-x_{i}\right) = 0 \\ &\theta_{0} = \frac{1}{N} \sum y_{i} - \theta_{1} \frac{1}{N} \sum x_{i} \\ &\theta_{1} \sum x_{i}^{2} = \sum y_{i}x_{i} - \theta_{0} \sum x_{i} \\ &\theta_{1} = \frac{\sum y_{i}x_{i} - \frac{1}{N} \sum y_{i} \sum x_{i}}{\sum x_{i} \sum x_{i}} \end{split}$$

#### Properties of the Solution

- •Setting  $\theta^*$  as before gives least squared error
- Define error on each data point as:

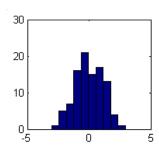
$$e_{i} = y_{i} - \theta_{_{1}}^{*} x_{_{i}} - \theta_{_{0}}^{*}$$



•Note property #1:

$$\frac{\partial R}{\partial \theta_0} = \frac{1}{N} \sum_{i=1}^{N} \left( y_i - \theta_1 x_i - \theta_0 \right) = 0$$

...average error is zero  $\frac{1}{N}\sum e_i=0$ 

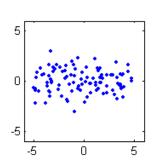


•Note property #2:

$$\frac{\partial R}{\partial \theta_1} = \frac{1}{N} \sum\nolimits_{i=1}^N \left( y_i - \theta_1 x_i - \theta_0 \right) x_i = 0$$

...error not correlated with data

$$\frac{1}{N}\sum e_i x_i = \frac{1}{N}e^T x = 0$$



- •More elegant/general to do  $\nabla_{\bf p} R = 0$  with linear algebra
- •Rewrite empirical risk in vecţor-matrix notation:

$$R(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \left( y_i - \theta_1 x_i - \theta_0 \right)^2$$

$$= \frac{1}{2N} \sum\nolimits_{i=1}^{N} \left[ y_i - \left[ \begin{array}{cc} 1 & x_i \end{array} \right] \left[ \begin{array}{c} \theta_0 \\ \theta_1 \end{array} \right] \right]^2$$

$$= \frac{1}{2N} \left\| \begin{array}{c} y_1 \\ \vdots \\ y_N \end{array} \right] - \left[ \begin{array}{cc} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{array} \right] \left[ \begin{array}{c} \theta_0 \\ \theta_1 \end{array} \right] \right\|^2$$

$$= rac{1}{2N} \left\| \mathbf{y} - \mathbf{X} \mathbf{\theta} \right\|^2$$

- •More elegant/general to do  $\nabla_{\bf p} R = 0$  with linear algebra
- •Rewrite empirical risk in vecţor-matrix notation:

$$R(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \left( y_i - \theta_1 x_i - \theta_0 \right)^2$$

$$=rac{1}{2N}{\sum}_{i=1}^{N}egin{bmatrix} y_i - \left[ egin{array}{cc} 1 & x_i \end{array} 
ight] \left[ egin{array}{cc} heta_0 \ heta_1 \end{array} 
ight] 
ight]^2$$

$$= \frac{1}{2N} \left[ \begin{array}{c} y_1 \\ \vdots \\ y_N \end{array} \right] - \left[ \begin{array}{cc} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{array} \right] \left[ \begin{array}{c} \theta_0 \\ \theta_1 \end{array} \right] \right]^2$$

$$= \frac{1}{2N} \left\| \mathbf{y} - \mathbf{X} \mathbf{\theta} \right\|^2$$

Can add more dimensions by adding columns to X matrix and rows to θ vector

- •More elegant/general to do  $\nabla_{_{\mathbf{n}}}R=0$  with linear algebra
- •Rewrite empirical risk in vector-matrix notation:

$$R(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \left( y_i - \theta_1 x_i - \theta_0 \right)^2$$

$$=rac{1}{2N}{\sum}_{i=1}^{N}egin{bmatrix} y_i - \left[ egin{array}{cc} 1 & x_i \end{array} 
ight] \left[ egin{array}{cc} heta_0 \ heta_1 \end{array} 
ight] ^2$$

$$=\frac{1}{2^N} \left[ \begin{array}{c} y_1 \\ \vdots \\ y_N \end{array} \right] - \left[ \begin{array}{ccc} 1 & x_1(1) & \dots & x_1(D) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N(1) & \dots & x_N(D) \end{array} \right] \left[ \begin{array}{c} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_D \end{array} \right] \left[ \begin{array}{c} \text{Can add more dimensions by adding columns to X matrix and} \end{array} \right]$$

$$=rac{1}{2N} \left\| \mathbf{y} - \mathbf{X} \mathbf{\theta} \right\|^2$$

rows to  $\theta$  vector

- More realistic dataset: many measurements
- •Have N apartments each with D measurements
- •Each row of X is [#rooms; latitude; longitude,...]

$$\mathbf{X} = \left[ \begin{array}{cccc} 1 & x_{_{\! 1}}\big(1\big) & \dots & x_{_{\! 1}}\big(D\big) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{_{\! N}}\big(1\big) & \dots & x_{_{\! N}}\big(D\big) \end{array} \right]$$



1	1212 Fifth Avenue PENTHOUSE	\$7,995,000
	Condo, Upper Carnegie Hill Listed by Nancy Packes Inc.	3 beds 3.5 baths 2,689 ft <sup>2</sup>
	210 East 73rd Street #PHB Co-op, Upper East Side Listed by Brown Harris Stevens	<b>\$3,495,000</b> 2 beds 3 baths
	66 East 11th Street Building, Greenwich Village Listed by Douglas Elliman	\$120,000,000
	150 West 56th Street #PH Condo, Midtown Listed by Douglas Elliman	<b>\$100,000,000</b> 6 beds 9 baths 8,000 ft <sup>2</sup>
	50 Central Park South #PH34/35 Condo, Central Park South Listed by Halstead Property	<b>\$95,000,000</b> 3 beds 3.5 baths
	15 Central Park West #35S Condo, Lincoln Square Listed by CORE	<b>\$95,000,000</b> 5 beds 5+ baths
	828 Fifth Avenue #XXX Co-op, Lenox Hill Listed by Stribling	<b>\$72,000,000</b> 8 beds 10.5 baths
	785 Fifth Avenue #PH1718 Co-op, Lenox Hill Listed by Corcoran	\$65,000,000 IN CONTRACT 7 beds 11 baths

•Solving gradient=0

$$\nabla_{\mathbf{p}}R=0$$

$$\nabla_{\theta} \left( \frac{1}{2N} \left\| \mathbf{y} - \mathbf{X} \theta \right\|^2 \right) = 0$$

•Solving gradient=0 
$$\nabla_{\boldsymbol{\theta}} R = 0$$
 
$$\nabla_{\boldsymbol{\theta}} \left( \frac{1}{2N} \left\| \mathbf{y} - \mathbf{X} \boldsymbol{\theta} \right\|^2 \right) = 0$$
 
$$\frac{1}{2N} \nabla_{\boldsymbol{\theta}} \left( \left( \mathbf{y} - \mathbf{X} \boldsymbol{\theta} \right)^T \left( \mathbf{y} - \mathbf{X} \boldsymbol{\theta} \right) \right) = 0$$

•Solving gradient=0 
$$\nabla_{\theta} R = 0$$
 
$$\nabla_{\theta} \left( \frac{1}{2N} \left\| \mathbf{y} - \mathbf{X} \theta \right\|^{2} \right) = 0$$
 
$$\frac{1}{2N} \nabla_{\theta} \left( \left( \mathbf{y} - \mathbf{X} \theta \right)^{T} \left( \mathbf{y} - \mathbf{X} \theta \right) \right) = 0$$
 
$$\frac{1}{2N} \nabla_{\theta} \left( \mathbf{y}^{T} \mathbf{y} - 2 \mathbf{y}^{T} \mathbf{X} \theta + \theta^{T} \mathbf{X}^{T} \mathbf{X} \theta \right) = 0$$

Solving gradient=0

$$\nabla_{\theta} R = 0$$

$$\nabla_{\theta} \left( \frac{1}{2N} \left\| \mathbf{y} - \mathbf{X} \theta \right\|^2 \right) = 0$$

$$\frac{1}{2N} \nabla_{\theta} \left( \left( \mathbf{y} - \mathbf{X} \theta \right)^{T} \left( \mathbf{y} - \mathbf{X} \theta \right) \right) = 0$$

$$\frac{1}{2N} \nabla_{\theta} \left( \mathbf{y}^{T} \mathbf{y} - 2 \mathbf{y}^{T} \mathbf{X} \theta + \theta^{T} \mathbf{X}^{T} \mathbf{X} \theta \right) = 0$$

$$\frac{1}{2N} \left( -2\mathbf{y}^T \mathbf{X} + 2\theta^T \mathbf{X}^T \mathbf{X} \right) = 0$$

$$\frac{\partial \vec{u}^T \vec{\theta}}{\partial \vec{\theta}} = \vec{u}^T$$

$$\frac{\partial \vec{\theta}^T \vec{\theta}}{\partial \vec{\theta}} = 2 \vec{\theta}^T$$

$$\frac{\partial \vec{\theta}^T A \vec{\theta}}{\partial \vec{\theta}} = \vec{\theta}^T \left( A + A^T \right)$$

•Solving gradient=0

$$\nabla_{\theta} R = 0$$

$$\nabla_{\theta} \left( \frac{1}{2N} \left\| \mathbf{y} - \mathbf{X} \theta \right\|^{2} \right) = 0$$

$$\frac{1}{2N} \nabla_{\theta} \left( \left( \mathbf{y} - \mathbf{X} \theta \right)^{T} \left( \mathbf{y} - \mathbf{X} \theta \right) \right) = 0$$

$$\frac{1}{2N} \nabla_{\theta} \left( \mathbf{y}^{T} \mathbf{y} - 2 \mathbf{y}^{T} \mathbf{X} \theta + \theta^{T} \mathbf{X}^{T} \mathbf{X} \theta \right) = 0$$

$$\frac{1}{2N} \left( -2 \mathbf{y}^{T} \mathbf{X} + 2 \theta^{T} \mathbf{X}^{T} \mathbf{X} \right) = 0$$

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^T\mathbf{y}$$

$$\frac{\partial \vec{u}^T \vec{\theta}}{\partial \vec{\theta}} = \vec{u}^T$$

$$\frac{\partial \vec{\theta}^T \vec{\theta}}{\partial \vec{\theta}} = 2 \vec{\theta}^T$$

$$\frac{\partial \vec{\theta}^T A \vec{\theta}}{\partial \vec{\theta}} = \vec{\theta}^T \left( A + A^T \right)$$

Solving gradient=0

$$\nabla_{_{\boldsymbol{\theta}}} R = 0$$

$$\nabla_{\theta} \left( \frac{1}{2N} \left\| \mathbf{y} - \mathbf{X} \theta \right\|^2 \right) = 0$$

$$\frac{1}{2N} \nabla_{\theta} \left( \left( \mathbf{y} - \mathbf{X} \theta \right)^{T} \left( \mathbf{y} - \mathbf{X} \theta \right) \right) = 0$$

$$\frac{1}{2N}\nabla_{\theta}\left(\mathbf{y}^{T}\mathbf{y}-2\mathbf{y}^{T}\mathbf{X}\mathbf{\theta}+\mathbf{\theta}^{T}\mathbf{X}^{T}\mathbf{X}\mathbf{\theta}\right)=0$$

$$\frac{1}{2N} \left( -2\mathbf{y}^T \mathbf{X} + 2\theta^T \mathbf{X}^T \mathbf{X} \right) = 0$$

$$\mathbf{X}^T \mathbf{X} \mathbf{\theta} = \mathbf{X}^T \mathbf{y}$$

$$\theta^* = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

•In Matlab: "t=pinv(X)\*y'' or " $t=X\setminus y''$  or "t=inv(X'\*X)\*X'\*y''

$$rac{\partial ec{u}^T ec{ heta}}{\partial ec{ heta}} = ec{u}^T$$

$$rac{\partial ec{ heta}^T ec{ heta}}{\partial ec{ heta}} = 2 ec{ heta}^T$$

$$\frac{\partial \vec{\theta}^T A \vec{\theta}}{\partial \vec{\theta}} = \vec{\theta}^T \left( A + A^T \right)$$

Solving gradient=0

$$\mathbf{X}^{T}\mathbf{X}\mathbf{\theta} = \mathbf{X}^{T}\mathbf{y}$$
 $\mathbf{\theta}^{*} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{y}$ 

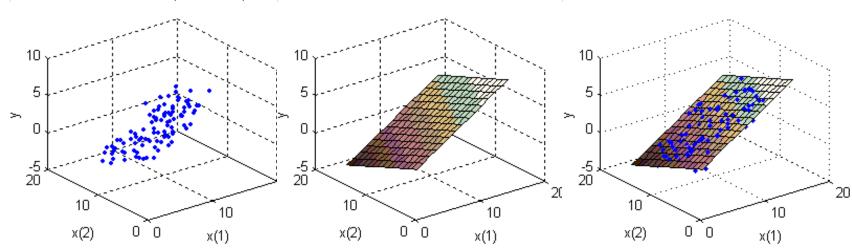
- •In Matlab: "t=pinv(X)\*y'' or " $t=X\setminus y''$  or "t=inv(X'\*X)\*X'\*y''
- •If the matrix X is skinny, the solution is probably unique
- •If X is fat (more dimensions than points) we get multiple solutions for theta which give zero error.
- •The pseudeoinverse (pinv(X)) returns the theta with zero error and which has the smallest norm.

$$\min_{\theta} \|\theta\|^2 \quad such \quad that \quad \mathbf{X}\theta = \mathbf{y}$$

#### 2D Linear Regression

•Once best  $\theta^*$  is found, we can plug it into the function:

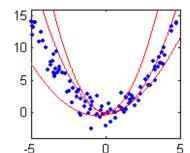
$$f(x; \theta^*) = \theta_2^* x(2) + \theta_1^* x(1) + \theta_0^*$$



•What would a fat X look like?

#### Polynomial Function Classes

- Back to 1-dim x (D=1) BUT Nonlinear
- •Polynomial:  $f(x;\theta) = \sum_{p=1}^{P} \theta_p x^p + \theta_0$

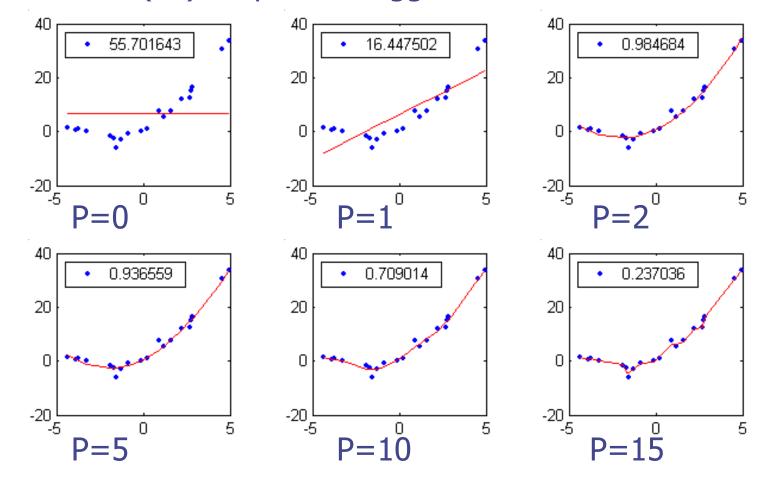


•Writing Risk: 
$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 & x_1^1 & \dots & x_1^P \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N^1 & \dots & x_N^P \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_P \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_P \end{bmatrix}$$

- •Order-P polynomial regression fitting for 1D variable is same as P-dimensional linear regression!
- •Construct a multidim  $\mathbf{x}_i = \left[ \begin{array}{ccc} x_i^0 & x_i^1 & x_i^2 & x_i^3 \end{array} \right]^T$  x-vector from x scalar
- $\bullet \text{More generally any} \quad \mathbf{x}_{i} = \left[ \begin{array}{ccc} \varphi_{0} \left( x_{i} \right) & \varphi_{1} \left( x_{i} \right) & \varphi_{2} \left( x_{i} \right) & \varphi_{3} \left( x_{i} \right) \end{array} \right]^{T}$

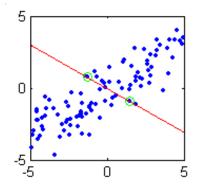
#### Underfitting/Overfitting

- •Try varying P. Higher P fits a more complex function class
- •Observe  $R(\theta^*)$  drops with bigger P



## **Evaluating The Regression**

- Unfair to use empirical to find best order P
- •High P (vs. N) can overfit, even linear case!
- •min  $R(\theta^*)$  not on training but on future data
- •Want model to Generalize to future data



True loss: 
$$R_{true}\left(\theta\right)=\int P\left(x,y\right)L\left(y,f\left(x;\theta\right)\right)dx\,dy$$

One approach: split data into training / testing portion

$$\left\{\!\left(\boldsymbol{x}_{\!\scriptscriptstyle 1},\boldsymbol{y}_{\!\scriptscriptstyle 1}\right)\!,\ldots,\!\left(\boldsymbol{x}_{\!\scriptscriptstyle N},\boldsymbol{y}_{\!\scriptscriptstyle N}\right)\!\right\}$$

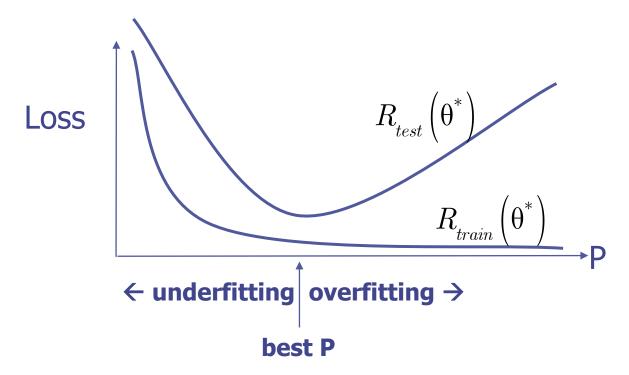
$$\left\{ \! \left( x_{\!_{1}}, y_{\!_{1}} \right), \ldots, \! \left( x_{\!_{N}}, y_{\!_{N}} \right) \! \right\} \qquad \qquad \left\{ \! \left( x_{\!_{N+1}}, y_{\!_{N+1}} \right), \ldots, \! \left( x_{\!_{N+M}}, y_{\!_{N+M}} \right) \! \right\}$$

•Estimate  $\theta^*$  with training loss:  $R_{train}\left(\theta\right) = \frac{1}{N}\sum_{i=1}^{N}L\left(y_i,f\left(x_i;\theta\right)\right)$ 

• Evaluate P with testing loss: 
$$R_{test}\left(\theta\right) = \frac{1}{M}\sum_{i=N+1}^{N+M}L\left(y_i,f\left(x_i;\theta\right)\right)$$

#### Crossvalidation

- Try fitting with different polynomial order P
- •Select P which gives lowest  $R_{test}(\theta^*)$



- Think of P as a measure of the complexity of the model
- Higher order polynomials are more flexible and complex