Machine Learning

4771

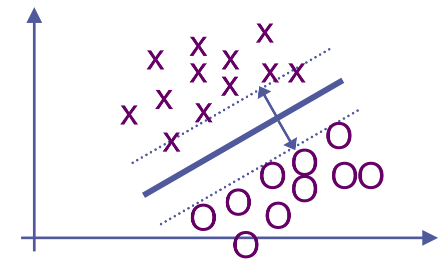
Instructor: Tony Jebara

Topic 2

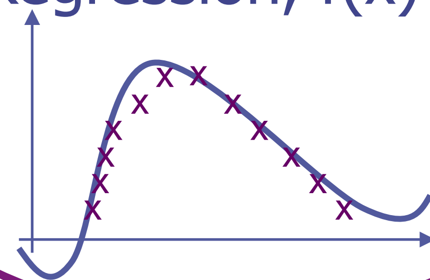
- Regression
- Empirical Risk Minimization
- Least Squares
- Higher Order Polynomials
- Under-fitting / Over-fitting
- Cross-Validation

Regression

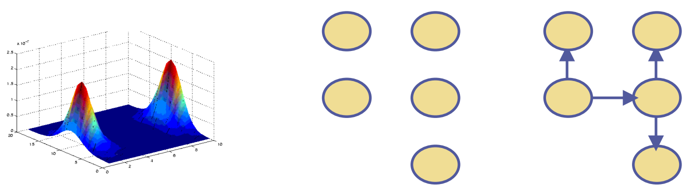
Classification



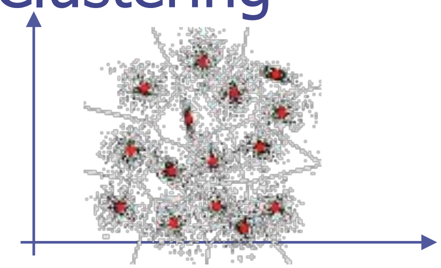
Regression, f(x)=y



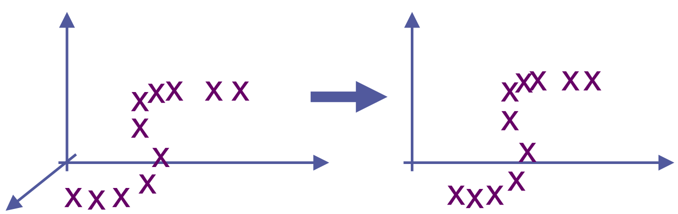
Density/Structure Estimation



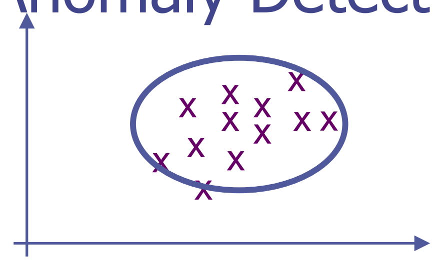
Clustering



Feature Selection



Anomaly Detection



Supervised

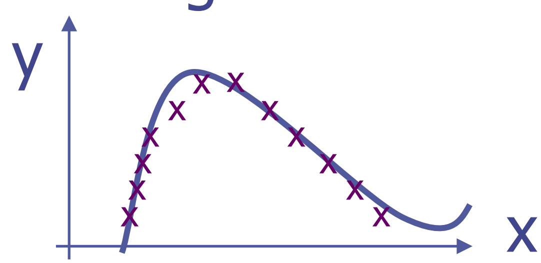
Unsupervised

Function Approximation

- Start with training dataset

$$\mathcal{X} = \left\{ \left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_N, y_N\right) \right\} \quad x \in \mathbb{R}^D = \begin{bmatrix} x(1) \\ x(2) \\ \ldots \\ x(D) \end{bmatrix} \quad y \in \mathbb{R}^1$$

- 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:

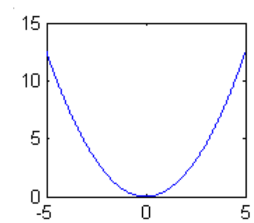
$$L\left(y, f\left(x\right)\right)$$

- Empirical Risk = the average loss over the dataset

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

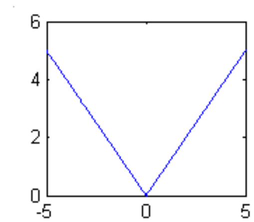
- Simplest loss: squared error from y value

$$L\left(y_i, f\left(x_i\right)\right) = \frac{1}{2}\left(y_i - f\left(x_i\right)\right)^2$$



- Other possible loss: absolute error

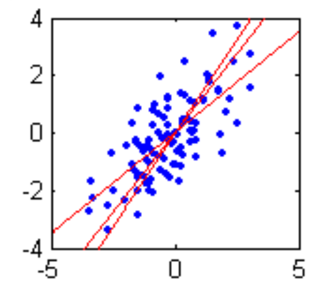
$$L\left(y_i, f\left(x_i\right)\right) = \left|y_i - f\left(x_i\right)\right|$$



Linear Function Classes

- Linear is simplest class of functions to search over:

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

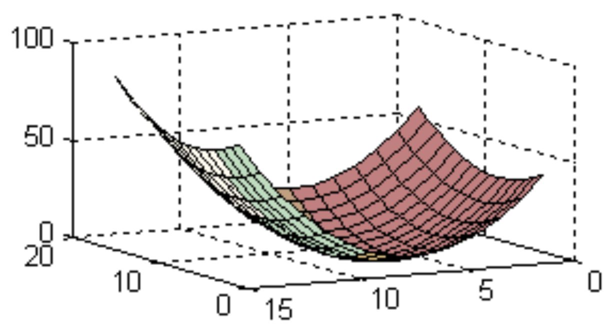


- Start with x being 1-dimensional (D=1):

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

- Plug in the above & minimize empirical risk over θ

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



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

Min by Gradient=0

- Gradient=0 means the partial derivatives are all 0

$$\nabla_\theta R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Take partials of empirical risk:

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

Min by Gradient=0

- Gradient=0 means the partial derivatives are all 0

$$\nabla_\theta R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Take partials of empirical risk:

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

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

Min by Gradient=0

- Gradient=0 means the partial derivatives are all 0

$$\nabla_\theta R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Take partials of empirical risk:

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

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

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

Min by Gradient=0

- Gradient=0 means the partial derivatives are all 0

$$\nabla_\theta R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Take partials of empirical risk:

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

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

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

$$\theta_0 = \frac{1}{N}\sum y_i - \theta_1 \frac{1}{N}\sum x_i$$

Min by Gradient=0

- Gradient=0 means the partial derivatives are all 0

$$\nabla_\theta R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Take partials of empirical risk:

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

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

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

$$\theta_0 = \tfrac{1}{N}\sum y_i - \theta_1 \tfrac{1}{N}\sum x_i$$

$$\theta_1 \sum x_i^2 = \sum y_i x_i - \theta_0 \sum x_i$$

Min by Gradient=0

- Gradient=0 means the partial derivatives are all 0

$$\nabla_{\theta} R = \begin{bmatrix} \frac{\partial R}{\partial \theta_0} \\ \frac{\partial R}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Take partials of empirical risk:

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

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

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

$$\theta_0 = \frac{1}{N}\sum y_i - \theta_1 \frac{1}{N}\sum x_i$$

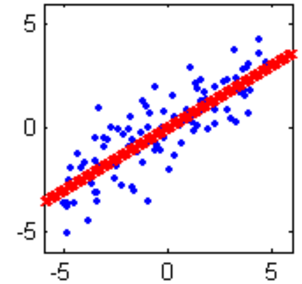
$$\theta_1 \sum x_i^2 = \sum y_i x_i - \theta_0 \sum x_i$$

$$\theta_1 = \frac{\sum y_i x_i - \frac{1}{N}\sum y_i \sum x_i}{\sum x_i^2 - \frac{1}{N}\sum x_i \sum x_i}$$

Properties of the Solution

- Setting θ^* as before gives least squared error
- Define error on each data point as:

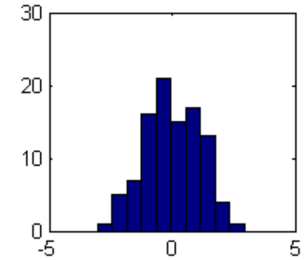
$$e_i = y_i - \theta_1^* x_i - \theta_0^*$$



- 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 $\frac{1}{N}\sum e_i = 0$

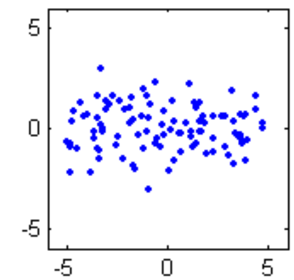


- Note property #2:

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

...error not correlated with data

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



Multi-Dimensional Regression

- More elegant/general to do $\nabla_\theta R = 0$ with linear algebra
- Rewrite empirical risk in vector-matrix notation:

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

$$= \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \begin{bmatrix} 1 & x_i \end{bmatrix}\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}\right)^2$$

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

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

Multi-Dimensional Regression

- More elegant/general to do $\nabla_\theta R = 0$ with linear algebra
- Rewrite empirical risk in vector-matrix notation:

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

$$= \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \begin{bmatrix} 1 & x_i \end{bmatrix}\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}\right)^2$$

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

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

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

Multi-Dimensional Regression

- More elegant/general to do $\nabla_\theta R = 0$ with linear algebra
- Rewrite empirical risk in vector-matrix notation:

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

$$= \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \begin{bmatrix} 1 & x_i \end{bmatrix}\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}\right)^2$$

$$= \frac{1}{2N}\left\|\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 & x_1(1) & \dots & x_1(D) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N(1) & \dots & x_N(D) \end{bmatrix}\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_D \end{bmatrix}\right\|^2$$

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

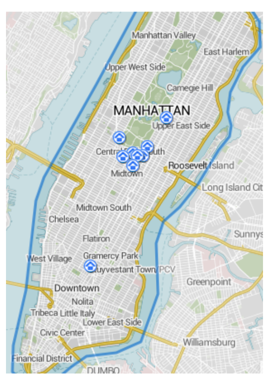
$$= \frac{1}{2N}\left\|\mathbf{y} - \mathbf{X}\theta\right\|^2$$

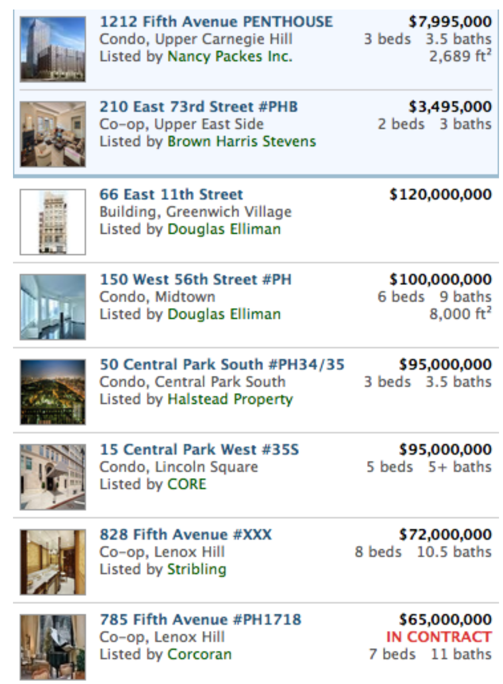
Tony Jebara, Columbia University

Multi-Dimensional Regression

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

$$\mathbf{X} = \begin{bmatrix} 1 & x_1(1) & \ldots & x_1(D) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N(1) & \ldots & x_N(D) \end{bmatrix}$$





Multi-Dimensional Regression

- Solving gradient=0 $\nabla_\theta R = 0$

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

Multi-Dimensional Regression

- Solving gradient=0

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

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

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

Multi-Dimensional Regression

- Solving gradient=0

$$\nabla_\theta R = 0$$

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

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

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

Multi-Dimensional Regression

- Solving gradient=0

$$\nabla_\theta R = 0$$

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

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

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

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

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

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

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

Multi-Dimensional Regression

- Solving gradient=0

$$\nabla_\theta R = 0$$

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

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

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

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

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

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

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

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

Multi-Dimensional Regression

- Solving gradient=0

$$\nabla_\theta R = 0$$

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

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

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

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

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

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

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

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

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

- In Matlab: “t=pinv(X)*y” or “t=X\y” or “t=inv(X’*X)*X’*y”

Multi-Dimensional Regression

- Solving gradient=0

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

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

- In Matlab: "t=pinv(X)*y" or "t=X\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.

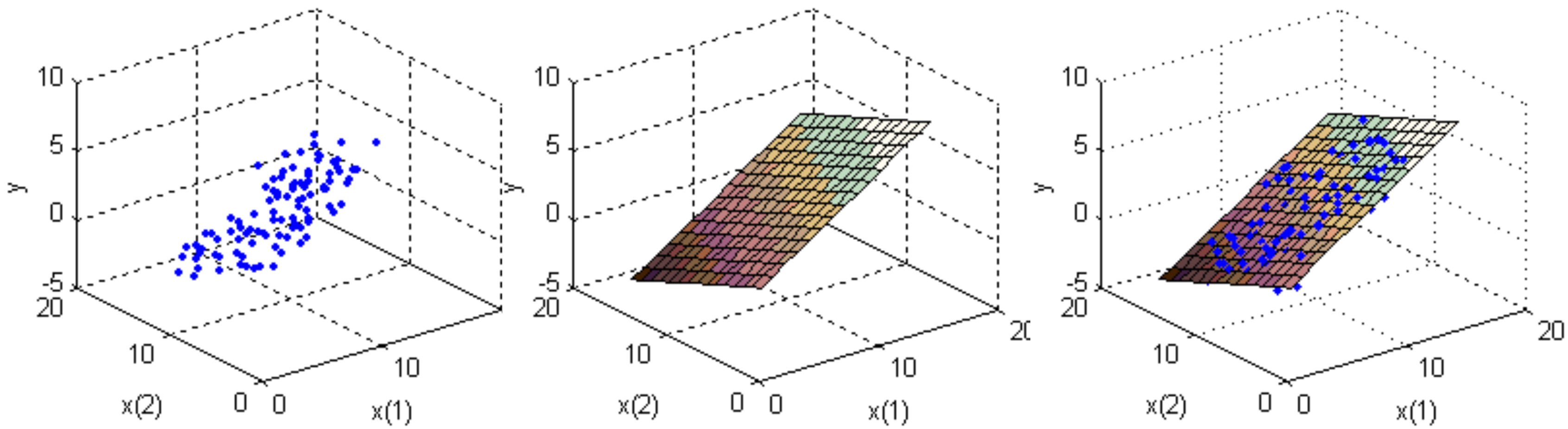
- The pseudeoinverse (pinv(X)) returns the theta with zero error and which has the smallest norm.

$$\min_\theta \left\|\theta\right\|^2 \; such\ that\ \mathbf{X}\theta = \mathbf{y}$$

2D Linear Regression

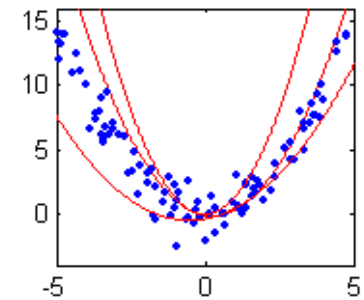
- Once best θ* is found, we can plug it into the function:

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



- What would a fat X look like?

Polynomial Function Classes



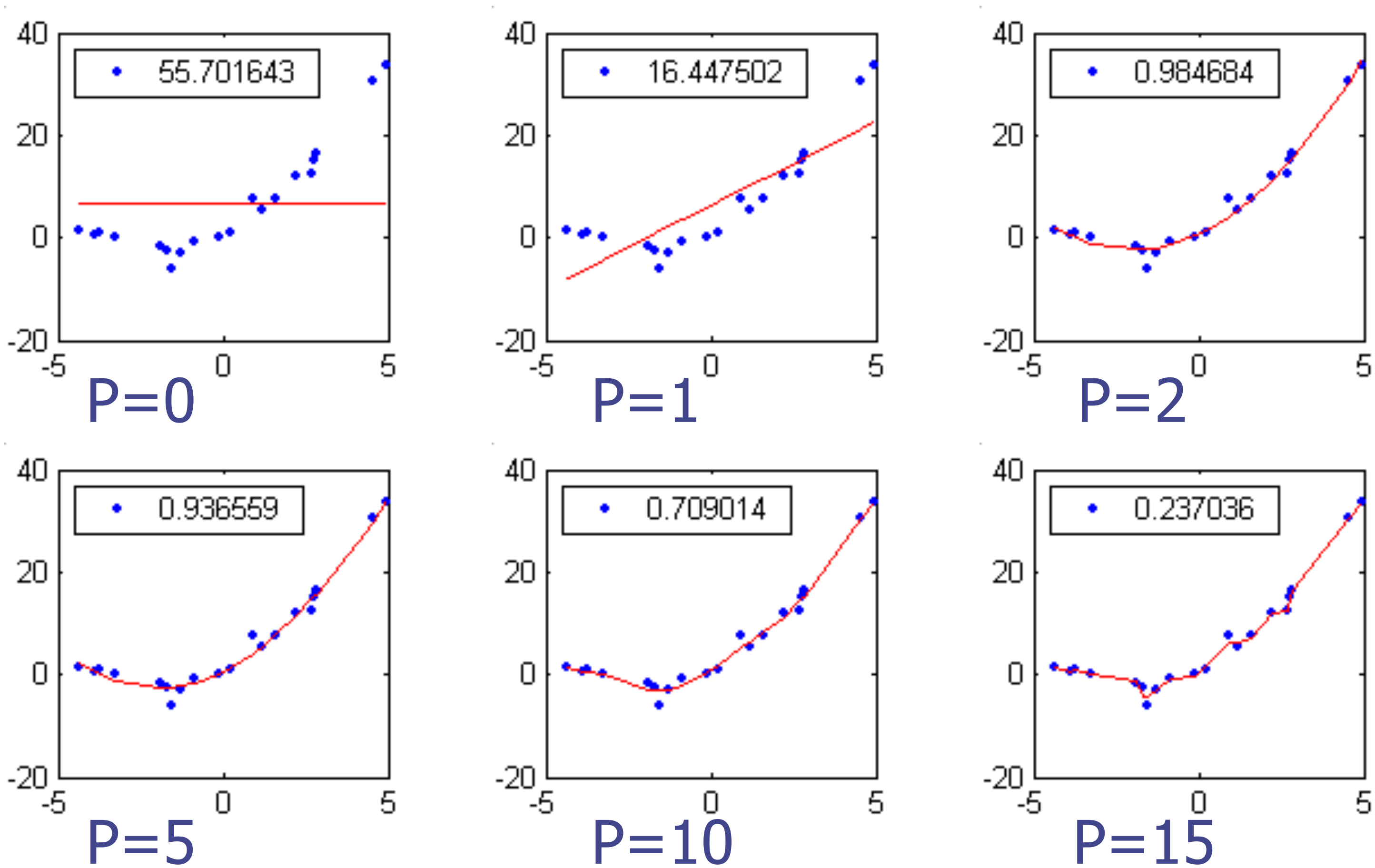
- Back to 1-dim x (D=1) BUT Nonlinear
- Polynomial: $f\left(x;\theta\right) = \sum_{p=1}^{P} \theta_p x^p + \theta_0$
- Writing Risk:

$$R\left(\theta\right) = \frac{1}{2N}\left\| \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 & x_1^1 & \dots & x_1^P \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N^1 & \dots & x_N^P \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_P \end{bmatrix} \right\|^2$$

- Order-P polynomial regression fitting for 1D variable is same as P-dimensional linear regression!
- Construct a multidim x-vector from x scalar $\mathbf{x}_i = \begin{bmatrix} x_i^0 & x_i^1 & x_i^2 & x_i^3 \end{bmatrix}^T$
- More generally any $\mathbf{x}_i = \begin{bmatrix} \phi_0\left(x_i\right) & \phi_1\left(x_i\right) & \phi_2\left(x_i\right) & \phi_3\left(x_i\right) \end{bmatrix}^T$

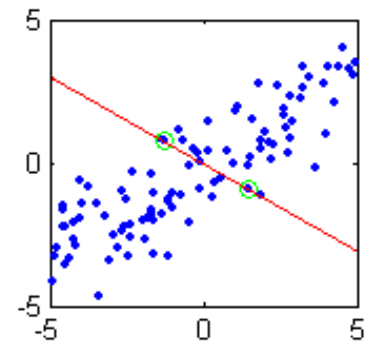
Underfitting/Overfitting

- Try varying P. Higher P fits a more complex function class
- Observe R(θ*) 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(θ*) not on training but on future data
- Want model to *Generalize* to future data



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

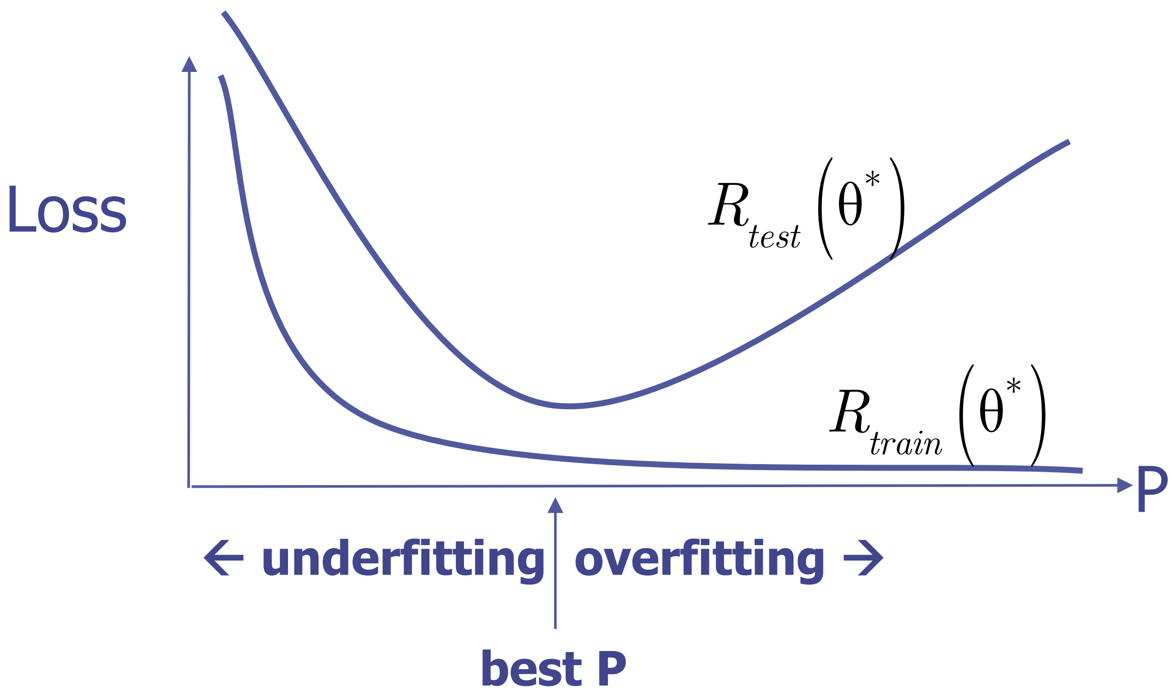
- One approach: split data into training / testing portion

$$\left\{\left(x_1,y_1\right),\ldots,\left(x_N,y_N\right)\right\} \qquad \left\{\left(x_{N+1},y_{N+1}\right),\ldots,\left(x_{N+M},y_{N+M}\right)\right\}$$

- Estimate θ* 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

Machine Learning

4771

Instructor: Tony Jebara

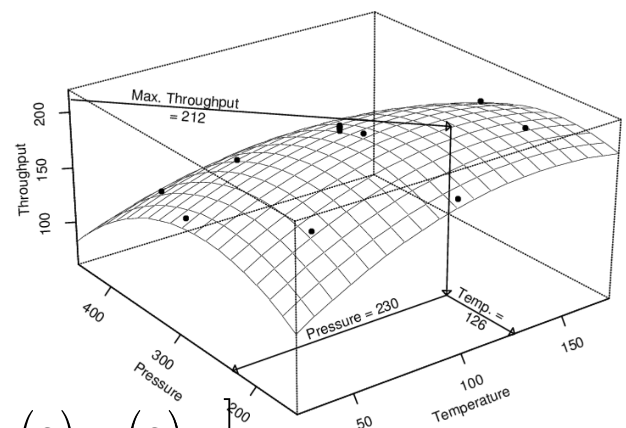
Topic 3

- Additive Models and Linear Regression
- Sinusoids and Radial Basis Functions
- Classification
- Logistic Regression
- Gradient Descent

Polynomial Basis Functions

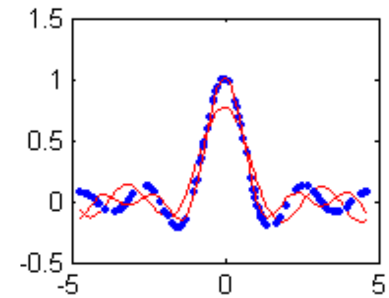
- To fit a P'th order polynomial function to multivariate data: concatenate columns of all monomials up to power P
- E.g. 2 dimensional data and 2nd order polynomial (quadratic)

$$\begin{bmatrix} x_1(1) & x_1(2) \\ \vdots & \vdots \\ x_i(1) & x_i(2) \\ \vdots & \vdots \\ x_N(1) & x_N(2) \end{bmatrix}$$



$$\begin{bmatrix} 1 & x_1(1) & x_1(2) & x_1(1)x_1(1) & x_1(1)x_1(2) & x_1(2)x_1(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_i(1) & x_i(2) & x_i(1)x_i(1) & x_i(1)x_i(2) & x_i(2)x_i(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N(1) & x_N(2) & x_N(1)x_N(1) & x_N(1)x_N(2) & x_N(2)x_N(2) \end{bmatrix}$$

Sinusoidal Basis Functions



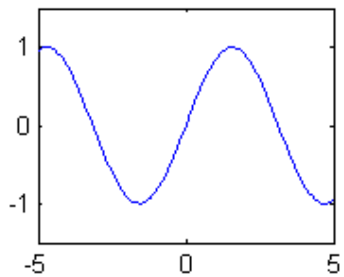
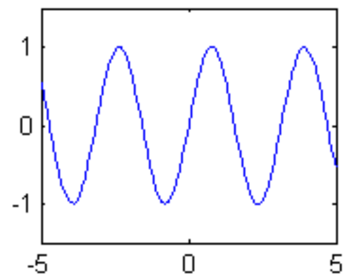
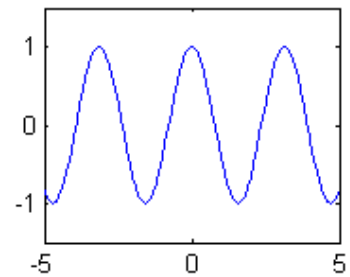
- More generally, we don't just have to deal with polynomials, use any set of basis fn's:

$$f\left(x;\theta\right) = \sum_{p=1}^{P} \theta_p \phi_p\left(x\right) + \theta_0$$

- These are generally called Additive Models
- Regression adds linear combinations of the basis fn's
- For example: Fourier (sinusoidal) basis

$$\phi_{2k}\left(x_i\right) = \sin\left(kx_i\right) \qquad \phi_{2k+1}\left(x_i\right) = \cos\left(kx_i\right)$$

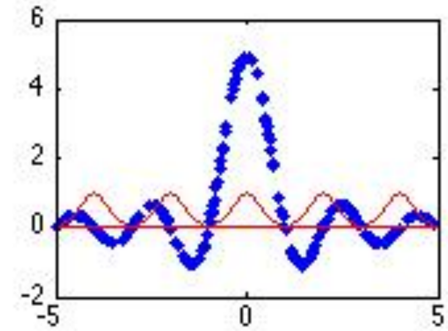
- Note, don't have to be a basis per se, usually subset

$\theta_1 \times$  $+\theta_2 \times$  $+\theta_3 \times$ 

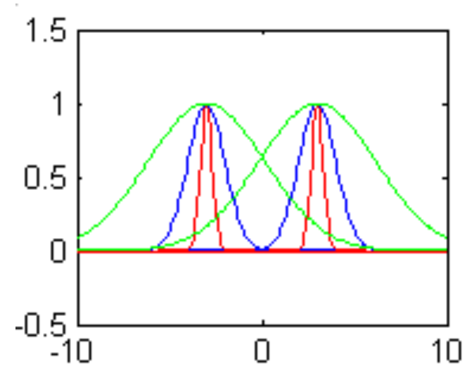
Radial Basis Functions

•Can act as prototypes of the data itself

$$f(\mathbf{x};\theta) = \sum_{k=1}^{N} \theta_k \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{x}-\mathbf{x}_k\right\|^2\right)$$

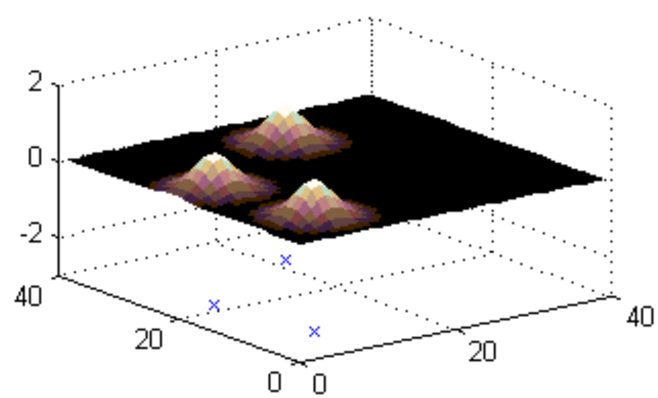


•Parameter σ = standard deviation
σ^2 = covariance



controls how wide bumps are
what happens if too big/small?

•Also works in multi-dimensions



•Called RBF for short

Radial Basis Functions

- Each training point leads to a bump function

$$f(\mathbf{x};\theta) = \sum_{k=1}^{N} \theta_k \exp\left(-\tfrac{1}{2\sigma^2}\left\|\mathbf{x}-\mathbf{x}_k\right\|^2\right)$$

- Reuse solution from linear regression: $\theta^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$
- Can view the data instead as X, a big matrix of size N x N

$$\mathbf{X} = \begin{bmatrix} \exp\left(-\tfrac{1}{2\sigma^2}\left\|\mathbf{x}_1-\mathbf{x}_1\right\|^2\right) & \exp\left(-\tfrac{1}{2\sigma^2}\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2\right) & \exp\left(-\tfrac{1}{2\sigma^2}\left\|\mathbf{x}_1-\mathbf{x}_3\right\|^2\right) \\ \exp\left(-\tfrac{1}{2\sigma^2}\left\|\mathbf{x}_2-\mathbf{x}_1\right\|^2\right) & \exp\left(-\tfrac{1}{2\sigma^2}\left\|\mathbf{x}_2-\mathbf{x}_2\right\|^2\right) & \exp\left(-\tfrac{1}{2\sigma^2}\left\|\mathbf{x}_2-\mathbf{x}_3\right\|^2\right) \\ \exp\left(-\tfrac{1}{2\sigma^2}\left\|\mathbf{x}_3-\mathbf{x}_1\right\|^2\right) & \exp\left(-\tfrac{1}{2\sigma^2}\left\|\mathbf{x}_3-\mathbf{x}_2\right\|^2\right) & \exp\left(-\tfrac{1}{2\sigma^2}\left\|\mathbf{x}_3-\mathbf{x}_3\right\|^2\right) \end{bmatrix}$$

- For RBFs, X is square and symmetric, so solution is just

$$\nabla_\theta R = 0 \rightarrow \mathbf{X}^T\mathbf{X}\theta = \mathbf{X}^T\mathbf{y} \rightarrow \mathbf{X}\theta = \mathbf{y} \rightarrow \theta^* = \mathbf{X}^{-1}\mathbf{y}$$

Evaluating Our Learned Function

- We minimized empirical risk to get θ^*
- How well does $f(x;\theta^*)$ perform on future data?
- It should *Generalize* and have low True Risk:

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

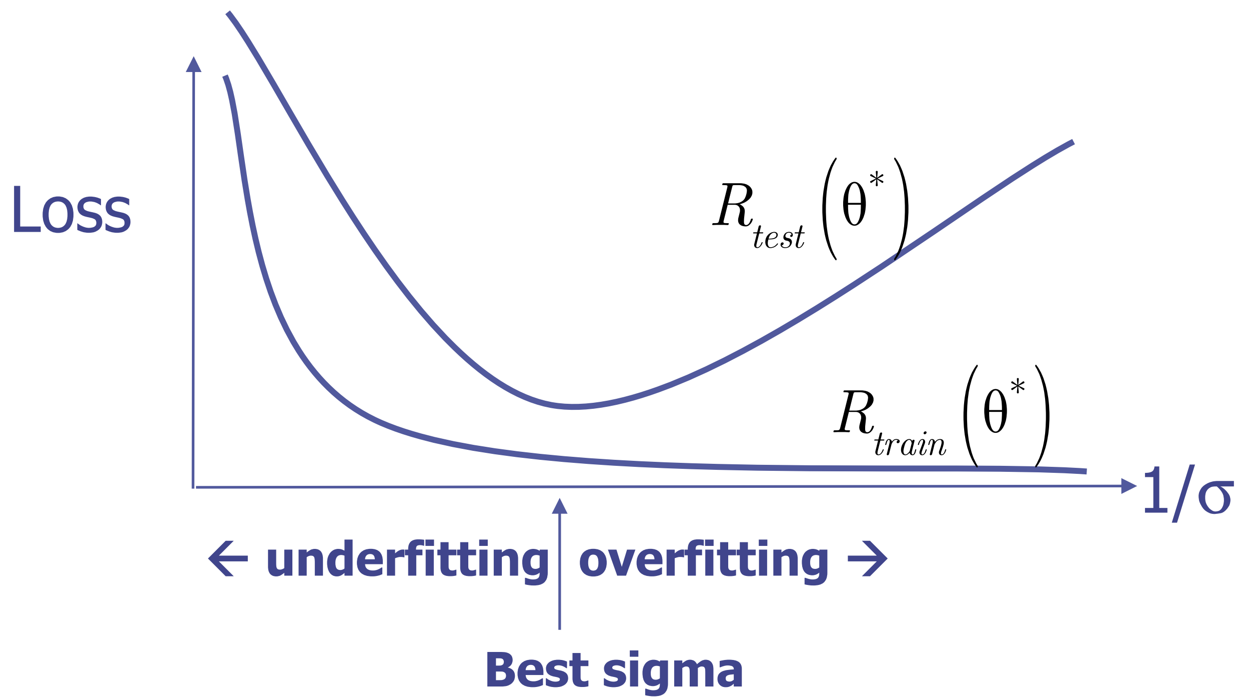
- Can't compute true risk, instead use Testing Empirical Risk
- We randomly split data into training and testing portions

$$\{(x_1,y_1),\ldots,(x_N,y_N)\} \qquad \{(x_{N+1},y_{N+1}),\ldots,(x_{N+M},y_{N+M})\}$$

- Find θ^* with training data: $R_{train}(\theta) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i;\theta))$
- Evaluate it with testing data: $R_{test}(\theta) = \frac{1}{M}\sum_{i=N+1}^{N+M} L(y_i, f(x_i;\theta))$

Crossvalidation

- Try fitting with different sigma radial basis function widths
- Select sigma which gives lowest $R_{test}(\theta^*)$



- Think of sigma as a measure of the simplicity of the model
- Thinner RBFs are more flexible and complex

Regularized Risk Minimization

- Empirical Risk Minimization gave overfitting & underfitting
- We want to add a penalty for using too many theta values
- This gives us the Regularized Risk

$$R_{regularized}\left(\theta\right) = R_{empirical}\left(\theta\right) + Penalty\left(\theta\right)$$
$$= \tfrac{1}{N}\sum_{i=1}^{N} L\left(y_i, f\left(x_i;\theta\right)\right) + \tfrac{\lambda}{2N}\left\|\theta\right\|^2$$

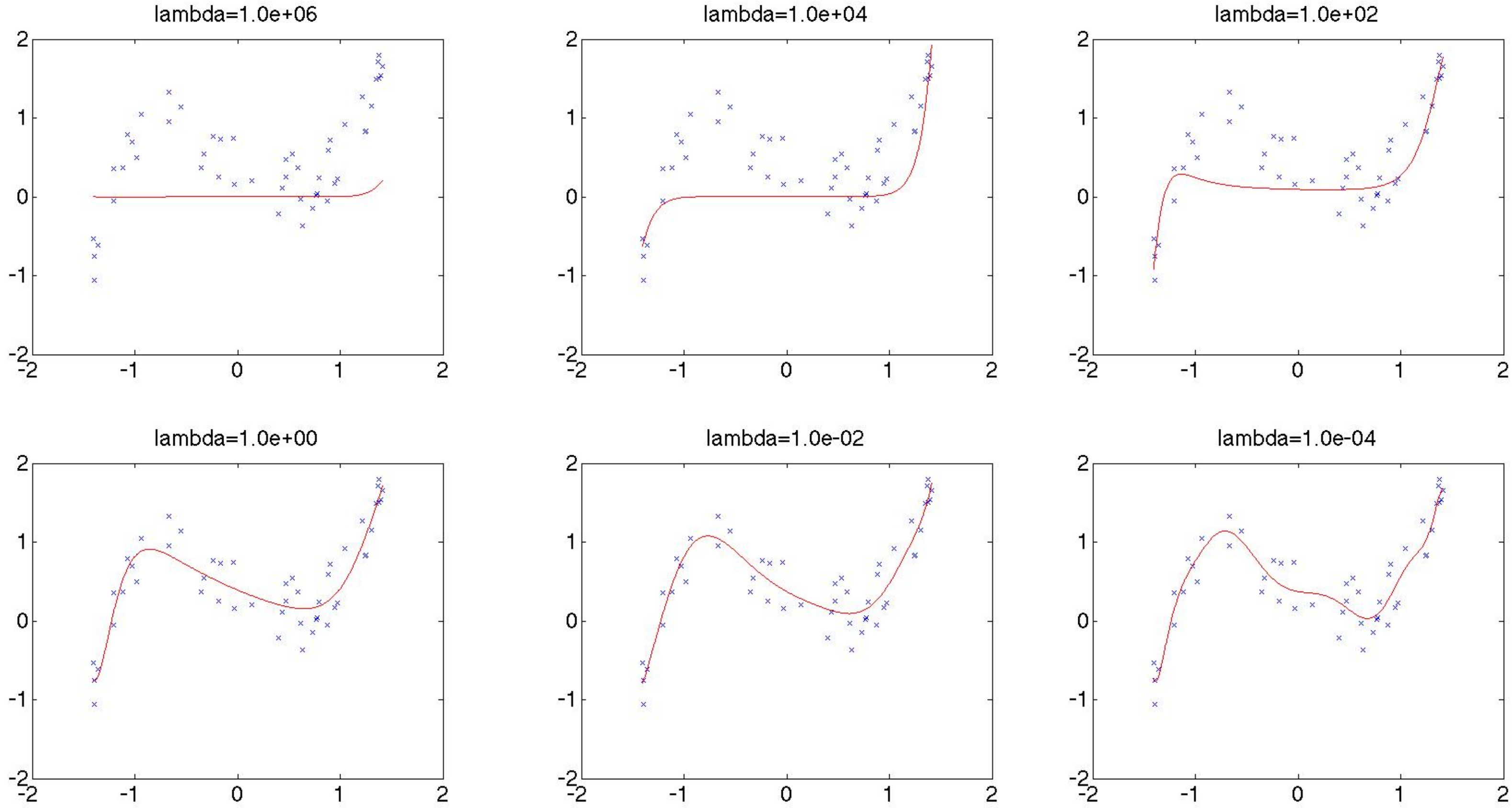
- Solution for Regularized Risk with Least Squares Loss:

$$\nabla_\theta R_{regularized} = 0 \quad \Rightarrow \quad \nabla_\theta\left(\tfrac{1}{2N}\left\|\mathbf{y} - \mathbf{X}\theta\right\|^2 + \tfrac{\lambda}{2N}\left\|\theta\right\|^2\right) = 0$$

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

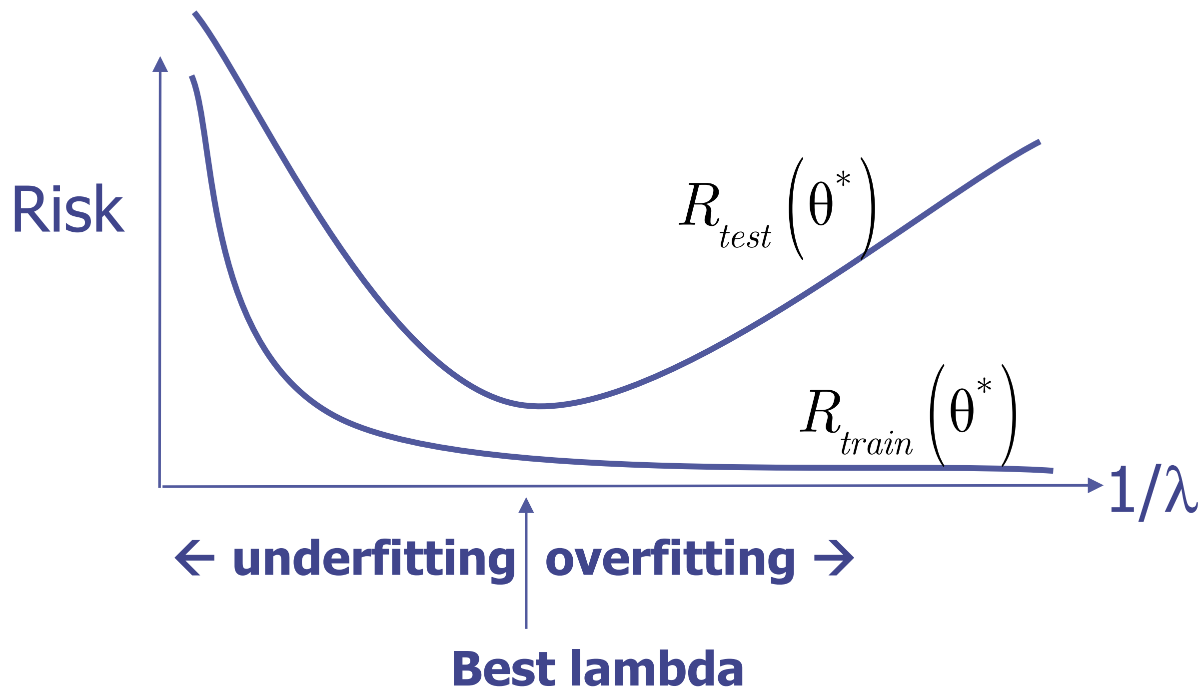
Regularized Risk Minimization

- Have D=16 features (or P=15 throughout)
- Try minimizing $R_{regularized}(\theta)$ to get θ^* with different λ
- Note that λ=0 give back Empirical Risk Minimization



Crossvalidation

- Try fitting with different lambda regularization levels
- Select lambda which gives lowest $R_{test}(\theta^*)$



- Lambda measures simplicity of the model
- Models with low lambda are more flexible

From Regression To Classification

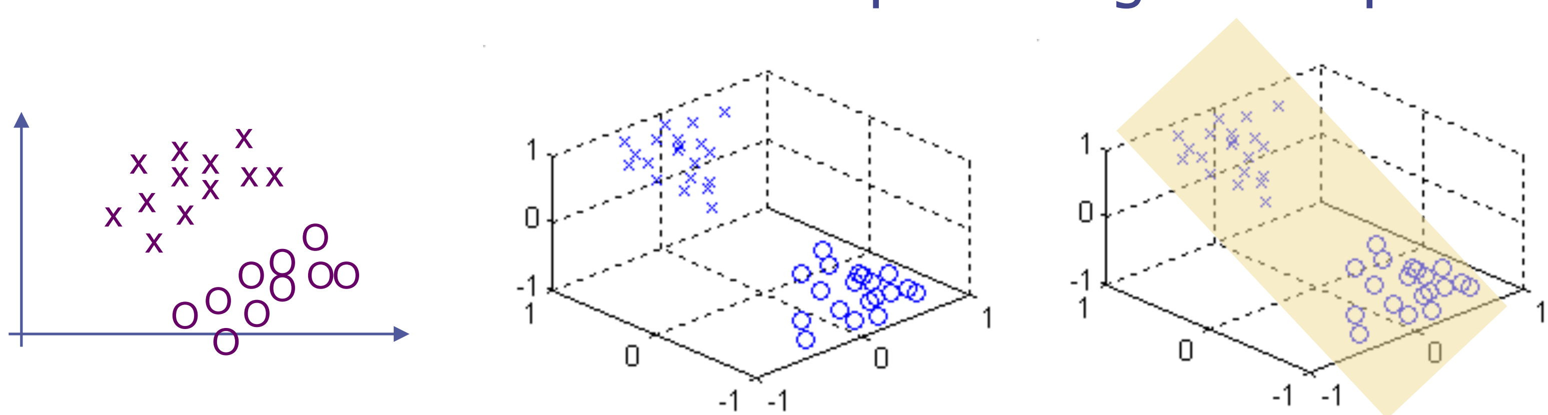
- Classification is another important learning problem

Regression $\mathcal{X} = \left\{ \left(\mathbf{x}_1, y_1\right), \left(\mathbf{x}_2, y_2\right), \ldots, \left(\mathbf{x}_N, y_N\right) \right\} \quad \mathbf{x} \in \mathbb{R}^D \quad y \in \mathbb{R}^1$

Classification $\mathcal{X} = \left\{ \left(\mathbf{x}_1, y_1\right), \left(\mathbf{x}_2, y_2\right), \ldots, \left(\mathbf{x}_N, y_N\right) \right\} \quad \mathbf{x} \in \mathbb{R}^D \quad y \in \left\{0, 1\right\}$

- E.g. Given x = [tumor size, tumor density]
  Predict y in {benign,malignant}

- Should we solve this as a least squares regression problem?



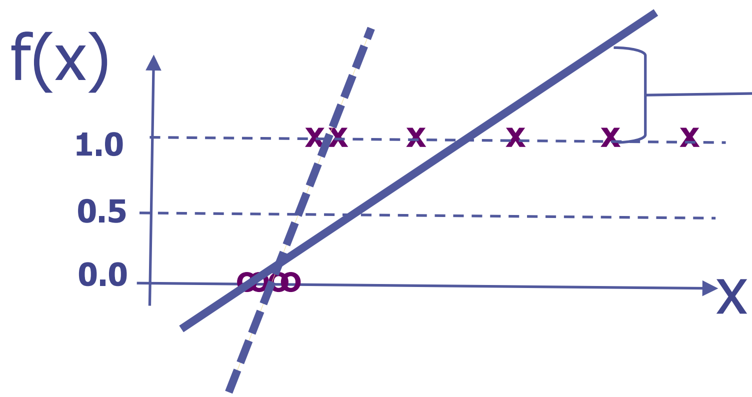
Classification vs. Regression

a) Classification needs binary answers like {0,1}
b) Least squares is an unfair measure of risk here
   e.g. Why penalize a correct but large positive y answer?
   e.g. Why penalize a correct but large negative y answer?

•Example: not good to use regression output for a decision

$f(x)>0.5 \rightarrow$ Class 1 $\qquad$ $f(x)<0.5 \rightarrow$ Class 0

if $f(x)=-3.8$ & correct class=0, squared error penalizes it…



We pay a hefty squared error loss here even if we got the correct classification result. The thick solid line model makes two mistakes while the dashed model is perfect

Classification vs. Regression

We will consider the following four steps to improve from naïve regression to get better classification learning:

1) Fix functions f(x) to give binary output (logistic neuron)
2) Fix our definition of the Risk we will minimize so that we get good classification accuracy (logistic loss)

…and later on…

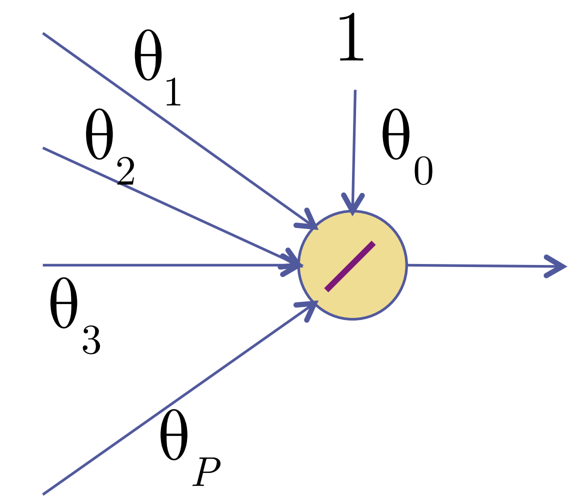
3) Make an even better fix on f(x) to binarize (perceptron)
4) Make an even better risk (perceptron loss)

Logistic Neuron (McCullough-Pitts)

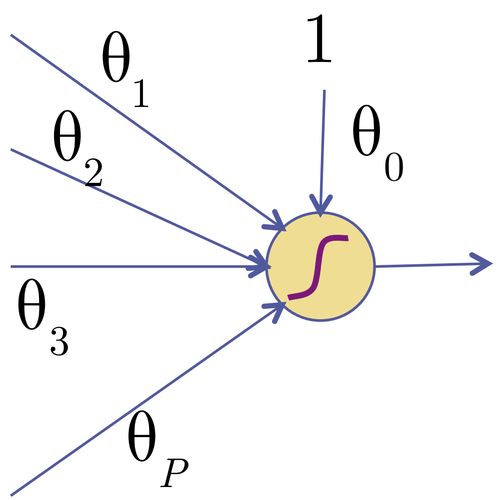
- To output binary, use squashing function g().

$$f\left(\mathbf{x};\theta\right) = \theta^T \mathbf{x}$$

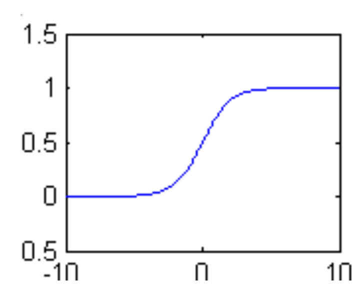
$$f\left(\mathbf{x};\theta\right) = g\left(\theta^T \mathbf{x}\right)$$
$$g\left(z\right) = \left(1 + \exp\left(-z\right)\right)^{-1}$$



Linear neuron



Logistic Neuron



- This squashing is called sigmoid or logistic function

Logistic Regression

- Given a classification problem with binary outputs

$$\mathcal{X} = \left\{ \left(\mathbf{x}_1, y_1\right), \left(\mathbf{x}_2, y_2\right), \ldots, \left(\mathbf{x}_N, y_N\right) \right\} \quad \mathbf{x} \in \mathbb{R}^D \quad y \in \left\{0,1\right\}$$

- Use this function and output 1 if f(x)>0.5 and 0 otherwise

$$f\left(\mathbf{x}; \theta\right) = \left(1 + \exp\left(-\theta^T \mathbf{x}\right)\right)^{-1}$$

Short hand for Linear Functions

- What happened to adding the intercept?

$$f\left(\mathbf{x};\theta\right) = \theta^T\mathbf{x} + \theta_0$$

$$= \begin{bmatrix} \theta(1) \\ \theta(2) \\ \vdots \\ \theta(D) \end{bmatrix}^T \begin{bmatrix} \mathbf{x}(1) \\ \mathbf{x}(2) \\ \vdots \\ \mathbf{x}(D) \end{bmatrix} + \theta_0 = \begin{bmatrix} \theta_0 \\ \theta(1) \\ \theta(2) \\ \vdots \\ \theta(D) \end{bmatrix}^T \begin{bmatrix} 1 \\ \mathbf{x}(1) \\ \mathbf{x}(2) \\ \vdots \\ \mathbf{x}(D) \end{bmatrix} = \vec{\theta}^T\vec{\mathbf{x}}$$

Logistic Regression

- Given a classification problem with binary outputs

$$\mathcal{X} = \left\{ \left(\mathbf{x}_1, y_1\right), \left(\mathbf{x}_2, y_2\right), \ldots, \left(\mathbf{x}_N, y_N\right) \right\} \quad \mathbf{x} \in \mathbb{R}^D \quad y \in \left\{0,1\right\}$$

- Fix#1: use f(x) below, output 1 if f(x)>0.5 and 0 otherwise

$$f\left(\mathbf{x};\theta\right) = \left(1 + \exp\left(-\theta^T \mathbf{x}\right)\right)^{-1}$$

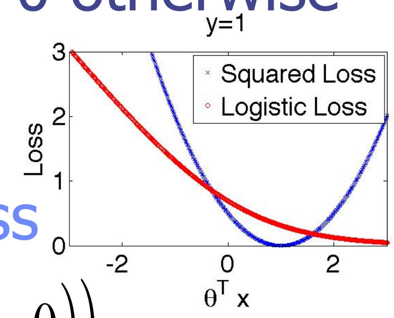
Logistic Regression

- Given a classification problem with binary outputs

$$\mathcal{X} = \left\{ \left(\mathbf{x}_1, y_1\right), \left(\mathbf{x}_2, y_2\right), \ldots, \left(\mathbf{x}_N, y_N\right) \right\} \quad \mathbf{x} \in \mathbb{R}^D \quad y \in \left\{0,1\right\}$$

- Fix#1: use f(x) below, output 1 if f(x)>0.5 and 0 otherwise

$$f\left(\mathbf{x};\theta\right) = \left(1 + \exp\left(-\theta^T \mathbf{x}\right)\right)^{-1}$$



- Fix#2: instead of squared loss, use Logistic Loss

$$L_{log}\left(y, f\left(\mathbf{x};\theta\right)\right) = \left(y_i - 1\right)\log\left(1 - f\left(\mathbf{x};\theta\right)\right) - y_i \log\left(f\left(\mathbf{x};\theta\right)\right)$$

- This method is called Logistic Regression.
- But Empirical Risk Minimization has no closed-form sol'n:

$$R_{emp}\left(\theta\right) = \tfrac{1}{N}\sum\nolimits_{i=1}^{N} \left(y_i - 1\right)\log\left(1 - f\left(\mathbf{x}_i;\theta\right)\right) - y_i \log\left(f\left(\mathbf{x}_i;\theta\right)\right)$$

Logistic Regression

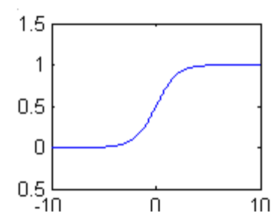
- With logistic squashing function, minimizing R(θ) is harder

$$R_{emp}(\theta) = \tfrac{1}{N}\sum_{i=1}^{N} (y_i - 1)\log\left(1 - f(\mathbf{x}_i;\theta)\right) - y_i \log\left(f(\mathbf{x}_i;\theta)\right)$$

$$\nabla_\theta R = \tfrac{1}{N}\sum_{i=1}^{N} \left(\frac{1-y_i}{1-f(\mathbf{x}_i;\theta)} - \frac{y_i}{f(\mathbf{x}_i;\theta)}\right) f'(\mathbf{x}_i;\theta) = 0 \quad ???$$

- Can't minimize risk and find best theta analytically!
- Let's try finding best theta numerically.
- Use the following to compute gradient

$$f(\mathbf{x};\theta) = \left(1 + \exp\left(-\theta^T\mathbf{x}\right)\right)^{-1} = g\left(\theta^T\mathbf{x}\right)$$

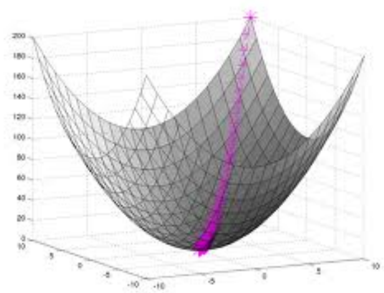


- Here, g() is the logistic squashing function

$$g(z) = \left(1 + \exp(-z)\right)^{-1} \quad g'(z) = g(z)\left(1 - g(z)\right)$$

Gradient Descent

- Useful when we can't get minimum solution in closed form
- Gradient points in direction of fastest increase
- Take step in the opposite direction!



- Gradient Descent Algorithm

choose scalar step size η, & tolerance ε
initialize θ^0 = small random vector

$$\theta^1 = \theta^0 - \eta \nabla_\theta R_{emp}\Big|_{\theta^0}, \quad t = 1$$

while $\left\| \theta^t - \theta^{t-1} \right\| \geq \in$ *{*

$$\theta^{t+1} = \theta^t - \eta \nabla_\theta R_{emp}\Big|_{\theta^t}, \quad t = t + 1$$

}

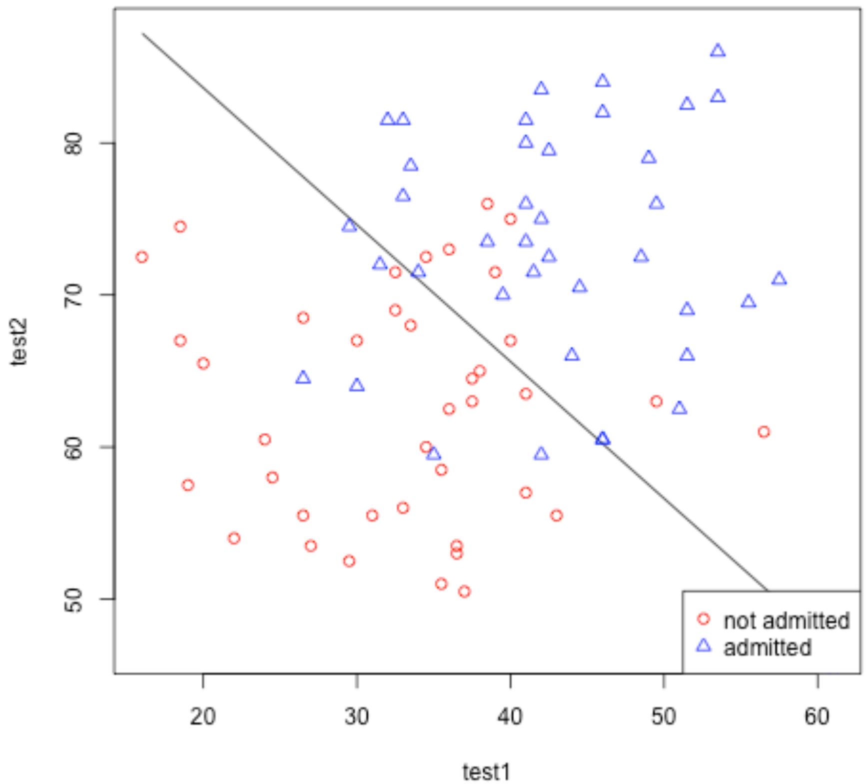
- For appropriate η, this will converge to local minimum

Logistic Regression

- Logistic regression gives better classification performance
- Its empirical risk is

$$R_{emp}\left(\theta\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - 1\right)\log\left(1 - f\left(\mathbf{x}_i;\theta\right)\right) - y_i \log\left(f\left(\mathbf{x}_i;\theta\right)\right)$$

- This R(θ) is convex so gradient descent always converges to the same solution



- Make predictions using

$$f\left(\mathbf{x};\theta\right) = \left(1 + \exp\left(-\theta^T \mathbf{x}\right)\right)^{-1}$$

- Output 1 if f > 0.5
- Output 0 otherwise