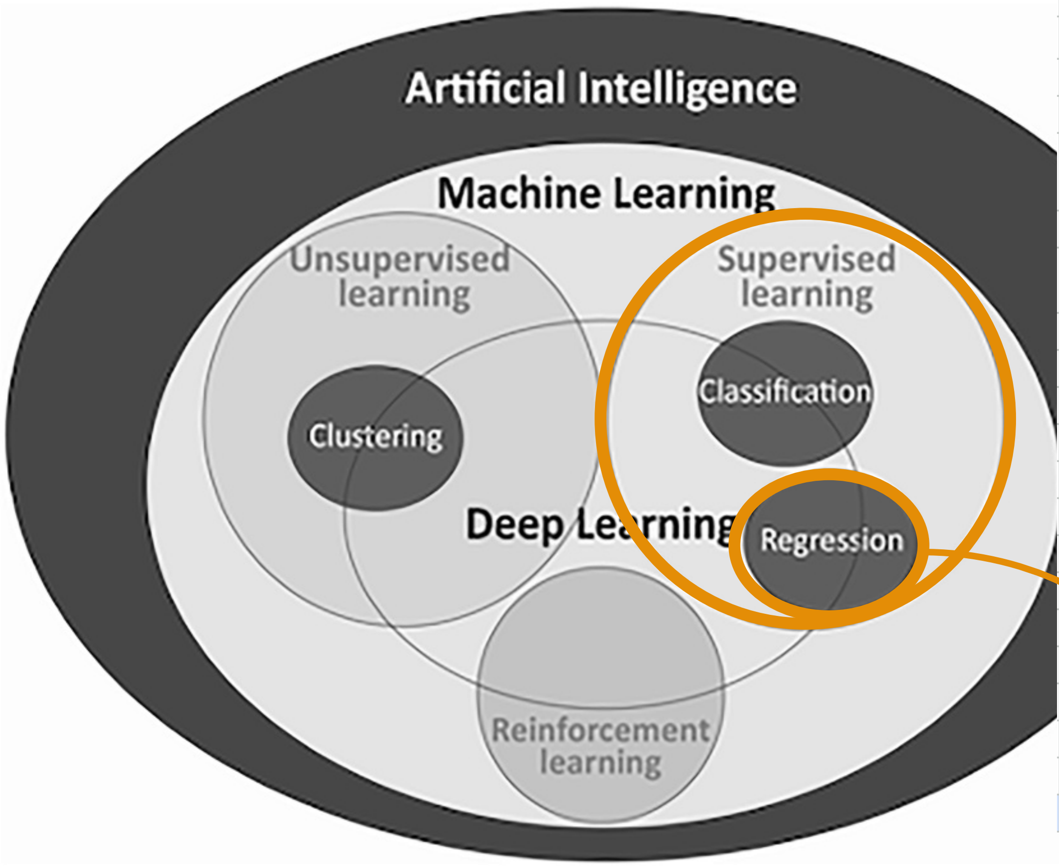
COSC 522 – Machine Learning

Regression

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
https://www.eecs.utk.edu/people/hairong-qi/
Email: hqi@utk.edu



Date & Lecture Recordings	Topics	
	Part 1: Statistical Methods	
	Baysian Learning	
08/20 (T)		Introduction
08/22 (R)		Baysian Decision Theory and Parametric Learning
08/27 (T)		Baysian Decision Theory and Non-Parametric Learning
08/29 (R)		Case Study: Representation for Natural Language (taught by An Cozma)
09/03 (T)		Parametric vs. Non-Parametric Learning: Some In-Depth Discus
09/05 (R)		Homework and Project Discussion (taught by Fanqi Wang)
	Neural Networks	
09/10 (T)		Biological Neuron and Perceptron
09/12 (R)		Perceptron
09/17 (T)		Back Propagation and Gradient Descent
09/19 (R)		Back Propagation
09/20 (F)		TRUST-AI Seminar
09/24 (T)		Kernel Methods and Review
09/26 (R)	Test 1	
10/01 (T)		Kernel Methods and Support Vector Machine
	Regression	
10/03 (R)		Logistic Regression
~~10/08 (T)~~	~~Fall Break (No Class)~~	

Questions

- Classification vs. Regression vs. Generation

- Linear regression and various basis functions
- What is global vs. local basis function?

- Maximum likelihood and least-square solution
- Least-square with regularization

PART I: REGRESSION AND BASIS FUNCTIONS

Linear regression (Linear function in w)

Generally

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

where $\phi_j(\mathbf{x})$ are known as *basis functions*.

Typically, $\phi_0(\mathbf{x}) = 1$, so that w_0 acts as a bias.

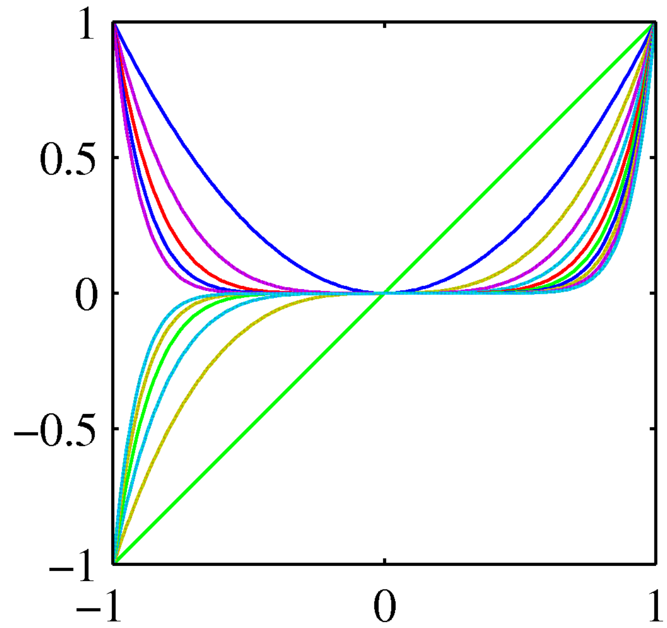
In the simplest case, we use linear basis functions : $\phi_d(\mathbf{x}) = x_d$.

Basis function - Polynomial

Polynomial basis functions:

$$\phi_j(x) = x^j.$$

These are global; a small change in x affect all basis functions.

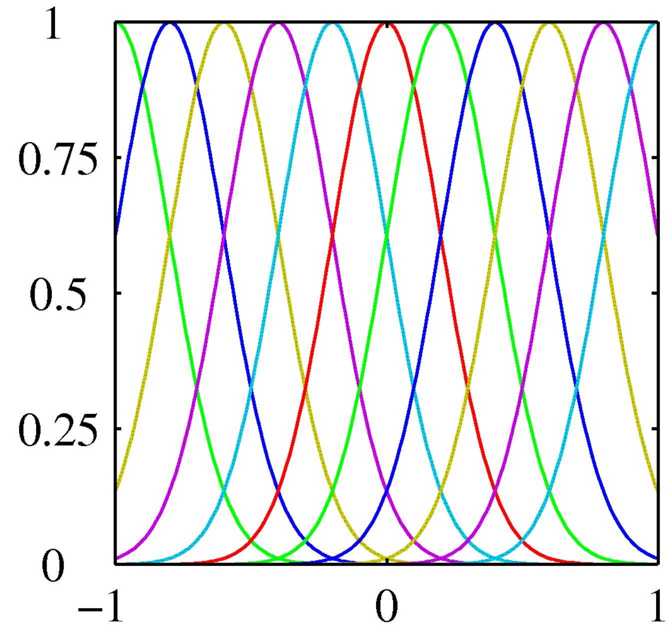


Polynomial regression

Basis function - Gaussian

Gaussian basis functions:

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$



These are local; a small change in x only affect nearby basis functions. μ and s control location and scale (width).

Basis function - Sigmoid

AICIP RESEARCH

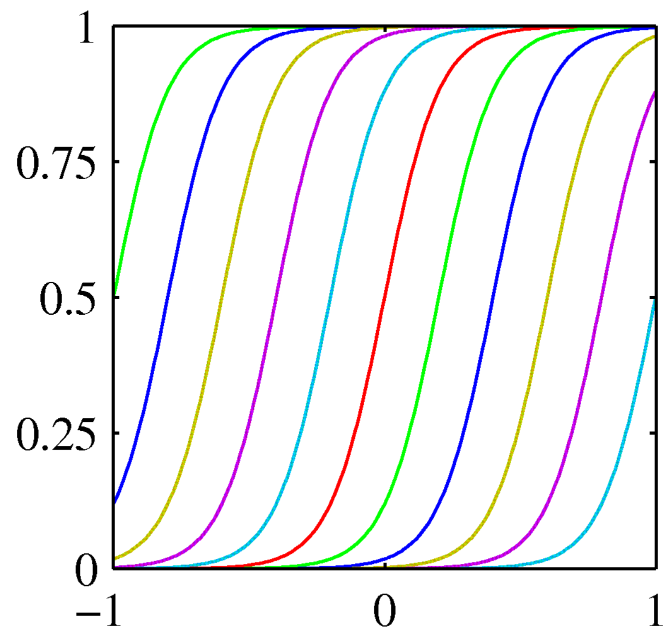
Sigmoidal basis functions:

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$

Also these are local; a small change in x only affect nearby basis functions. μ and s control location and scale (slope).



Logistic regression

AICIP RESEARCH

- The logistic function

$$\phi(x_k) = p(x_k) = \sigma(\frac{x_k - \mu}{s}) = \frac{1}{1 + e^{-\frac{x_k - \mu}{s}}}$$

- The log loss for the *k*th point

$$\begin{cases} -\ln \phi_k & \text{if } y_k = 1 \\ -\ln(1 - \phi_k) & \text{if } y_k = 0 \end{cases}$$

- The cost function: cross entropy

$$l(\beta_0, \beta_1) = \sum_k -y_k \log p_k - (1 - y_k) \log(1 - p_k)$$

- Find μ and s that best predict the probability of x belonging to a certain category

PART II: HOW TO SOLVE?

Maximum Likelihood and Least Squares (1)

AICIP RESEARCH

Assume observations from a deterministic function with added Gaussian noise:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \quad \text{where} \quad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1})$$

which is the same as saying,

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}).$$

Given observed inputs, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and targets, $\mathbf{t} = [t_1, \ldots, t_N]^{\mathrm{T}}$, we obtain the likelihood function

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}).$$

Maximum Likelihood and Least Squares (2)

Taking the logarithm, we get

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{w},\beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned}$$

where

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

is the sum-of-squares error.

Maximum Likelihood and Least Squares (3)

Computing the gradient and setting it to zero yields

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\} \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} = \mathbf{0}.$$

Solving for ⍰, we get

$$\mathbf{w}_{\mathrm{ML}} = \overbrace{\left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}}} \mathbf{t}$$

The Moore-Penrose pseudo-inverse, $\boldsymbol{\Phi}^{\dagger}$.

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

Least squares with regularization terms

- L2 norm (sum of square or Weight decay): Ridge regression
- L1 norm (LASSO regression), q=1 (sparsity)
- L12 norm (ElasticNet regression), q=1 and 2, M=2.

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}.$$

$$\mathbf{w} = \left(\lambda\mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}.$$

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

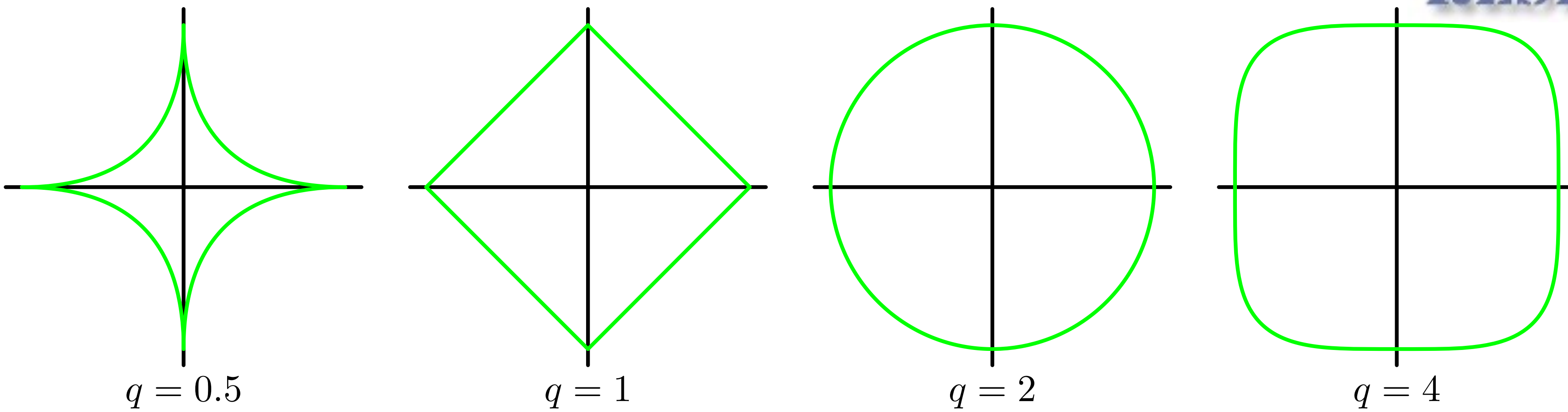
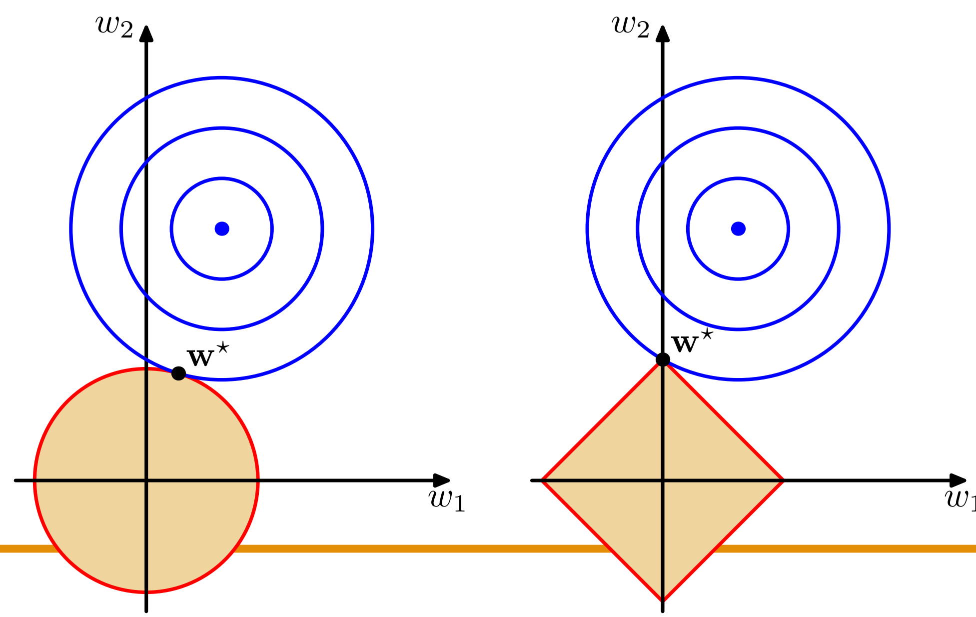


Figure 3.3 Contours of the regularization term in (3.29) for various values of the parameter q.

Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^{\star}$. The lasso gives a sparse solution in which $w_1^{\star} = 0$.



AICIP RESEARCH

Linear regression - Summary

- Linear regression $$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$
  - Simple linear regression (d=1)
  - Multiple linear regression (d>1)

    $\phi_{\square}$ □□□□□□ □$_{\square}$
  - Polynomial regression $\phi_j(x) = x^j.$
  - Logistic regression $$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right)$$
- Solving linear regression with maximum likelihood
  - Unconstrained formulation leads to least-squares solution $$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$
  - Constrained formulation with regularization terms $$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$
    - L2 norm → Ridge regression (q=2)
    - L1 norm → LASSO regression (q=1)
    - L12 norm → ElasticNet regression (q=1 and q=2)