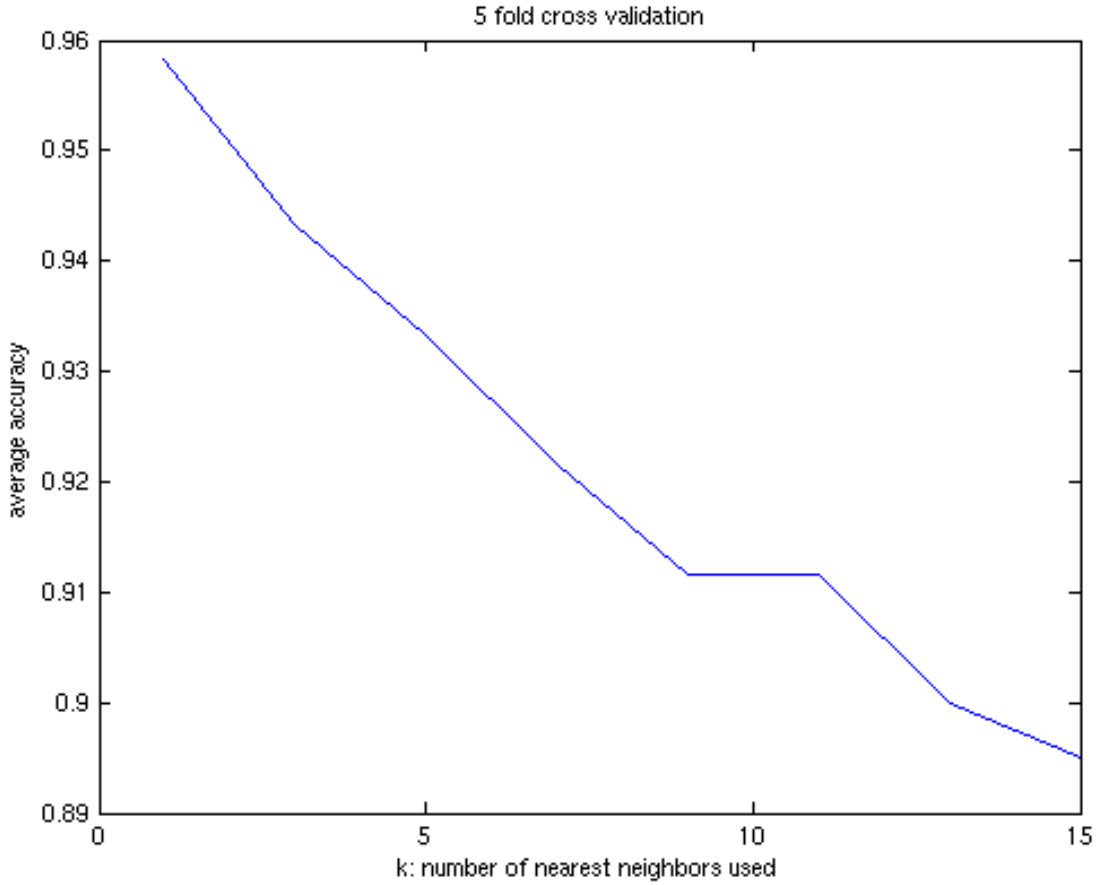
Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2011
Prof. Erik Sudderth

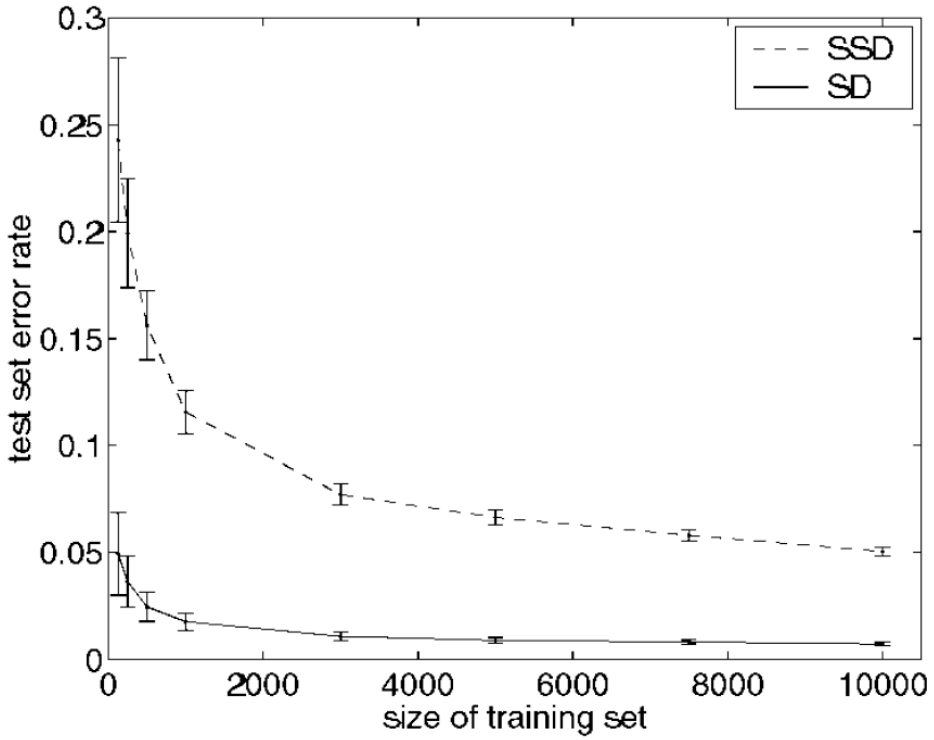
Lecture 9: Bayesian Regression,
Logistic Regression

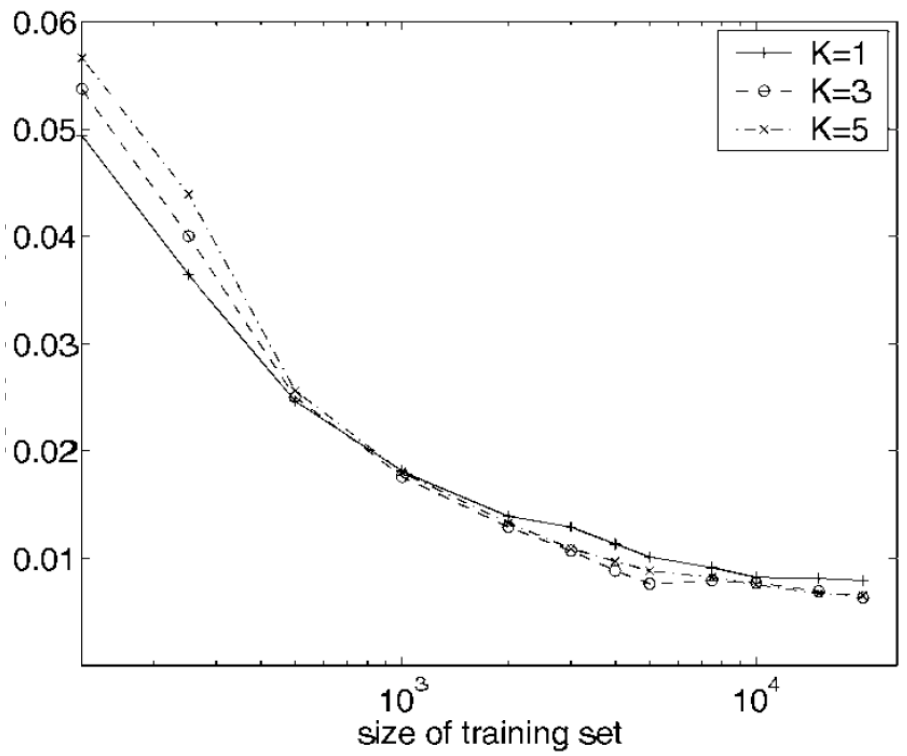
Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

K-NN Cross-Validation: MNIST Digits (Homework 3)



K-NN Performance: Shape Contexts



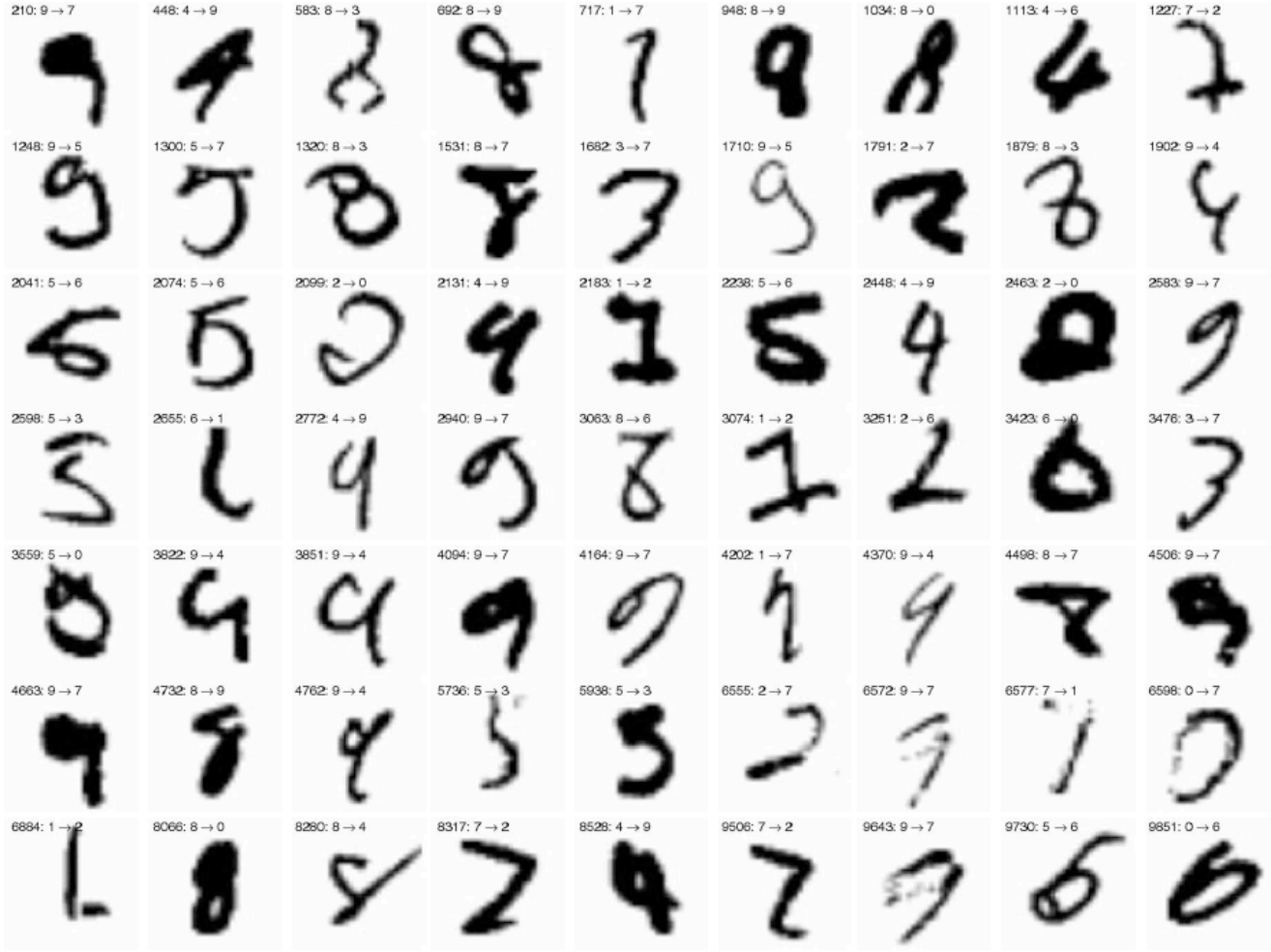


Alternative Distance Measures

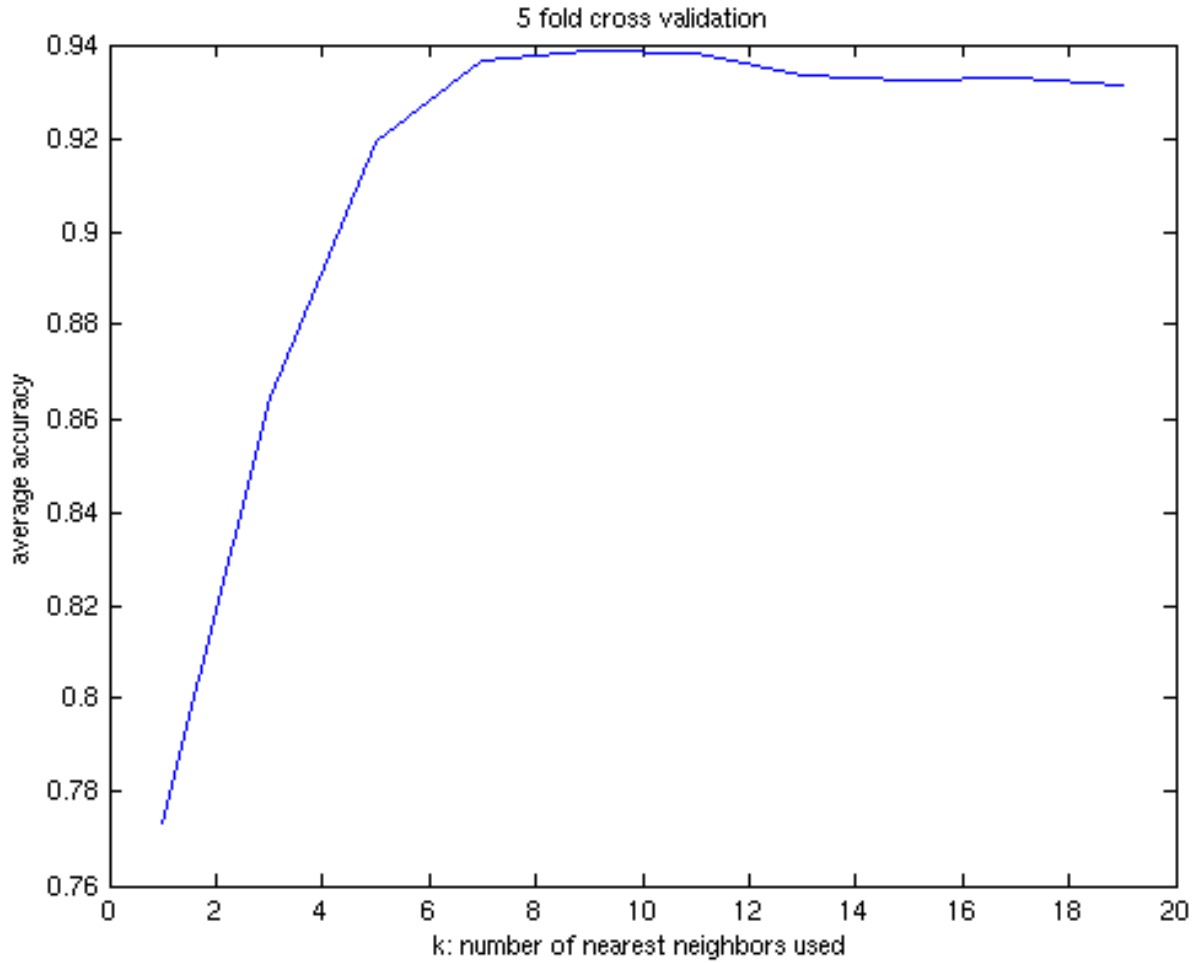
Choice of Neighborhood Size K

Belongie, Malik, & Puzicha, PAMI 2002

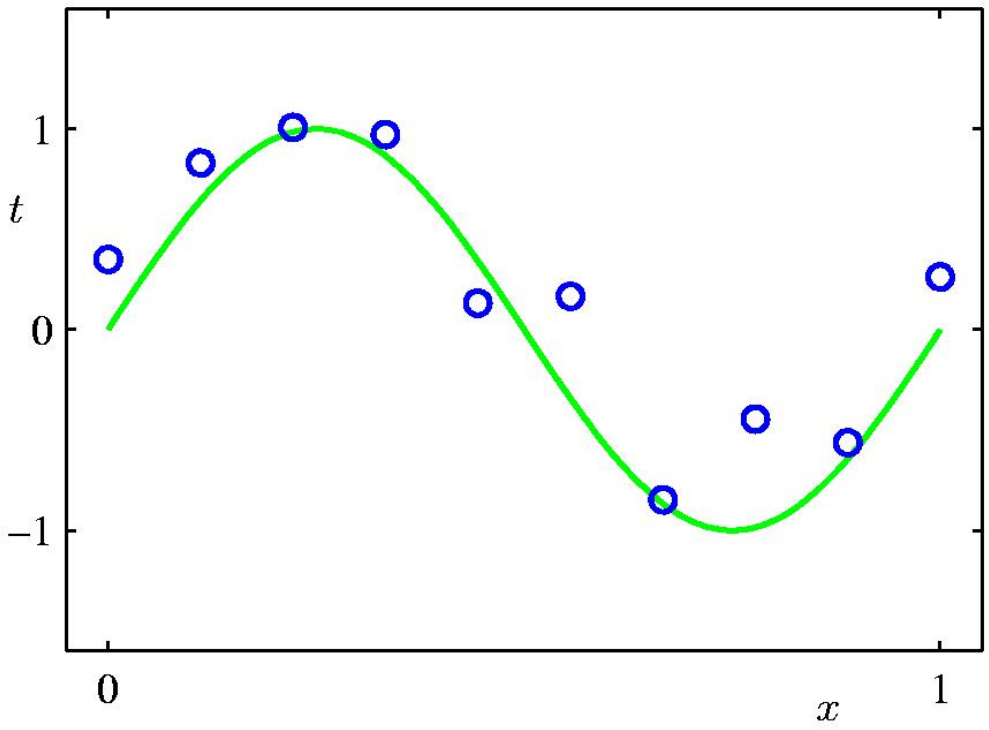
MNIST: Shape Context Errors



K-NN Cross-Validation: Nursery



Polynomial Curve Fitting



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Slides adapted from Bishop's "Pattern Recognition and Machine Learning"
Notation differs from Murphy's "Machine Learning: A Probabilistic Perspective"

Regularized Least Squares (1)

- Consider the error function:

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

Data term + Regularization term

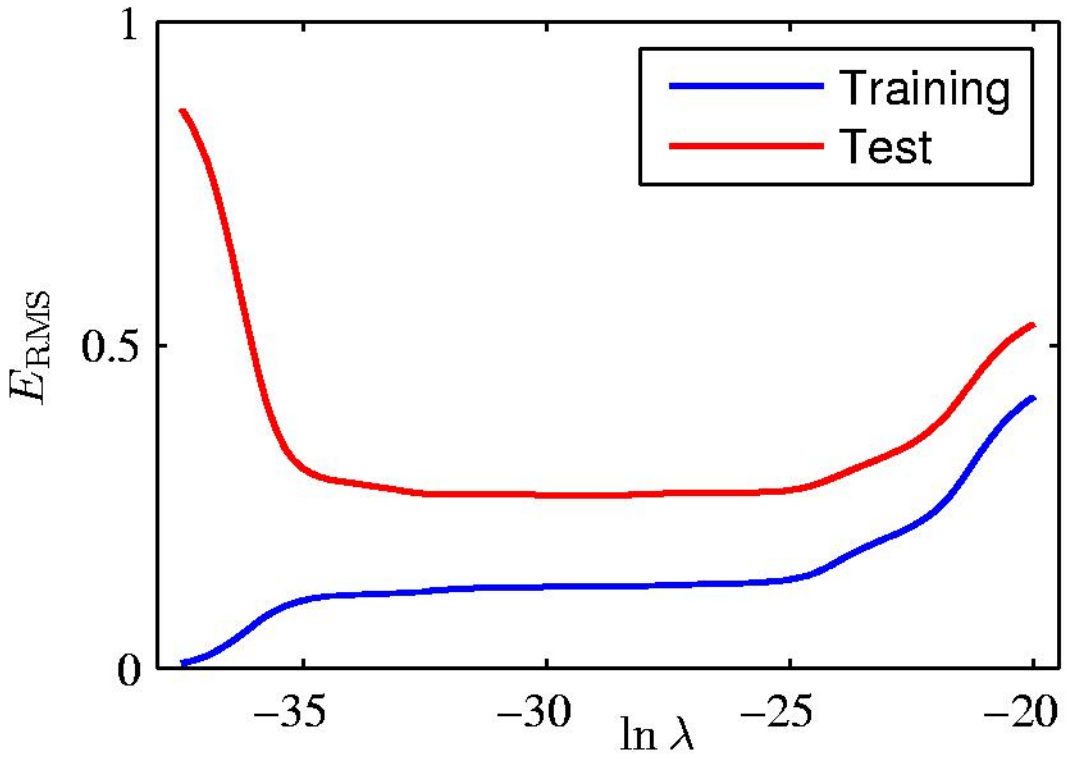
- With the sum-of-squares error function and a quadratic regularizer, we get

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

- which is minimized by

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

Regularization: E_{RMS} vs. $\ln \lambda$



Bayesian Linear Regression

- Gaussian prior and likelihood imply Gaussian posteriors:

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0).$$

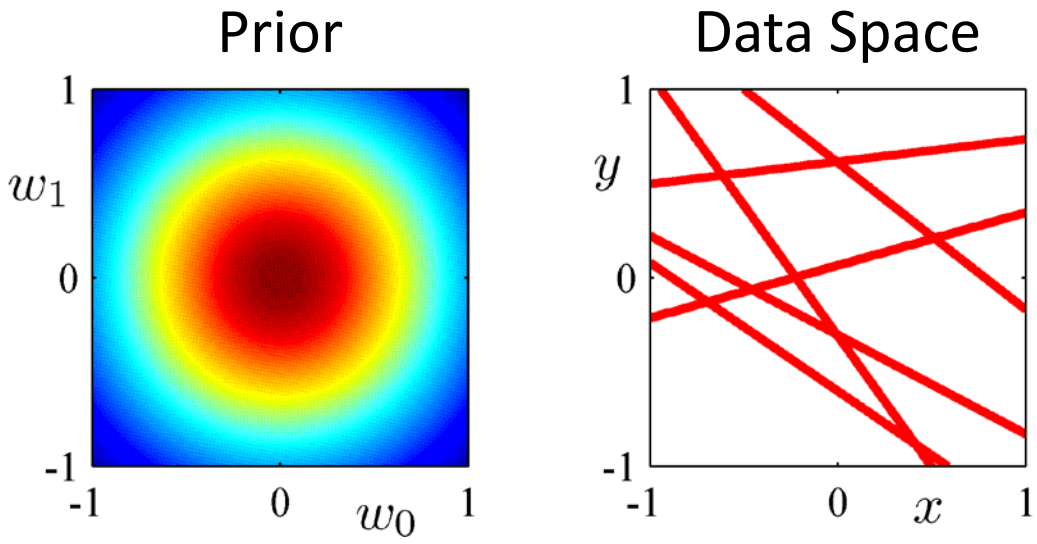
$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta \mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \right) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}. \end{aligned}$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}. \end{aligned}$$

Bayesian Linear Regression (3)

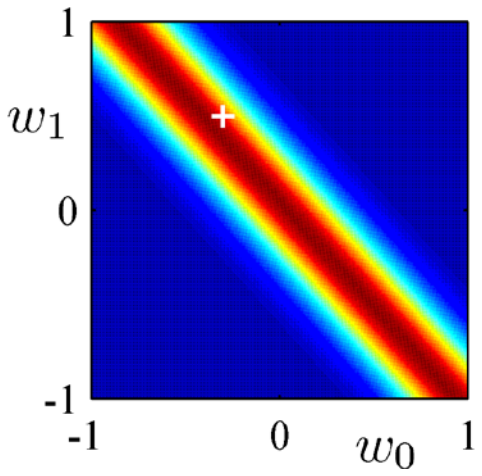
0 data points observed



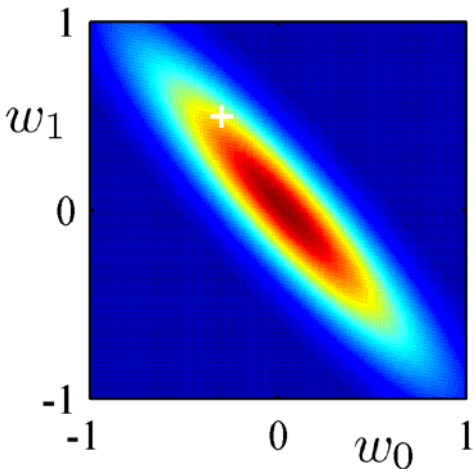
Bayesian Linear Regression (4)

1 data point observed

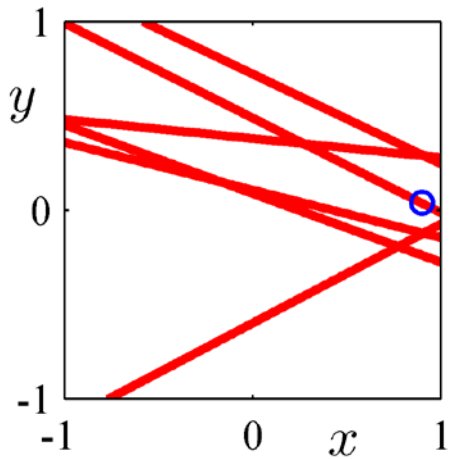
Likelihood



Posterior



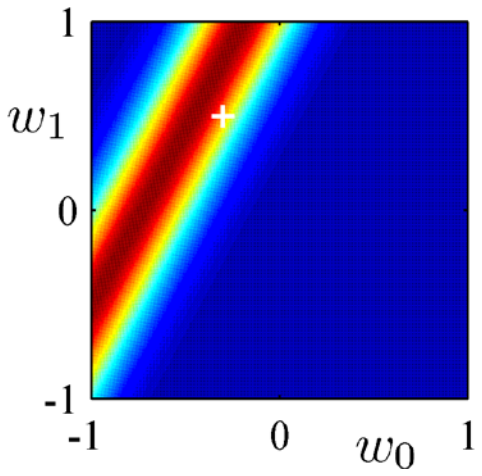
Data Space



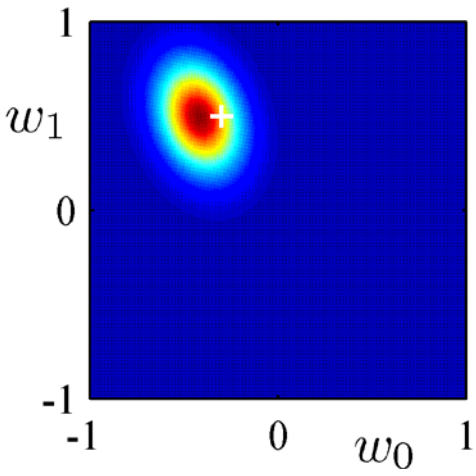
Bayesian Linear Regression (5)

2 data points observed

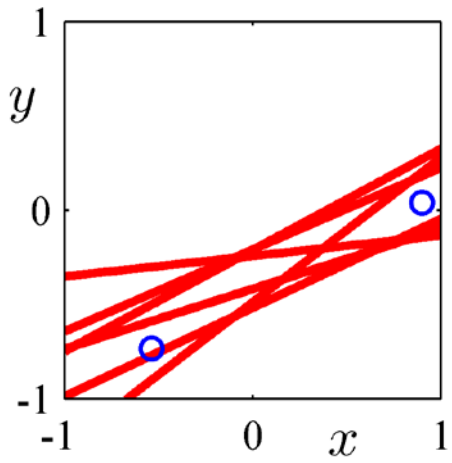
Likelihood



Posterior



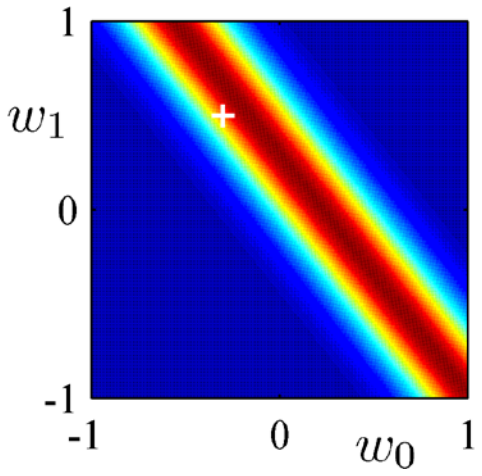
Data Space



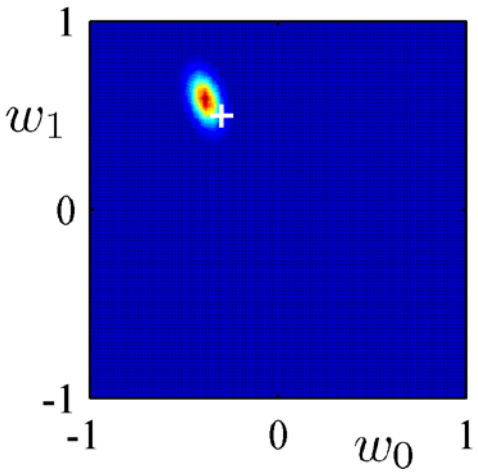
Bayesian Linear Regression (6)

20 data points observed

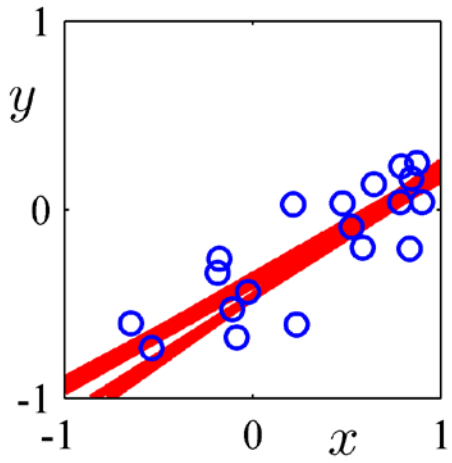
Likelihood



Posterior



Data Space



Predictive Distribution (1)

- Predict t for new values of x by integrating over w:

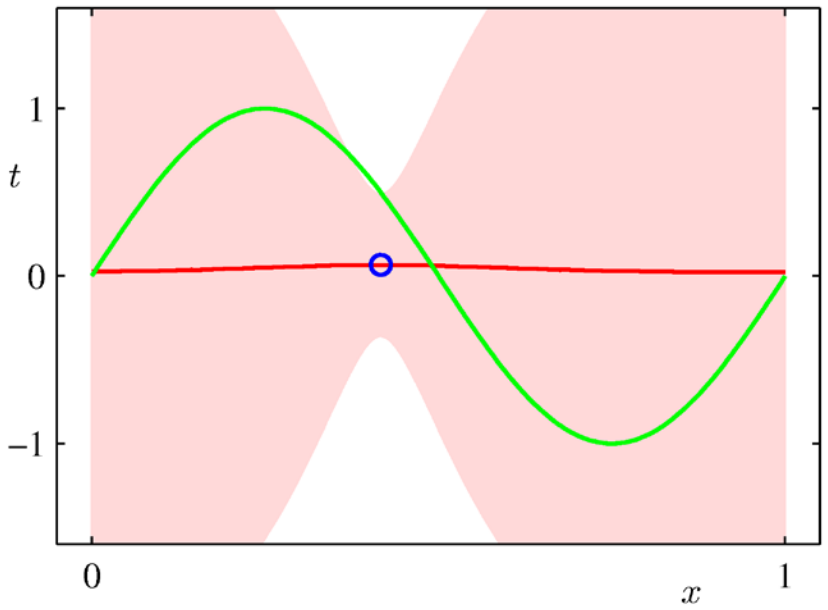
$$p(t|\mathbf{t},\alpha,\beta) = \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)\,\mathrm{d}\mathbf{w}$$
$$= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\sigma_N^2(\mathbf{x}))$$

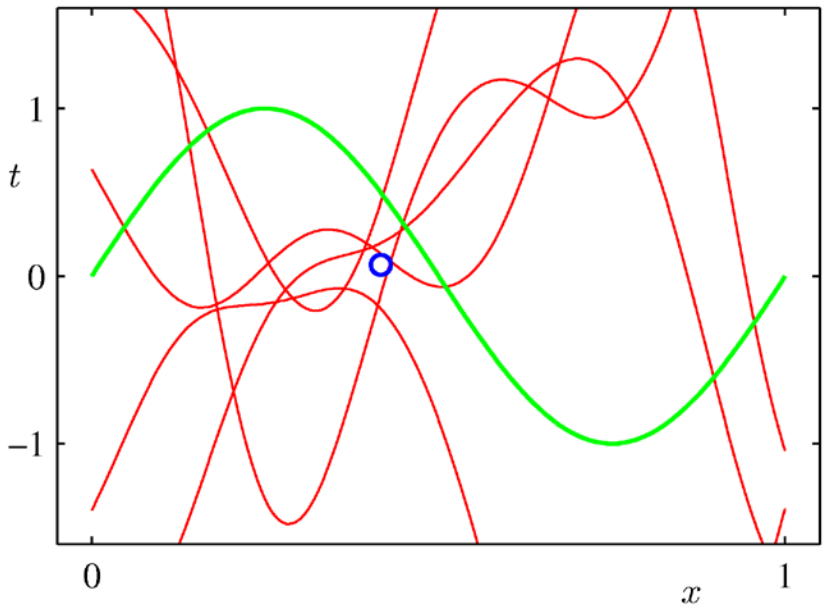
- where

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}).$$

Predictive Distribution (2)

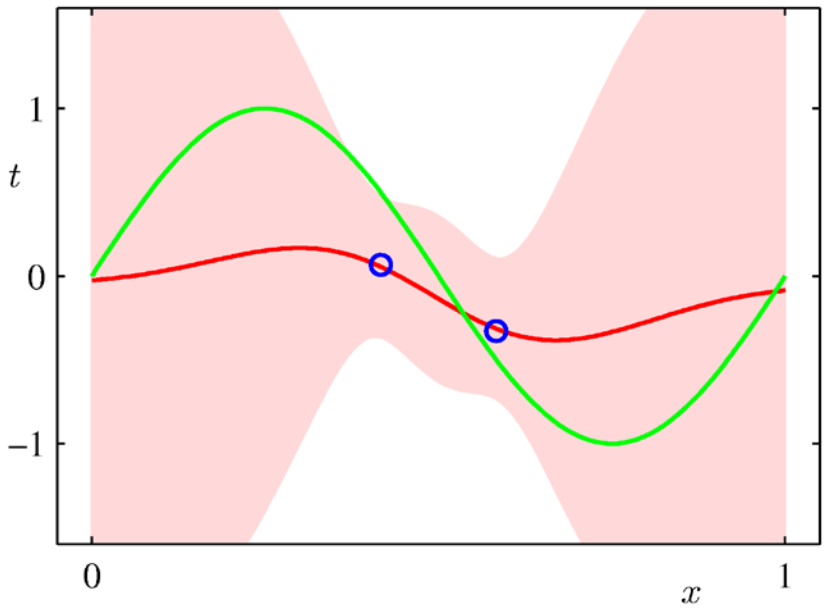
- Example: Sinusoidal data, 9 Gaussian basis functions, 1 data point

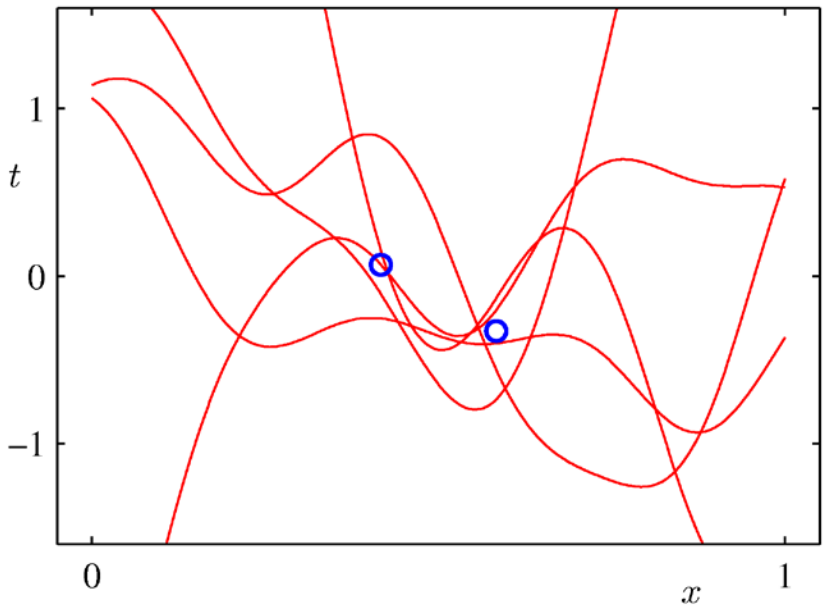




Predictive Distribution (3)

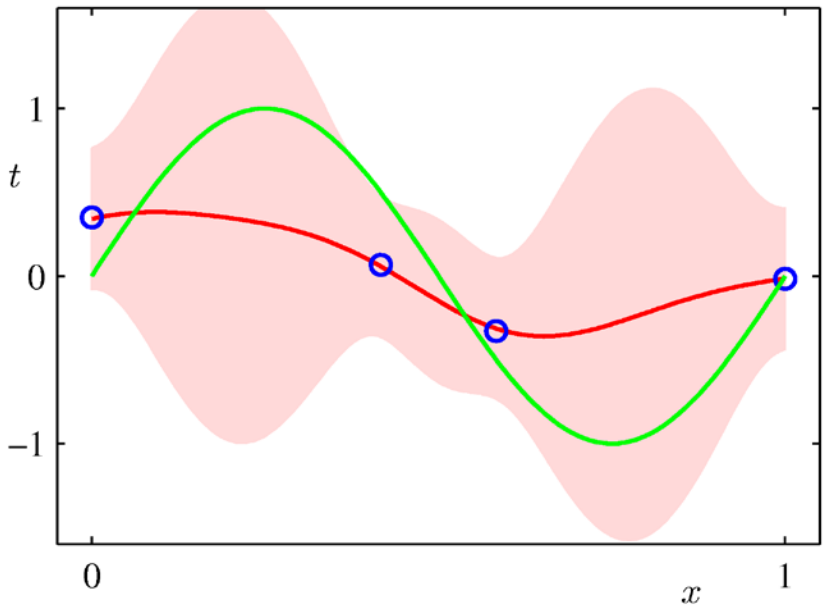
- Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points

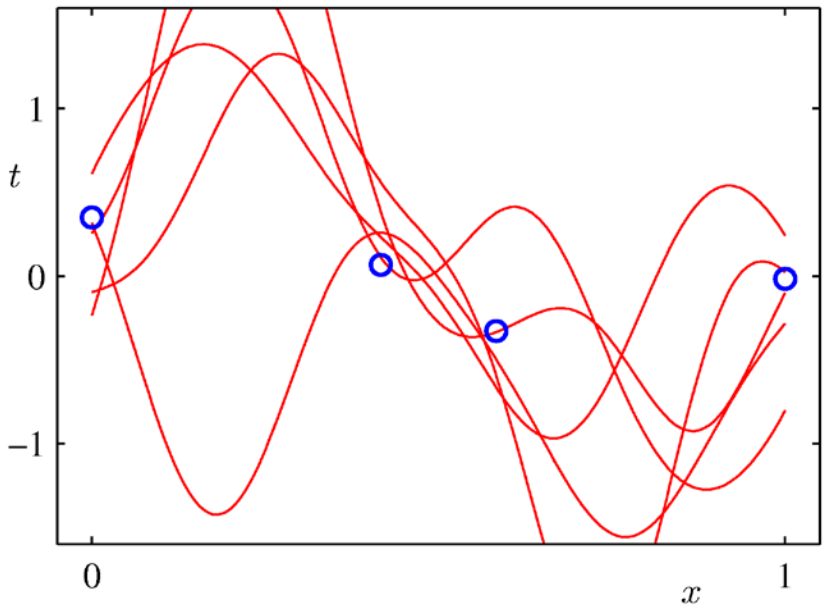




Predictive Distribution (4)

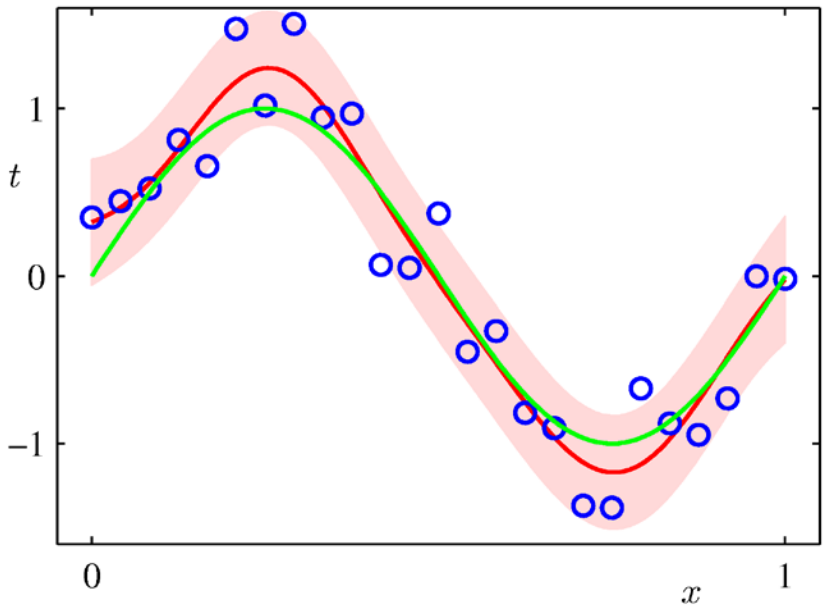
- Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points

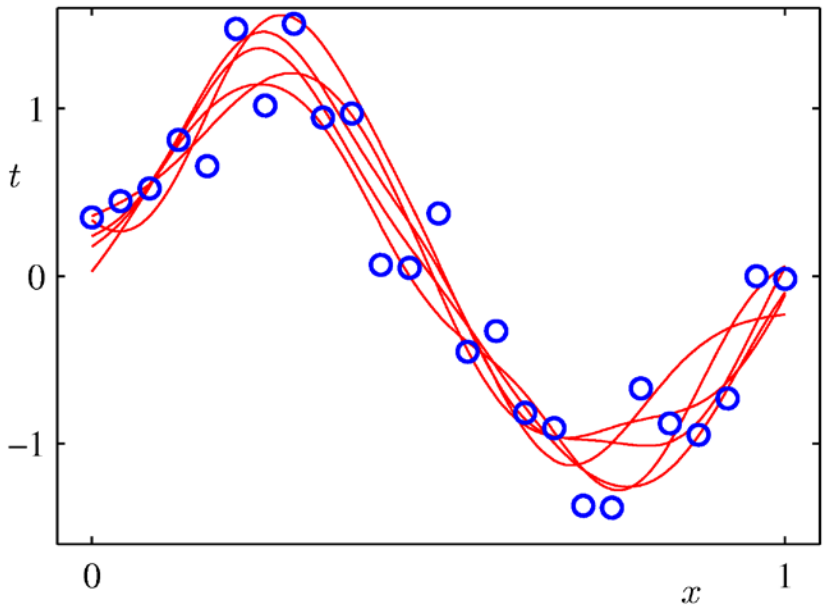




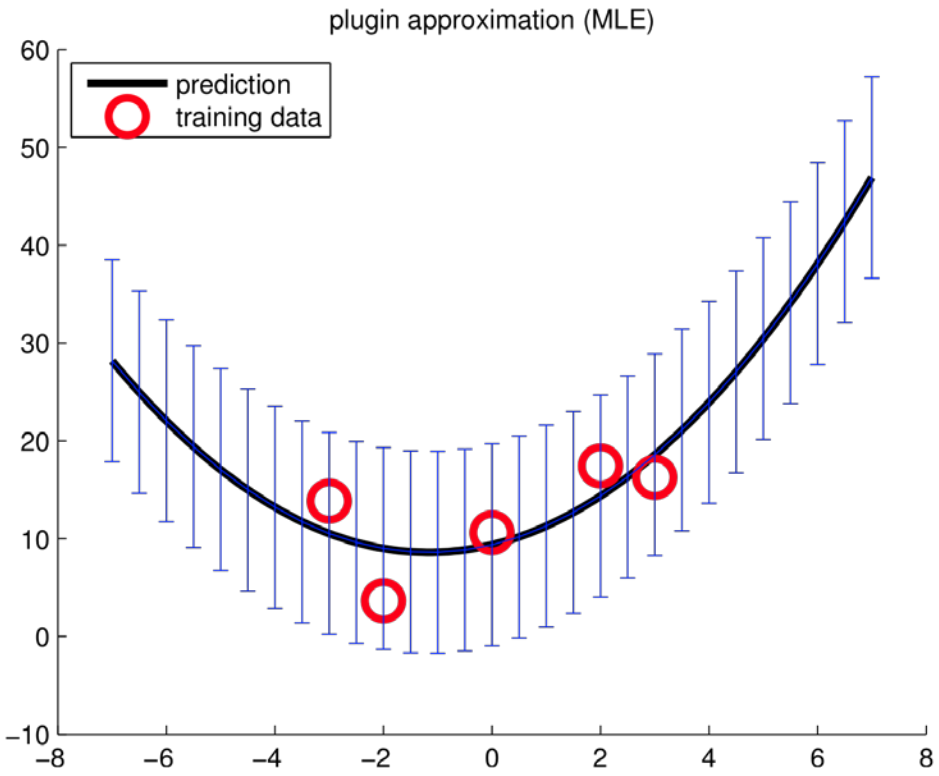
Predictive Distribution (5)

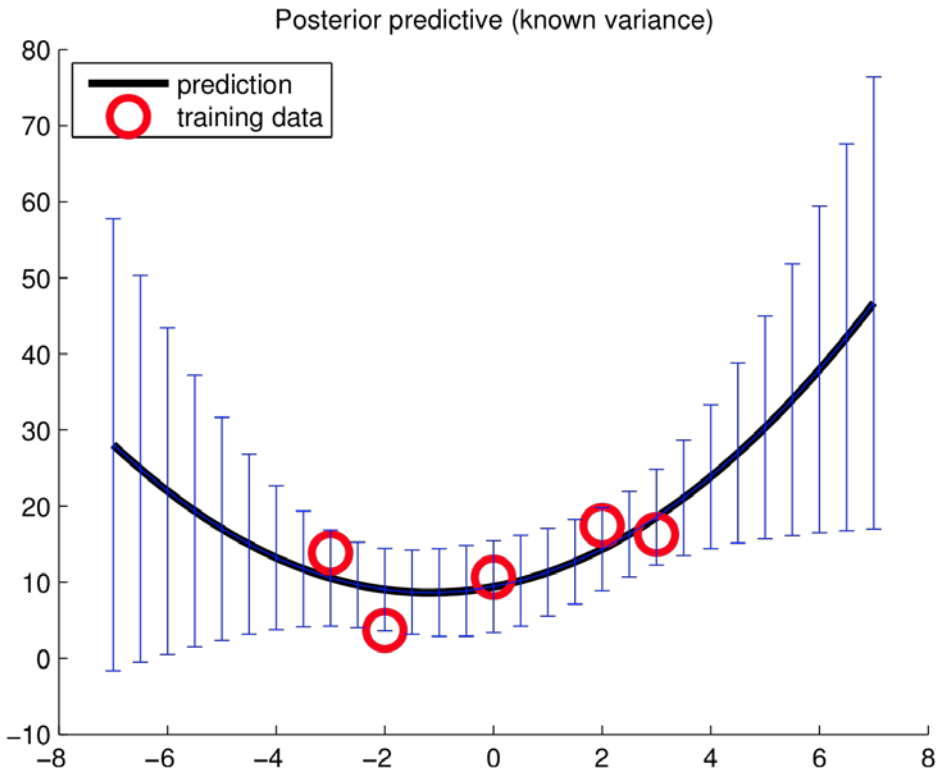
- Example: Sinusoidal data, 9 Gaussian basis functions, 25 data points



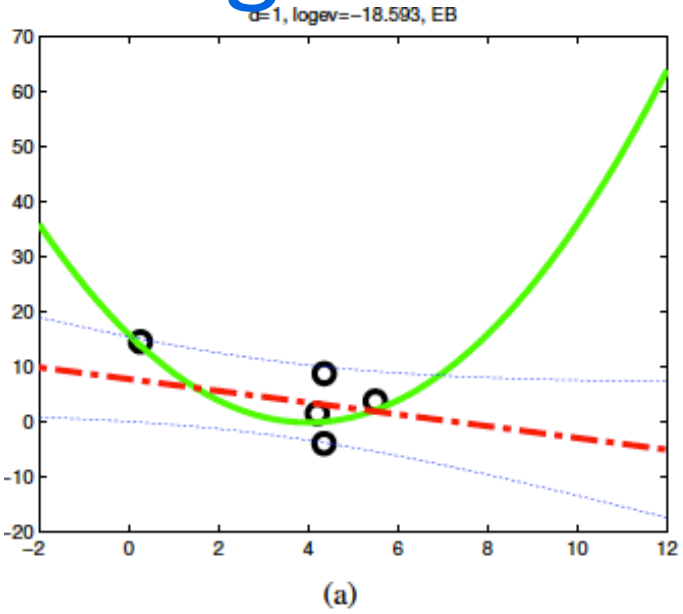


Estimation vs. Predictive Distributions

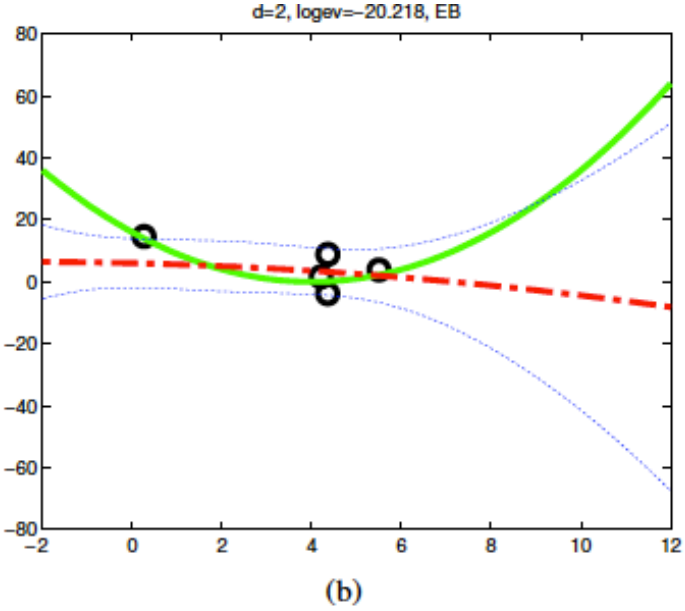




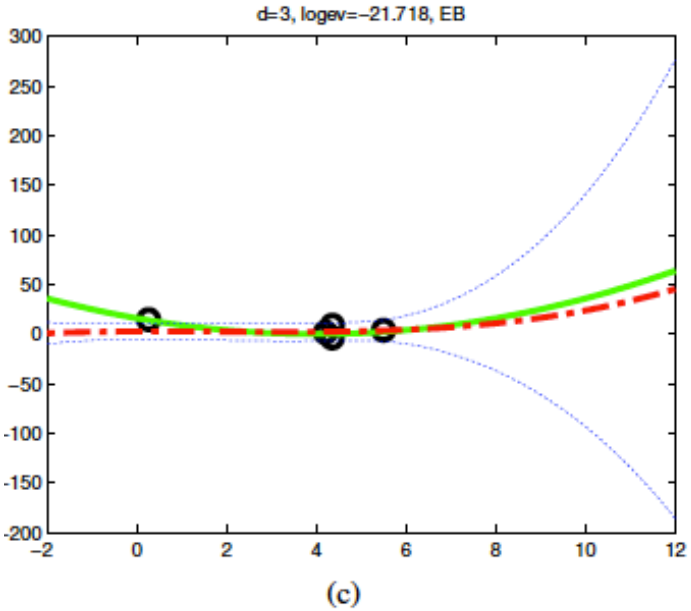
Marginal Data Likelihood: N=5



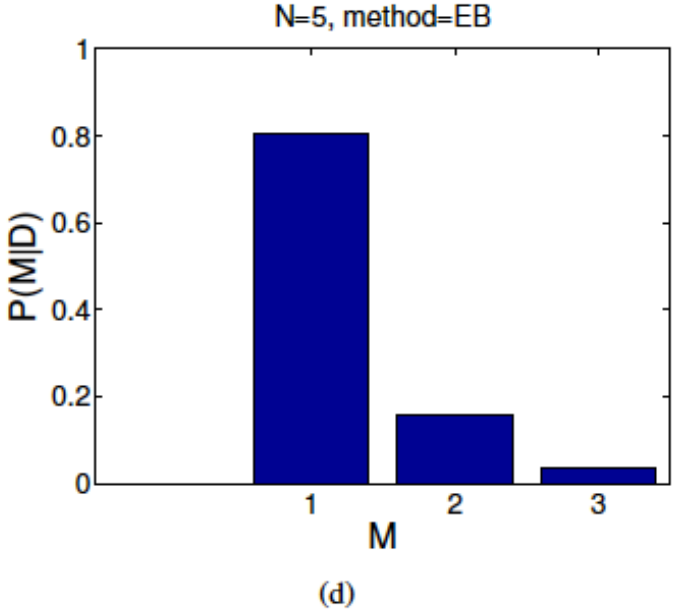
(a)



(b)

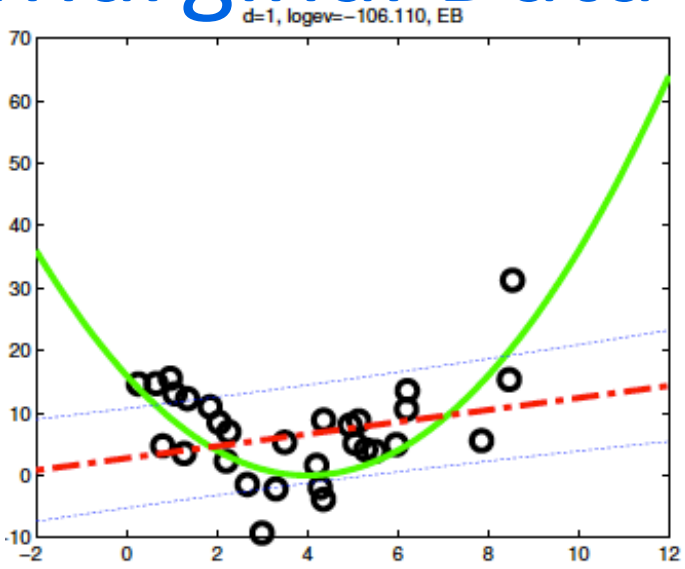


(c)

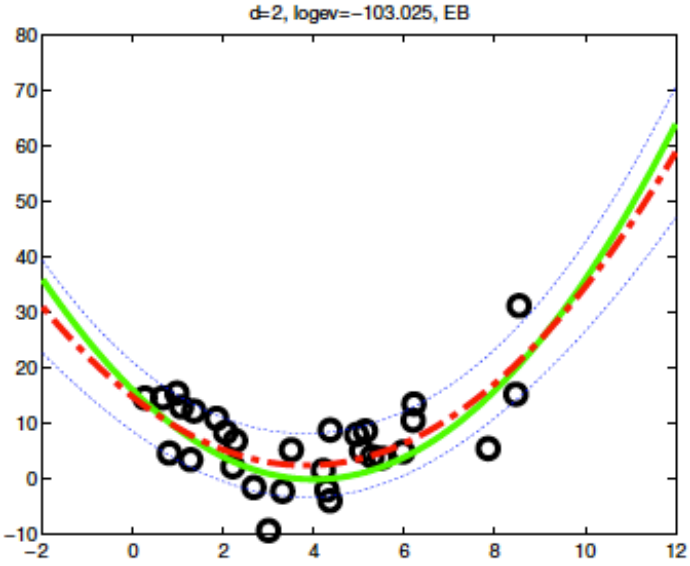


(d)

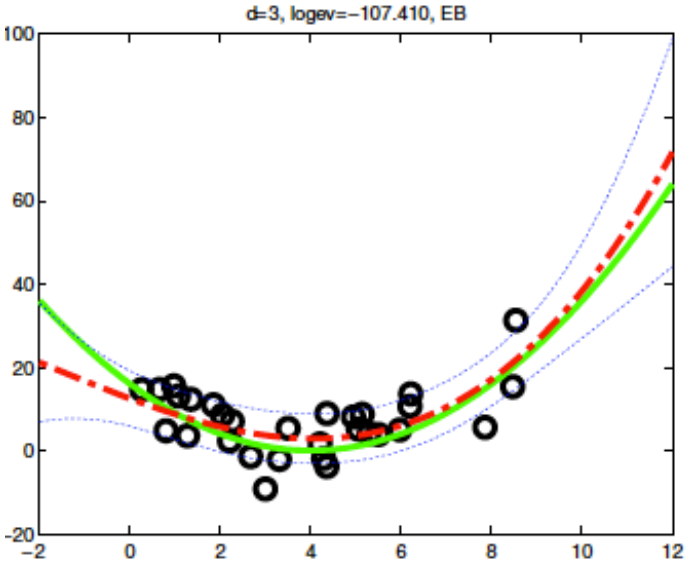
Marginal Data Likelihood: N=30



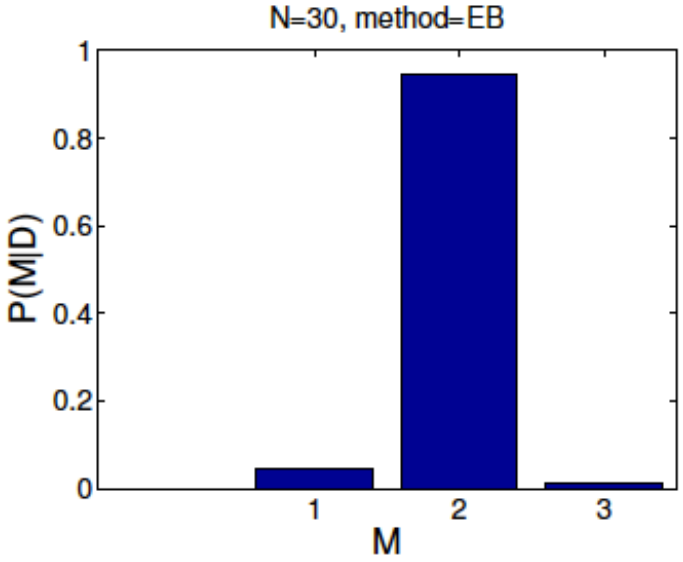
(a)



(b)



(c)



(d)

Bayesian Ockham's Razor

"Plurality must never be posited without necessity."

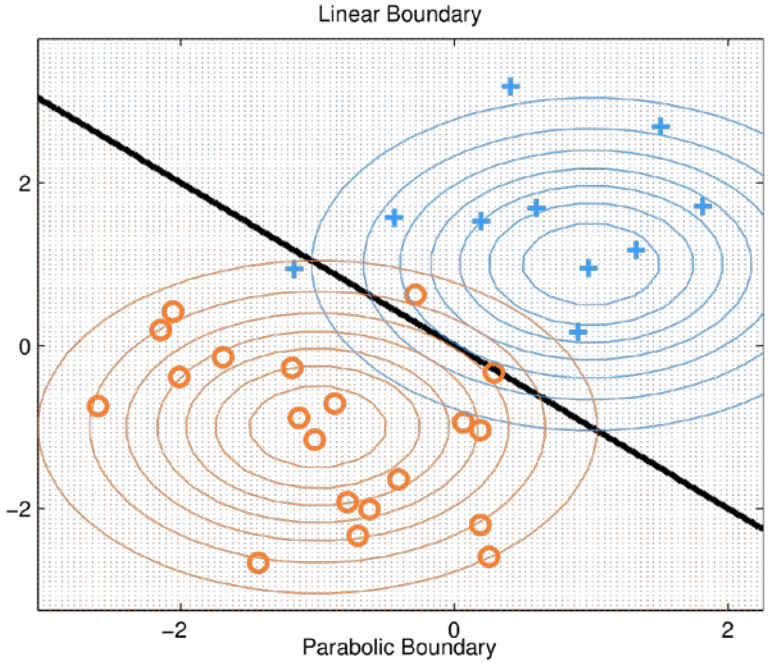
William of Ockham

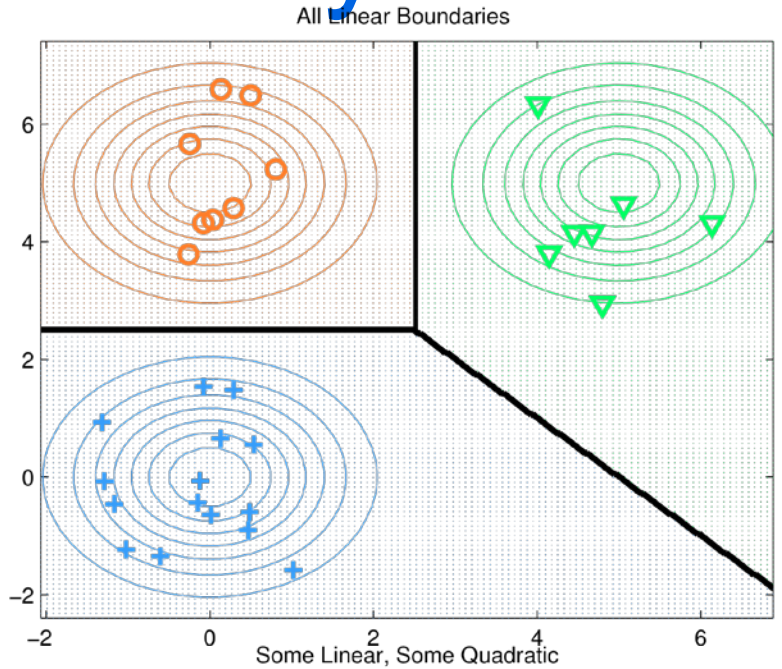
Back to Classification

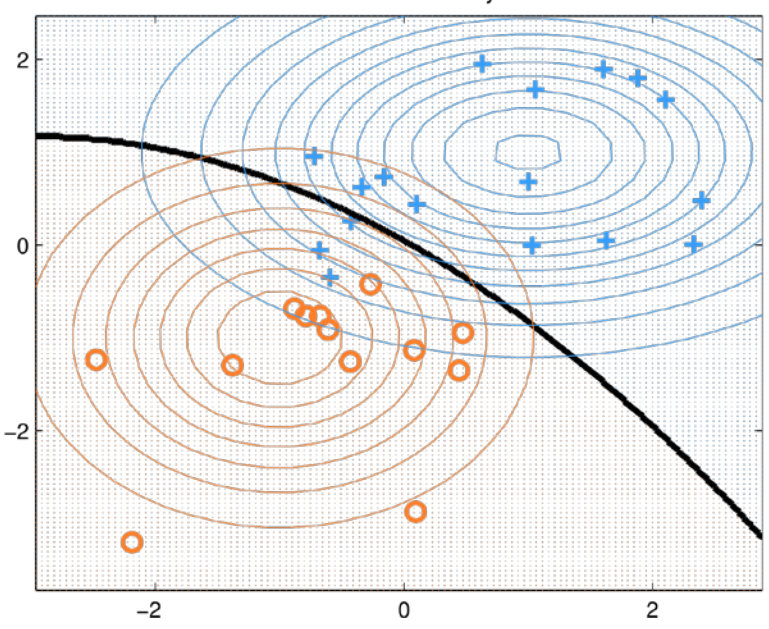
	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

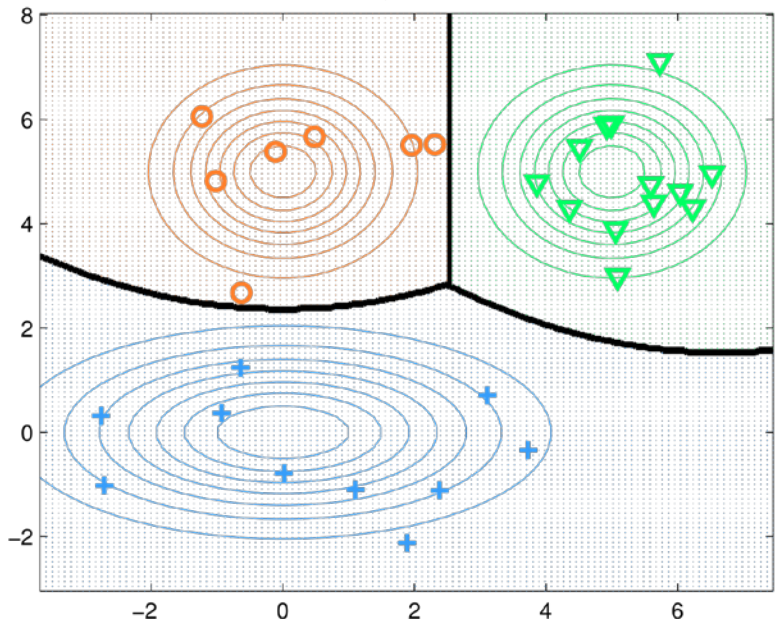
- Predict label/response y from feature x
- Generative classification: Apply Bayes' rule to learned p(x,y)
- Discriminative or conditional regression & classification: directly learn a model of p(y | x), assuming x always given

Discriminant Analysis



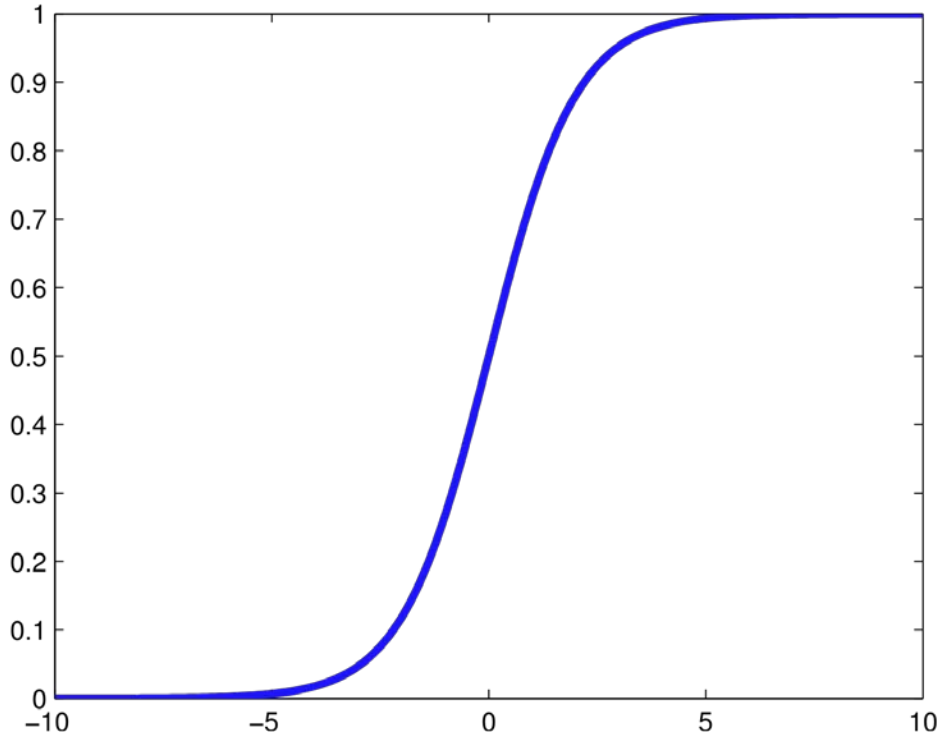


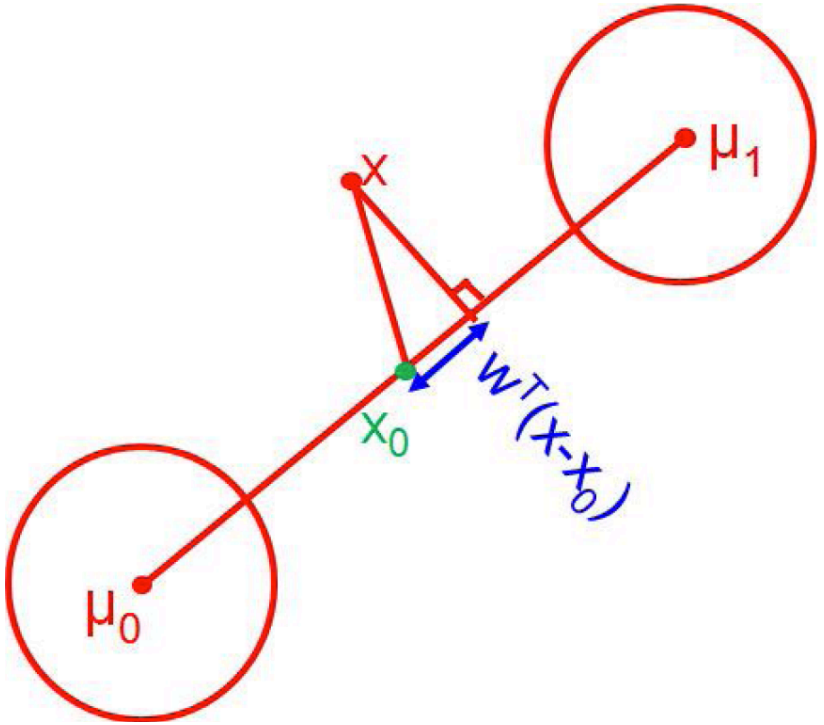




Binary Discriminant Analysis

Logistic Function





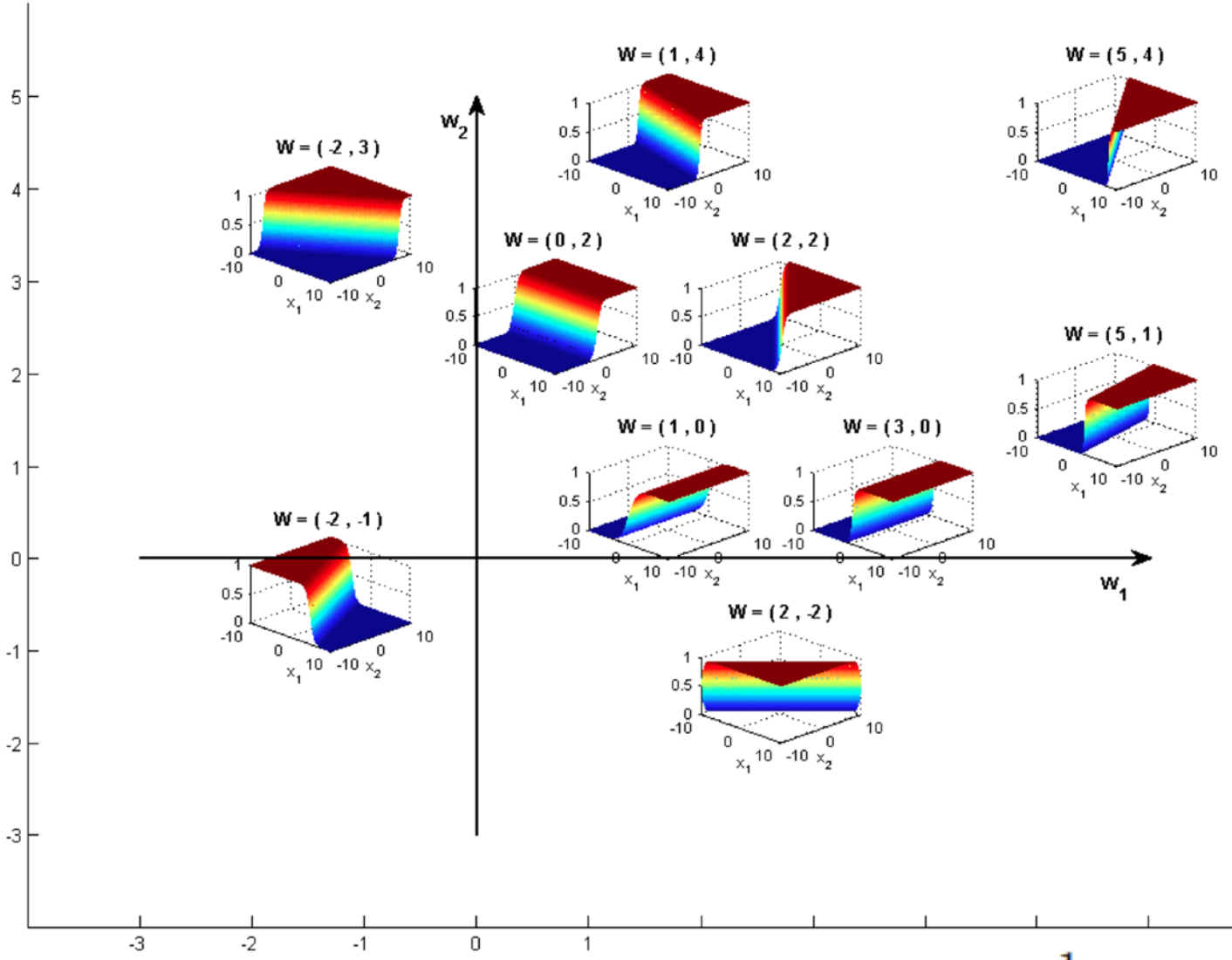
$$\text{sigm}(\eta) := \frac{1}{1+\exp(-\eta)} = \frac{e^{\eta}}{e^{\eta}+1}$$

$$p(y=1|\mathbf{x},\boldsymbol{\theta}) = \sigma(\mathbf{w}^T(\mathbf{x}-\mathbf{x}_0))$$

$$\mathbf{w} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0 = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\frac{\log(\pi_1/\pi_0)}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}$$

Logistic Regression

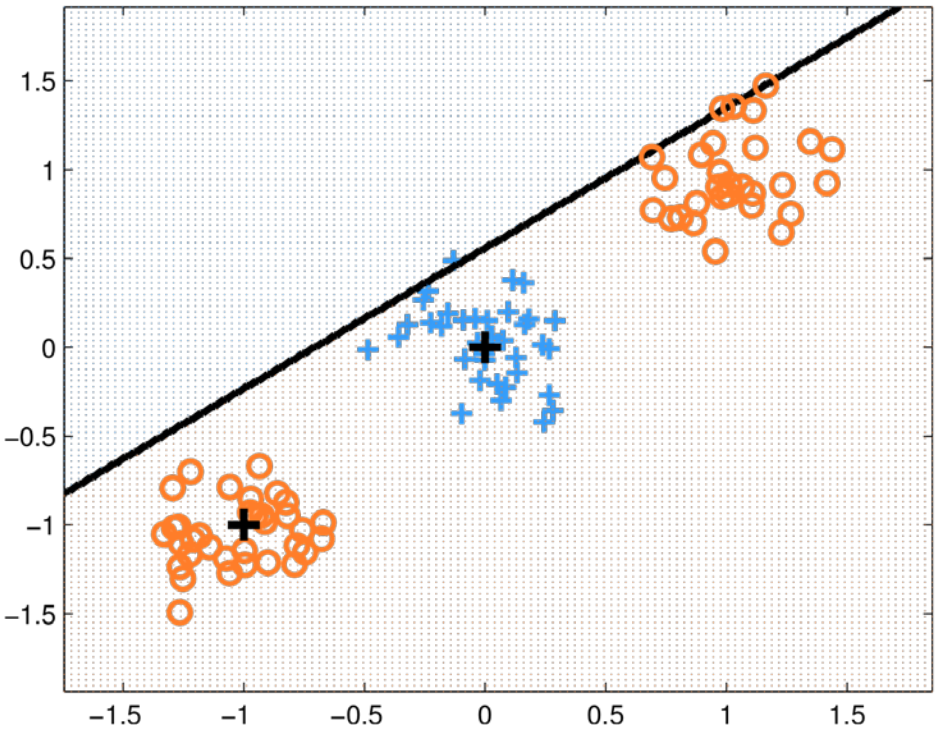


$$p(y|\mathbf{x}, \mathbf{w}) = \mathrm{Ber}(y|\mathrm{sigm}(\mathbf{w}^T\mathbf{x}))$$

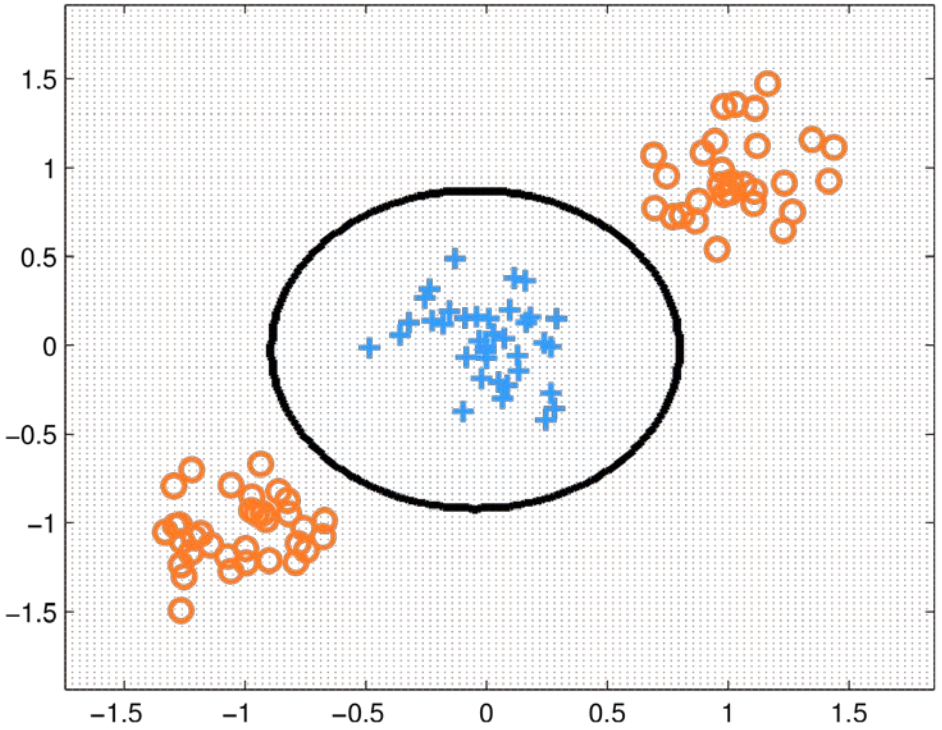
$$\mathrm{sigm}(\eta) := \frac{1}{1+\exp(-\eta)} = \frac{e^{\eta}}{e^{\eta}+1}$$

Decision Boundaries

By assumption, a linear function of input features.

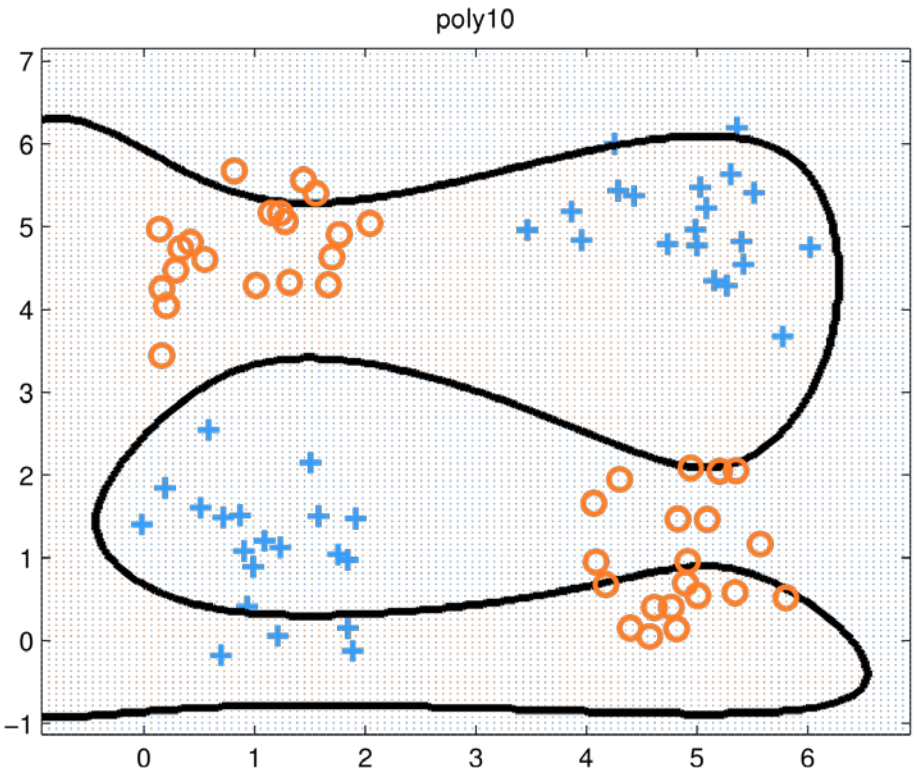


Raw features.

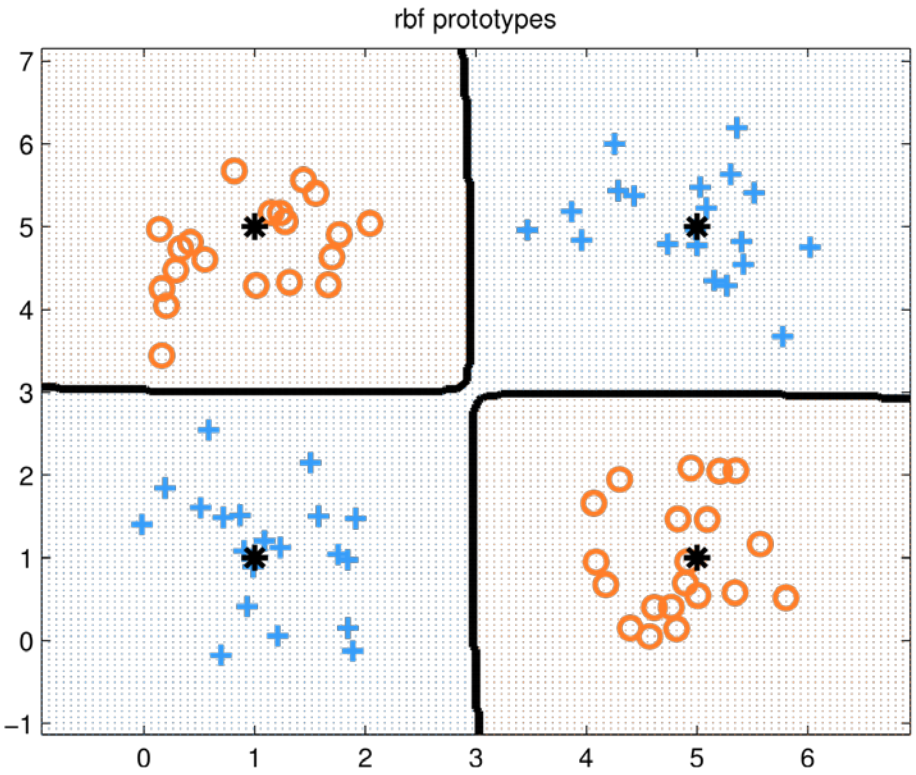


2nd-order polynomial expansion.

Importance of Features: XOR



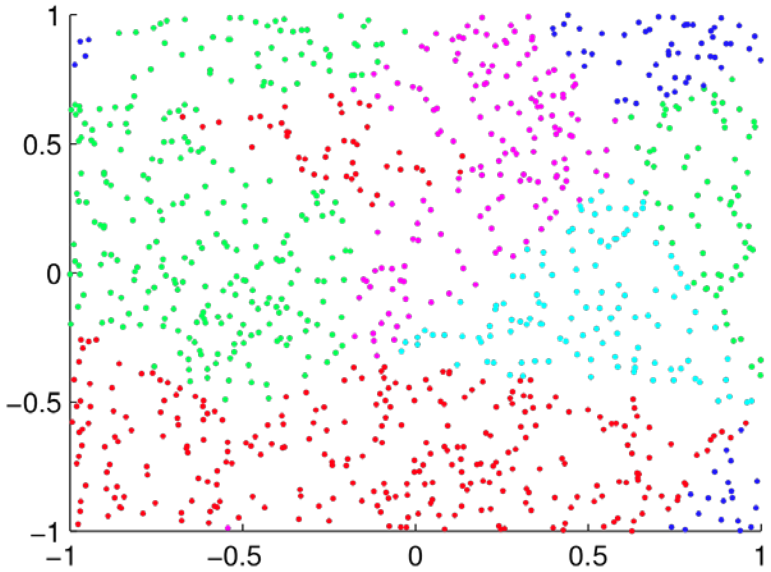
10^{th}-order polynomial.



4 radial basis functions.

Avoid RBF placement via kernel methods…

Multinomial Logistic Regression



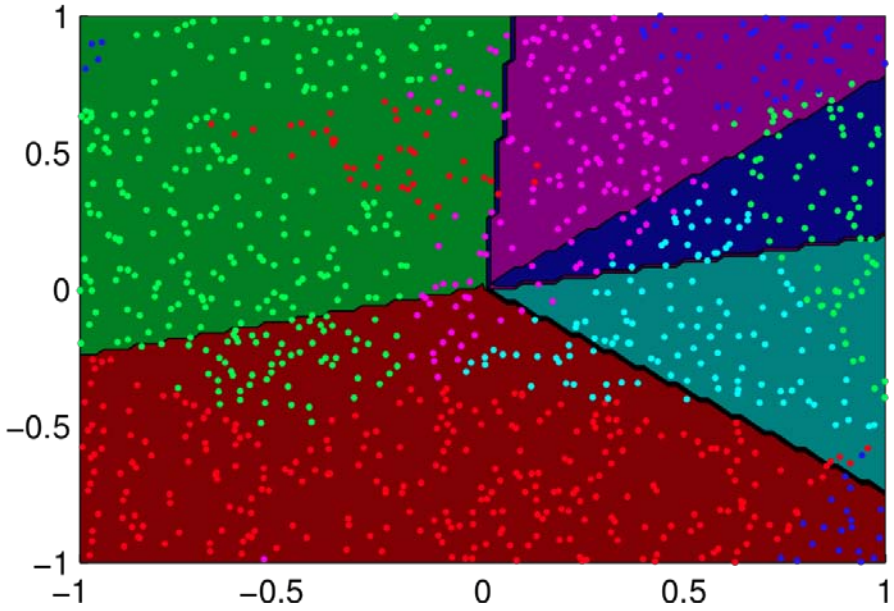
$$p(y|\mathbf{x}, \mathbf{W}) = \mathrm{Cat}(y|\mathcal{S}(\mathbf{W}^T\mathbf{x}))$$

$$\mathcal{S}(\eta)_c = \frac{e^{\eta_c}}{\sum_{c'=1}^{C} e^{\eta_{c'}}}$$

as $T \to 0$

$$\mathcal{S}(\eta/T)_c = \begin{cases} 1.0 & \text{if } c = \arg\max_{c'} \eta_{c'} \\ 0.0 & \text{otherwise} \end{cases}$$

Linear Multinomial Logistic Regression



Kernel−RBF Multinomial Logistic Regression