Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2011

Prof. Erik Sudderth

Lecture 7: Gaussian Distributions in Classification and Regression

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

A Learning Toolbox

Tuning Model Parameters to Data

- Maximum likelihood (ML) estimation: Frequency matching
- Bayesian MAP estimation: Regularized ML
- Bayesian posterior predictive distribution, posterior means, …

Model Selection

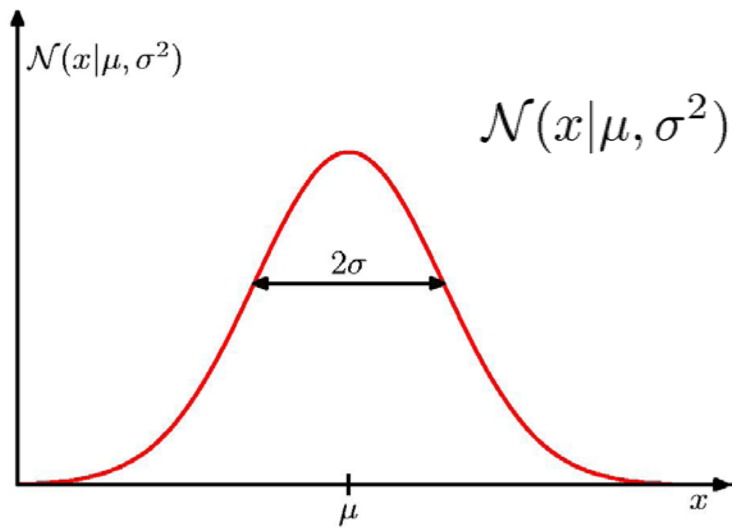
- Empirical performance on held-out data: Cross-validation
- Model complexity penalization: Bayesian marginal likelihood
- Approximations coming later: BIC, AIC, MDL, …

Probability Distributions

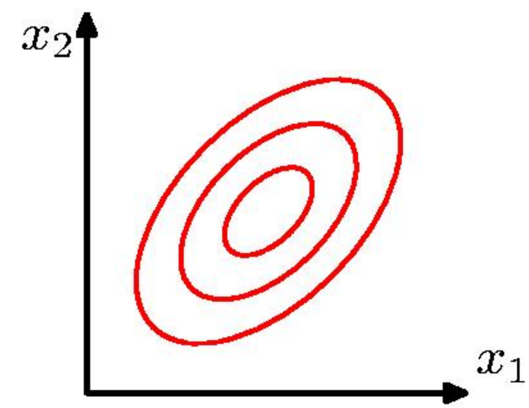
- Dirichlet, beta, multinoulli, bernoulli: Discrete data
- Multivariate normal, gamma: Continuous data

Useful Model Families: Remainder of course…

The Gaussian Distribution

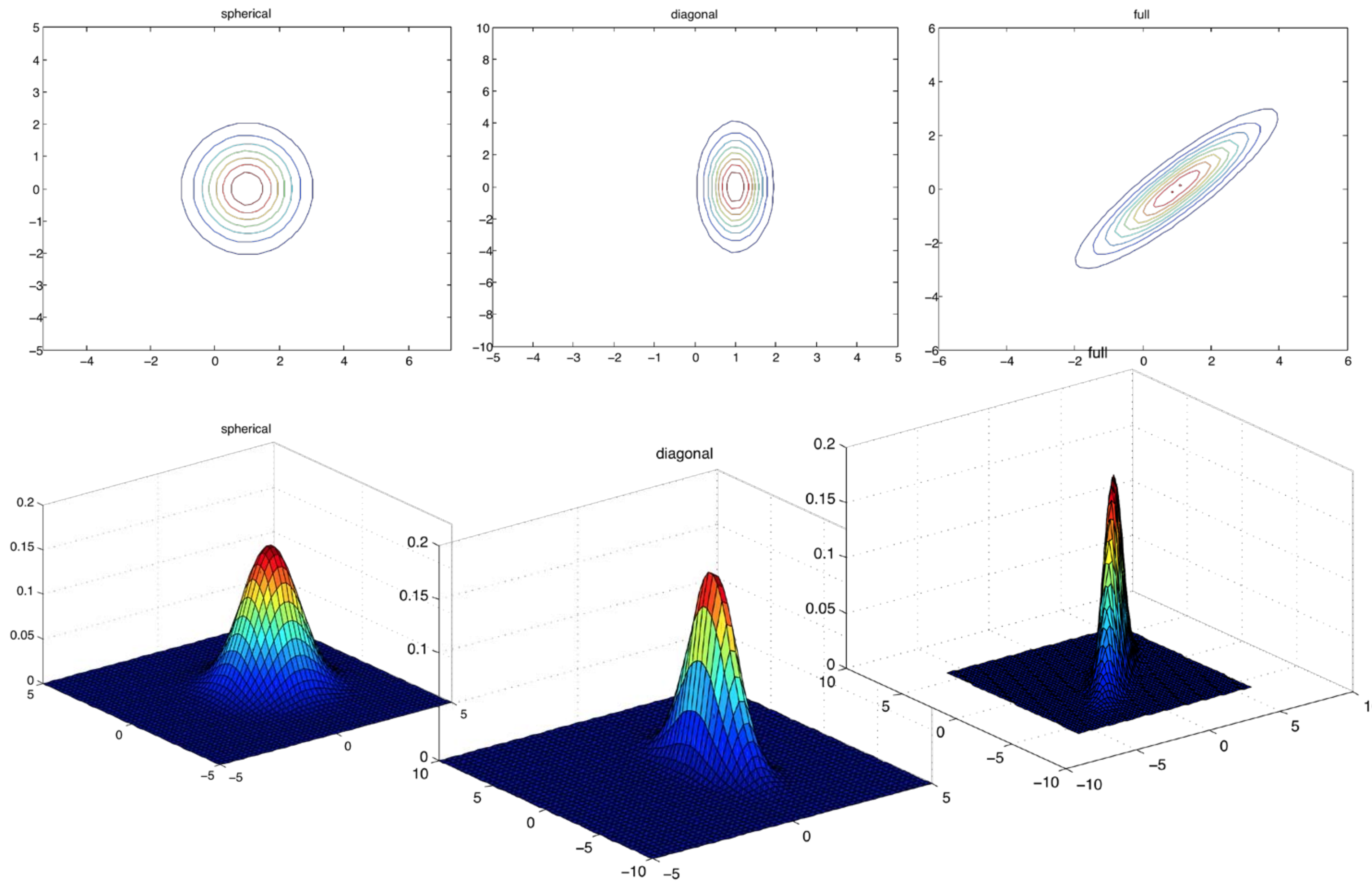


$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

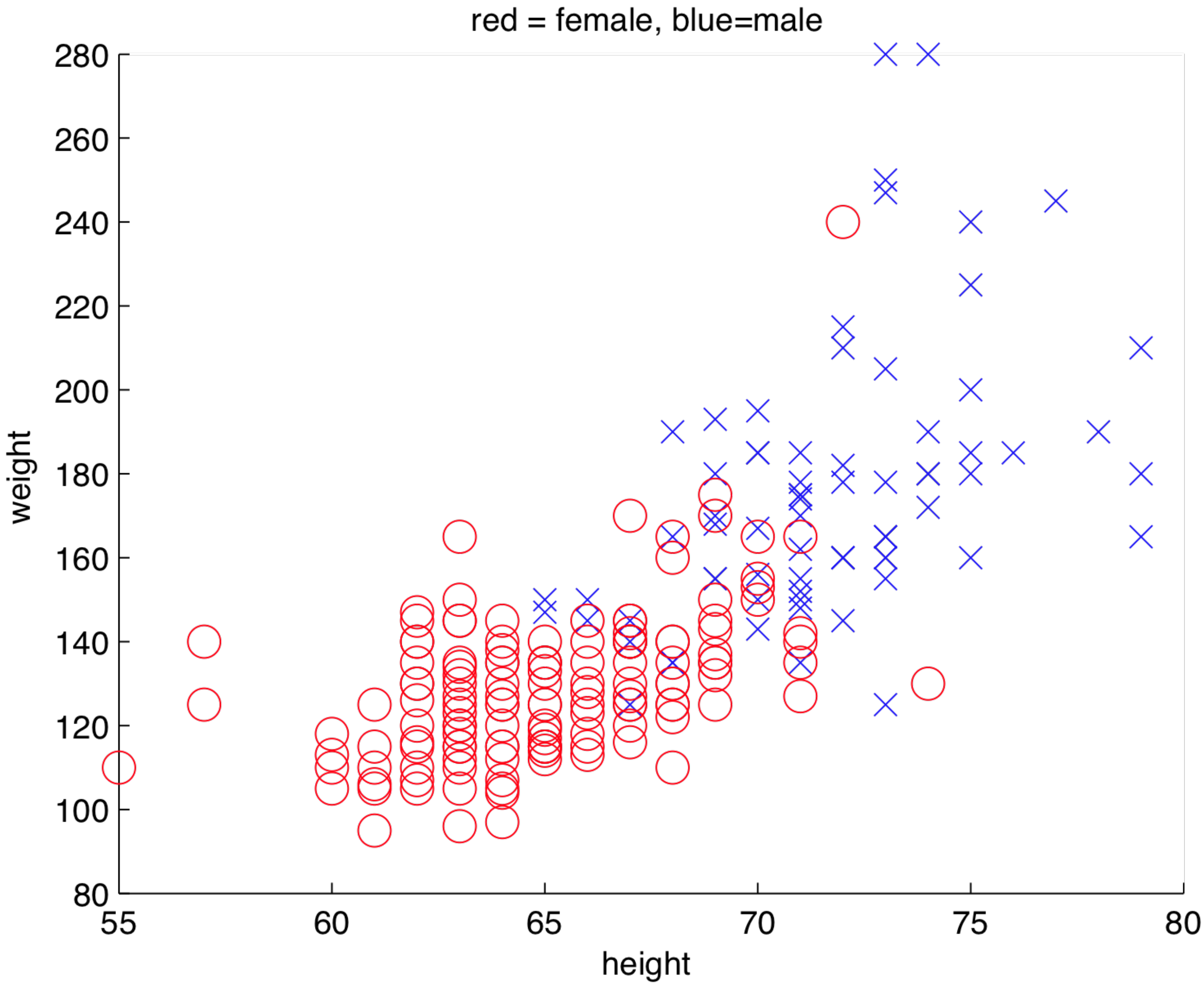


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

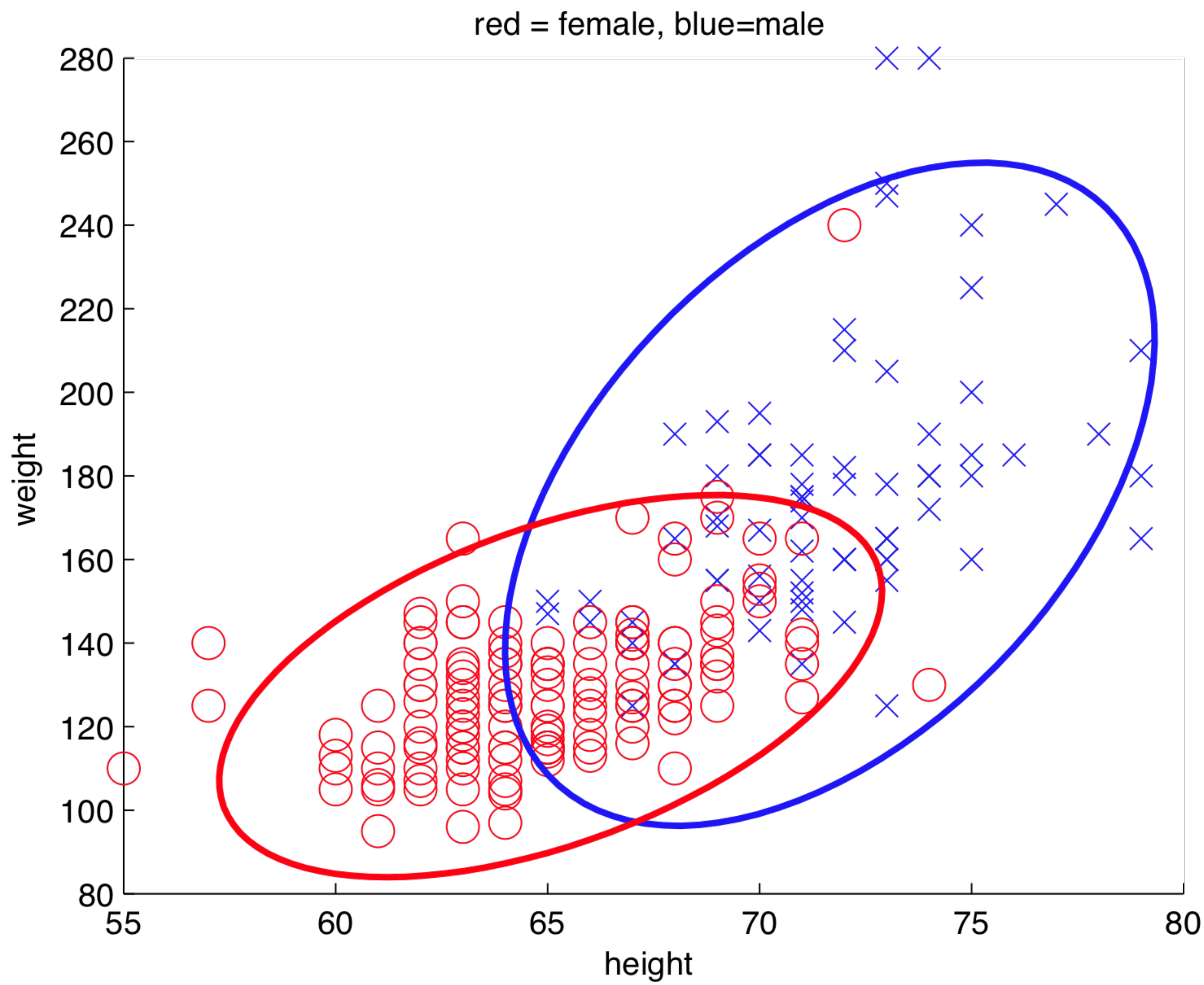
Two-Dimensional Gaussians



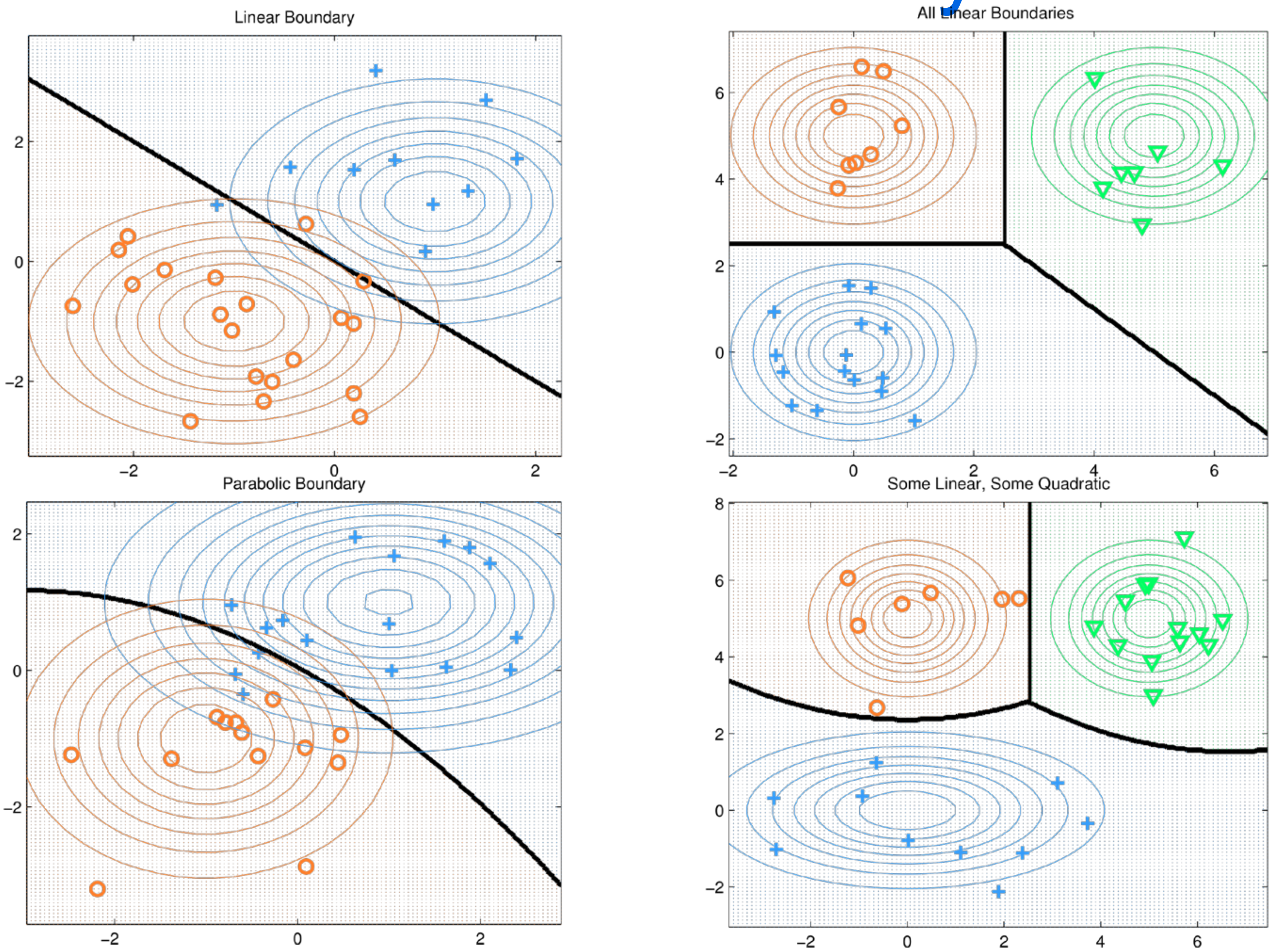
Class-Specific Gaussians



Class-Specific Gaussians

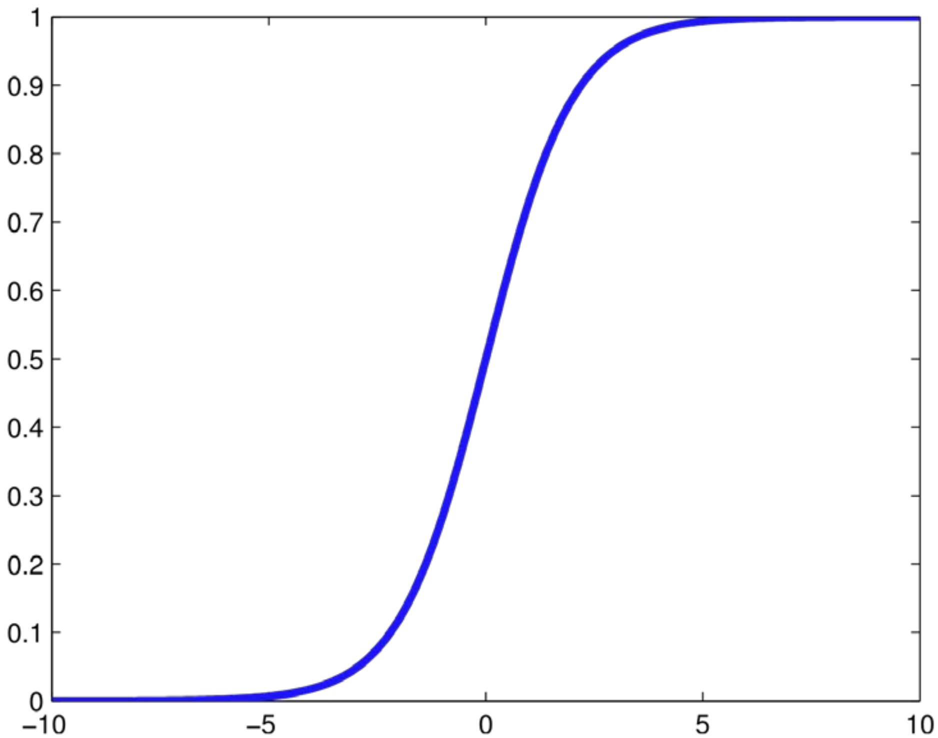


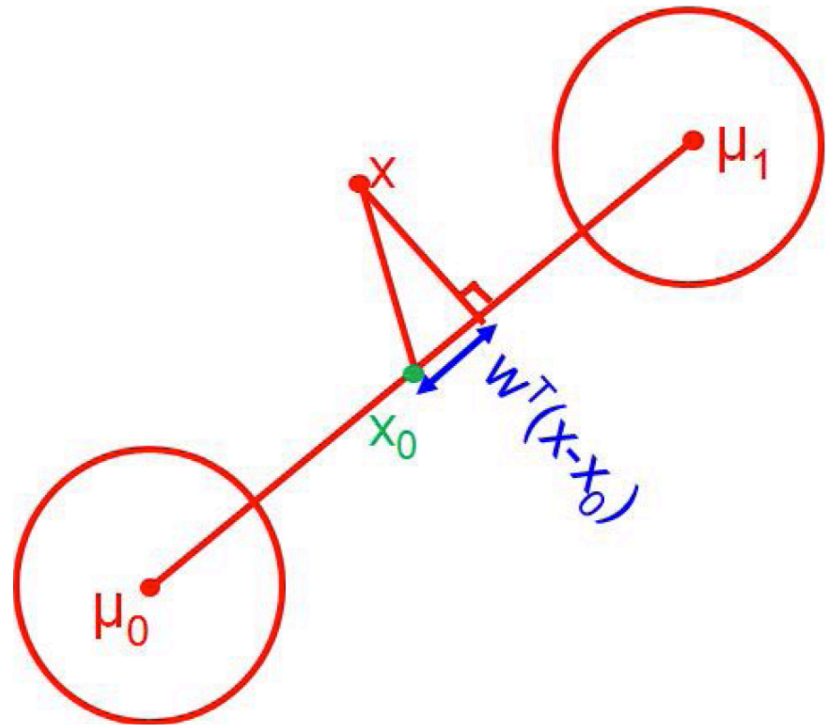
Discriminant Analysis



Binary Discriminant Analysis

Logistic Function





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

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

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

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

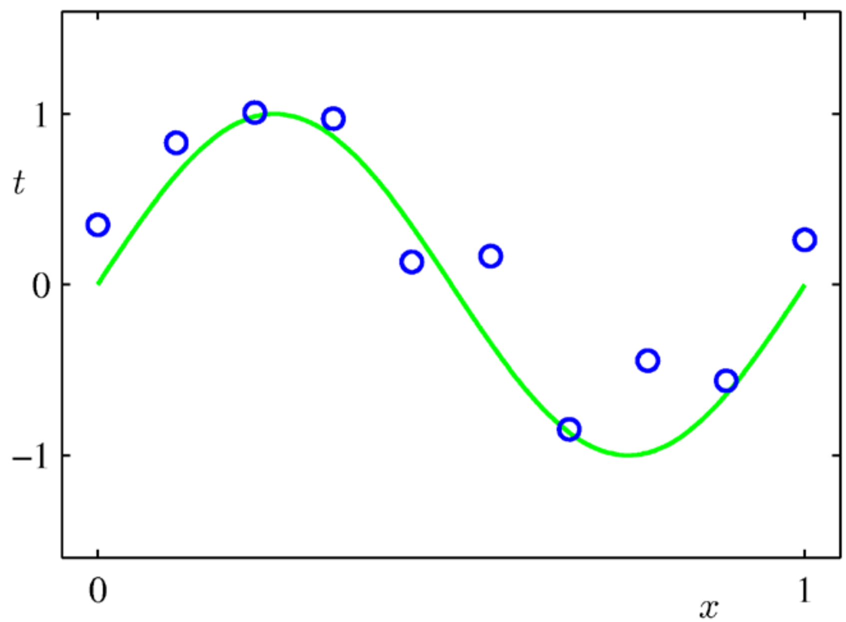
A Change in Direction

	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

Linear Basis Function Models (1)

- Example: Polynomial Curve Fitting



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

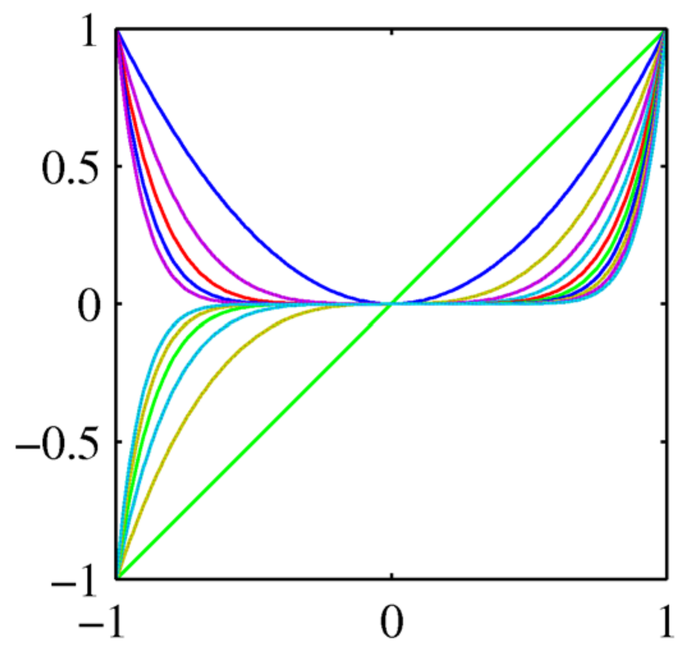
Slides adapted from Bishop's "Pattern Recognition and Machine Learning"
Notation differs from Murphy's "Machine Learning: A Probabilistic Perspective"

Linear Basis Function Models (2)

- Polynomial basis functions:

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

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

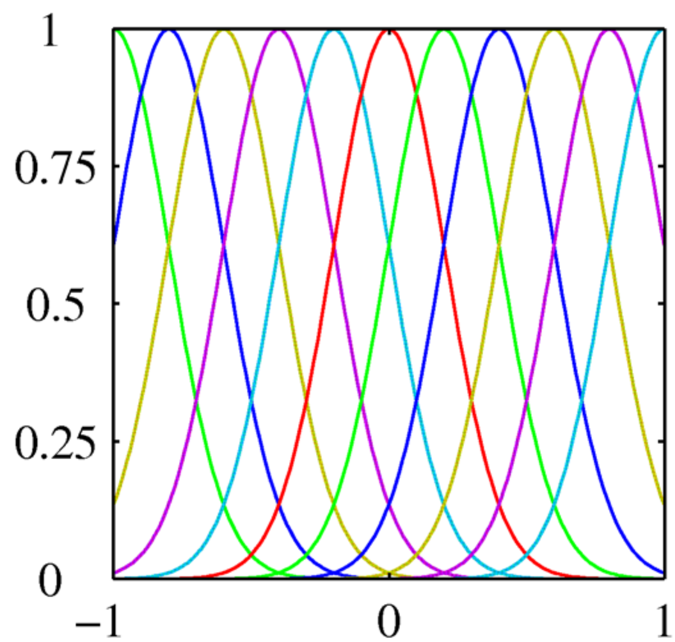


Linear Basis Function Models (3)

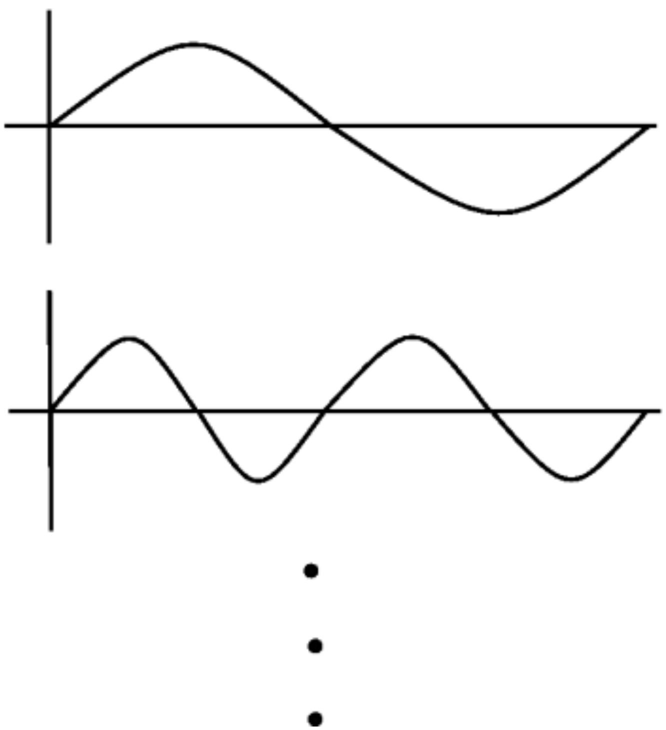
- Gaussian basis functions:

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

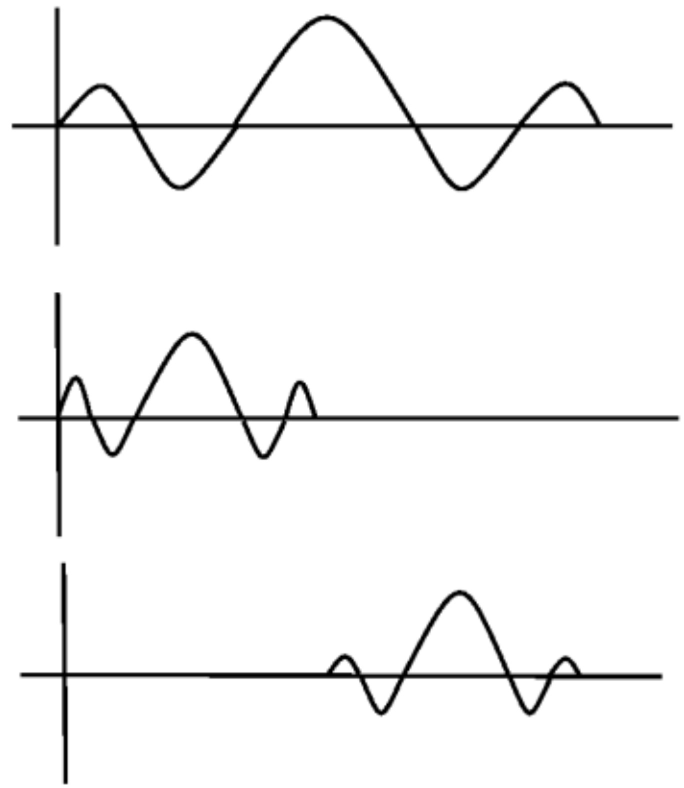
- These are local; a small change in x only affect nearby basis functions. Parameters control location and scale (width).



Linear Basis Function Models (4)

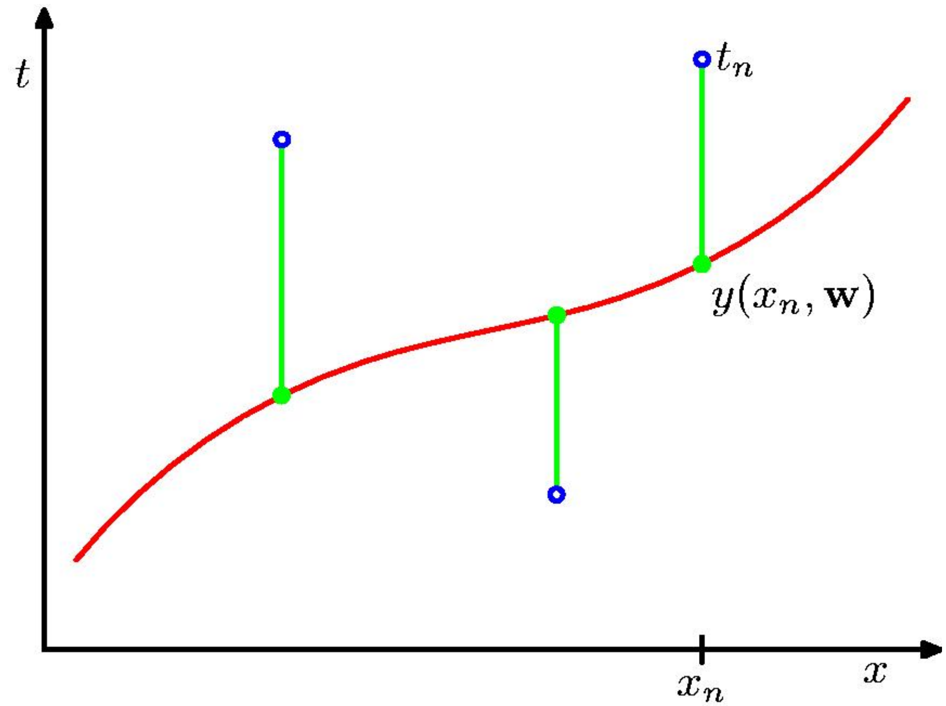


Fourier Basis



Wavelet Basis

Sum-of-Squares Error Function



$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$