Introduction to Machine Learning

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

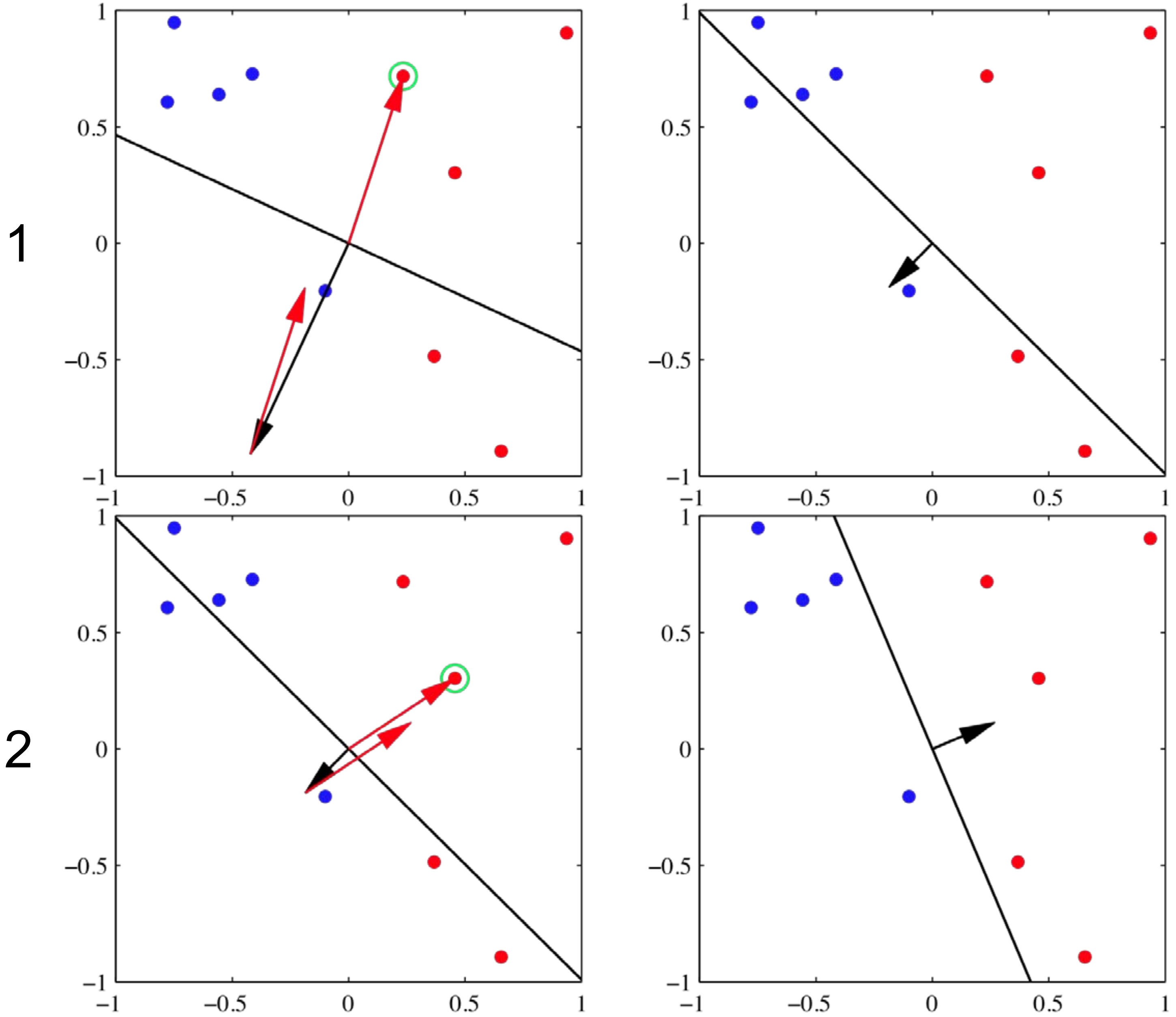
Lecture 16: Perceptron Algorithm, GP Classification, Support Vector Machines

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

Perceptron MARK 1 Computer

Frank Rosenblatt, late 1950s

Perceptron Algorithm Convergence



C. Bishop, Pattern Recognition & Machine Learning

Perceptron Algorithm Properties

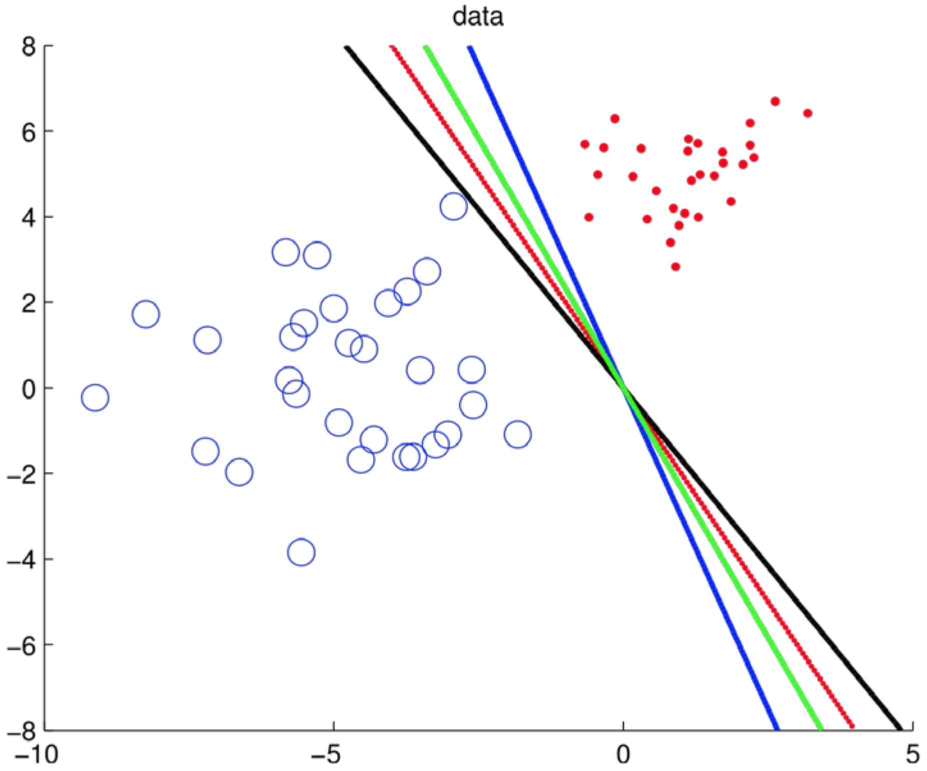
Strengths

- Guaranteed to converge if data linearly separable (in feature space; reduces angle to true separators)
- Easy to construct kernel representation of algorithm

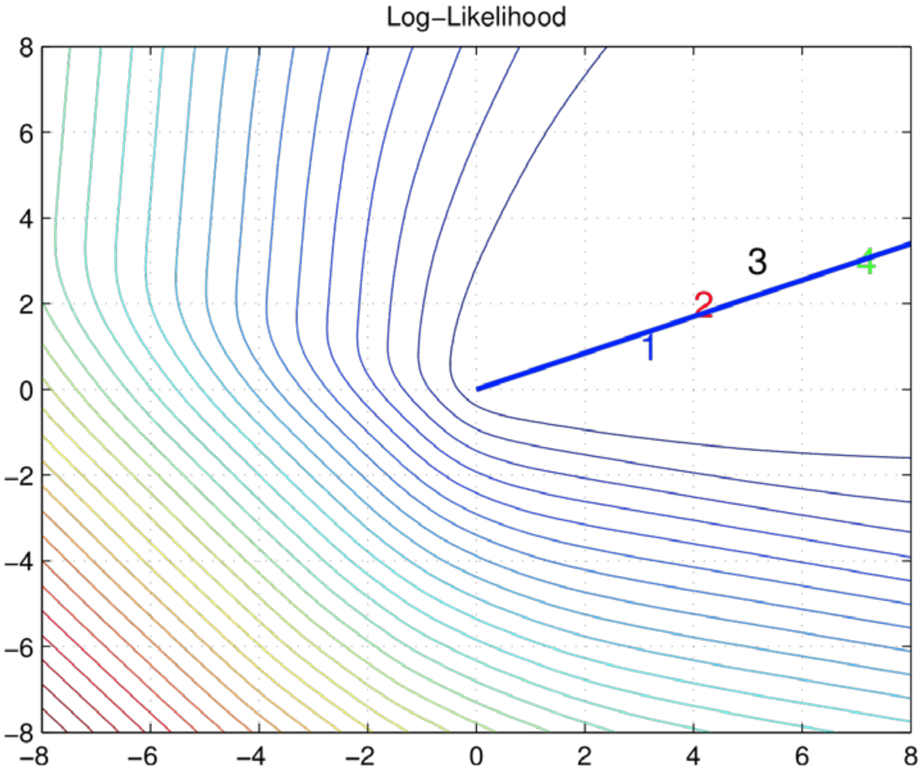
Weaknesses

- May be slow to converge (worst-case performance poor)
- If data not linearly separable, will never converge
- Solution depends on order data visited; no notion of a “best” separating hyperplane
- Non-probabilistic: No measure of confidence in decisions, difficult to generalize to other problems

Logistic Regression Likelihood



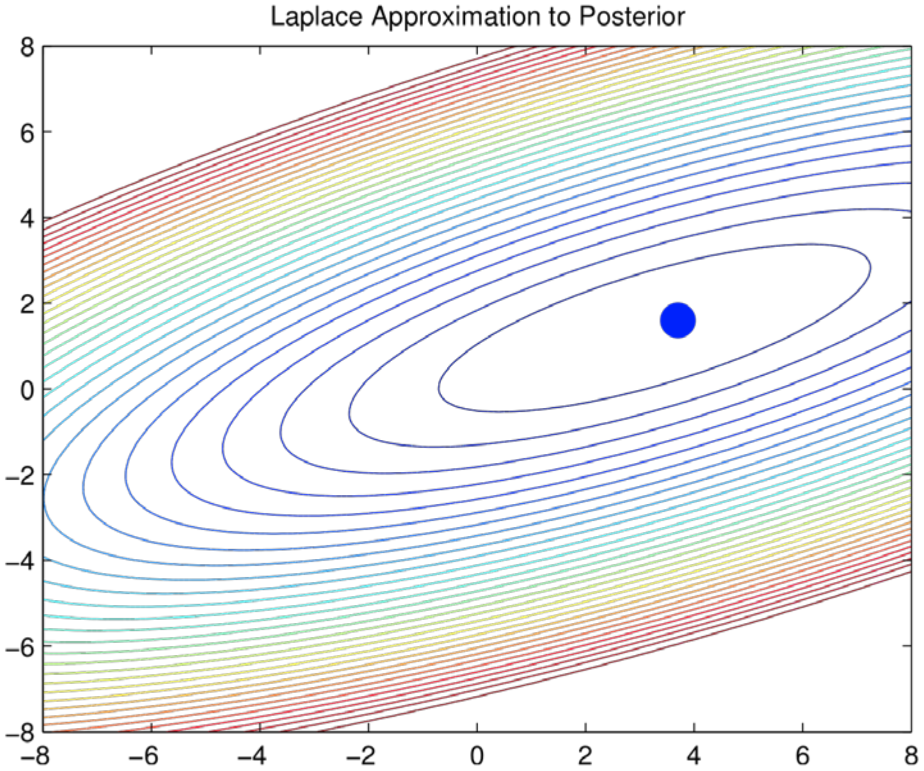
Linearly Separable Data



Log-likelihood Function

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \prod_{i:y_i=1} \frac{e^{\mathbf{w}^T\mathbf{x}_i}}{1+e^{\mathbf{w}^T\mathbf{x}_i}} \prod_{i:y_i=0} \frac{1}{1+e^{\mathbf{w}^T\mathbf{x}_i}} = \exp(\mathbf{w}^T \sum_i y_i \mathbf{x}_i) \prod_{i=1}^{N} (1+e^{\mathbf{w}^T\mathbf{x}_i})^{-1}$$

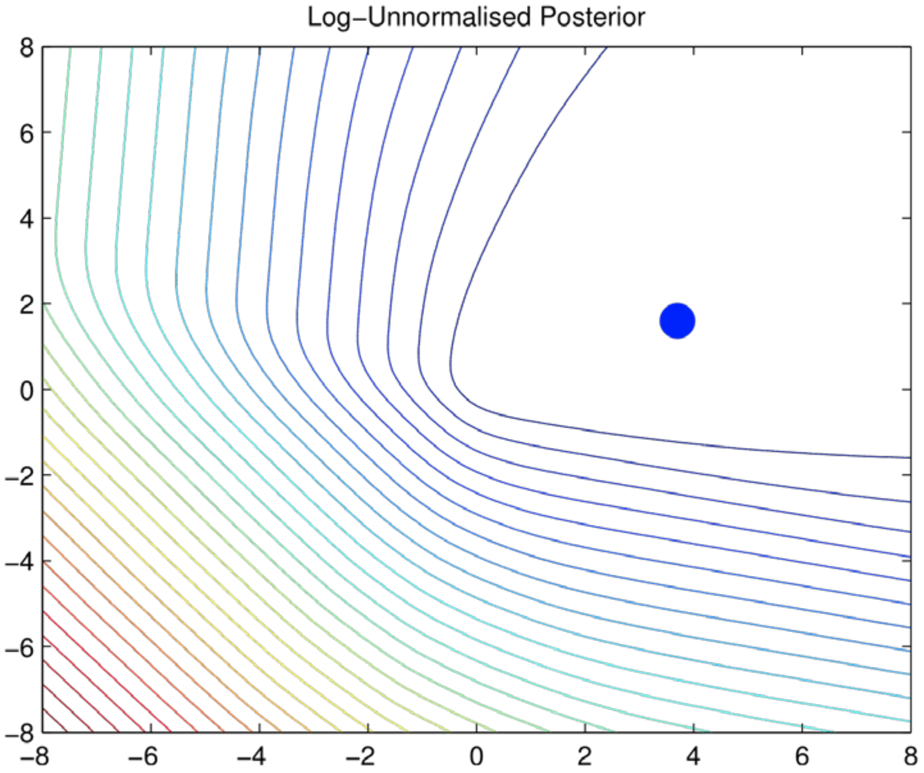
Laplace Approximation of LR Posterior



Laplace Approximation

$$p(\mathbf{w}|\mathcal{D}) \quad \approx \quad \mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}, \mathbf{H}^{-1})$$

$$\mathbf{H} = -\nabla^2 E(\mathbf{w})|_{\hat{\mathbf{w}}}$$

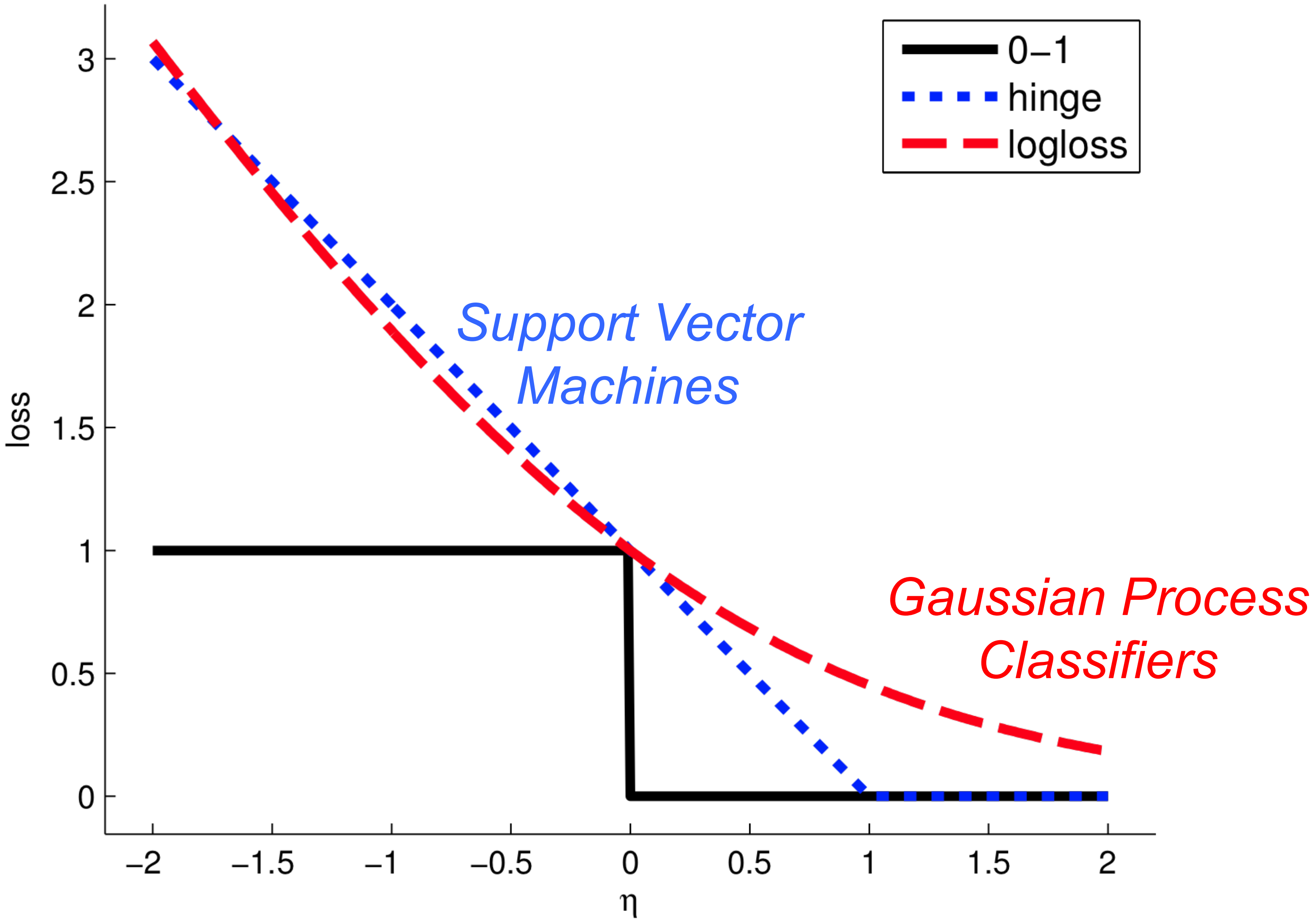


Log Posterior Distribution

$$E(\mathbf{w}) = -(\log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w}))$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} E(\mathbf{w})$$

Losses for Binary Classification



Support Vector Machines

Gaussian Process Classifiers