Slides modified from:

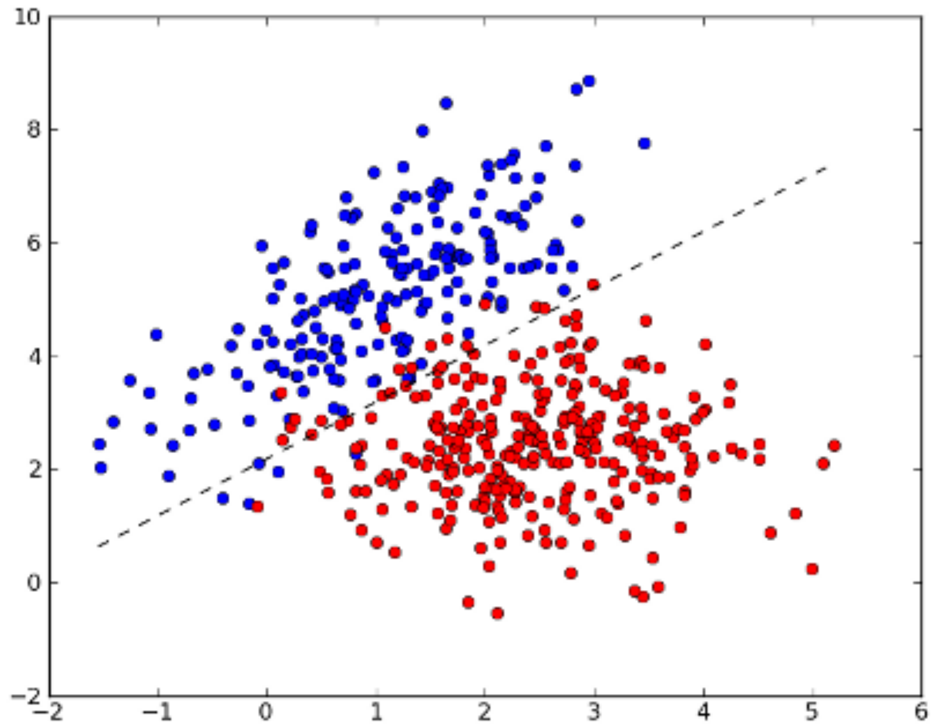
PATTERN RECOGNITION AND MACHINE LEARNING

CHRISTOPHER M. BISHOP

and:

Computer vision: models, learning and inference.

Classification



Example: Gender Classification

Incremental logistic regression

$$Pr(w_i|\mathbf{x}_i) = \text{Bern}_{w_i}\left[\frac{1}{1+\exp[-\phi_0 + \sum_{k=1}^{K}\phi_k \mathrm{f}[\mathbf{x}_i, \boldsymbol{\xi}_k]]}\right]$$

300 arc tan basis functions: $\mathrm{f}[\mathbf{x}_i, \boldsymbol{\xi}_k] = \arctan[\boldsymbol{\xi}_k^T \mathbf{x}_i]$

Results: 87.5% (humans=95%)

Regression vs. Classification

Regression:

$$x \in [-\infty, \infty], t \in [-\infty, \infty]$$

Classification (two classes):

$$x \in [-\infty, \infty], t \in \{0, 1\}$$

Regression vs. Classification

Linear regression model prediction (y real)

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

Classification: y in range (0,1) (posterior probabilities)

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0\right)$$

f: Activation function (nonlinear)

Decision surface: $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \mathrm{constant}$

(Generalized linear models)

Decision Theory

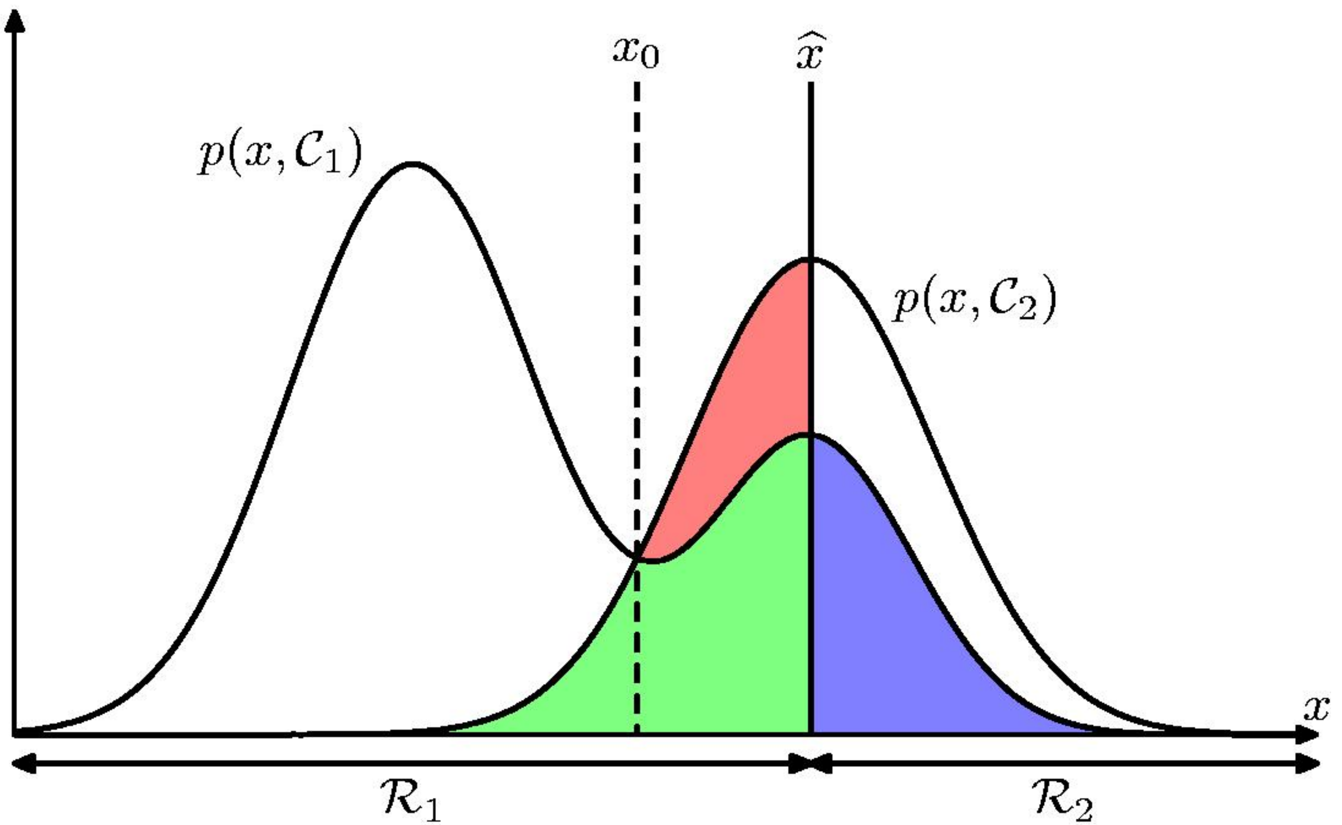
Inference step

Determine either $p(t|\mathbf{x})$ or $p(\mathbf{x}, t)$.

Decision step

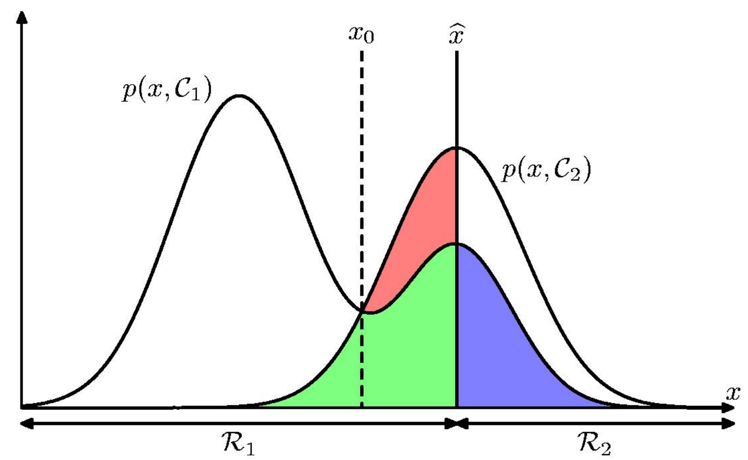
For given x, determine optimal t.

Minimum Misclassification Rate



Minimum Misclassification Rate

$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}.
\end{aligned}
$$



We are free to choose the decision rule that assigns each point x to one of the two classes. This defines the decision regions Rk.

To minimize integrand: $p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$ must be small

Assign x to class for which the posterior $p(\mathcal{C}_k|\mathbf{x})$ is larger!

Minimum Expected Loss

Example: classify medical images as ‘cancer’ or ‘normal’

$$
\begin{array}{ccc}
 & & \text{Decision} \\
 & & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\
\text{Truth} & \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}
\end{array}
$$

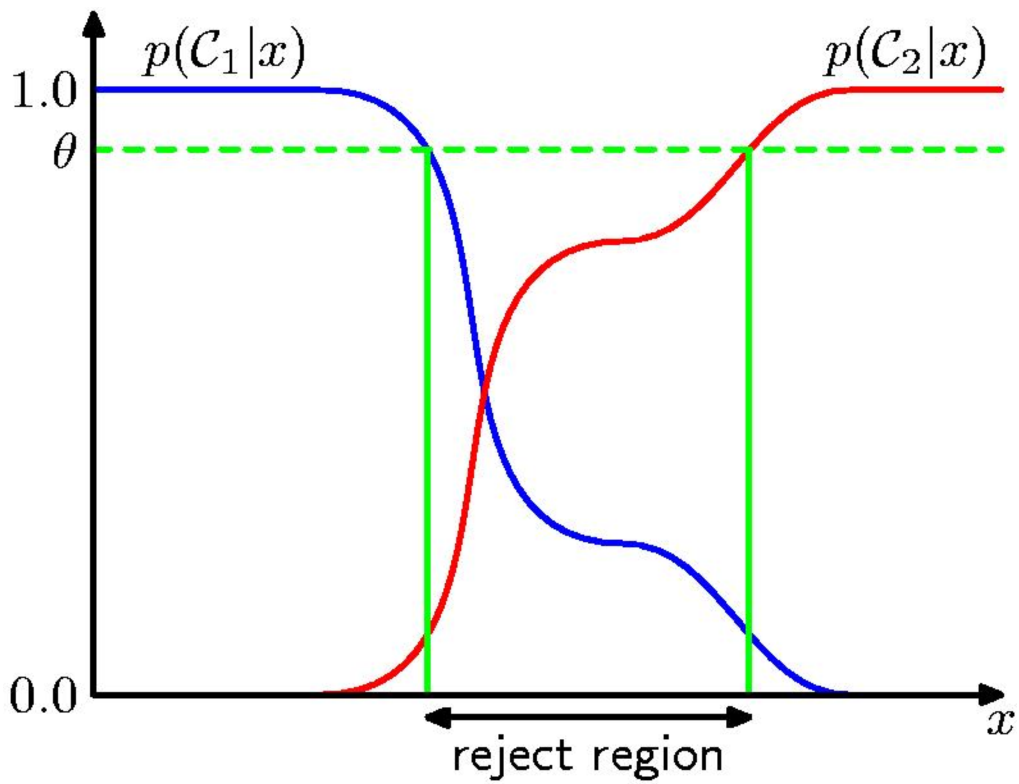
Minimum Expected Loss

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

Regions $\mathcal{R}_j$ are chosen to minimize

$$\mathbb{E}[L] = \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$$

Reject Option



Three strategies

1. Modeling the class-conditional density for each class C_k, and prior, then use Bayes

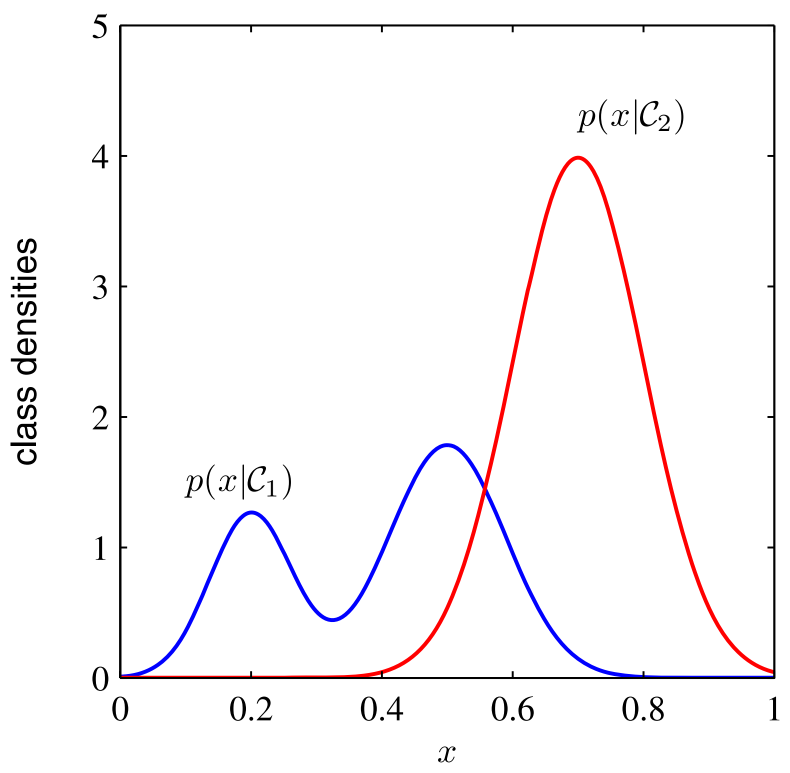
$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

   Equivalently, model joint distribution p(x,C_k) (Models of distribution of input and output are generative models)

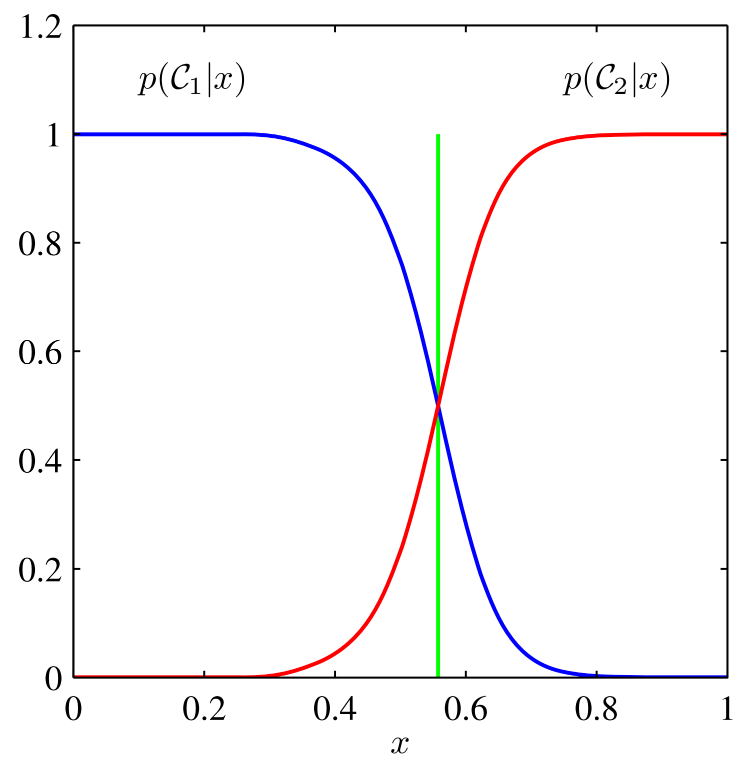
2. First solve the inference problem of determining the posterior class probabilities p(C_k|x), and then subsequently use decision theory to assign each new x to one of the classes (discriminative models)
3. Find discriminant function that directly maps x to class label

Class-conditional density vs. posterior

Class-conditional densities



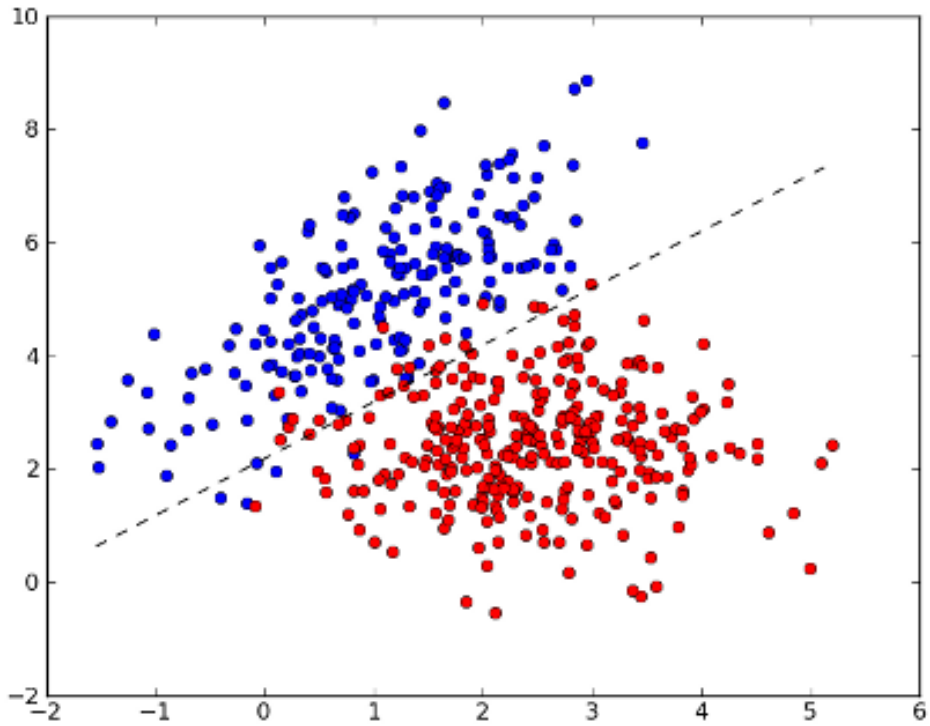
Posterior probabilities



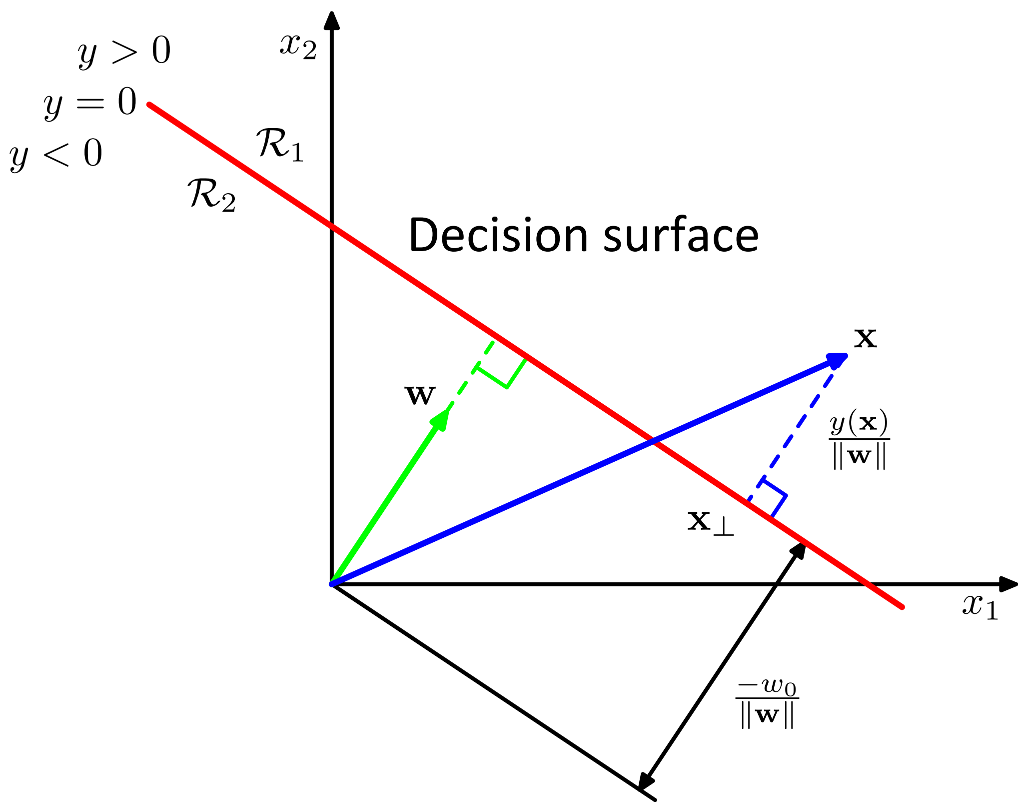
Why Separate Inference and Decision?

- Minimizing risk (loss matrix may change over time)
- Reject option
- Unbalanced class priors
- Combining models

Several dimensions



Several dimensions



$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

weight vector

bias

$\mathcal{C}_1$ if $y(\mathbf{x}) \geqslant 0$

$\mathcal{C}_2$ otherwise.

Perceptron 1

A linear discriminant model by Rosenblatt (1962)

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)$$

with
$$f(a) = \begin{cases} +1, & a \geqslant 0 \\ -1, & a < 0 \end{cases}$$

and feature vector $\boldsymbol{\phi}(\mathbf{x})$

Perceptron 2

Perceptron criterion

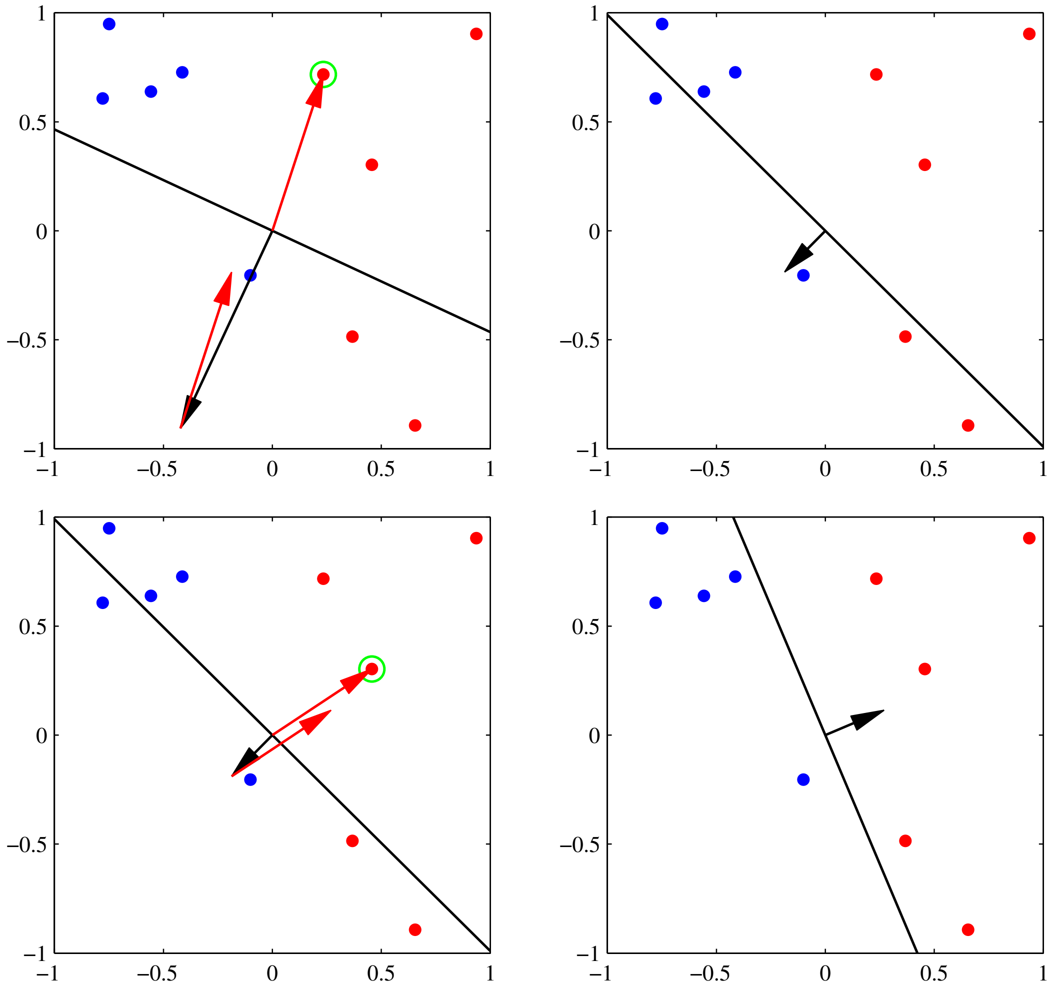
$$E_{\mathrm{P}}(\mathbf{w}) = -\sum_{n \in \mathcal{M}} \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}_n t_n$$

M is set of misclassified patterns

Learning:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_{\mathrm{P}}(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta \boldsymbol{\phi}_n t_n$$

Perceptron 3

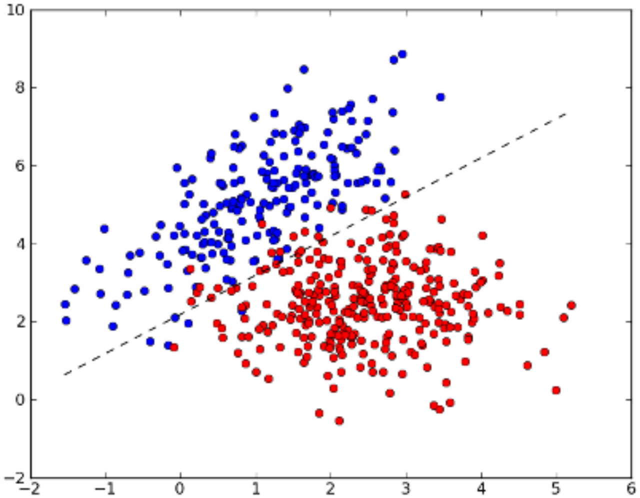


Fisher's linear discriminant 1

Projecting data down to one dimension

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

But how?



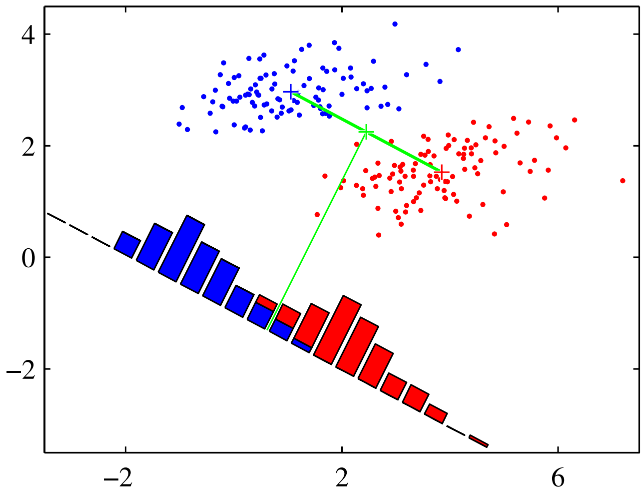
Fisher's linear discriminant 2

Define class means

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n, \qquad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n$$

Try maximize

$$m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)$$



Fisher's linear discriminant 3

Instead, consider: ratio of between class variance to within class variance

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

With

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

Called Fisher criterion. Maximize it!

Fisher's linear discriminant 4

Fisher criterion

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

Rewrite

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}$$

Between-class cov.

$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$$

Within-class cov.

$$\mathbf{S}_{\mathrm{W}} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}$$

Fisher's linear discriminant 5

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}$$

Differentiate with respect to w

$$(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})\mathbf{S}_{\mathrm{W}}\mathbf{w} = (\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})\mathbf{S}_{\mathrm{B}}\mathbf{w}$$

S_B**w** is proportional to ($\mathbf{m}_2$-$\mathbf{m}_1$)

Thus (Fisher's linear discriminant):

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

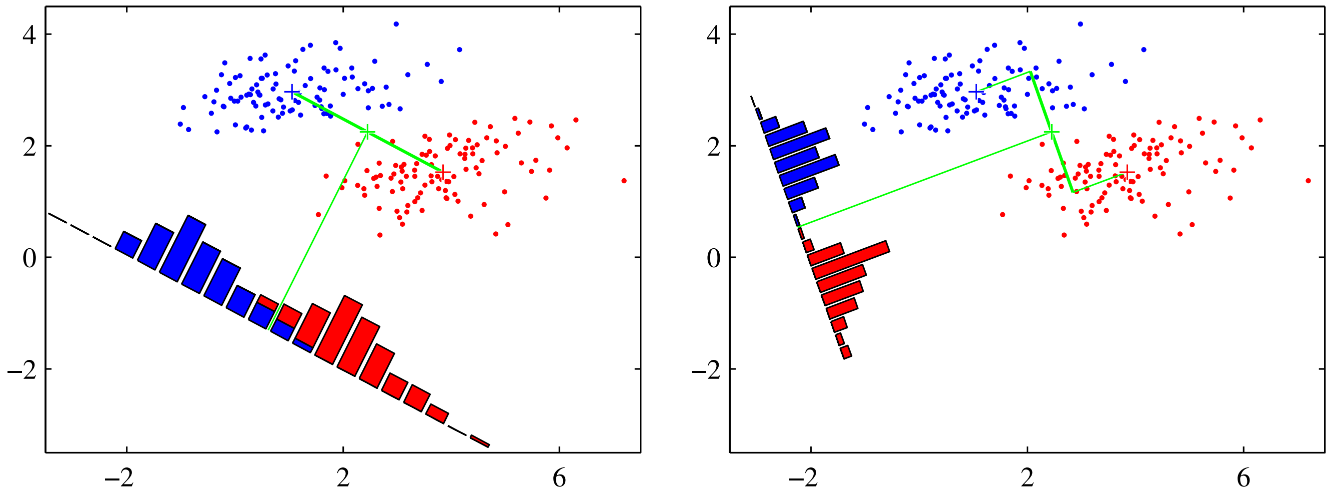
Fisher's linear discriminant 6

Fisher's linear discriminant

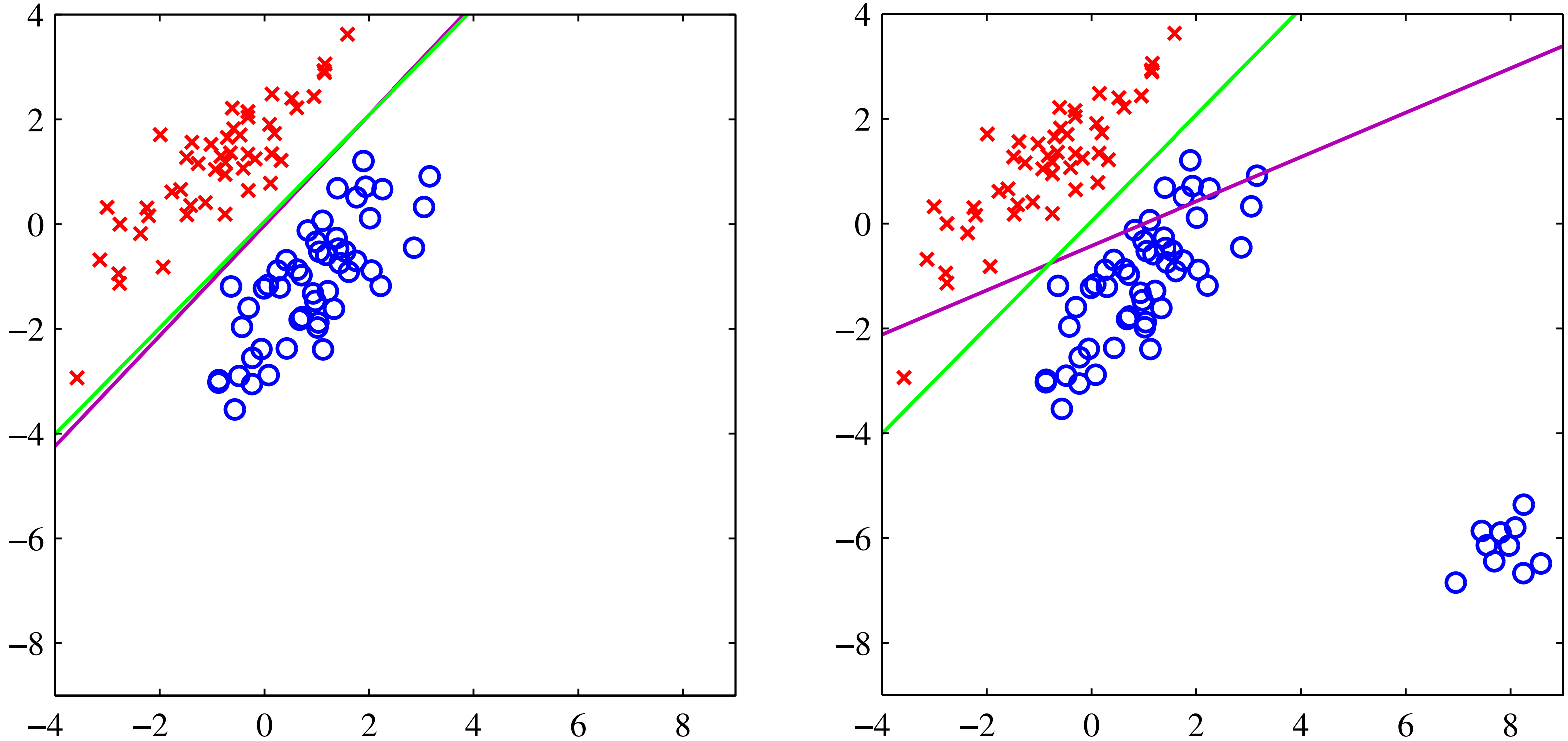
$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

Fisher Criterion

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$



Least squares for classification fails



Use logistic regression instead!

Probabilistic generative models

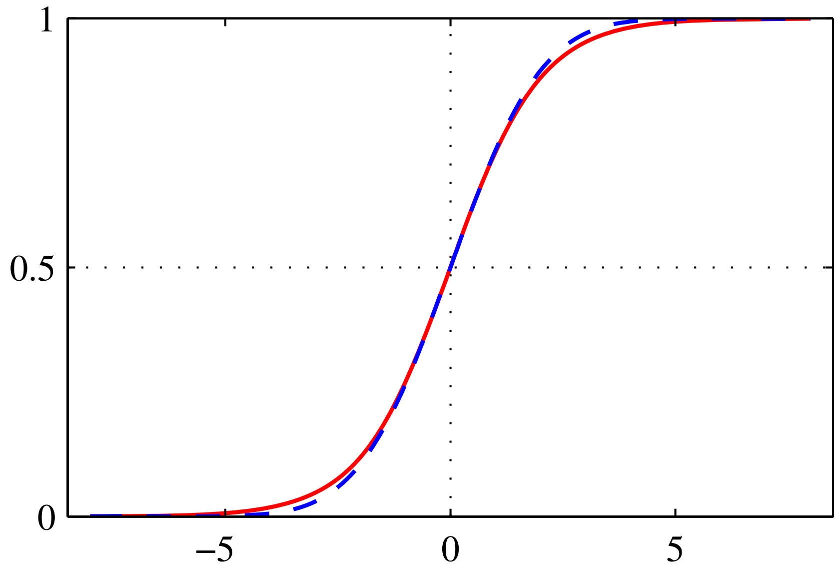
Posterior probability for class C_1 can be written

$$\begin{aligned} p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\ &= \frac{1}{1+\exp(-a)} = \sigma(a) \end{aligned}$$

with $a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$

and the logistic sigmoid function $\sigma(a) = \frac{1}{1+\exp(-a)}$

Logistic sigmoid function



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

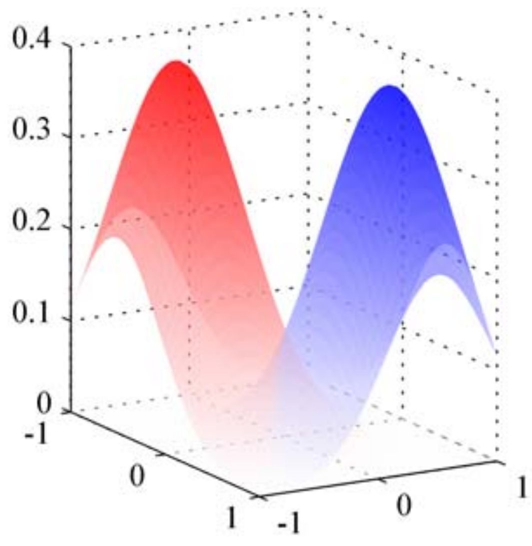
Gaussian class-conditional densities (different means, but equal variances)

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

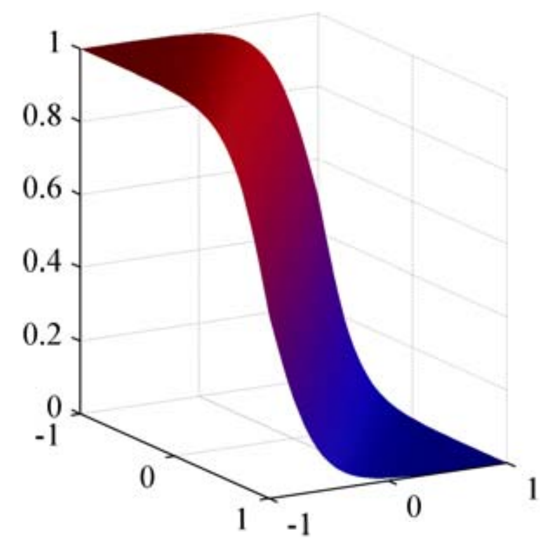
Yields $$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$$

with
$$\begin{aligned}
\mathbf{w} &= \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}
\end{aligned}$$

Linear function of x in argument of logistic sigmoid class-conditional densities



posterior probability $p(C_1|x)$\



Probabilistic discriminative models: Logistic regression - A model of classification

Posterior probability of class C_1 can be written as a logistic sigmoid acting on a linear function of the feature vector ϕ

$$p(\mathcal{C}_1|\boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right)$$

$\sigma(\bullet)$ is sigmoid function $$\sigma(a) = \frac{1}{1+\exp(-a)}$$

Also: $$p(\mathcal{C}_2|\boldsymbol{\phi}) = 1 - p(\mathcal{C}_1|\boldsymbol{\phi})$$

M parameter (M(M+5)/2+1 for generative model)

Maximum likelihood logistic regression (1)

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n}$$

With $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$ and $y_n = p(\mathcal{C}_1|\phi_n)$

for a data set $\{\phi_n, t_n\}$, where $t_n \in \{0,1\}$ and $\phi_n = \phi(x_n)$, with $n = 1, \ldots, N$

Error function

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

with $y_n = \sigma(\mathbf{w}^T \phi_n)$

Maximum likelihood logistic regression (2)

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

Gradient of the error function with respect to **w**

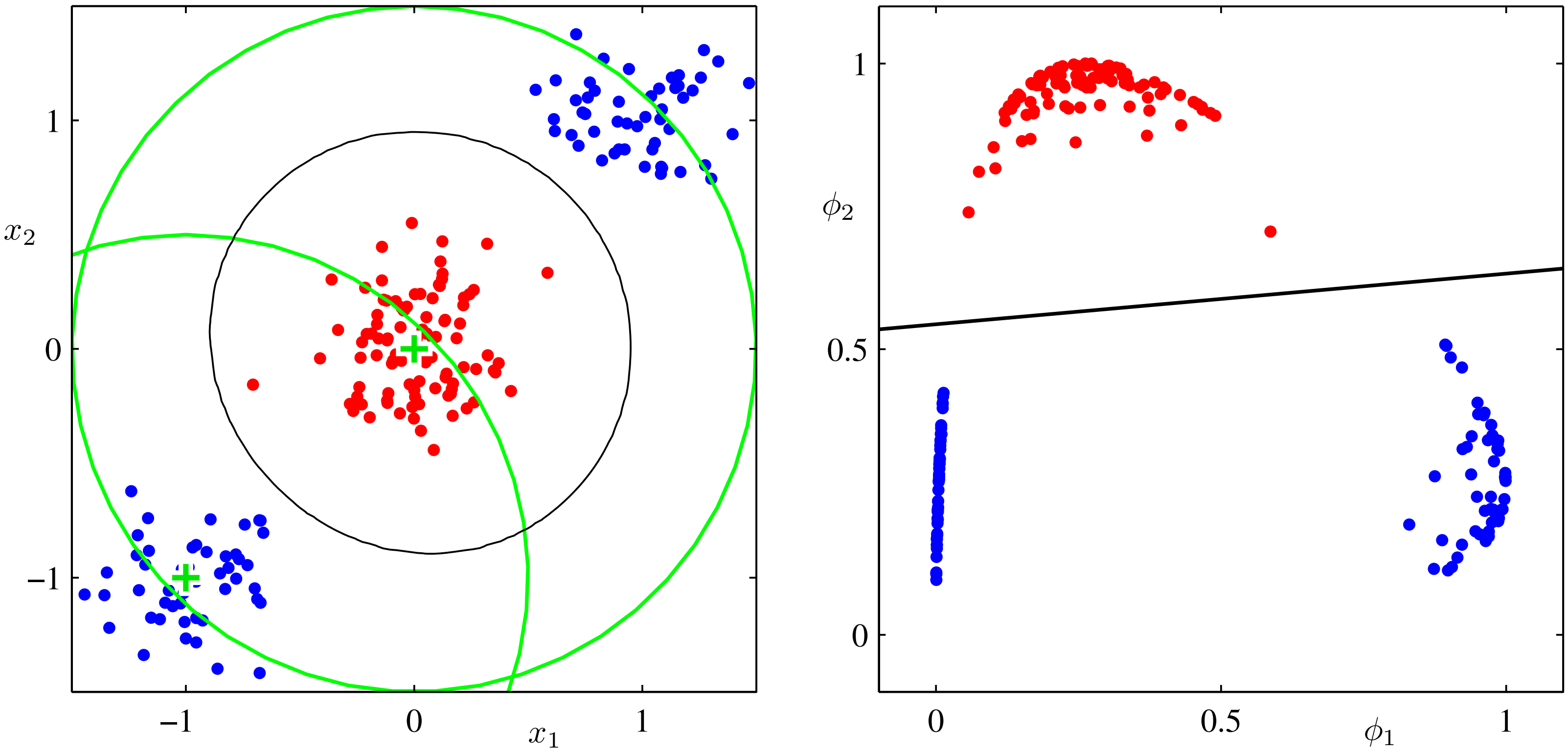
$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi}_n$$

Stochastic gradient update rule

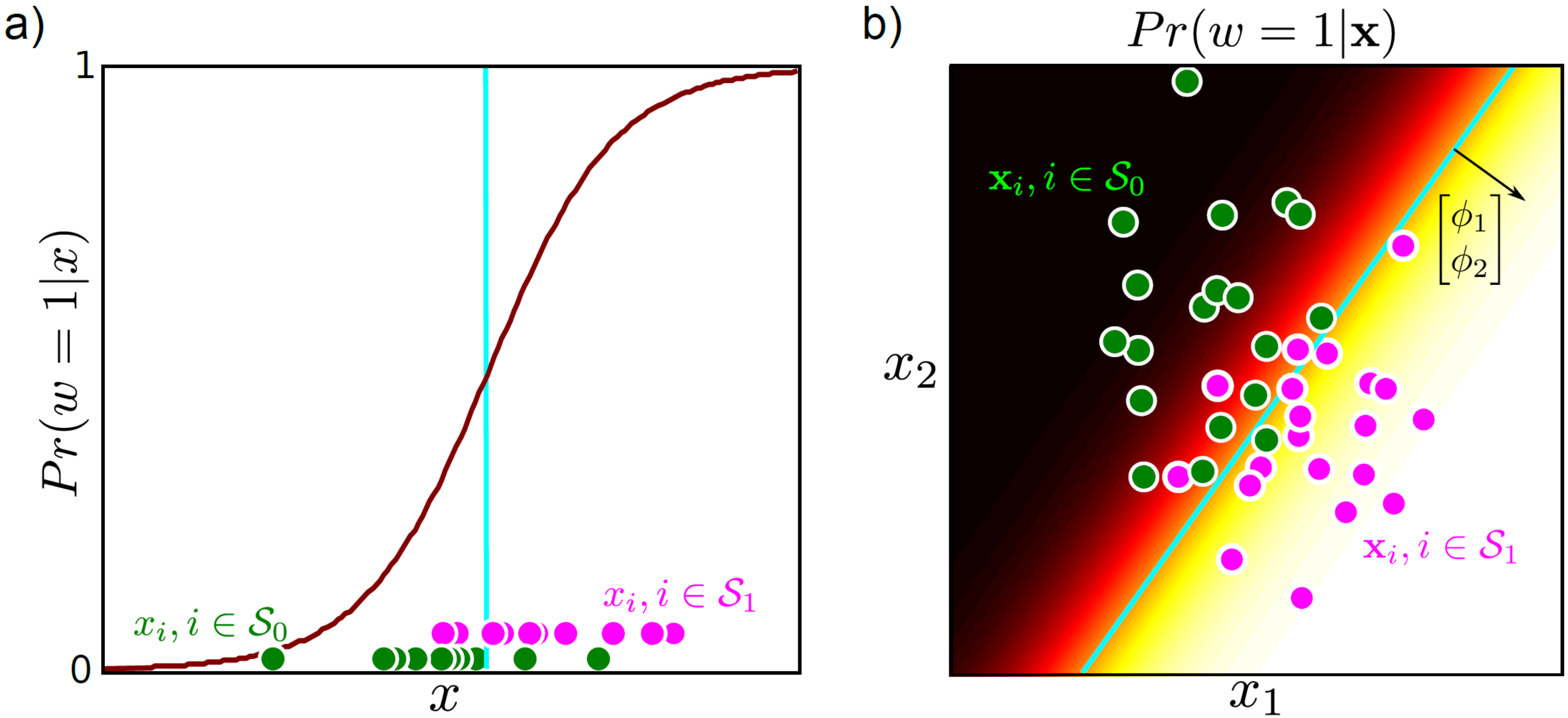
$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n$$

τ: iteration number; η: learning rate parameter

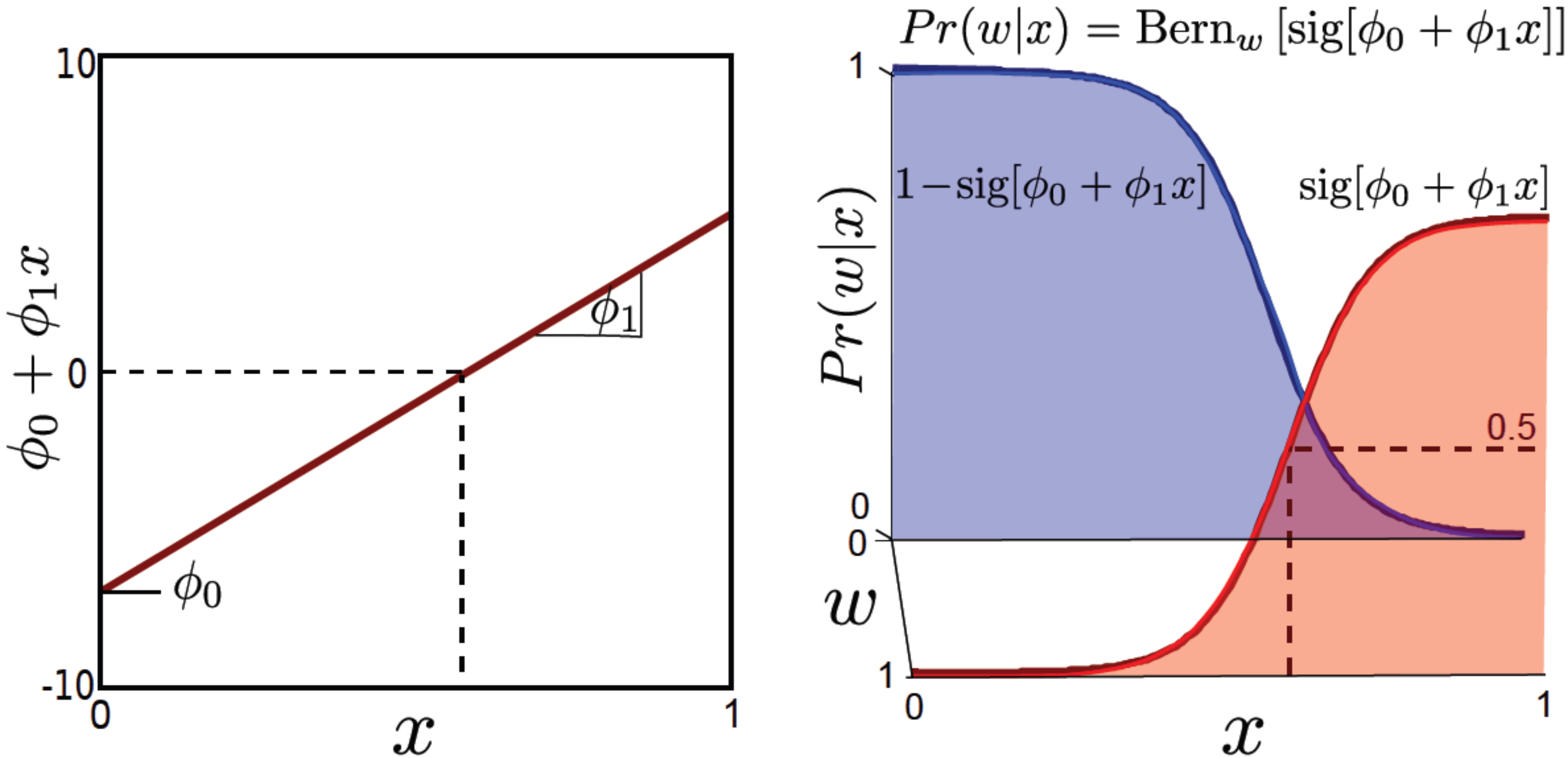
Example



Logistic regression



$$Pr(w|\boldsymbol{\phi}, \mathbf{x}) = \text{Bern}_w \left[\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}]} \right]$$



Two parameters
$$\boldsymbol{\theta} = \{\phi_0, \phi_1\}$$

Learning by standard methods (ML,MAP, Bayesian)
Inference: Just evaluate Pr(w|x)

Neater Notation

$$Pr(w|\phi_0, \boldsymbol{\phi}, \mathbf{x}) \quad = \quad \text{Bern}_w\left[\text{sig}[a]\right]$$

To make notation easier to handle, we

- Attach a 1 to the start of every data vector

$$\mathbf{x}_i \leftarrow [1 \quad \mathbf{x}_i^T]^T$$

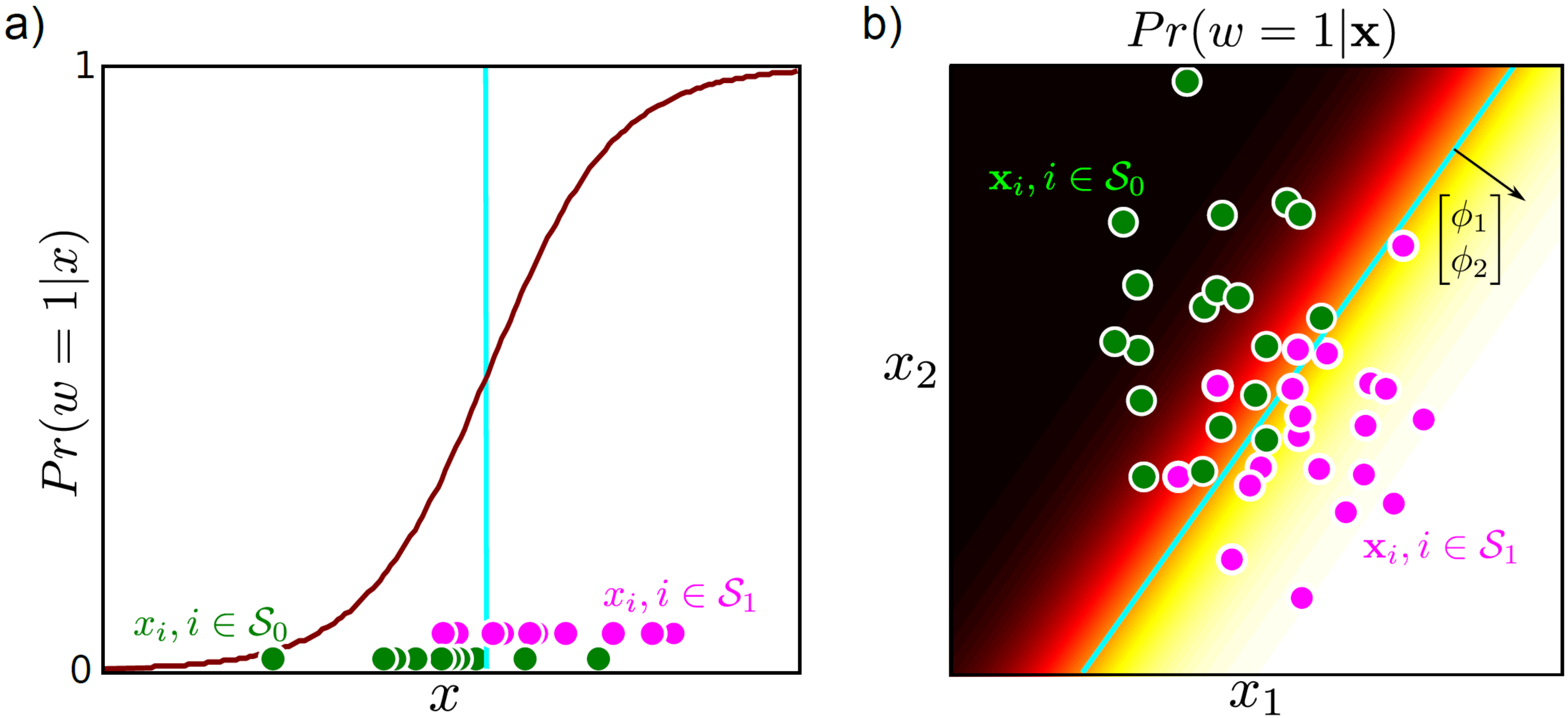
- Attach the offset to the start of the gradient vector $\boldsymbol{\phi}$

$$\boldsymbol{\phi} \leftarrow [\phi_0 \quad \boldsymbol{\phi}^T]^T$$

New model:

$$Pr(w|\boldsymbol{\phi}, \mathbf{x}) = \text{Bern}_w\left[\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}]}\right]$$

Logistic regression



$$Pr(w|\boldsymbol{\phi}, \mathbf{x}) = \text{Bern}_w\left[\frac{1}{1+\exp[-\boldsymbol{\phi}^T\mathbf{x}]}\right]$$