Bayesian Model Comparison (1)

How do we choose the 'right' model?

Assume we want to compare models M_i, i=1, …,L, using data D; this requires computing

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i).$$

Posterior Prior *Model evidence* or *marginal likelihood*

Bayes Factor: ratio of evidence for two models

$$\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}$$

Bayesian Model Comparison (2)

Having computed p(M_ijD), we can compute the predictive (mixture) distribution

$$p(t|\mathbf{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i|\mathcal{D}).$$

A simpler approximation, known as *model selection*, is to use the model with the highest evidence.

Bayesian Model Comparison (3)

For a model with parameters w, we get the model evidence by marginalizing over w

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)\,\mathrm{d}\mathbf{w}.$$

Note that

$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)}$$

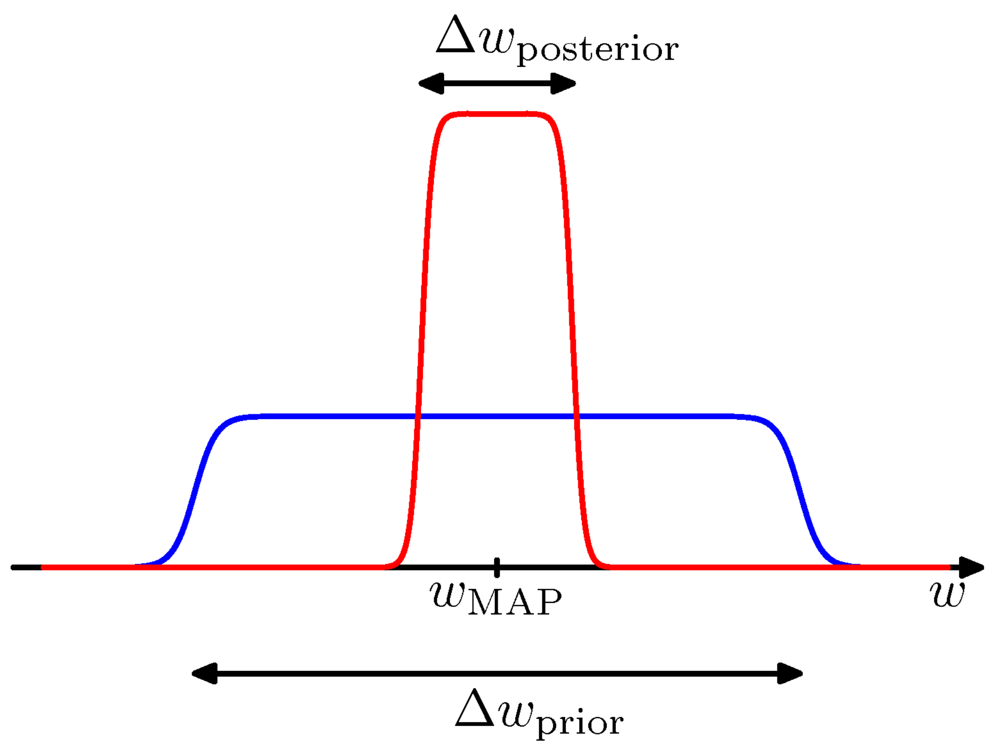
Bayesian Model Comparison (4)

For a given model with a single parameter, w, consider the approximation

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\,\mathrm{d}w$$

$$\simeq \quad p(\mathcal{D}|w_{\mathrm{MAP}})\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}$$

where the posterior is assumed to be sharply peaked.



Bayesian Model Comparison (5)

Taking logarithms, we obtain

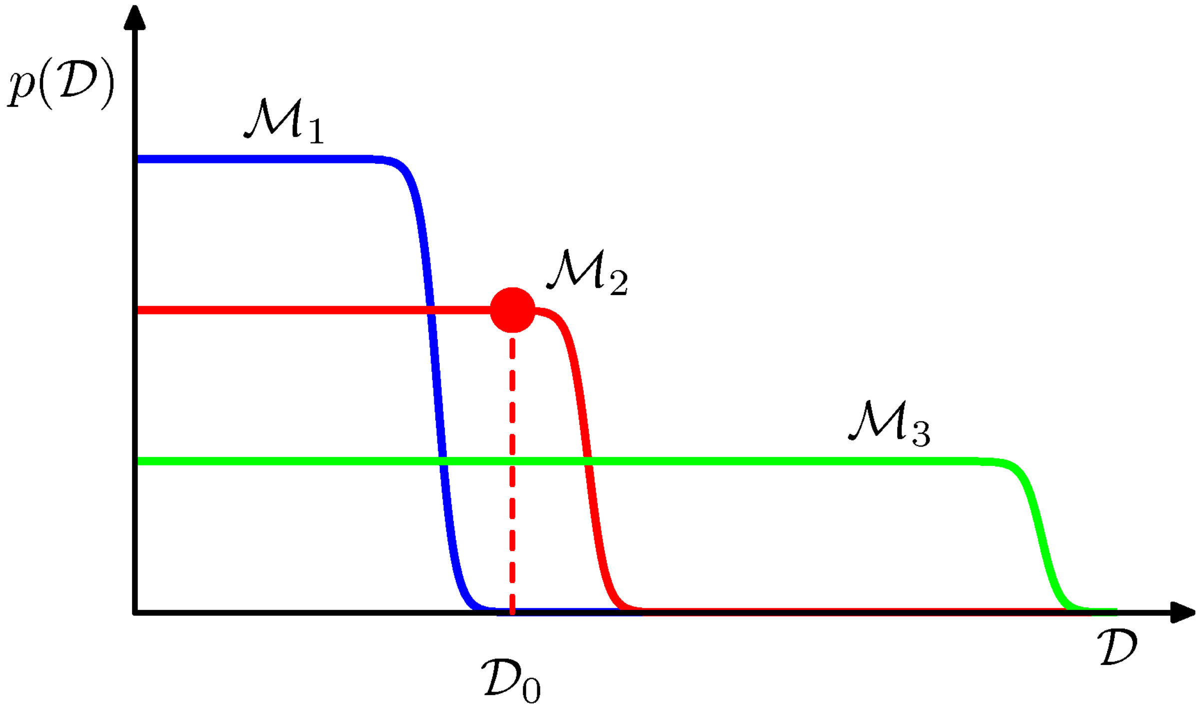
$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\mathrm{MAP}}) + \underbrace{\ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right)}_{\text{Negative}}.$$

With M parameters, all assumed to have the same ratio $\Delta w_{\mathrm{posterior}}/\Delta w_{\mathrm{prior}}$, we get

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathbf{w}_{\mathrm{MAP}}) + \underbrace{M\ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right)}_{\text{Negative and linear in M.}}.$$

Bayesian Model Comparison (6)

Matching data and model complexity



The Evidence Approximation (1)

The fully Bayesian predictive distribution is given by

$$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta)p(\alpha, \beta|\mathbf{t})\,\mathrm{d}\mathbf{w}\,\mathrm{d}\alpha\,\mathrm{d}\beta$$

but this integral is intractable. Approximate with

$$p(t|\mathbf{t}) \simeq p\left(t|\mathbf{t}, \widehat{\alpha}, \widehat{\beta}\right) = \int p\left(t|\mathbf{w}, \widehat{\beta}\right) p\left(\mathbf{w}|\mathbf{t}, \widehat{\alpha}, \widehat{\beta}\right)\,\mathrm{d}\mathbf{w}$$

where $\left(\widehat{\alpha}, \widehat{\beta}\right)$ is the mode of $p(\alpha, \beta|\mathbf{t})$, which is assumed to be sharply peaked; a.k.a. *empirical Bayes, type II* or *generalized maximum likelihood,* or *evidence approximation*.

The Evidence Approximation (2)

From Bayes' theorem we have

$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta)p(\alpha, \beta)$$

and if we assume p(α,β) to be flat we see that

$$\begin{aligned} p(\alpha, \beta|\mathbf{t}) &\propto p(\mathbf{t}|\alpha, \beta) \\ &= \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w}. \end{aligned}$$

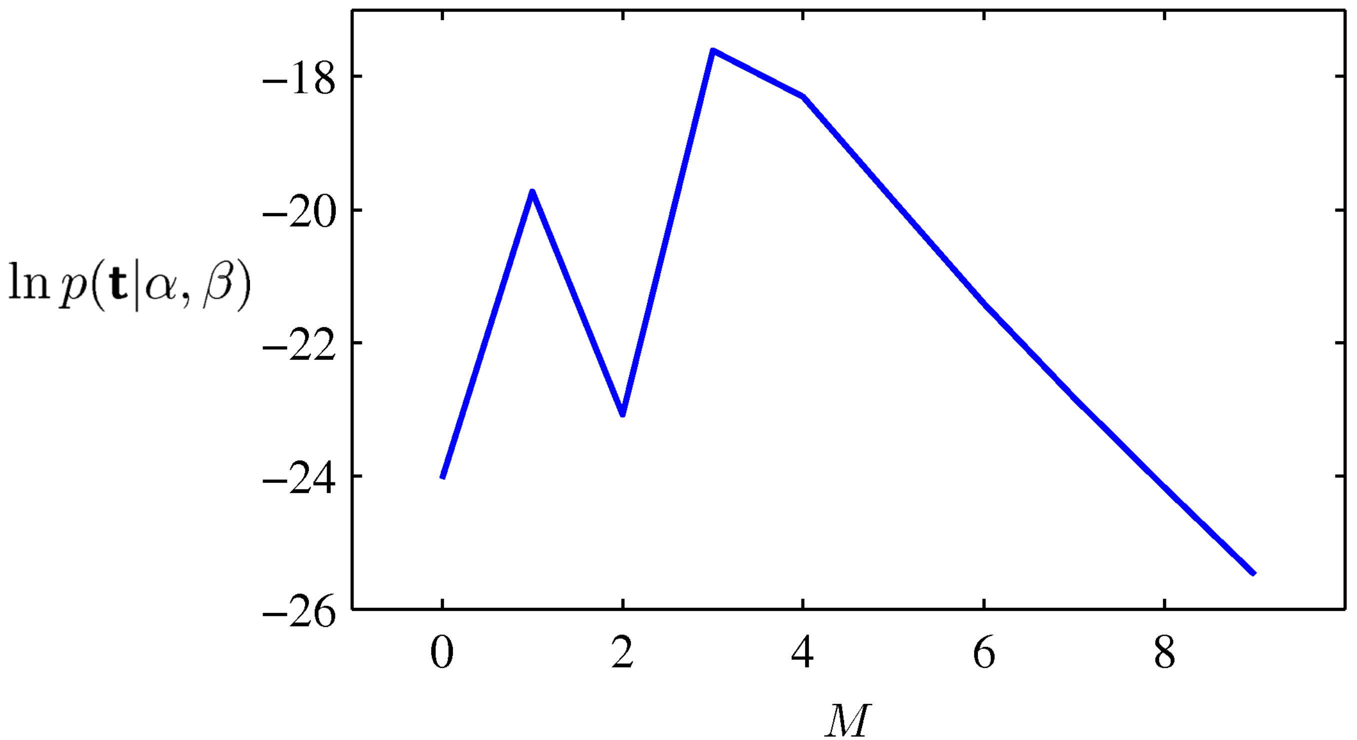
General results for Gaussian integrals give

$$\ln p(\mathbf{t}|\alpha, \beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) + \frac{1}{2}\ln|\mathbf{S}_N| - \frac{N}{2}\ln(2\pi).$$

The Evidence Approximation (3)

Example: sinusoidal data, M th degree polynomial,

$\alpha = 5 \times 10^{-3}$



Regression vs. Classification

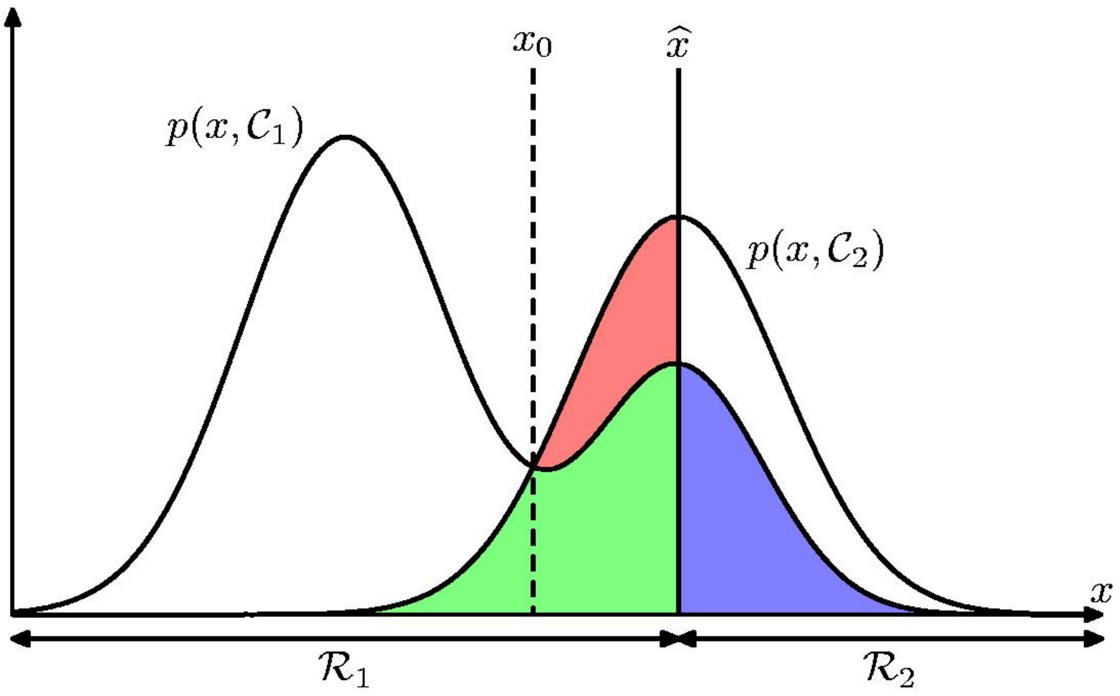
Regression:

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

Classification:

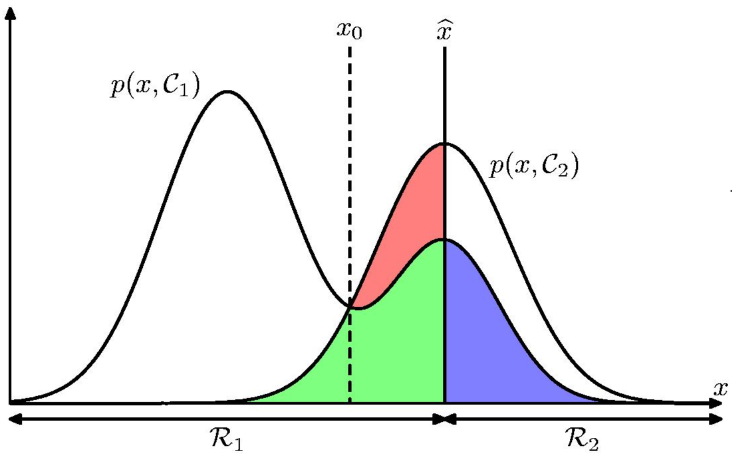
$$x \in [-\infty, \infty], t \in \{0, 1\}$$

Minimum Misclassification Rate



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

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.

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!

Three strategies

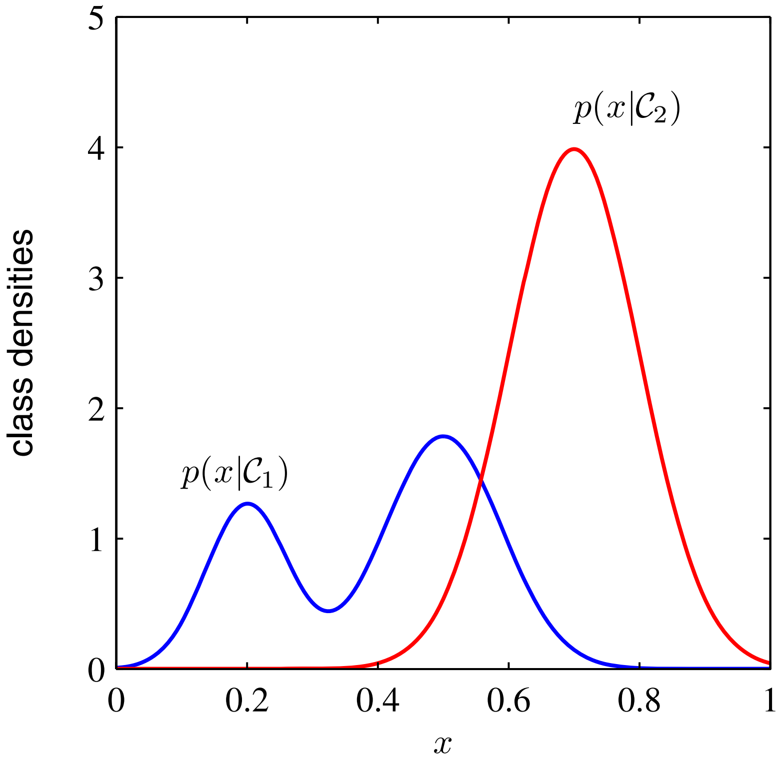
1. Modeling the class-conditional density for each class C_k, and prior, then use Bayes

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

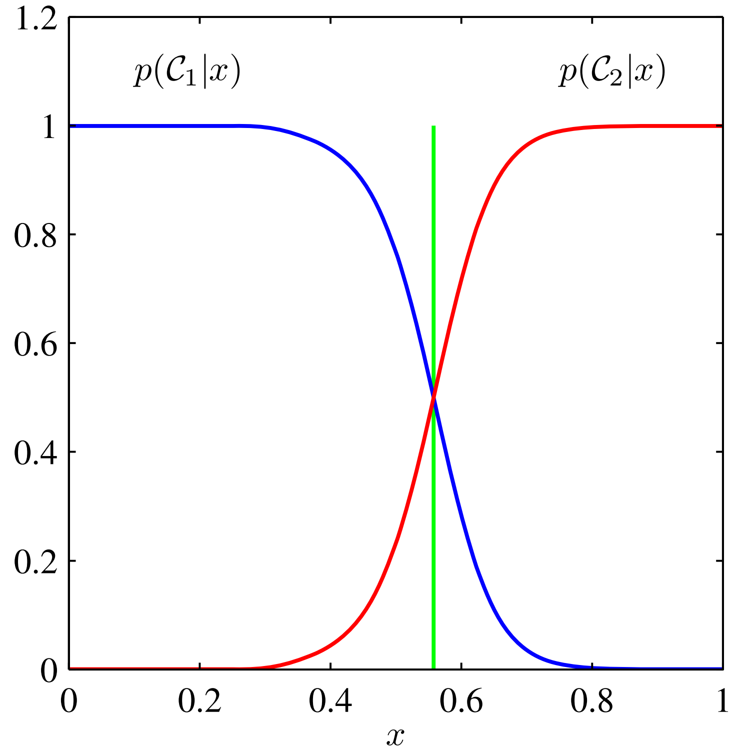
2. First solve the inference problem of determining the posterior class probabilities p(C_k|x), and then subsequently use decision theory to assign each new x to one of the classes
3. Find discriminant function that directly maps x to class label

Class-conditional density vs. posterior

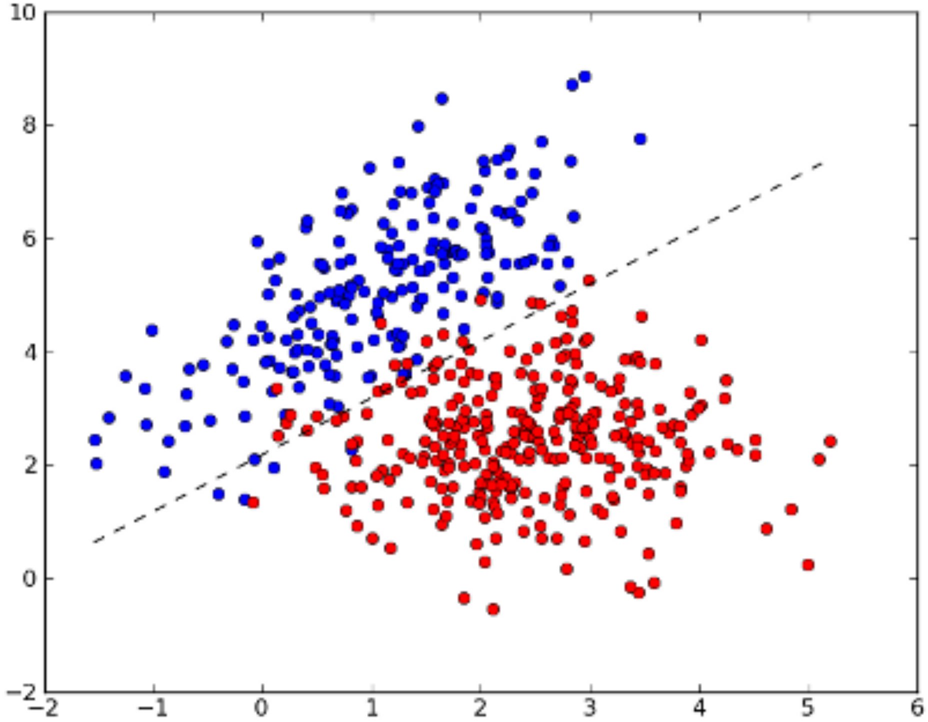
Class-conditional densities



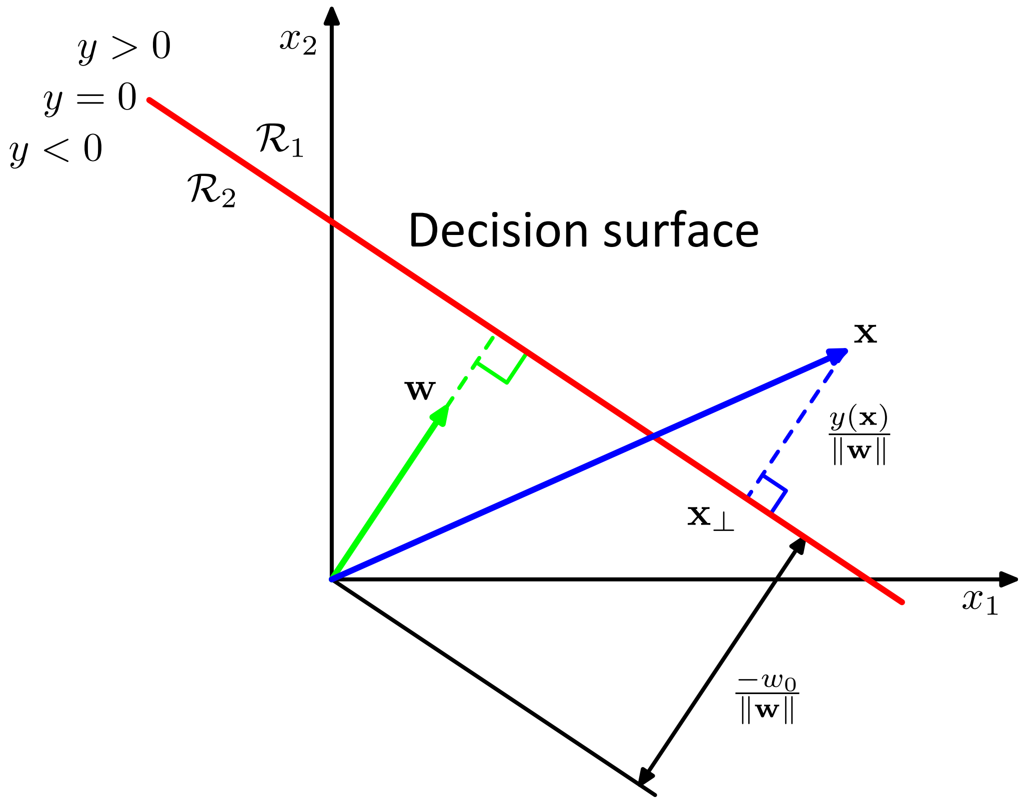
Posterior probabilities



Several dimensions



Several dimensions



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

weight vector

bias

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

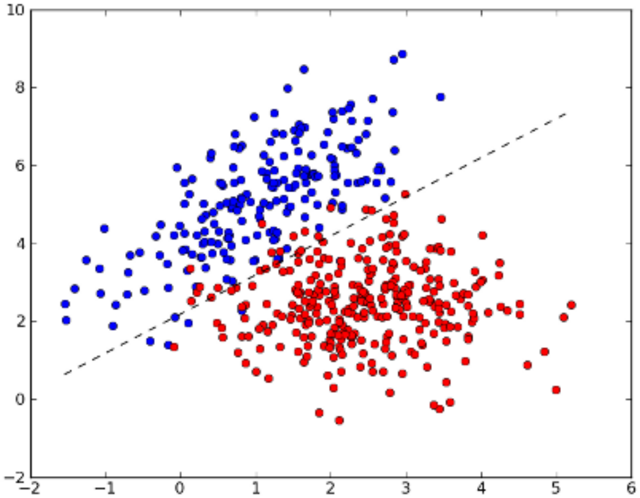
$\mathcal{C}_2$ otherwise.

Fisher's linear discriminant 1

Projecting data down to one dimension

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

But how?



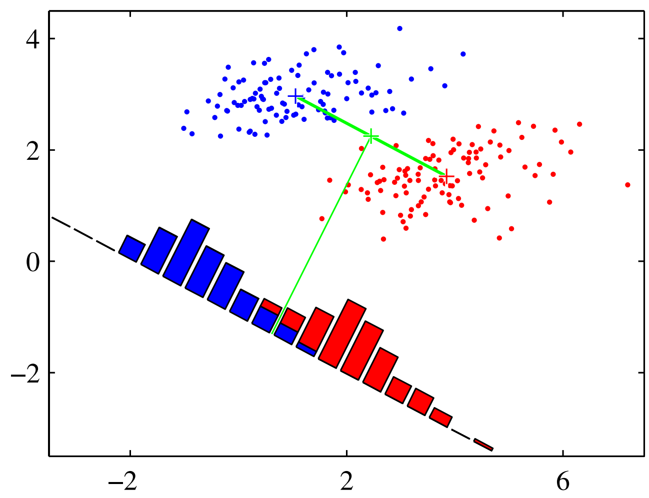
Fisher's linear discriminant 2

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

Maximizing the Fisher Criterion we obtain

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

with the total within class covariance

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

This is called Fisher's linear discriminant

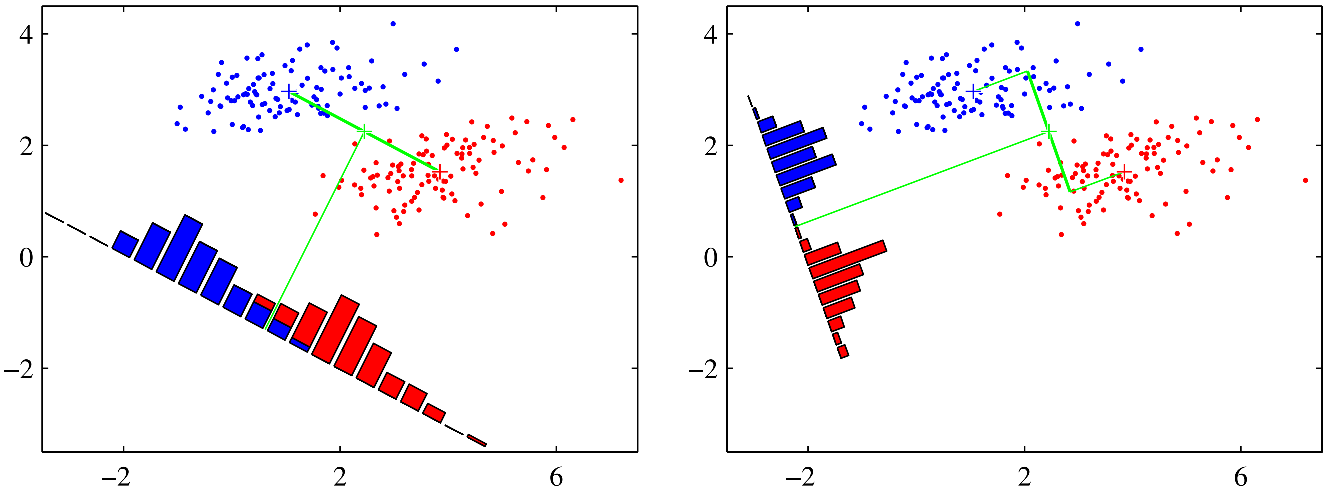
Fisher's linear discriminant 4

Fisher's linear discriminant

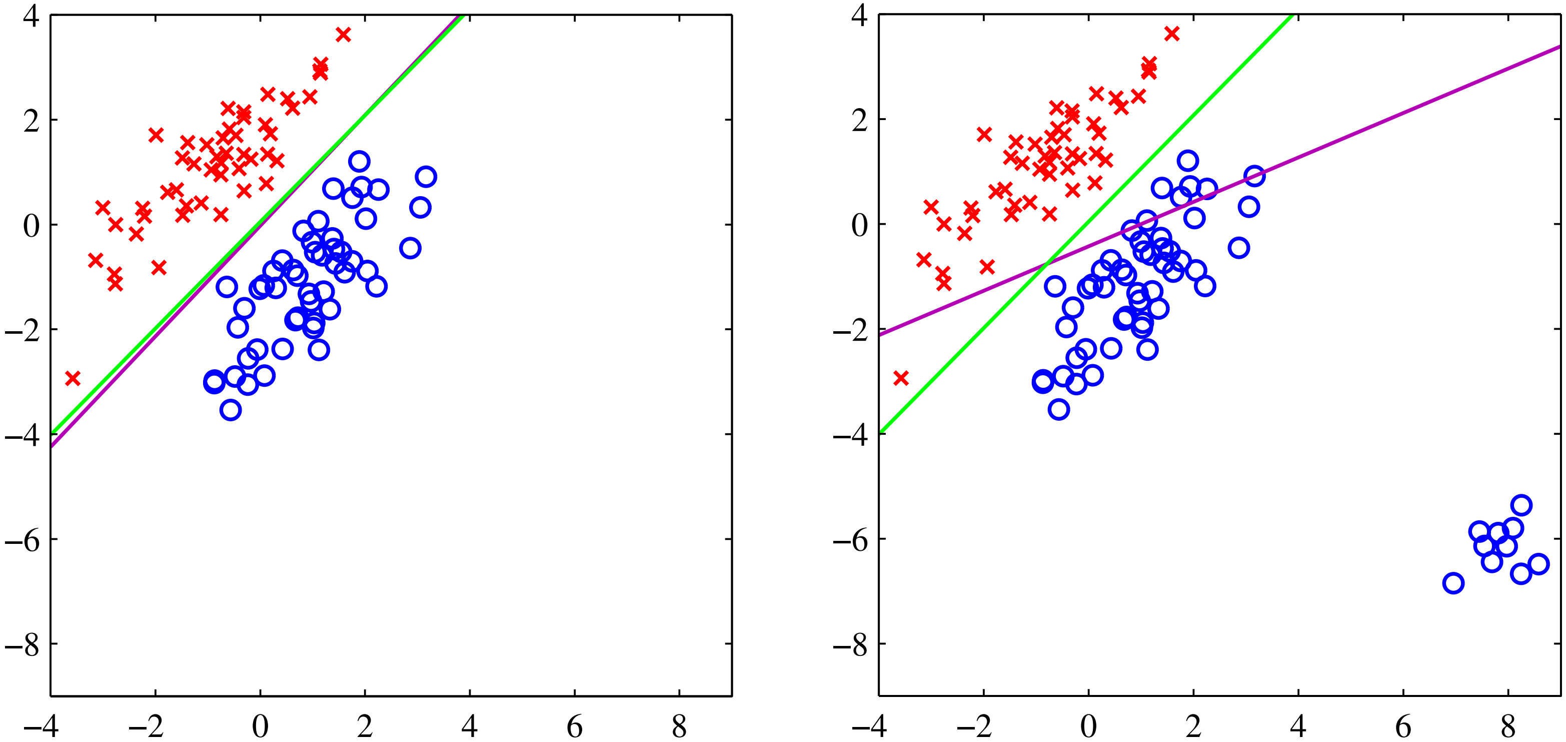
$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

Fisher Criterion

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

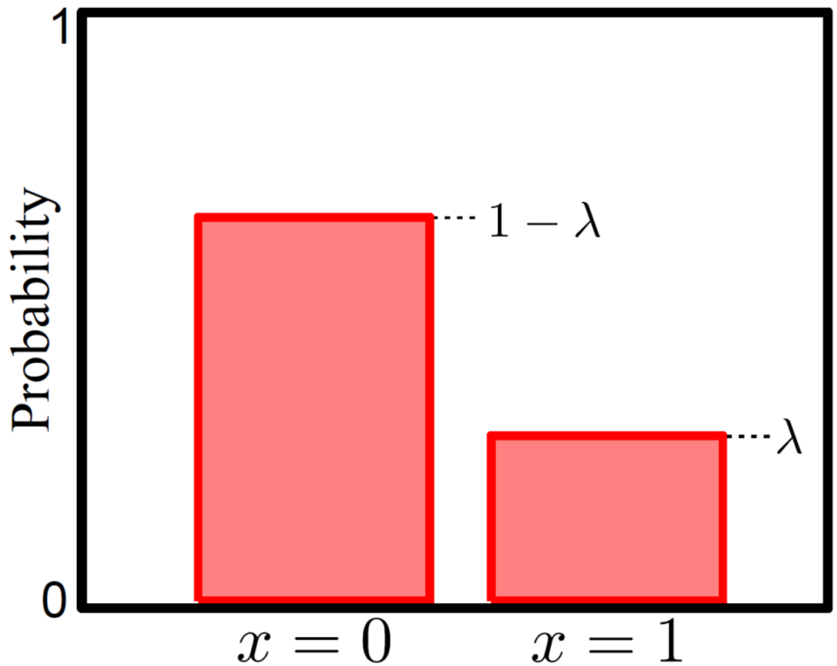


Least squares for classification fails



Use logistic regression instead!

Bernoulli Distribution



$$Pr(x=0) = 1-\lambda$$
$$Pr(x=1) = \lambda.$$

or

$$Pr(x) = \lambda^x(1-\lambda)^{1-x}$$

For short we write:

$$Pr(x) = \text{Bern}_x[\lambda]$$

Bernoulli distribution describes situation where only two possible outcomes y=0/y=1 or failure/success

Takes a single parameter $\lambda \in [0,1]$

Logistic Regression

Consider two class problem.

- Choose Bernoulli distribution over world.
- Make parameter λ a function of x

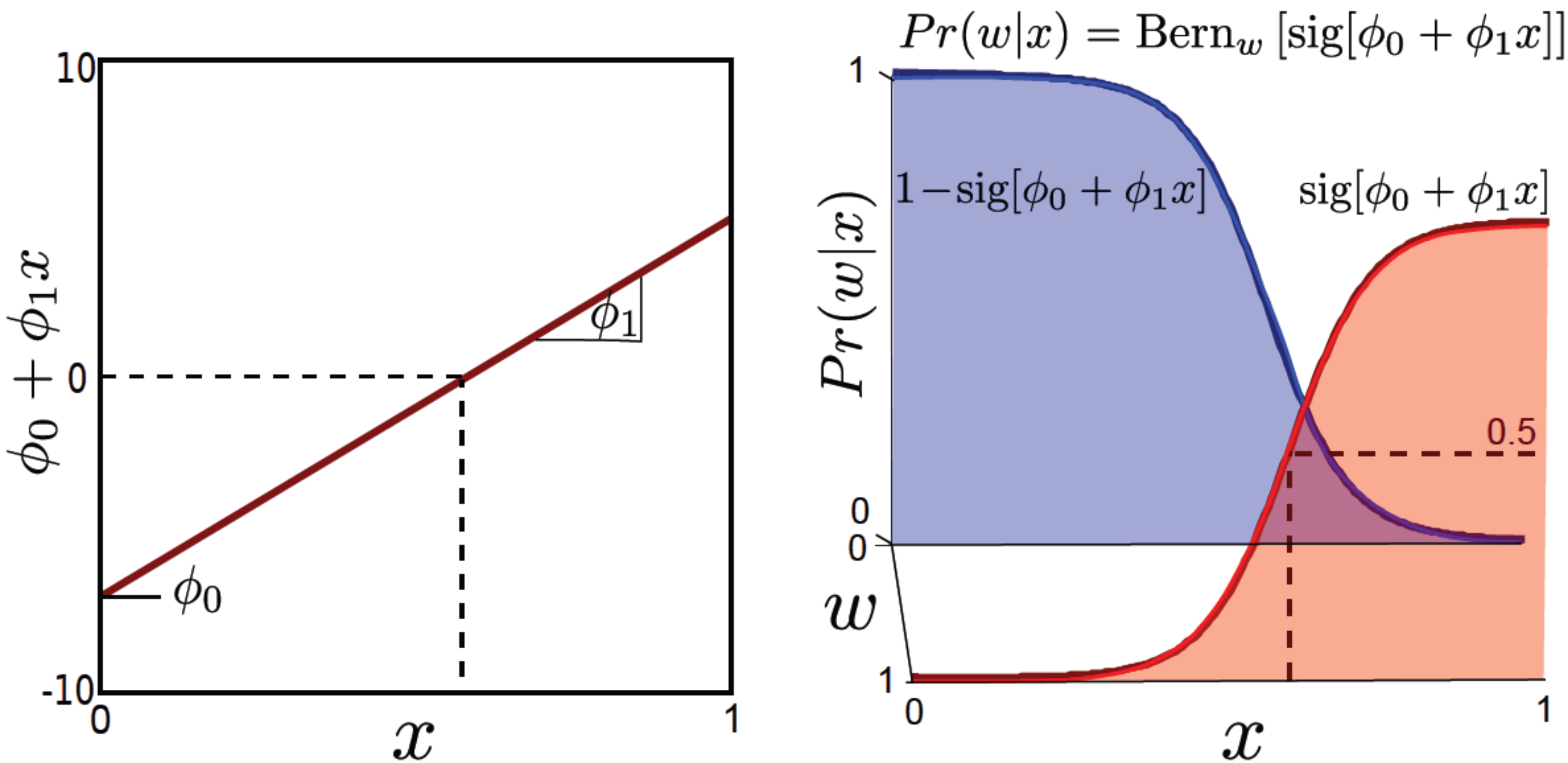
$$Pr(w|\phi_0, \boldsymbol{\phi}, \mathbf{x}) \quad = \quad \text{Bern}_w\left[\text{sig}[a]\right]$$

Model activation with a linear function

$$a = \phi_0 + \boldsymbol{\phi}^T \mathbf{x}$$

creates number between $[-\infty, \infty]$. Maps to $[0, 1]$ with

$$\text{sig}[a] = \frac{1}{1 + \exp[-a]}$$



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

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

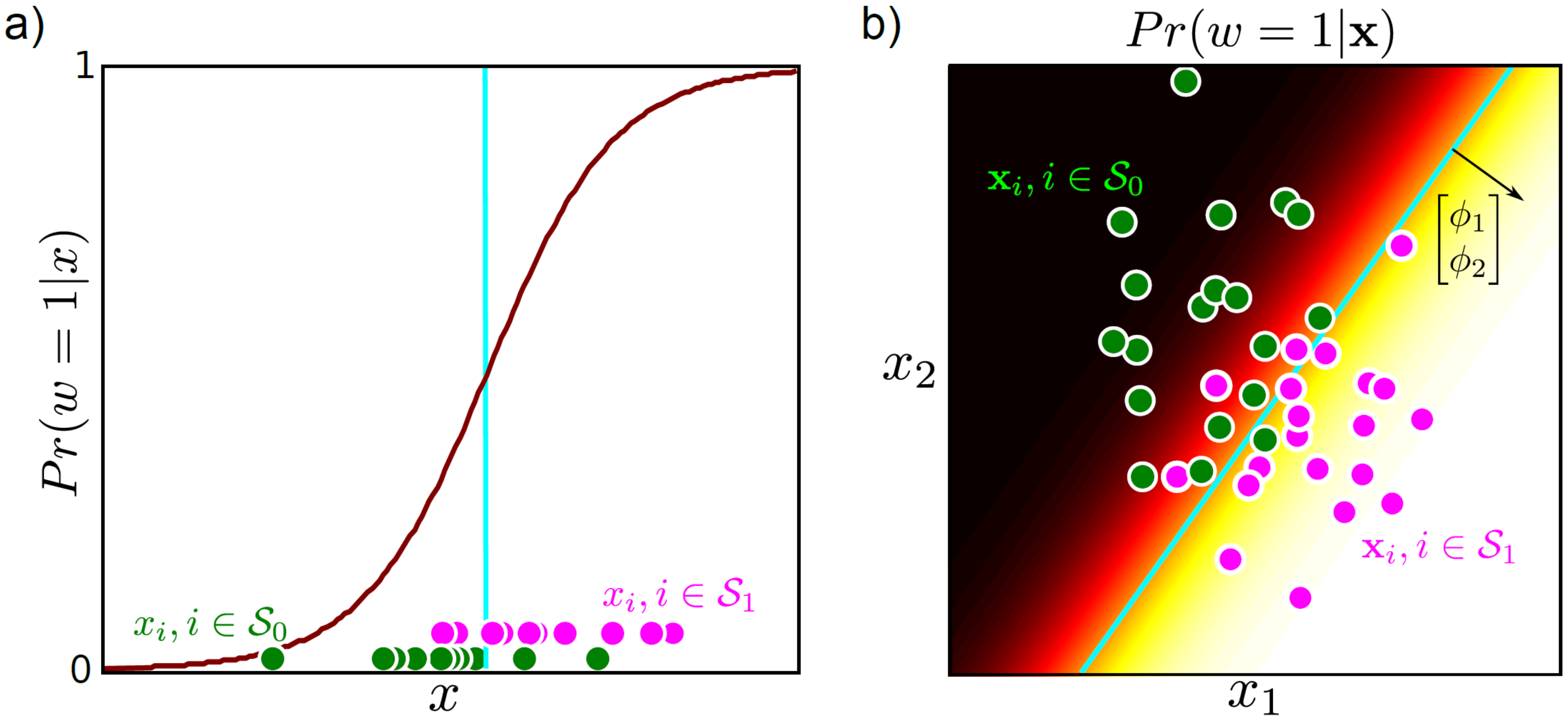
- Attach the offset to the start of the gradient vector $\boldsymbol{\phi}$

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

New model:

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

Logistic regression



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