CSci5521: Machine Learning Fundamentals

- Linear Discrimination

Catherine Qi Zhao
Computer Science and Engineering
University of Minnesota

Likelihood- vs. Discriminant-based Classification

- Likelihood-based: Assume a model for $p(x|C_i)$, use Bayes' rule to calculate $P(C_i|x)$

$$g_i(x) = \log P(C_i|x)$$

- Discriminant-based: Assume a model for $g_i(x|\Phi_i)$; no density estimation

Bayes' Rule: *K*>2 Classes

3

$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)}$$
$$= \frac{p(x|C_i)P(C_i)}{\sum_{k=1}^{K} p(x|C_k)P(C_k)}$$

$P(C_k) \geq 0$ and $\sum_{k=1}^{K} P(C_k) = 1$

choose C_i if $P(C_i|x) = \max_k P(C_k|x)$

Parametric Classification

- Discriminant function

$$g_i(x) = p(x|C_i)P(C_i)$$

or

$$g_i(x) = \log p(x|C_i) + \log P(C_i)$$

- Gaussian

$$p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp[-\frac{(x-\mu_i)^2}{2\sigma_i^2}]$$

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log\sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

Likelihood- vs. Discriminant-based Classification

- Likelihood-based: Assume a model for $p(x|C_i)$, use Bayes' rule to calculate $P(C_i|x)$

$$g_i(x) = \log P(C_i|x)$$

- Discriminant-based: Assume a model for $g_i(x|\Phi_i)$; no density estimation

- Estimating the boundaries is enough; no need to accurately estimate the densities inside the boundaries

Notations

- Discriminant: $g_i(x)$, $g_i(x|w_i, w_{i0})$

  $g(x)$

- Discriminant in pairwise separation:

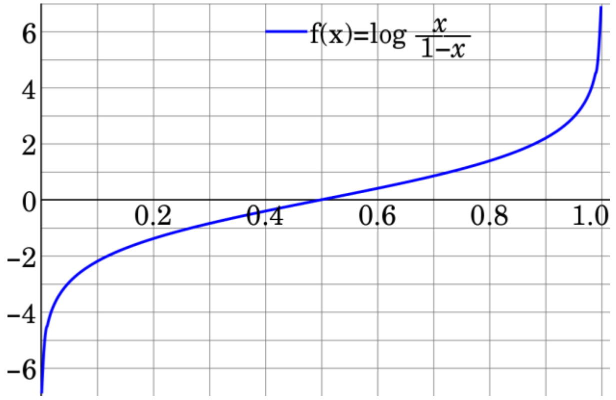
  $g_{ij}(x)$, $g_{ij}(x|w_{ij}, w_{ij0})$

- Error: E , $E(w)$, $E(w, w_0|X)$

- Log likelihood ratio: $\log \frac{p(x|C_1)}{p(x|C_2)}$

- Logit / log-odds:

$$logit(P(C_1|x)) = \log \frac{P(C_1|x)}{1 - P(C_{1|x})}$$



$\log \frac{p}{1-p}$

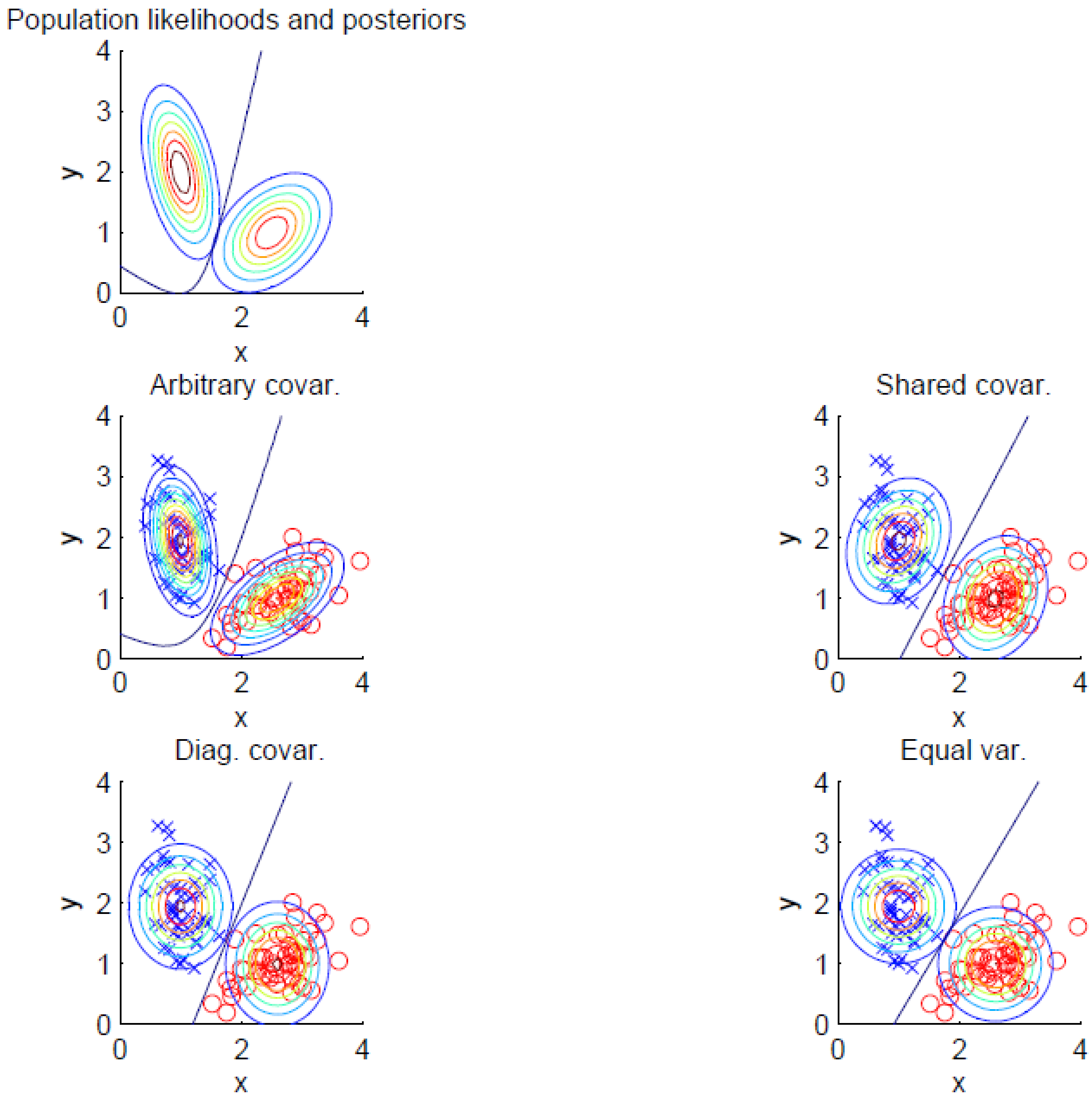
Linear Discriminant

- Linear discriminant:

$$g_i(x|w_i, w_{i0}) = w_i^T x + w_{i0} = \sum_{j=1}^{d} w_{ij} x_j + w_{i0}$$

$w_i = \left[\;\right]_d \qquad \left[\;\right]_d$

- Advantages:
  - Simple: O(*d*) space/computation
  - Intuitive and easy for knowledge extraction: Weighted sum of attributes; positive/negative weights, magnitudes
  - Optimal discriminant is linear when $p(x|C_i)$ are Gaussian with shared cov matrix; useful when classes are (almost) linearly separable


Population likelihoods and posteriors
Arbitrary covar.
Shared covar.
Diag. covar.
Equal var.
x
y
0
1
2
3
4

Common Covariance Matrix **S**

- Shared common sample covariance **S**

$$S = \sum_i \hat{P}(C_i) S_i$$

- Discriminant reduces to

$$g_i(x) = -\frac{1}{2}(x - m_i)^T S^{-1} (x - m_i) + \log \hat{P}(C_i)$$

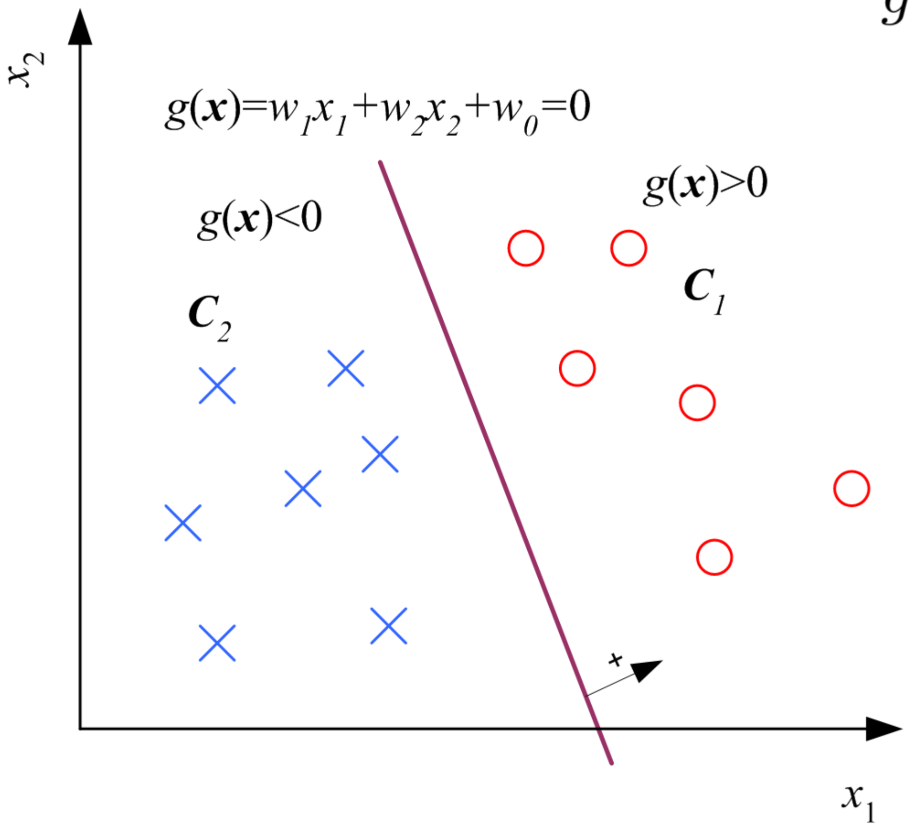
which is a linear discriminant

$$g_i(x) = w_i^T x + w_{i0}$$

where

$$w_i = S^{-1} m_i \qquad w_{i0} = -\frac{1}{2} m_i^T S_i^{-1} m_i + \log \hat{P}(C_i)$$

Two Classes



$$
\begin{aligned}
g(x) &= g_1(x) - g_2(x) \\
&= (w_1^T x + w_{10}) - (w_2^T x + w_{20}) \\
&= (w_1 - w_2)^T x + (w_{10} - w_{20}) \\
&= w^T x + w_0
\end{aligned}
$$

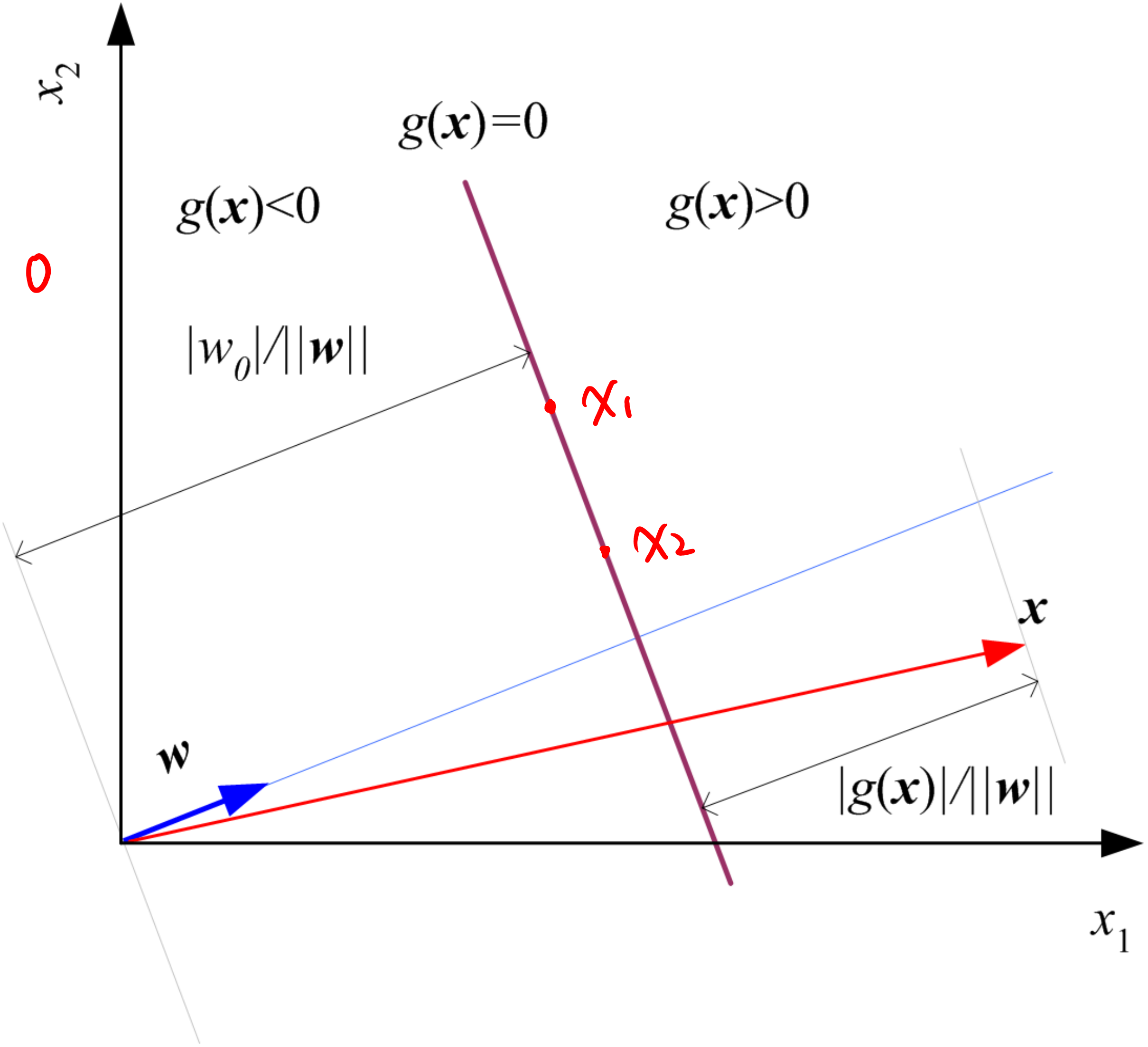
choose $\begin{cases} C_1, & \text{if } g(x) > 0 \\ C_2, & \text{otherwise} \end{cases}$

Geometry

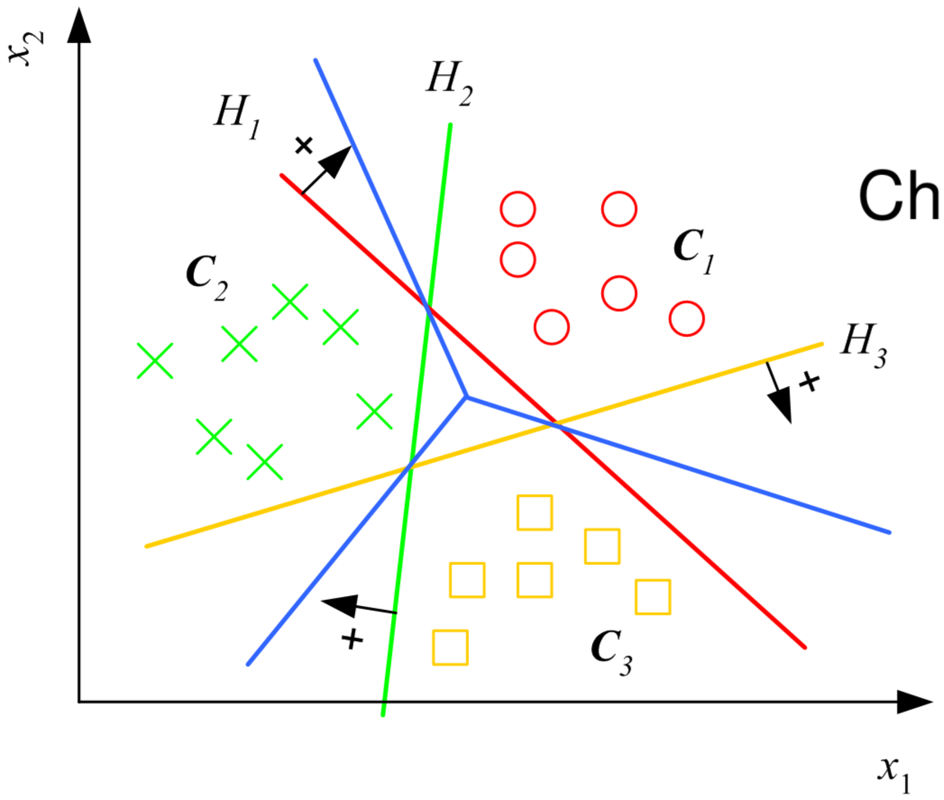
$g(x_1) = g(x_2) = 0$

$w^T x_1 + w_0 = w^T x_2 + w_0 = 0$

$w^T (x_1 - x_2) = 0$



Multiple Classes

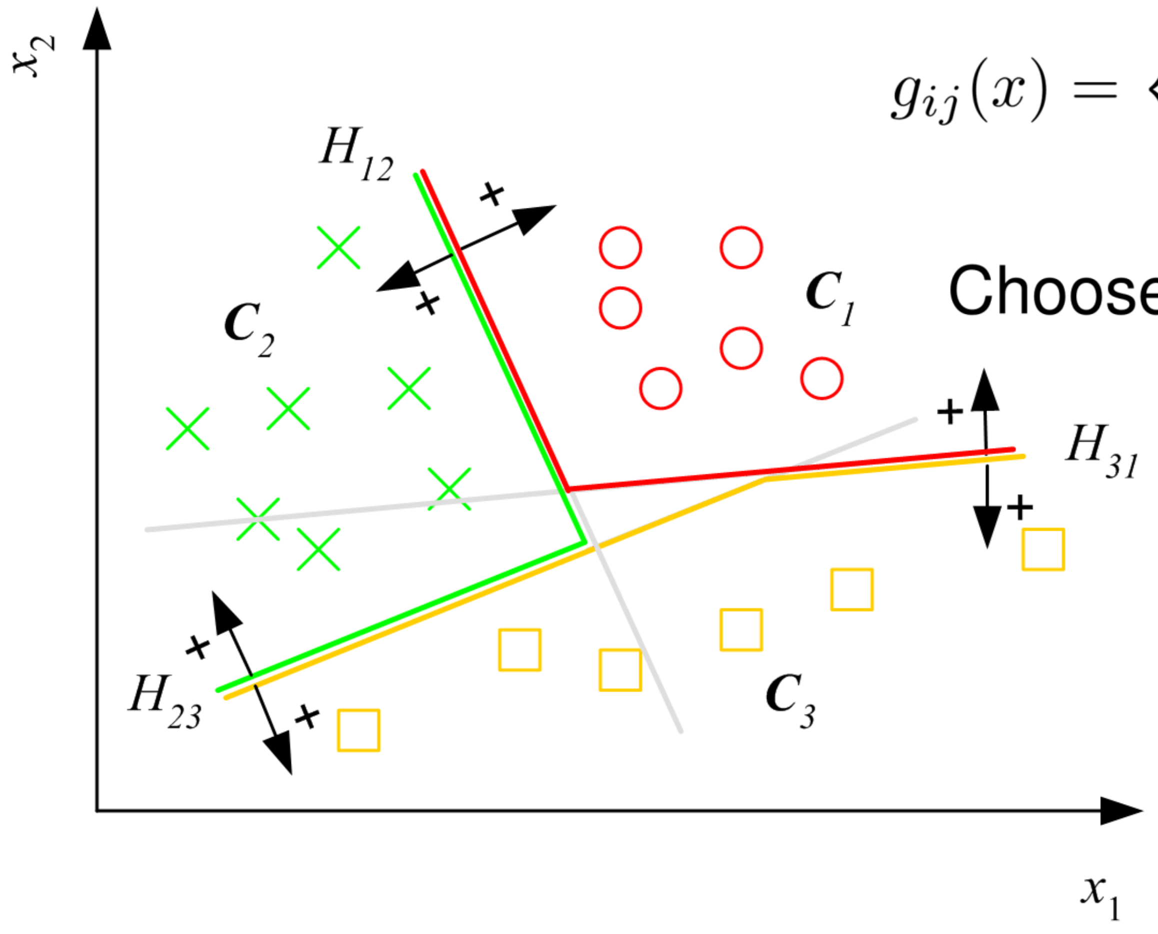


$$g_i(x|w_i, w_{i0}) = w_i^T x + w_{i0}$$

Choose C_i if $g_i(x) = \max_{j=1}^{k} g_j(x)$

Classes are linearly separable

Pairwise Separation



$$g_{ij}(x|w_{ij}, w_{ij0}) = w_{ij}^T x + w_{ij0}$$

$$g_{ij}(x) = \begin{cases} > 0, & \text{if } x \in C_i \\ \leq 0, & \text{if } x \in C_j \\ \text{do not care} & \text{otherwise} \end{cases}$$

Choose C_i if $\forall j \neq i, g_{ij}(x) > 0$

$$g_i(x) = \sum_{j \neq i} g_{ij}(x)$$

From Discriminants to Posteriors

When $p(x|C_i) \sim \mathcal{N}(\mu_i, \Sigma)$

$$g_i(x|w_i, w_{i0}) = w_i^T x + w_{i0}$$

$$w_i = \Sigma^{-1}\mu_i \qquad w_{i0} = -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \log P(C_i)$$

$y \equiv P(C_1|x)$ and $P(C_2|x) = 1 - y$

Choose C_1 if $\begin{cases} y > 0.5 \\ y/(1-y) > 1 \\ \log[y/(1-y)] > 0 \end{cases}$ and C_2 otherwise

$$
\begin{aligned}
logit(P(C_1|x)) &= \log \frac{P(C_1|x)}{\boxed{1-P(C_1|x)}} = \log \frac{P(C_1|x)}{\boxed{P(C_2|x)}} \\
&= \log \frac{p(x|C_1)}{p(x|C_2)} + \log \frac{P(C_1)}{P(C_2)} \\
&= \log \frac{(2\pi)^{-d/2}|\Sigma|^{-1/2} exp[-(1/2)(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))]}{(2\pi)^{-d/2}|\Sigma|^{-1/2} exp[-(1/2)(x-\mu_2)^T\Sigma^{-1}(x-\mu_2))]} + \log \frac{P(C_1)}{P(C_2)} \\
&= w^T x + w_0
\end{aligned}
$$

$p(C_1|x) + p(C_2|x) = 1$

where $w = \Sigma^{-1}(\mu_1 - \mu_2) \quad w_0 = -\frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)$

$\log \frac{p}{1-p} = \alpha$

$\Downarrow$

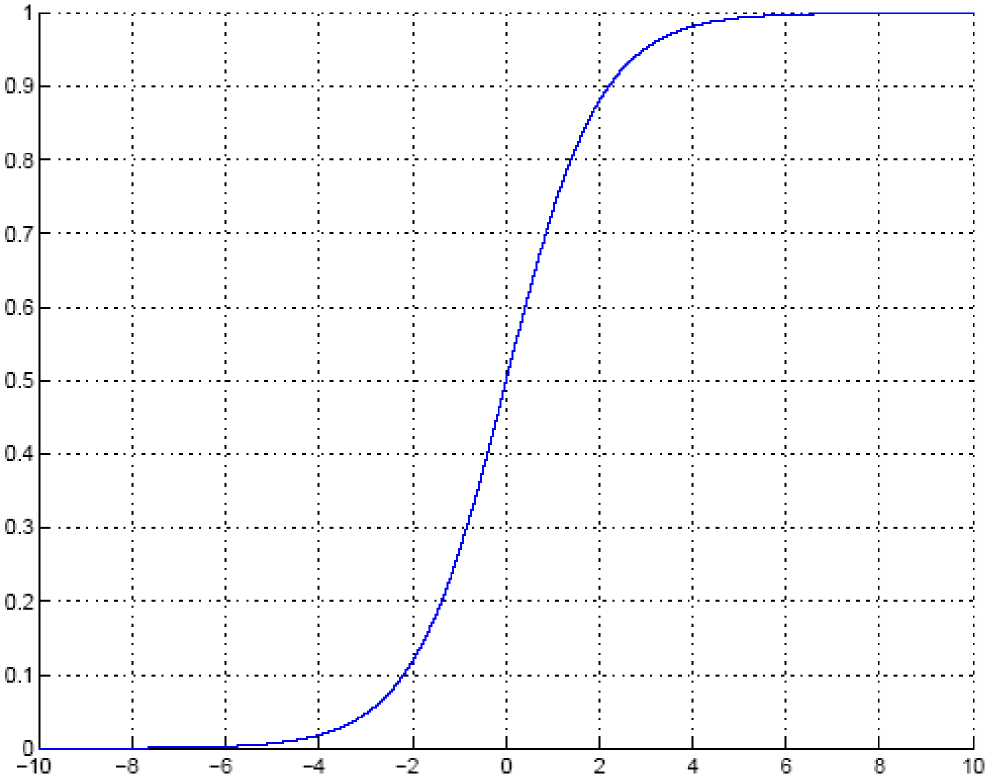
$p = sigmoid(\alpha)$

The Inverse of logit $\log \frac{\boxed{P(C_1|x)}}{1-P(C_1|x)} = w^T x + w_0$

$0 \sim 1$

$$P(C_1|x) = sigmoid(w^T x + w_0) = \frac{1}{1+exp[-(w^T x + w_0)]}$$

Sigmoid (Logistic) Function



- During training: estimate discriminant parameters

$X \longrightarrow \theta \ (w, w_0)$

- During testing:

Calculate $g(x) = w^T x + w_0$ and choose C_1 if $g(x) > 0$, or

Calculate $y = sigmoid(w^T x + w_0)$ and choose C_1 if $y > 0.5$

Finding parameters

- Likelihood-based classification:
  - Parameters: sufficient statistics of the likelihood
  - Parameter estimation: e.g., maximum likelihood
- Discriminant-based classification
  - Parameters: sufficient statistics of the discriminant
  - Parameters estimation: minimize training error

Gradient-Descent

- E(*w*|X) is error with parameters *w* on sample X

  $w^*=\arg\min_w E(w\,|X)$

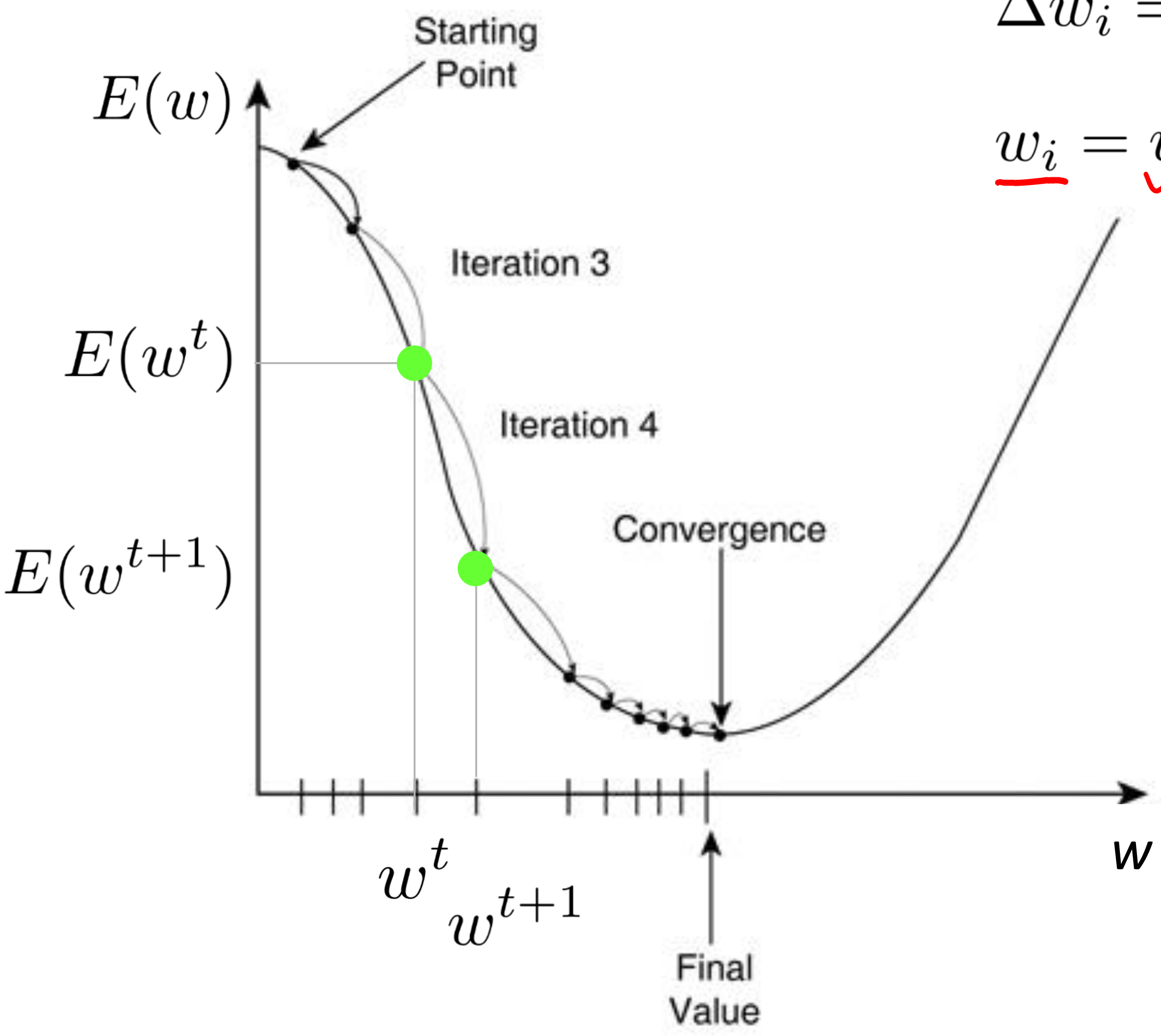
- Gradient

$$\nabla_w E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, ..., \frac{\partial E}{\partial w_d}\right]^T$$

- Gradient-descent:

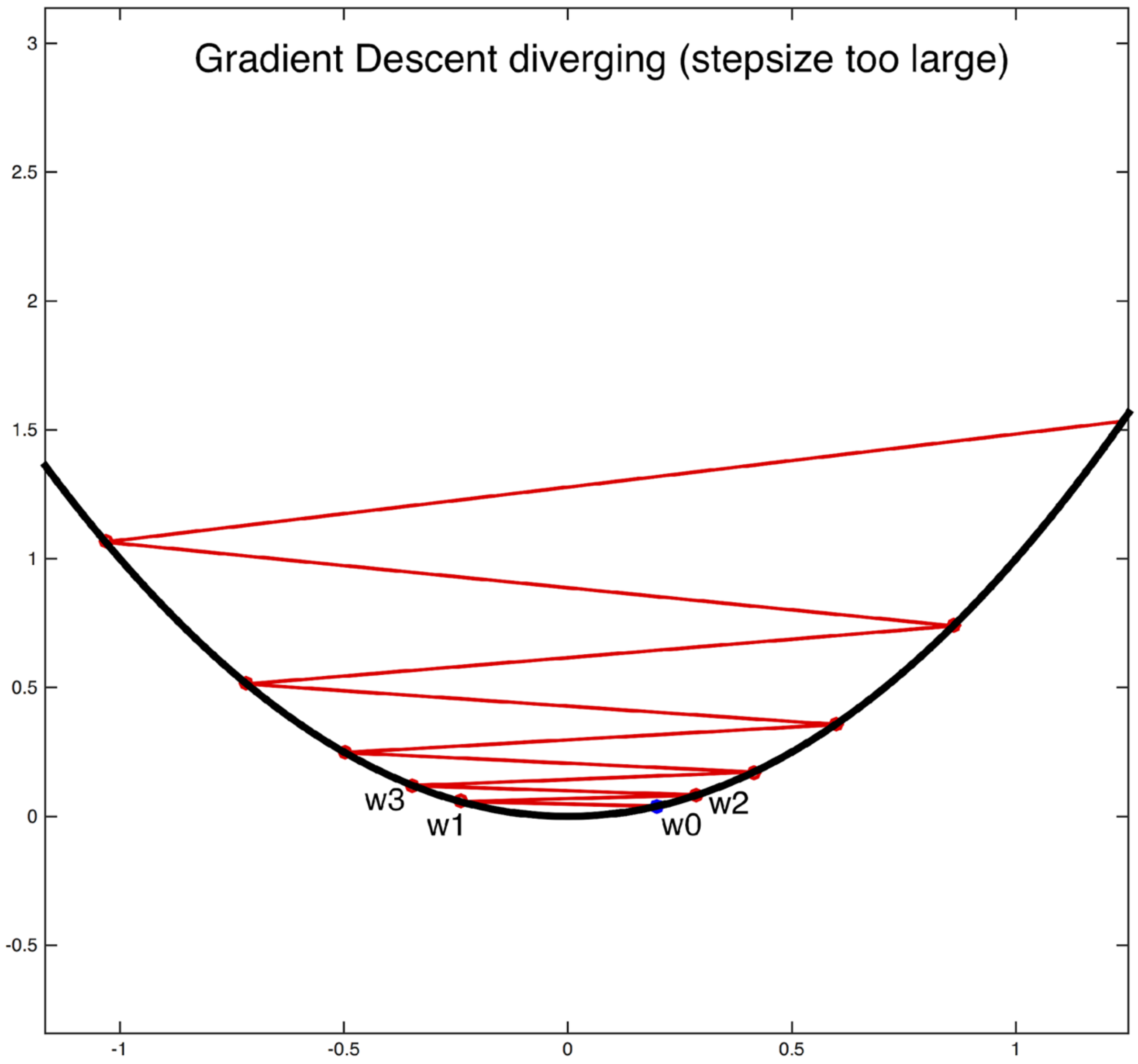
  Starts from random *w* and updates *w* iteratively in the negative direction of gradient

Gradient-Descent



$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i} \quad \forall i$$

$$w_i = w_i + \Delta w_i$$


Gradient Descent diverging (stepsize too large)
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-0.5
0
0.5
1
w3
w1
w0
w2

Logistic Discrimination

Instead of modeling the **class-conditional densities,** modeling **their ratio.**

Logistic Discrimination

Two classes: Assume log likelihood ratio is linear

$$\log \frac{p(x|C_1)}{p(x|C_2)} = w^T x + w_0^o$$

$$logit(P(C_1|x)) = \log \frac{P(C_1|x)}{1 - P(C_{1|x})} = \log \frac{p(x|C_1)}{p(x|C_2)} + \log \frac{P(C_1)}{P(C_2)}$$
$$= w^T x + w_0$$

where $w_0 = w_0^o + \log \frac{P(C_1)}{P(C_2)}$

$$y = P(C_1|x) = \frac{1}{1 + exp[-(w^T x + w_0)]}$$

Training: Two Classes

$$X = \{x^t, r^t\} \quad r^t | x^t \sim \text{Bernoulli}(y^t)$$

$$y = P(C_1|x) = \frac{1}{1 + exp[-(w^T x + w_0)]}$$

$$l(w, w_0|X) = \prod_t (y^t)^{r^t} (1 - y^t)^{(1-r^t)}$$

ln

$$E = -\log l$$

$$E(w, w_0|X) = -\sum_t (r^t \log y^t + (1 - r^t) \log(1 - y^t))$$

Bernoulli Distribution

- The r.v. *X* is a 0/1 indicator variable and takes the value 1 for a success outcome and is 0 otherwise. *p* is the probability that the result of trial is a success. Then

$$P\{X = 1\} = p \text{ and } P\{X = 0\} = 1 - p$$

- which can equivalently be written as

$$P\{X = i\} = p^i(1-p)^{1-i}, i = 0, 1$$

- If *X* is Bernoulli, its expected value and variance are

$$E[X] = p, Var(X) = p(1-p)$$

Training: Gradient-Descent

$$\frac{\partial y}{\partial w}$$

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial y}\boxed{\frac{\partial y}{\partial \alpha}\frac{\partial \alpha}{\partial w}}$$

$$E(w, w_0|X) = -\sum_t (r^t \log y^t + (1 - r^t)\log(1 - y^t))$$

if $y = sigmoid(\alpha)$ $\quad \dfrac{\partial y}{\partial \alpha} = y(1-y)$

$\alpha = w^T x + w_0$

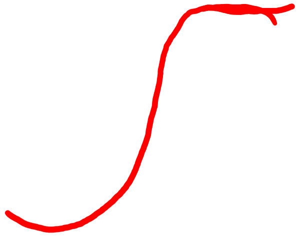
$$\frac{\partial E}{\partial w_0}$$

$$\Delta w_j = -\eta \frac{\partial E}{\partial w_j} = \eta \sum_t \left(\frac{r^t}{y^t} - \frac{1 - r^t}{1 - y^t}\right) y^t(1 - y^t) x_j^t$$

$$= \eta \sum_t (r^t - y^t) x_j^t, j = 1, ..., d$$

$$\Delta w_0 = -\eta \frac{\partial E}{\partial w_0} = \eta \sum_t (r^t - y^t) \eta$$

- How to initialize the weights?



- When to stop update?

Some materials credit to former 5521, Introduction to Machine Learning by Ethem Alpaydin and online resources