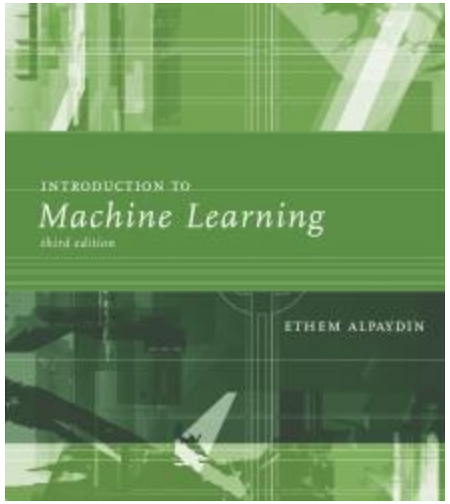


Linear Discrimination & Logistic Discrimination

Lecture notes by Ethem Alpaydın
Introduction to Machine Learning (Boğaziçi Üniversitesi)

Lecture notes by Kevyn Collins-Thompson
Applied Machine Learning (Coursera)

Lecture notes by Andrew NG
Machine Learning by Stanford University (Coursera)

Likelihood- vs. Discriminant-based Classification

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

$$g_i(\boldsymbol{x}) = \log P(C_i|\boldsymbol{x})$$

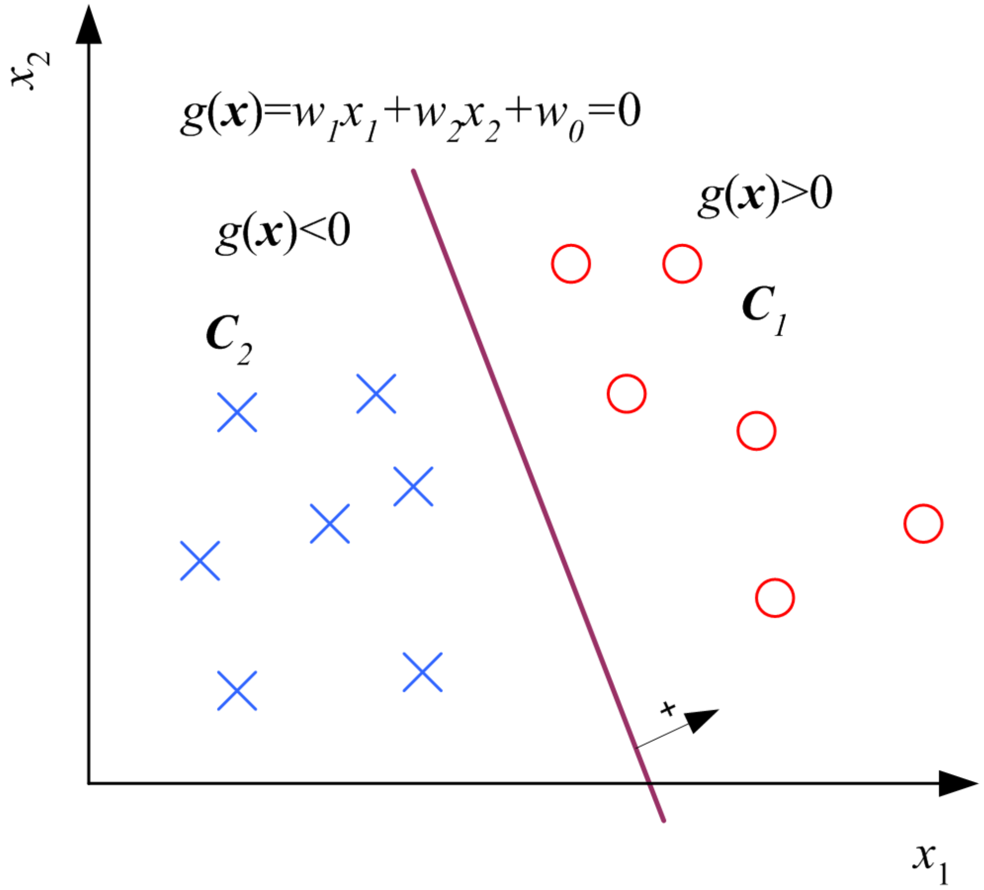
- Discriminant-based: Assume a model for $g_i(\boldsymbol{x}|\Phi_i)$; no density estimation
- Estimating the boundaries is enough; no need to accurately estimate the densities inside the boundaries

Linear Discriminant

- The *linear discriminant* is used frequently mainly due to its simplicity: both the space and time complexities are O*(d)*.
- The linear model is easy to understand: the final output is a weighted sum of the input attributes x_j

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

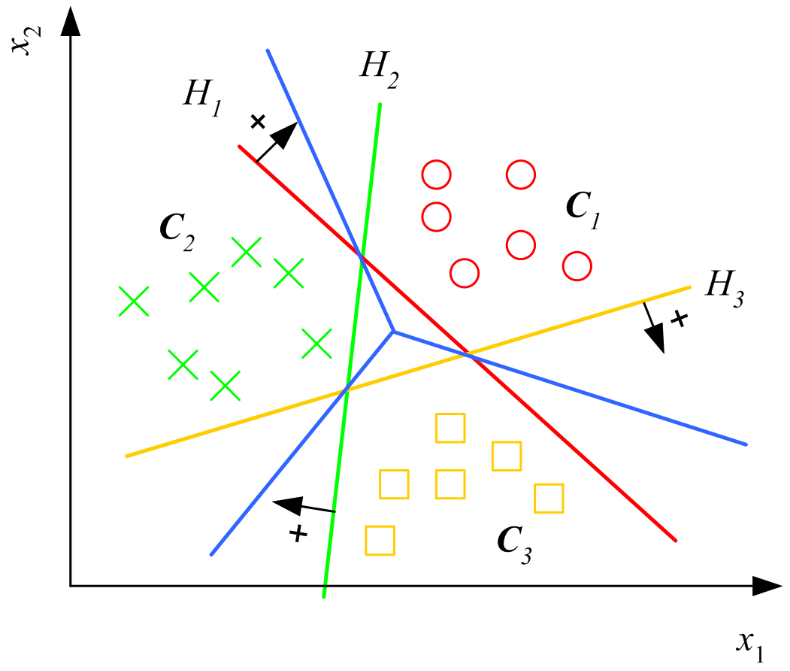
Two Classes



$$
\begin{aligned}
g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\
&= \left(\mathbf{w}_1^T\mathbf{x} + w_{10}\right) - \left(\mathbf{w}_2^T\mathbf{x} + w_{20}\right) \\
&= \left(\mathbf{w}_1 - \mathbf{w}_2\right)^T\mathbf{x} + \left(w_{10} - w_{20}\right) \\
&= \mathbf{w}^T\mathbf{x} + w_0
\end{aligned}
$$

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

Multiple Classes (one-vs-all)



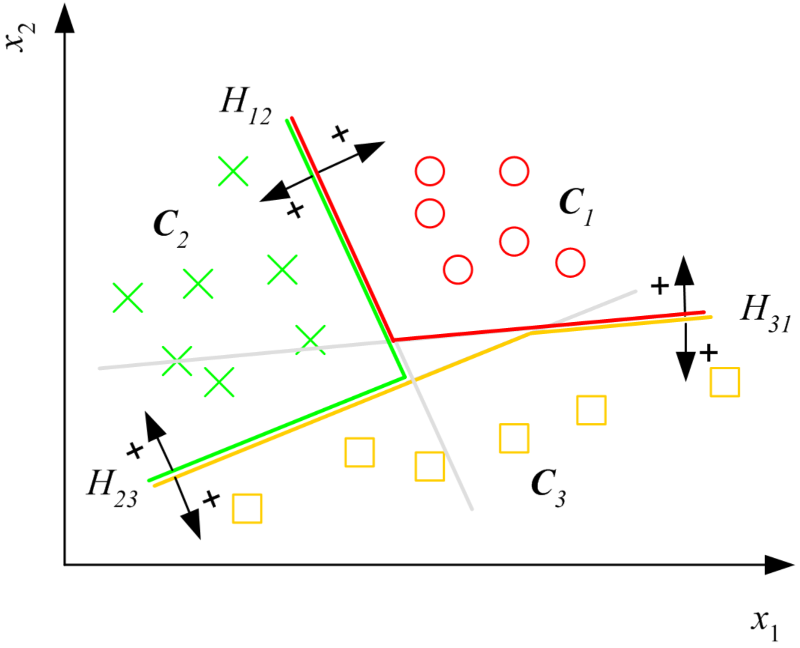
- In linear classification, each hyperplane H_i separates the examples of C_i from the examples of all other classes.
- Thus for it to work, the classes should be linearly separable.
- Blue lines are the induced boundaries of the linear classifier.

$$g_i(\mathbf{x}|\mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

Choose C_i if

$$g_i(\mathbf{x}) = \max_{j=1}^{K} g_j(\mathbf{x})$$

Pairwise Separation



- If the classes are not linearly separable, one approach is to divide it into a set of linear problems.
- One possibility is *pairwise separation* of classes
- It uses $K(K-1)/2$ linear discriminants, $g_{ij}(\mathbf{x})$, one for every pair of distinct classes

$$g_{ij}\left(\mathbf{x} \mid \mathbf{w}_{ij}, w_{ij0}\right) = \mathbf{w}_{ij}^T \mathbf{x} + w_{ij0}$$

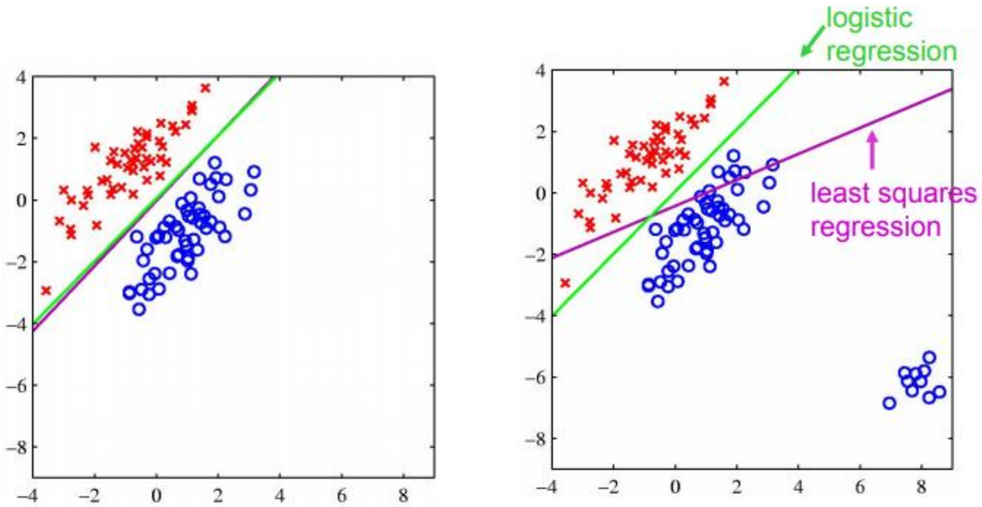
$$g_{ij}(\mathbf{x}) = \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i \\ \leq 0 & \text{if } \mathbf{x} \in C_j \\ \text{don't care} & \text{otherwise} \end{cases}$$

choose C_i if

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

Logistic Regression

- Logistic regression is used for classification problems
- For classification problems we cannot use the linear regression model because in that case, r could get any value in the interval $(-\infty, \infty)$



https://www.cs.toronto.edu/~urtasun/courses/CSC411_Fall16/04_prob_classif_handout.pdf

Logistic Regression

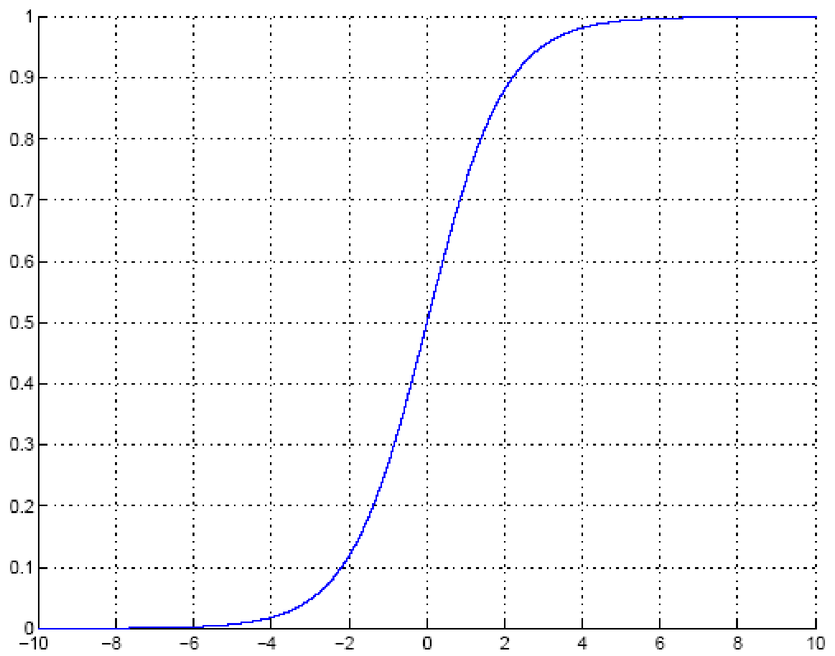
Change the form for the hypotheses

$$y = \sigma(w^T x + w_0) = \frac{1}{1 + e^{-(w^T x + w_0)}}$$

$\sigma(z)$ is called the sigmoid function or logistic function.

$\sigma(z)$ can output values in the interval $[0,1]$

Sigmoid Function (Logistic Function)



$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Logistic Regression

If we have a value between 0 and 1, let's use it to model class probability

$$y = \sigma(w^T x + w_0)$$

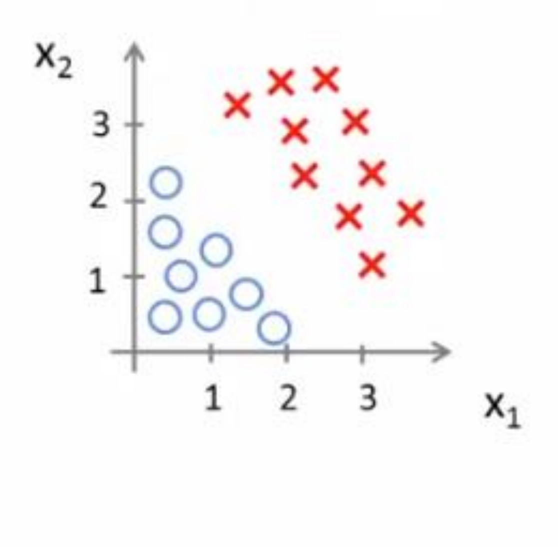
$y = P(C_1|x)$ and $P(C_2|x) = 1 - y$

choose C_1 if $y \geq 0.5$ ($\sigma(z) \geq 0.5 \; when \; z \geq 0$)
choose C_2 if $y < 0.5$ ($\sigma(z) < 0.5 \; when \; z < 0$)

Logistic Regression – Decision boundary

Decision boundary: $w^T x + w_0 = 0$

Example



Training: Two Classes

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_t \quad r^t \mid \mathbf{x}^t \sim \text{Bernoulli}(y^t)$$

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

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

$$E = -\log l$$

$$E(\mathbf{w}, w_0 \mid \mathcal{X}) = -\sum_t r^t \log y^t + (1 - r^t) \log (1 - y^t)$$ Cross entropy

Training: Gradient-Descent

$$E(\mathbf{w}, w_0 \mid \mathcal{X}) = -\sum_t r^t \log y^t + (1 - r^t) \log (1 - y^t)$$

$$\text{If } y = \text{sigmoid}(a) \quad \frac{dy}{da} = y(1-y)$$

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

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

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

Training: Gradient-Descent

For $j = 0, \ldots, d$

  $w_j \leftarrow$rand(-0.01,0.01)

Repeat

  For $j = 0, \ldots, d$

    $\Delta w_j \leftarrow 0$

  For $t = 1, \ldots, N$

    $o \leftarrow 0$

    For $j = 0, \ldots, d$

      $o \leftarrow o + w_j x_j^t$

    $y \leftarrow \text{sigmoid}(o)$

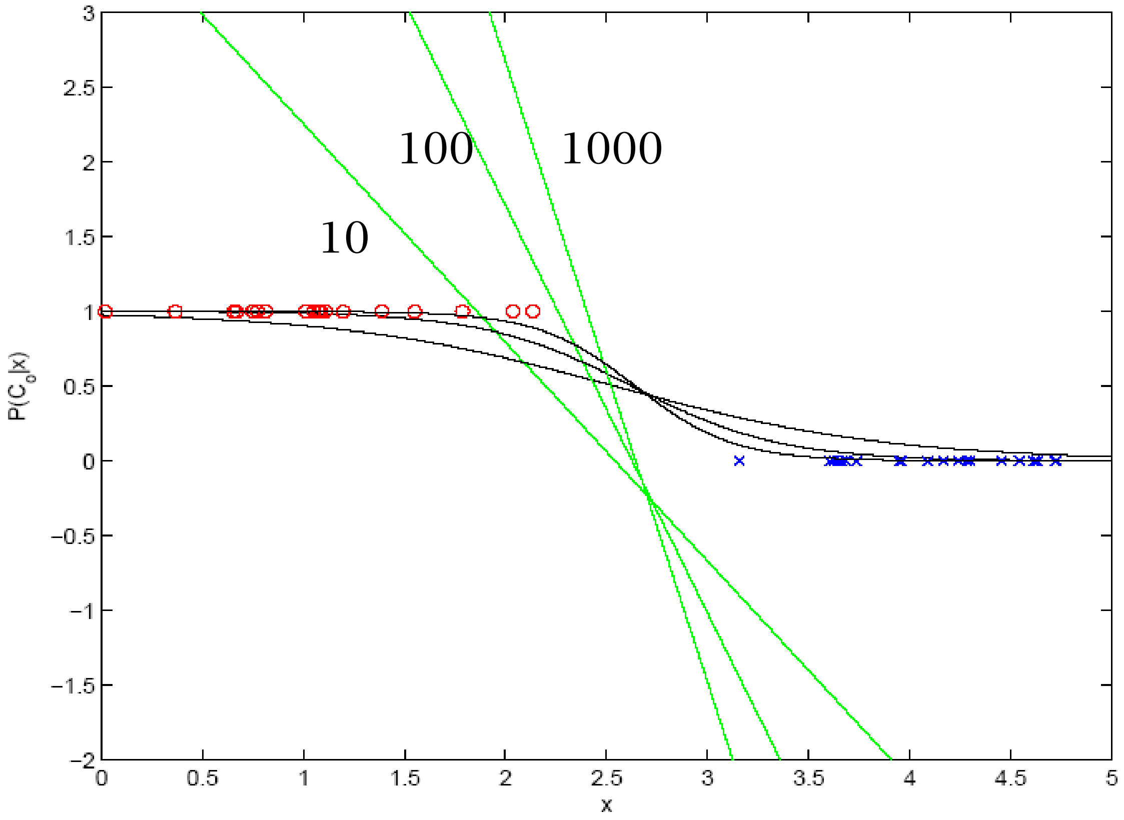
    $\Delta w_j \leftarrow \Delta w_j + (r^t - y) x_j^t$

  For $j = 0, \ldots, d$

    $w_j \leftarrow w_j + \eta \Delta w_j$

Until convergence

Training: Gradient-Descent



For a univariate two-class problem (shown with '∘' and '×'), the evolution of the line $wx + w_0$ and the sigmoid output after 10, 100, and 1,000 iterations over the sample.

K>2 classes

$$\mathcal{X} = \{\mathbf{x}^t, \mathbf{r}^t\}_t \quad r^t \mid \mathbf{x}^t \sim \text{Mult}_K(1, \mathbf{y}^t)$$

$$y = \hat{P}(C_i \mid \mathbf{x}) = \frac{\exp\left[\mathbf{w}_i^T \mathbf{x} + w_{i0}\right]}{\sum_{j=1}^{K} \exp\left[\mathbf{w}_j^T \mathbf{x} + w_{j0}\right]}, i = 1, \ldots, K$$

softmax

$$l(\{\mathbf{w}_i, w_{i0}\}_i \mid \mathcal{X}) = \prod_t \prod_i \left(y_i^t\right)^{\left(r_i^t\right)}$$

$$E(\{\mathbf{w}_i, w_{i0}\}_i \mid \mathcal{X}) = -\sum_t r_i^t \log y_i^t$$

$$\Delta \mathbf{w}_j = \eta \sum_t \left(r_j^t - y_j^t\right) \mathbf{x}^t \quad \Delta w_{j0} = \eta \sum_t \left(r_j^t - y_j^t\right)$$

Gradient Descent

For $i = 1, \ldots, K$, For $j = 0, \ldots, d$, $w_{ij} \leftarrow \text{rand}(-0.01, 0.01)$
Repeat
    For $i = 1, \ldots, K$, For $j = 0, \ldots, d$, $\Delta w_{ij} \leftarrow 0$
    For $t = 1, \ldots, N$
        For $i = 1, \ldots, K$
            $o_i \leftarrow 0$
            For $j = 0, \ldots, d$
                $o_i \leftarrow o_i + w_{ij} x_j^t$
        For $i = 1, \ldots, K$
            $y_i \leftarrow \exp(o_i) / \sum_k \exp(o_k)$
        For $i = 1, \ldots, K$
            For $j = 0, \ldots, d$
                $\Delta w_{ij} \leftarrow \Delta w_{ij} + (r_i^t - y_i) x_j^t$
    For $i = 1, \ldots, K$
        For $j = 0, \ldots, d$
            $w_{ij} \leftarrow w_{ij} + \eta \Delta w_{ij}$
Until convergence