Linear Discrimination

Berlin Chen
Graduate Institute of Computer Science & Information Engineering
National Taiwan Normal University

References:

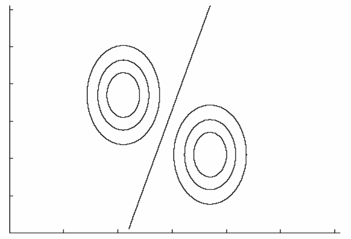
1. *Introduction to Machine Learning* , Chapter 10

2. Most of the slides were adopted from *Ethem Alpaydin's teaching materials*

Likelihood- vs. Discriminant-based Classification

- Classification
  - Define a set of discriminant functions $g_j(x),\ j=1,\ldots,K$
  - Choose a class C_i if

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



covariance matrices are identical between classes

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

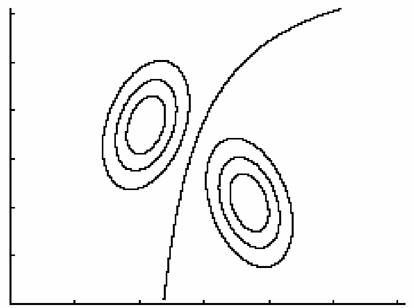
$$g_i(x) = \log p(C_i|x)$$

  - Discriminant-based: Assume a model for $g_i(x|\Phi_i)$ and no density estimation is needed
    - Estimate the boundaries is enough
    - No need to accurately estimate the densities inside the boundaries

Linear Discriminant

- Linear discriminant

$$g_i(\mathbf{x} \mid \mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} = \sum_{j=1}^{d} w_{ij} x_j + w_{i0}$$



covariance matrices are different between classes

- Advantages:
  - Simple: O(*d*) space/computation
  - Knowledge extraction
    - Weighted sum of attributes
      - Positive/negative weights, magnitudes
    - E.g., credit scoring
  - Optimal when $p(\boldsymbol{x}|C_i)$ are Gaussian with shared covariance matrix
    - Useful when classes are (almost) linearly separable

Generalized Linear Model

- Quadratic discriminant

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

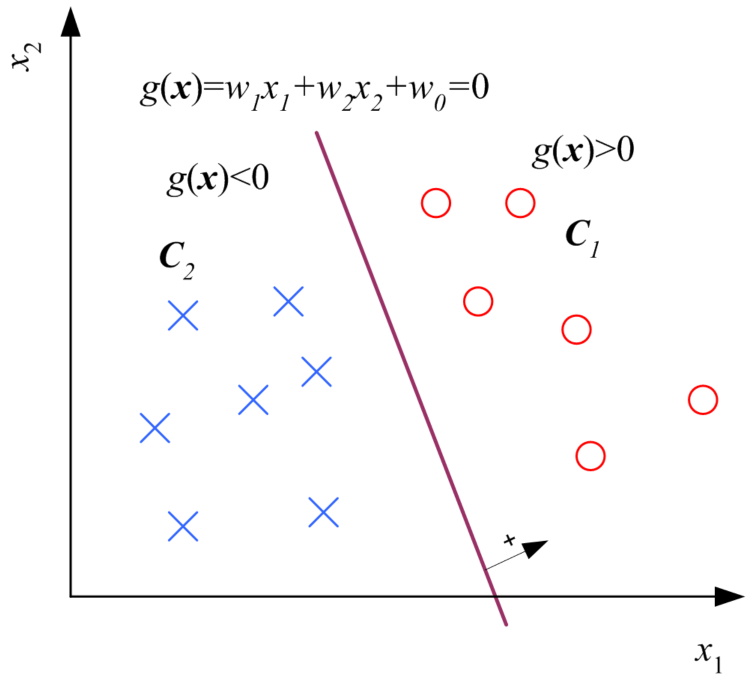
- Higher-order (product) terms

$$z_1 = x_1,\ z_2 = x_2,\ z_3 = x_1^2,\ z_4 = x_2^2,\ z_5 = x_1 x_2$$

  - Map from ***x*** to ***z*** using nonlinear basis functions and use a linear discriminant in ***z***-space

$$g_i(\mathbf{x}) = \sum_{j=1}^{k} w_j \phi_{ij}(\mathbf{x})$$

Two-Class Linear Classification

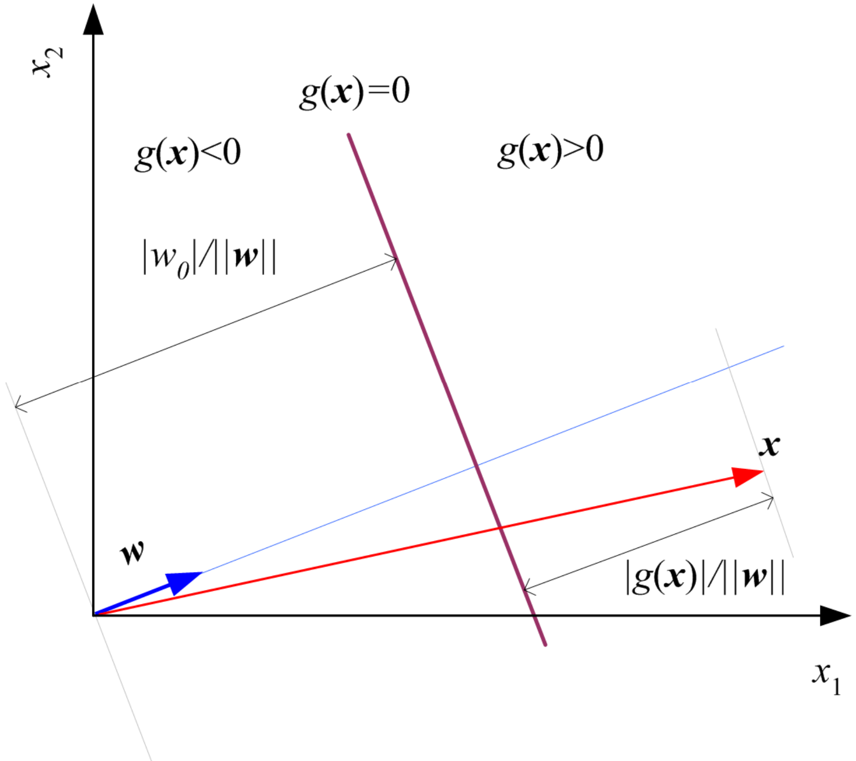


$$
\begin{aligned}
g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\
&= \left(\mathbf{w}_1^T\mathbf{x} + w_{10}\right) - \left(\mathbf{w}_2^T\mathbf{x} + w_{20}\right) \\
&= \left(\mathbf{w}_1 - \mathbf{w}_2\right)^T\mathbf{x} + \left(w_{10} - w_{20}\right) \\
&= \mathbf{w}^T\mathbf{x} + w_0
\end{aligned}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

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

Geometry of the Linear Discriminant

- **w** is normal to any vector lying on the decision hyperplane



$$\mathbf{x} = \mathbf{x}_p + r \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

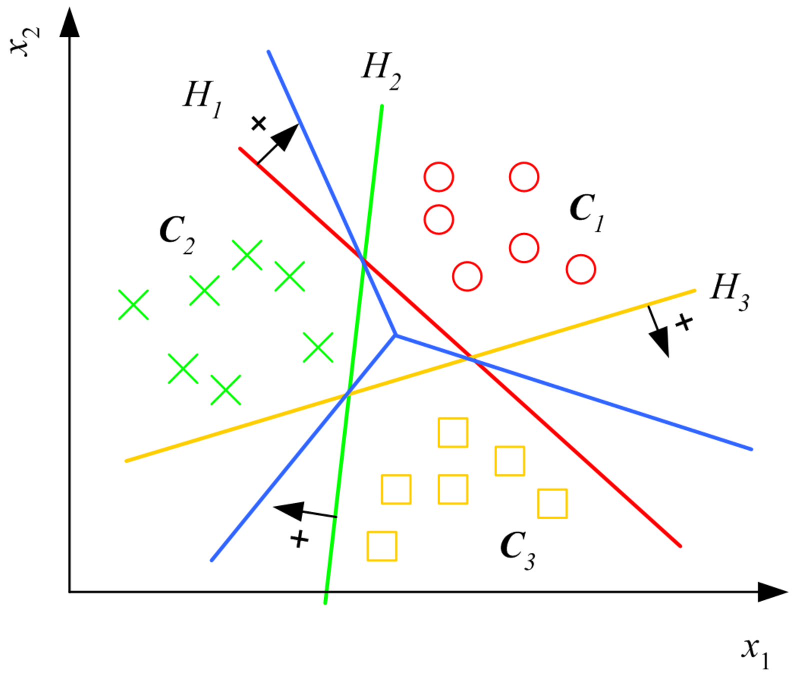
$\mathbf{x}_p$: normal projection of $\mathbf{x}$ on to the hyperplane

r : distance from $\mathbf{x}$ to the hyperplane , $r = \dfrac{g(\mathbf{x})}{\|\mathbf{w}\|}$

Multiple-Class Linear Classification (1/2)

- Assume that classes are linearly separable

$$\Rightarrow g_i(\mathbf{x} \mid \mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad \text{and} \quad g_i(\mathbf{x} \mid \mathbf{w}_i, w_{i0}) = \begin{cases} > 0 & \mathbf{x} \in C_i \\ \leq 0 & \text{otherwise} \end{cases}$$



ideal case

Classification:

Choose C_i if

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

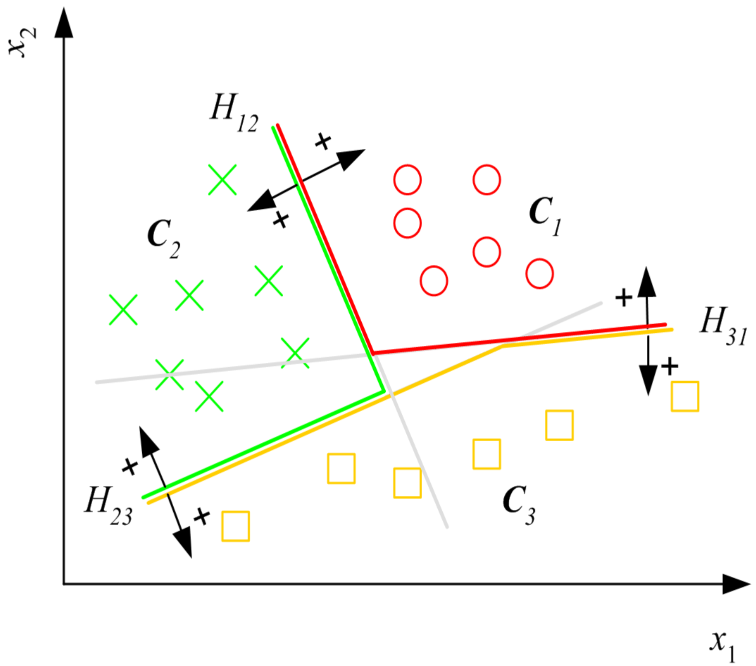
Multiple-Class Linear Classification (2/2)

- Assume that classes are pairwise separable

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

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

For *K*-class classification,
it needs *K*(*K*-1)/2 linear discriminants



Classification:

choose C_i if

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

However, in real-world applications, it's not always the case (linear separation)
- Using summation instead of conjunction

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

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

Parametric Discrimination Revisited

- The discriminant functions for Gaussian class densities $p(\boldsymbol{x} \mid C_i)$ sharing a common covariance matrix are linear

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

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

$$\mathbf{w}_i = \Sigma^{-1}\mu_i \quad w_{i0} = -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \log P(C_i)$$

- For two-class classification

$$y \equiv P(C_1 \mid \boldsymbol{x}) \text{ and } P(C_2 \mid \boldsymbol{x}) = 1 - y$$

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

– $\log[y/(1-y)]$ is known as the logit transformation/log odds of y

Posterior Probability and Sigmoid Function (1/2)

- For two normal (Gaussian) classes sharing a common covariance matrix

$$\text{logit}\left(P(C_1 \mid \mathbf{x})\right) = \log \frac{P(C_1 \mid \mathbf{x})}{1 - P(C_1 \mid \mathbf{x})} = \log \frac{P(C_1 \mid \mathbf{x})}{P(C_2 \mid \mathbf{x})} = \log \frac{p(\mathbf{x} \mid C_1)}{p(\mathbf{x} \mid C_2)} + \log \frac{P(C_1)}{P(C_2)}$$

$$= \log \frac{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-(1/2)(\mathbf{x} - \mu_1)^T \Sigma^{-1} (\mathbf{x} - \mu_1)\right]}{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-(1/2)(\mathbf{x} - \mu_2)^T \Sigma^{-1} (\mathbf{x} - \mu_2)\right]} + \log \frac{P(C_1)}{P(C_2)}$$

$$= \mathbf{w}^T \mathbf{x} + w_0$$

$$\text{, where } \mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2), \quad w_0 = -\frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) + \log \frac{P(C_1)}{P(C_2)}$$

$$\log \frac{x}{1-x} = y \Rightarrow \frac{x}{1-x} = \exp y \Rightarrow x = \frac{\exp y}{1 + \exp y} \Rightarrow \frac{1}{1 + \exp(-y)}$$

$$\text{The inverse of logit} \quad \left(\log \frac{P(C_1 \mid \mathbf{x})}{1 - P(C_1 \mid \mathbf{x})} = \mathbf{w}^T \mathbf{x} + w_0\right)$$

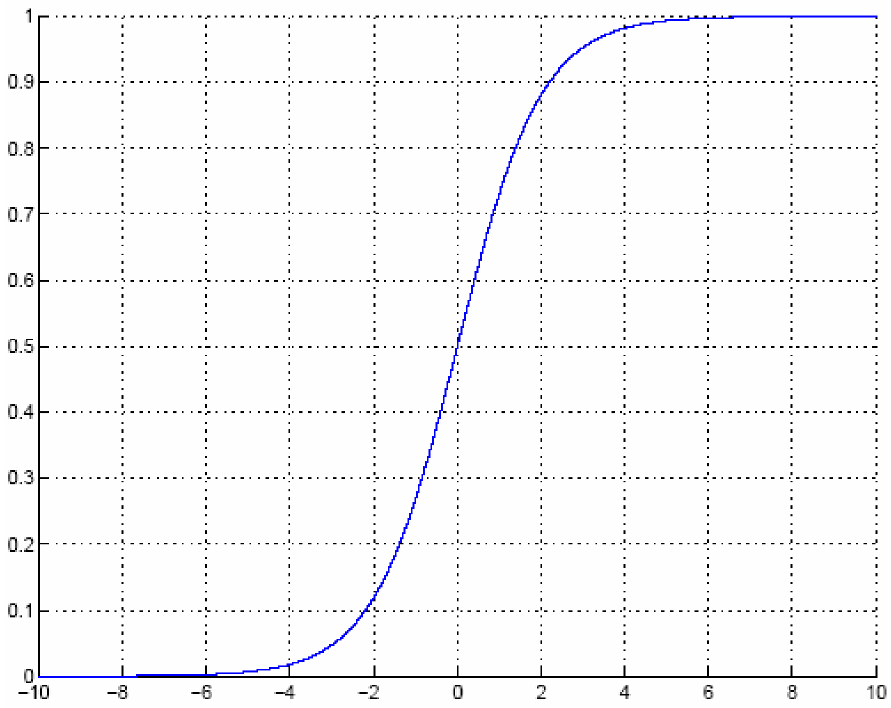
$$\Rightarrow P(C_1 \mid \mathbf{x}) = \text{sigmoid}\left(\mathbf{w}^T \mathbf{x} + w_0\right) = \frac{1}{1 + \exp\left[-\left(\mathbf{w}^T \mathbf{x} + w_0\right)\right]}$$

Sigmoid function

Posterior Probability and Sigmoid Function (2/2)

- Plot of Sigmoid (Logistic) Function

$y = \text{sigmoid}(g(\mathbf{x}))$



$g(\mathbf{x})$

1. Calculate $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$ and choose C_1 if $g(\mathbf{x}) > 0$, or
2. Calculate $y = \text{sigmoid}(\mathbf{w}^T \mathbf{x} + w_0)$ and choose C_1 if $y > 0.5$

Gradient Descent (1/2)

- For the discriminant-based approach, parameters are optimized to minimize the classification error on the training set
  - E(***w***|X) is error with parameters ***w*** (*d*-dimensional) on sample X

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w} \mid X)$$

  - The gradient vector composed of partial derivatives

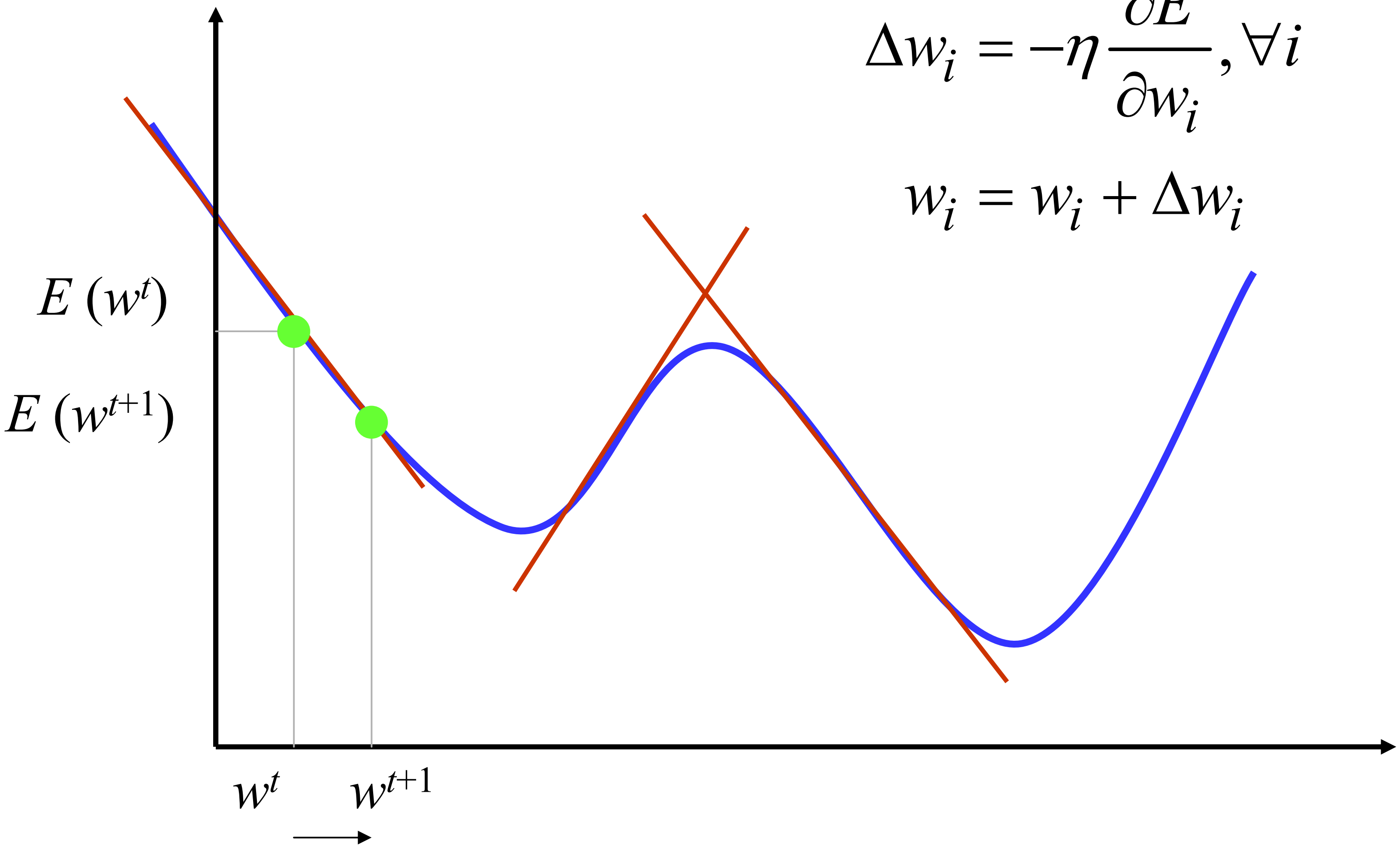
$$\nabla_w E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_d} \right]^T$$

- Gradient-descent:
  - Starts from random ***w*** and updates ***w*** iteratively in the negative direction of gradient

Gradient Descent (2/2)

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

$$w_i = w_i + \Delta w_i$$



Logistic Discrimination (1/8)

- Two classes: Assume log likelihood ratio is linear
  - Classes share a common covariance matrix

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

$$\operatorname{logit}(P(C_1 \mid \mathbf{x})) = \log \frac{P(C_1 \mid \mathbf{x})}{1 - P(C_1 \mid \mathbf{x})} = \log \frac{p(\mathbf{x} \mid C_1)}{p(\mathbf{x} \mid C_2)} + \log \frac{P(C_1)}{P(C_2)}$$

$$= \mathbf{w}^T \mathbf{x} + w_0$$

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

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

Logistic Discrimination (2/8)

- Training of discriminant parameters: the two-class case
  - Assume r^t, given $\mathbf{x}^t$, is Bernoulli with probability $y^t = P(C_1 \mid \mathbf{x}^t)$

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

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

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

$$E = -\log l$$

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

Logistic Discrimination (3/8)

- Training of discriminant parameters: the two-class case

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

$$\left[\text{Note that if } y = \text{sigmoid}(\mathrm{a}) \Rightarrow \frac{dy}{da} = y(1-y)\right]$$

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

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

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

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

$$\frac{dy}{da} = \frac{\exp(-a)}{[1+\exp(-a)]^2} = \frac{1}{1+\exp(-a)} \frac{\exp(-a)}{1+\exp(-a)}$$

$$= y \cdot (1-y)$$

$$a = \mathbf{w}^T \mathbf{x}^t + w_0$$

$$= \left(\sum_{j=1}^{d} w_j x_j^t \right) + w_0$$

Logistic Discrimination (4/8)

- Training of discriminant parameters: the two-class case

```
For j = 0, ..., d
    w_j ←rand(-0.01,0.01)
Repeat
    For j = 0, ..., d
        Δw_j ← 0
    For t = 1, ..., N
        o ← 0
        For j = 0, ..., d
            o ← o + w_j x_j^t
        y ← sigmoid(o)
        Δw_j ← Δw_j + (r^t − y) x_j^t
    For j = 0, ..., d
        w_j ← w_j + η Δw_j
Until convergence
```

Figure 10.6: Logistic discrimination algorithm implementing gradient-descent for the single output case with two classes. For w_0, we assume that there is an extra input x_0, which is always $+1$: $x_0^t \equiv +1, \forall t$.

Logistic Discrimination (5/8)

- Multiple classes: Take one of the classes, e.g., C_K, as the reference class and assume

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

$$\log \frac{p(\mathbf{x} \mid C_i)}{p(\mathbf{x} \mid C_K)} = \mathbf{w}_i^T \mathbf{x} + w_{i0}^o \Rightarrow \log \frac{p(C_i \mid \mathbf{x})}{p(C_K \mid \mathbf{x})} = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

softmax

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

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

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

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

Logistic Discrimination (6/8)

- Appendix

$$\log \frac{p(C_i \mid \mathbf{x})}{p(C_K \mid \mathbf{x})} = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

$$\Rightarrow \frac{p(C_i \mid \mathbf{x})}{p(C_K \mid \mathbf{x})} = \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right) \Rightarrow \sum_{i=1}^{K-1} \frac{p(C_i \mid \mathbf{x})}{p(C_K \mid \mathbf{x})} = \sum_{i=1}^{K-1} \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)$$

$$\Rightarrow \frac{1 - p(C_K \mid \mathbf{x})}{p(C_K \mid \mathbf{x})} = \sum_{i=1}^{K-1} \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)$$

$$\Rightarrow p(C_K \mid \mathbf{x}) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)}$$

$$\Rightarrow p(C_i \mid \mathbf{x}) = \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right) \cdot p(C_K \mid \mathbf{x}) = \frac{\exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)}{1 + \sum_{i=1}^{K-1} \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)}$$

$$= \frac{\exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)}{\sum_{i=1}^{K} \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right) + \left(1 - \exp\left(\mathbf{w}_K^T \mathbf{x} + w_{i0}\right)\right)} = \frac{\exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)}{\sum_{i=1}^{K} \exp\left(\mathbf{w}_i^T \mathbf{x} + w_{i0}\right)}$$

Note that

$$\exp\left(\mathbf{w}_K^T \mathbf{x} + w_{i0}\right) = \frac{p(C_K \mid \mathbf{x})}{p(C_K \mid \mathbf{x})} = 1$$

Logistic Discrimination (7/8)

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

Figure 10.8: Logistic discrimination algorithm implementing gradient-descent for the case with $K > 2$ classes. For generality, we take $x_0^t \equiv 1, \forall t$.

Logistic Discrimination (8/8)

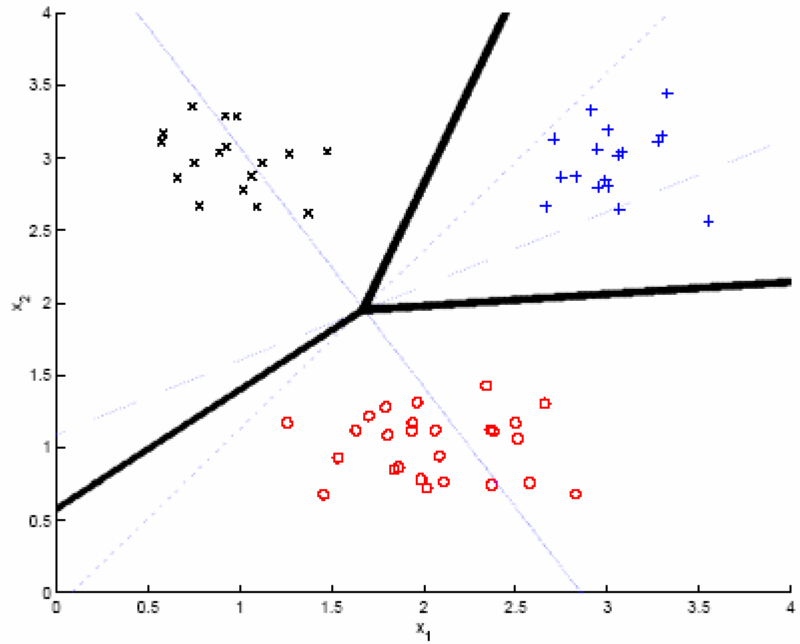


Figure 10.9: For a two-dimensional problem with three classes, the solution found by logistic discrimination. Thin lines are where $g_i(\boldsymbol{x}) = 0$, and the thick line is the boundary induced by the linear classifier choosing the maximum.

$$\text{thin line: } g_i(\mathbf{x}|\mathbf{w}_i, w_{i0}) = \begin{cases} > 0 & \mathbf{x} \in C_i \\ \leq 0 & \text{otherwise} \end{cases} \quad \text{(ideal case)}$$

$$\text{thick line: } p(C_i \mid \mathbf{x}) = \frac{\exp\left[\mathbf{w}_i^T \mathbf{x} + w_{i0}\right]}{\sum_{j=1}^{K} \exp\left[\mathbf{w}_j^T \mathbf{x} + w_{j0}\right]}$$

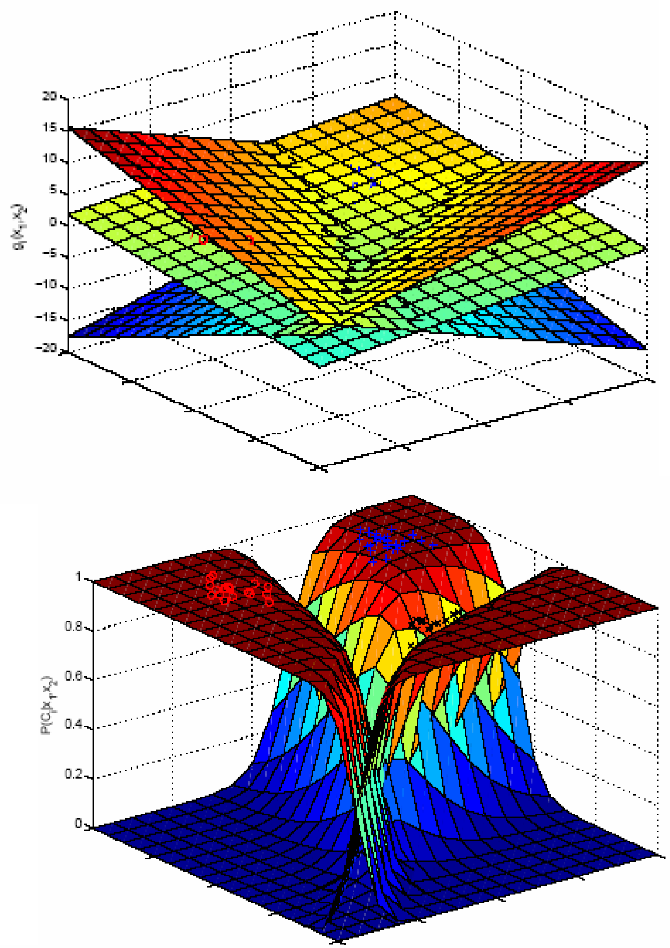


Figure 10.10: For the same example in figure 10.9, the linear discriminants (top), and the posterior probabilities after the softmax (bottom).

Generalizing the Linear Model

- Quadratic

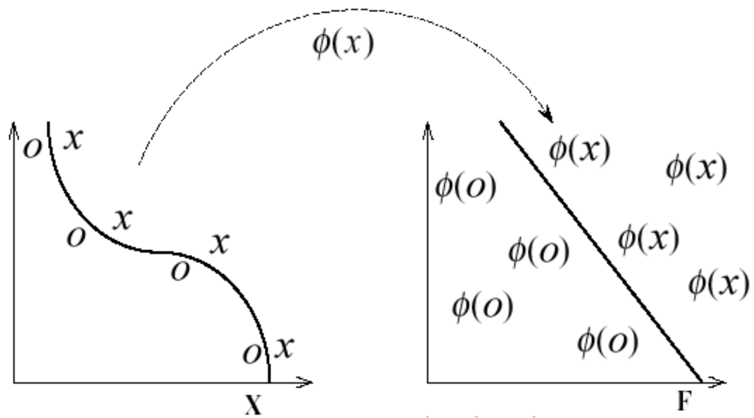
$$\log \frac{p(\mathbf{x} \mid C_i)}{p(\mathbf{x} \mid C_K)} = \mathbf{x}^T \mathrm{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

- Sum of basis functions

$$\log \frac{p(\mathbf{x} \mid C_i)}{p(\mathbf{x} \mid C_K)} = \mathbf{w}_i^T \phi(\mathbf{x}) + w_{i0}$$

where $\phi(\mathbf{x})$ are basis functions

- Kernels in SVM
- Hidden units in neural networks



Discrimination by Regression

- Classes are NOT mutually exclusive and exhaustive
  - An instance can belong to different classes with different probabilities

$$\mathbf{r}^t = \mathbf{y}^t + \boldsymbol{\varepsilon} \text{ where } \quad \boldsymbol{\varepsilon} \sim N\left(0, \sigma^2 \mathbf{I}_d\right)$$

$$\therefore \ \mathbf{r}^t \sim N\left(\mathbf{y}^t, \sigma^2 \mathbf{I}_d\right)$$

$$y_i^t = \text{sigmoid}\left(\mathbf{w}_i^T \mathbf{x}^t + w_{i0}\right) = \frac{1}{1 + \exp\left[-\left(\mathbf{w}_i^T \mathbf{x}^t + w_{i0}\right)\right]}$$

$$l\left(\mathbf{w}, w_0 \mid \mathcal{X}\right) = \prod_t \frac{1}{(2\pi)^{d/2} \left|\sigma^2 \mathbf{I}_d\right|^{1/2}} \exp\left[-\left(\mathbf{r}^t - \mathbf{y}^t\right)^T \left(\sigma^2 \mathbf{I}_d\right)^{-1} \left(\mathbf{r}^t - \mathbf{y}^t\right)\right]$$

$$\Rightarrow \ E\left(\{\mathbf{w}_i, w_{i0}\}_i \mid \mathcal{X}\right) = \frac{1}{2} \sum_t \left\|\mathbf{r}^t - \mathbf{y}^t\right\|^2 = \frac{1}{2} \sum_t \sum_i \left(r_i^t - y_i^t\right)^2$$

$$\Delta \mathbf{w}_i = \eta \sum_t \left(r_i^t - y_i^t\right) y_i^t \left(1 - y_i^t\right) \mathbf{x}^t$$

$$\Delta w_{i0} = \eta \sum_t \left(r_i^t - y_i^t\right) y_i^t \left(1 - y_i^t\right)$$

Equivalent to minimizing the sum of square errors (sharing a common diagonal covariance matrix)

Optimal Separating Hyperplane (1/5)

- Cortes and Vapnik, 1995; Vapnik, 1995

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_t \text{ where } r^t = \begin{cases} +1 & \text{if } \mathbf{x}^t \in C_1 \\ -1 & \text{if } \mathbf{x}^t \in C_2 \end{cases}$$

$$\text{find } \mathbf{w} \text{ and } w_0 \text{ such that}$$

$$\mathbf{w}^T \mathbf{x}^t + w_0 \geq +1 \text{ for } r^t = +1$$

$$\mathbf{w}^T \mathbf{x}^t + w_0 \leq +1 \text{ for } r^t = -1$$

$$\text{which can be rewritten as}$$

$$r^t \left(\mathbf{w}^T \mathbf{x}^t + w_0 \right) \geq +1$$

  - We do not only want the instances to be on the right side of the hyperplane, but we also want them some distance away

Optimal Separating Hyperplane (2/5)

- Margin
  - Definition: Margin is the distance from the discriminant to the closest instances on either side
  - Distance of $\mathbf{x}^t$ to the hyperplane is $\dfrac{\left|\mathbf{w}^T\mathbf{x}^t + w_0\right|}{\|\mathbf{w}\|}$
    - Can also be rewritten as $\dfrac{r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right)}{\|\mathbf{w}\|}$, where $r^t \in \{-1, +1\}$
  - We would like that $\dfrac{r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right)}{\|\mathbf{w}\|} \geq \rho, \forall t$
    - For a unique solution, fix $\rho \cdot \|\mathbf{w}\|$ and to maximize the margin

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t$$

– Have to do with the input dimension *d*

Optimal Separating Hyperplane (3/5)

- Margin

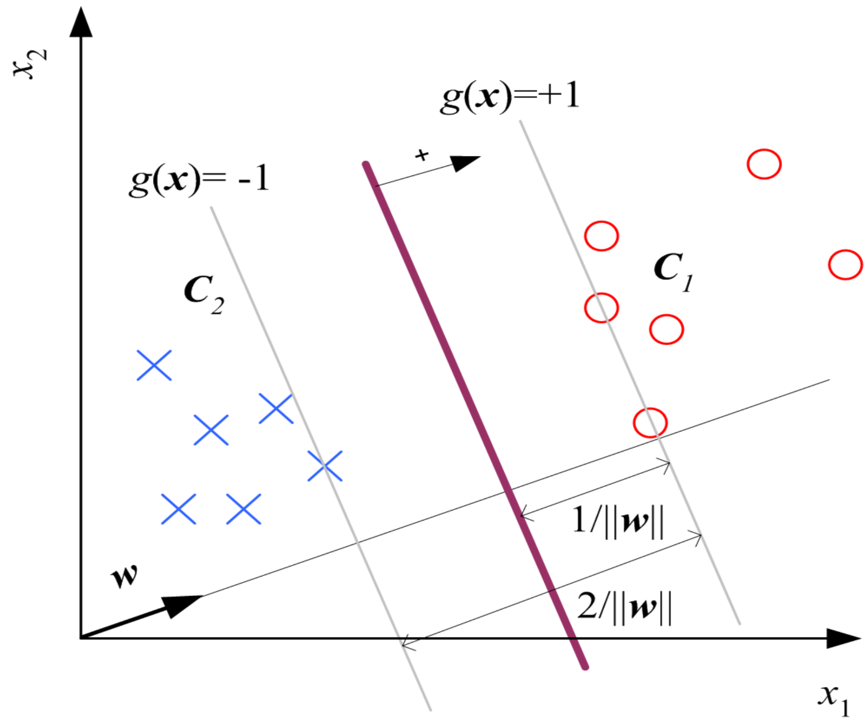


Figure 10.11: On both sides of the optimal separating hyperplance, the instances are at least $1/\|\boldsymbol{w}\|$ away and the total margin is $2/\|\boldsymbol{w}\|$.

Optimal Separating Hyperplane (4/5)

- Primal Problem
  - Minimize L_p with respect to $\mathbf{w}, w_0$

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t$$

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N}\alpha^t\left[r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) - 1\right]$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N}\alpha^t r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) + \sum_{t=1}^{N}\alpha^t$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{t=1}^{N}\alpha^t r^t \mathbf{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{t=1}^{N}\alpha^t r^t = 0$$

Optimal Separating Hyperplane (5/5)

- Dual Problem (Karush-Kuhn-Tucker)
  - Maximize L_d with respect to α^t

$$L_d = \frac{1}{2}\left(\mathbf{w}^T\mathbf{w}\right) - \mathbf{w}^T\sum_t \alpha^t r^t \mathbf{x}^t - w_0 \sum_t \alpha^t r^t + \sum_t \alpha^t$$

$$= \frac{1}{2}\left(\mathbf{w}^T\mathbf{w}\right) + \sum_t \alpha^t$$

scalar

Quadratic optimization

$$= \frac{1}{2}\sum_t\sum_s \alpha^t \alpha^s r^t r^s \left(\mathbf{x}^t\right)^T \mathbf{x}^s + \sum_t \alpha^t$$

$$\text{subject to } \sum_{\mathrm{t}} \alpha^t r^t = 0 \text{ and } \alpha^t \geq 0, \forall t$$

- Most α^{t*} are 0 and only a small number have $\alpha^{t*} > 0$; they are support vectors
- Have to do with the number of training instances, but not the input dimension

Kernel Machines

- Preprocess input ***x*** by basis functions

$$\mathbf{z} = \boldsymbol{\varphi}\left(\mathbf{x}^t\right)$$

$$g(\mathbf{z}) = \mathbf{w}^T\mathbf{z} \qquad (\text{assume } z_0 = 1)$$

$$g(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x})$$

- The SVM solution

$$\mathbf{w} = \sum_t \alpha^t r^t \mathbf{z}^t = \sum_t \alpha^t r^t \boldsymbol{\varphi}\left(\mathbf{x}^t\right)$$

$$g(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) = \sum_t \alpha^t r^t \boldsymbol{\varphi}\left(\mathbf{x}^t\right)^T \boldsymbol{\varphi}(\mathbf{x})$$

$$g(\mathbf{x}) = \sum_t \alpha^t r^t K\left(\mathbf{x}^t, \mathbf{x}\right)$$

Kernel Functions

- Polynomials of degree *q*: $K(\boldsymbol{x}^t, \boldsymbol{x}) = (\boldsymbol{x}^T \boldsymbol{x}^t + 1)^q$

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^2$$
$$= (x_1 y_1 + x_2 y_2 + 1)^2$$
$$= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2$$
$$\phi(\boldsymbol{x}) = \left[1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2\right]^T$$

- Radial-basis functions: $K(\boldsymbol{x}^t, \boldsymbol{x}) = \exp\left[-\frac{\|\boldsymbol{x}^t - \boldsymbol{x}\|^2}{\sigma^2}\right]$

- Sigmoidal functions: $K(\boldsymbol{x}^t, \boldsymbol{x}) = \tanh\left(2\boldsymbol{x}^T \boldsymbol{x}^t + 1\right)$