Embeddings into Feature Space

Given a set of labelled vectors belonging to some domain set $\mathcal{X}$ with labels in $\mathcal{Y} = \{1, -1\}$

$$S := (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)$$

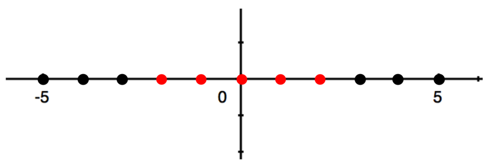
it is unlikely that these vectors will be separable by a hyperplane. The halfspace hypothesis space is rather restrictive in real-world applications of machine learning. For example, consider the vector space $\mathbb{R}$ and training set composed of vectors

$$-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5$$

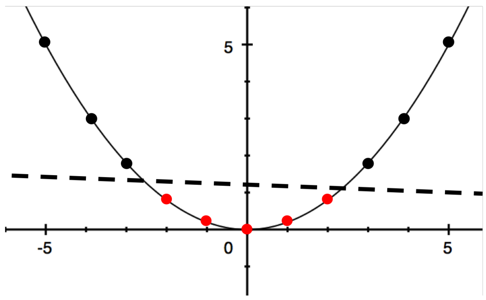
with labels defined as follows.

$$y = \begin{cases} 1 & |x| > 2 \\ -1 & \text{otherwise} \end{cases}$$

This scenario is illustrated in Figure 1. Clearly, these items are not linearly separable.



(a)



(b)

Figure 1: (a) A set of 11 items in $\mathbb{R}$. Red denotes $y = 1$ and black denotes $y = -1$. (b) The items projected into $\mathbb{R}^2$ by the function $\psi(x) = [x, x^2]$. Now they are linearly separable.

One solution that will allow us to learn a linear classifier on non-separable vectors is to project the vectors into a new space (usually of higher dimension) where they are linearly separable. We consider the mapping ψ of vectors in $\mathcal{X}$ into a higher-dimensional space $\mathcal{F}$ called the **feature-space**:

$$\psi : \mathcal{X} \rightarrow \mathcal{F}$$

Note that for any probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, we can define its image probability distribution $\mathcal{D}^{\psi}$ over $\mathcal{F} \times \mathcal{Y}$ as follows:

$$P_{\mathcal{D}^{\psi}}(\mathbf{v}, y) = \sum_{\mathbf{x}:\psi(\mathbf{x})=\mathbf{v}} P_{\mathcal{D}}(\mathbf{x}, y)$$

Finally, the generalization error over $\mathcal{D}$ is defined as

$$L_{\mathcal{D}}(h) = E_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(h \circ \psi, \mathbf{x}, y)]$$

Example: Polynomial Mapping

Given vector space $\mathbb{R}^n$, we define a k-degree polynomial mapping from $\mathbb{R}$ to $\mathbb{R}$ as

$$p(x) = \sum_{j=0}^{k} w_j x^j \tag{1}$$

We see that we can formulate the projection function

$$\psi(x) = [1, x, x^2, \ldots, x^k]$$

and consider the vector

$$\mathbf{w} = [w_0, w_1, \ldots, w_k]$$

for which Equation 1 can be viewed as the dot-product between $\mathbf{w}$ and $\psi(x)$. That is

$$\langle \mathbf{w}, \psi(x) \rangle = \sum_{j=0}^{k} w_j \psi_j(x)$$

$$= \sum_{j=0}^{k} w_j x^j$$

Thus, if we find a hyperplane defined by $\mathbf{w} \in \mathbb{R}^k$ that separates the $\psi(x)$ vectors, this will be the equivalent of finding a polynomial decision boundary in $\mathbb{R}$.

This process can be generalized into any $\mathbb{R}^n$ space by defining the multivariate polynomial as

$$p(\mathbf{x}) = \sum_{r=0}^{n} \sum_{j \in \{0,1,\ldots,n\}^r} w_J \prod_{i=1}^{r} x_{J_i}$$

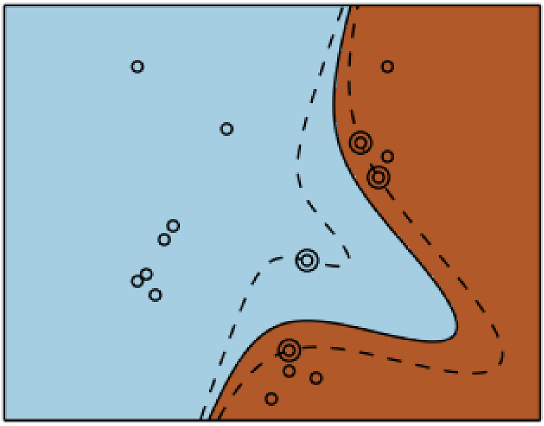


Figure 2: Finding the maximum-margin hyperplane in the polynomial-based projected space equates to finding a polynomial in $\mathbb{R}^2$